УДК 519.6

# НЕКОТОРЫЕ РЕЗУЛЬТАТЫ ОЦЕНКИ ПОГРЕШНОСТИ ВЫЧИСЛЕНИЙ С ОКРУГЛЕНИЯМИ НА ЭВМ НА ПРИМЕРЕ РЕШЕНИЯ ОДНОЙ ЗАДАЧИ ЛИНЕЙНОЙ АЛГЕБРЫ

А. С. Сухих, В. И. Федянин, В. Ф. Юдинцев
(РФЯЦ-ВНИИЭФ)

**На примерах расчетов двумерной задачи теплопроводности с известным точным решением сделан подбор двух параметров формулы, предложенной для оценки относительной погрешности вычислений с округлениями на ЭВМ.**

## Введение

Для оценок вычислительных погрешностей, накапливаемых из-за округлений, авторами была предложена и численно подтверждена на некоторых методических расчетах критпараметра $\lambda$ следующая формула относительной погрешности $d$ результата, который получается за $N$ арифметических операций, выполняемых с округлениями на ЭВМ:

$$d = 10^K \sqrt{N} C_1 \Delta_1. \qquad (1)$$

Здесь $\Delta_1$ — предельная относительная погрешность одной операции округления, константа, зависящая от длины основного слова ЭВМ. Так, для компьютера HP Work Station X400, на котором проводились расчеты с одинарной и двойной точностью, длина $m_1$ одинарного слова для размещения чисел REAL*4 равна 32 двоичным разрядам, $\Delta_1 = 10^{-7}$; длина $m_1$ двойного слова для размещения чисел REAL*8 равна 64 двоичным разрядам, $\Delta_1 = 10^{-16}$ (подробнее примеры расчета $\Delta_1$ приведены в разд. 2).

Выражение $\sqrt{N}$ в (1) при больших $N$ обусловлено вероятностным механизмом сложения относительных погрешностей аргументов арифметических операций, выполняемых с округлениями, а множитель $10^K$ связан с теми особенностями численной методики и проводимого конкретного расчета, которые могут резко увеличивать накопленную погрешность. Одной из таких особенностей является наличие в алгоритмах операций вычитания близких величин, увеличивающих относительную погрешность результата. Наконец, коэффициент $C_1$ корректирует завышенность* как теоретической оценки $\sqrt{N}$, так и величины $\Delta_1$. Дело в том, что в современных компьютерах арифметические сопроцессоры могут использовать и увеличенную (например, $m_1 = 80$ вместо стандартной $m_1 = 64$) разрядную сетку для чисел типа REAL при совершении части арифметических операций. Подобного рода аппаратные и программистские возможности ЭВМ приводят к существенному уменьшению величины $\Delta_1$ в части арифметических операций. Коэффициент $C_1$ дает возможность эффективно учесть это явление.

В выполненных авторами расчетах трех задач переноса были получены следующие значения параметров формулы (1):
$10^{K_1} = 343, \quad 10^{K_2} = 0{,}03, \quad 10^{K_3} = 5289;$
$C_1 \approx 0{,}1$ и $N \approx 10^8$ во всех задачах. Отметим высокую точность (доли процента) определения параметра $10^K$ с помощью использованной авторами процедуры и низкую точность определения параметра $C_1$ (приблизительно в 2—3 раза завышен коэффициент $C_1 = 0{,}1$) из-за недостаточной точности фиксации параметров $d$ и $N$ в расчетах.

В данной работе решается одна стационарная плоская двумерная задача теплопроводности с известным точным решением, так что накапливаемую в расчетах погрешность $d$ можно фиксировать точно. Другой особенностью задачи оказывается факт близости в расчетах параметра

*Можно показать, что в предельном случае, когда все $N$ операций являются операциями умножения или деления и после каждой выполняется округление, накапливаемая вероятностным механизмом часть погрешности равна $\sqrt{N} C_1^* \Delta_1$, где $C_1^* \approx 1$.

$10^K$ к 1,0 ($10^K \approx 0{,}995$), так что величина $d$ с хорошей точностью совпадает с погрешностью, накапливаемой вероятностным механизмом. Наконец, и параметр $N$ в расчетах удается отследить довольно точно. Поэтому задача может служить хорошим тестовым примером для оценки значений параметра $C_1$ в расчетах. Такая оценка и выполняется в настоящей работе (разд. 1).

Решение выбранной задачи сводится к решению систем конечно-разностных линейных уравнений большой размерности, но с сильно разреженными матрицами. Решение таких систем находилось двумя итерационными методами, реализованными в библиотеке стандартных программ: быстро сходящимся методом сопряженных градиентов (CG-методом) и очень медленно сходящимся методом Ричардсона (R-методом). По обоим методам расчеты проводились на нескольких измельчающихся пространственных сетках, чем достигалось варьирование параметра $N$. Основной полученный результат таков.

В расчетах по R-методу, в которых значение $N$ достигало $10^{12}$, зафиксировано неплохое (с точностью $\sim 5\,\%$) выполнение соотношения (1) с коэффициентом $C_1 \approx 0{,}01$. В расчетах по CG-методу, в которых параметр $N$ достиг значения порядка $10^{10}$, соотношение (1) выполняется заметно хуже (с точностью $\sim 30\,\%$), с коэффициентом $C_1 \approx 0{,}001$. Тем не менее полученные результаты укрепляют уверенность в применимости формулы (1) для априорных оценок вычислительной погрешности от округлений, накапливающейся за $N \geq 10^{12}$ арифметических операций. В частности, в расчетах больших задач вполне достижимыми могут являться следующие значения параметров:

$$K \approx 6 \div 7; \quad C_1 \approx 0{,}001 \div 0{,}1.$$

Завершая Введение, отметим еще некоторые обстоятельства. Данная работа была инициирована вопросами, обсуждавшимися в начале 90-х годов на семинарах математического отделения РФЯЦ-ВНИИЭФ, проводимых под руководством И. Д. Софронова. На одном из этих семинаров Н. А. Дмитриевым было высказано следующее утверждение: относительная погрешность результата, получаемого за $N$ арифметических операций с округлениями, при больших $N$ должна быть пропорциональной $\sqrt{N}$ (из вероятностных соображений). Попытка доказать это утверждение, учитывая особенности некоторых разработанных в отделении численных методик, и привела в 2002 г. авторов настоящей статьи к формуле (1).

Отдельные фрагменты формулы встречаются в литературе. Так, значения параметра $10^K$ могут служить оценкой числа обусловленности в задачах линейной алгебры ([1], с. 304). Зависимость $\sqrt{N}$ фактически получена в работе [2], с. 75. В работе [3] преодолевалась ошибка от округлений, которая в обозначениях, принятых в этой работе, достигала значения $10^K\sqrt{N}C_1 = 10^6$. Но в полном виде (1) зависимость для $d$ авторам не встречалась.

## 1. Результаты численных расчетов

**1.** Приведем кратко постановку решаемой задачи.

В области $0 \leq x \leq 1$; $0 \leq y \leq 1$ ищется решение $T\,(x, y)$ уравнения

$$-\frac{\partial}{\partial x} D_x \frac{\partial T}{\partial x} - \frac{\partial}{\partial y} D_y \frac{\partial T}{\partial y} = 0 \qquad (2)$$

с граничным условием

$$T_{\text{г}} = 1{,}0 \qquad (3)$$

и коэффициентами $D_x = D_y = 0{,}1$.

Для решения берется равномерная сетка: $h_x = 1/p$, $h_y = 1/q$. В серединах счетных ячеек сетки вводятся искомые значения $T_{i+1/2,j+1/2}$, удовлетворяющие хорошо известному пятиточечному конечно-разностному уравнению, аппроксимирующему в каждой ячейке уравнение (2). Таким образом, задача сводится к решению системы линейных уравнений

$$AT = f,$$

где $T$ — вектор искомых температур; $f$ — вектор правых частей, легко вычисляемый по условию (3).

Точное решение задачи $T_{i+1/2,j+1/2} = 1{,}0$ находилось двумя итерационными методами. Значение $T^{\nu=0}_{i+1/2,j+1/2}$ на начальной итерации задавалось равным $f$, т. е. нулю во всех ячейках области, кроме "приграничных".

**2.** Результаты выполненных расчетов сведены в табл. 1, 2 для CG-метода и табл. 3, 4 для R-метода.

Сделаем некоторые пояснения к таблицам. Расчеты выполнялись с одинарной ($m_1 = 32$, $\Delta_1 = 10^{-7}$) и двойной ($m_1 = 64$, $\Delta_1 = 10^{-16}$)

*Таблица 1*

**Погрешности $\alpha_T = \tilde{T} - 1{,}0$ решения задачи методом сопряженных градиентов на сетке $p \times q = 128 \times 128$ в зависимости от вносимой в правую часть системы относительной погрешности $\delta_f$ и от числа итераций $\nu$**

| $\delta_f$ | $\alpha_T$: $\Delta_1 = 10^{-7}$; $\nu = 280$ | $\Delta_1 = 10^{-16}$; $\nu = 280$ | $\Delta_1 = 10^{-16}$; $\nu = 350$ | $\Delta_1 = 10^{-16}$; $\nu = 400$ |
|---|---|---|---|---|
| $10^{-2}$ | $0_2 9817600$* | $0_2 9816926$ | — | — |
| $10^{-3}$ | $0_3 9822845$ | $-0_3 9817014$ | — | — |
| $10^{-4}$ | $-0_4 9912252$ | $-0_4 9817225$ | — | — |
| $10^{-5}$ | $0_4 1060963$ | $0_5 9817397$ | $-0_5 9817007$ | — |
| $10^{-6}$ | $-0_5 2205372$ | $0_6 9820410$ | $-0_6 9817002$ | — |
| $10^{-7}$ | $-0_5 1549721$ | $0_7 9846883$ | $-0_7 9817002$ | — |
| $10^{-8}$ | $-0_5 1370907$ | $0_8 9848003$ | $0_8 9816973$ | — |
| $10^{-9}$ | $-0_5 1370907$ | $0_9 9861469$ | $0_9 9817036$ | — |
| $10^{-10}$ | — | $0_9 1009333$ | $0_{10} 9817214$ | — |
| $10^{-11}$ | — | $0_{10} 1812062$ | $0_{11} 9817924$ | — |
| $10^{-12}$ | — | $-0_{10} 1225930$ | $0_{12} 9841017$ | — |
| $10^{-13}$ | — | $-0_{10} 1189426$ | $0_{12} 1003642$ | $0_{12} 1003642$ |
| $10^{-14}$ | — | $-0_{10} 1185396$ | $0_{13} 1754152$ | $0_{13} 1310063$ |
| $10^{-15}$ | — | $-0_{10} 1182376$ | — | $0_{14} 4551914$ |
| $10^{-16}$ | — | $-0_{10} 1183775$ | — | $0_{14} 5884182$ |
| 0,0 | $-0_5 1370907$ | $-0_{10} 1181266$ | $0_{14} 7105427$ | $0_{14} 5995204$ |

*Таблица 2*

**Погрешности $\alpha_T = \tilde{T} - 1{,}0$ решения задачи методом сопряженных градиентов на сетке $p \times q$ без возмущения правой части системы ($\delta_f = 0$) в зависимости от числа итераций $\nu$**

| $p \times q$ | $\Delta_1 = 10^{-7}$: $\nu$ | $\alpha_T$ | $\Delta_1 = 10^{-16}$: $\nu$ | $\alpha_T$ |
|---|---|---|---|---|
| $500 \times 500$ | 600 | $-0_2 3246903$ | 600 | $-0_2 3254881$ |
| | 800 | $0_3 1732111$ | 800 | $-0_5 2854855$ |
| | 900 | $0_4 3755093$ | 900 | $-0_7 3472526$ |
| | 990 | $0_5 6437302$ | 990 | $0_9 5746150$ |
| | 996 | $0_5 5602873$ | 1000 | $0_9 4377601$ |
| | 998 | $-0_5 5543232$ | 1200 | $-0_{11} 1332712$ |
| | 1000 | $-0_5 5602837$ | 1400 | $-0_{13} 1509903$ |
| | 1200 | $-0_5 5602837$ | 1500 | $0_{14} 9992007$ |
| | — | — | 1500 | $0_6 9952881 (\delta_f = 10^{-6})$ |
| | — | — | 2000 | $0_{14} 9992007$ |
| $1000 \times 1000$ | 1000 | $-0_1 1547474$ | 1000 | $-0_1 1626337$ |
| | 2000 | $-0_4 2896786$ | 2000 | $0_9 2713381$ |
| | 2200 | $-0_4 1287460$ | 2500 | $-0_{11} 1247225$ |
| | 2300 | $-0_4 1233816$ | 2800 | $-0_{13} 4096723$ |
| | 2400 | $-0_4 1233816$ | 3000 | $-0_{13} 3819167$ |
| | — | — | 3000 | $0_6 9976417 (\delta_f = 10^{-6})$ |
| | — | — | 4000 | $-0_{13} 3819167$ |

---

*Здесь и далее в тексте для краткости используется запись десятичных дробей $0_n N \equiv 0{,}N \cdot 10^{-n}$.

*Таблица 3*

**Погрешности $\alpha_T = \tilde{T} - 1{,}0$ решения задачи методом Ричардсона на сетке $p \times q = 128 \times 128$ в зависимости от вносимой в правую часть системы относительной погрешности $\delta_f$ и от числа итераций $\nu$**

| $\delta_f$ | $\alpha_T$: $\Delta_1 = 10^{-7}$; $\nu = 60\,000$ | $\Delta_1 = 10^{-16}$; $\nu = 60\,000$ | $\Delta_1 = 10^{-16}$; $\nu = 600\,000$ |
|---|---|---|---|
| 0,1 | $-0_1 9824061$ | — | — |
| 0,01 | $-0_2 9887993$ | — | — |
| $10^{-3}$ | $-0_2 1052737$ | $-0_3 9817014$ | — |
| $10^{-4}$ | $-0_3 1682639$ | $-0_4 9817080$ | — |
| $10^{-5}$ | $-0_4 8201599$ | $-0_5 9817745$ | $0_5 9817007$ |
| $10^{-6}$ | $-0_4 7331371$ | $-0_6 9824390$ | — |
| $10^{-7}$ | — | $-0_7 9890838$ | $0_7 9817007$ |
| $10^{-8}$ | — | $-0_7 3078888$ | — |
| $10^{-10}$ | — | $-0_7 3031738$ | — |
| $10^{-12}$ | — | $-0_7 3031674$ | $-0_{12} 9819923$ |
| $10^{-14}$ | — | $-0_7 3031673$ | $0_{13} 1021405$ |
| $10^{-16}$ | — | $-0_7 3031673$ | $0_{14} 3552714$ |
| 0,0 | $-0_4 7265806$ | $-0_7 3031673$ | $0_{14} 3774758$ |
| 0,0 | $-0_4 7265806$ ($\nu = 100\,000$) | $-0_{12} 1945111$ ($\nu = 100\,000$) | $0_{14} 3774758$ ($\nu = 900\,000$) |

*Таблица 4*

**Погрешности $\alpha_T = \tilde{T} - 1{,}0$ решения задачи методом Ричардсона на сетках $p \times q$ без возмущения правой части системы ($\delta_f = 0$) в зависимости от числа итераций $\nu$**

| $p \times q$ | $\Delta_1 = 10^{-7}$: $\nu$ | $\alpha_T$ | $\Delta_1 = 10^{-16}$: $\nu$ | $\alpha_T$ |
|---|---|---|---|---|
| $100 \times 100$ | $21 \cdot 10^3$ | $-0_4 4601479$ | $22 \cdot 10^4$ | $-0_{14} 5440093$ |
| | $22 \cdot 10^3$ | $-0_4 4523993$ | $23 \cdot 10^4$ | $-0_{14} 5107026$ |
| | $60 \cdot 10^3$ | $-0_4 4523993$ | $30 \cdot 10^4$ | $-0_{14} 5107026$ |
| $200 \times 200$ | $7 \cdot 10^4$ | $-0_3 2927184$ | $2{,}35 \cdot 10^5$ | $-0_{12} 5129230$ |
| | $7{,}5 \cdot 10^4$ | $-0_3 1770258$ | $2{,}4 \cdot 10^5$ | $-0_{12} 3566257$ |
| | $10 \cdot 10^4$ | $-0_3 1770258$ | $3{,}0 \cdot 10^5$ | $-0_{12} 3568257$ |
| $400 \times 400$ | $24 \cdot 10^4$ | $-0_3 7736683$ | $7 \cdot 10^5$ | $-0_9 7587611$ |
| | $25 \cdot 10^4$ | $-0_3 7053614$ | $9 \cdot 10^5$ | $-0_{11} 1345590$ |
| | $30 \cdot 10^4$ | $-0_3 7053614$ | $14 \cdot 10^5$ | $-0_{11} 1345590$ |
| $600 \times 600$ | $4 \cdot 10^5$ | $-0_2 6709456$ | $18 \cdot 10^5$ | $-0_{10} 3395162$ |
| | $4{,}5 \cdot 10^5$ | $-0_2 3522038$ | $21 \cdot 10^5$ | $-0_{11} 2951861$ |
| | — | — | $24 \cdot 10^5$ | $-0_{11} 2951861$ |
| | $6 \cdot 10^5$ | $-0_2 1582801$ | — | — |
| | $8 \cdot 10^5$ | $-0_2 1582801$ | — | — |

точностью. Все расчеты в табл. 1, 3 и два расчета в табл. 2 выполнены с внесением в компоненты вектора $f$ "пилообразных" возмущений, определяемых формулой

$$\widetilde{f} = f \times (1 + \delta_f, 1 - \delta_f, 1 + \delta_f, 1 - \delta_f, \ldots) . \quad (4)$$

Формулу (4) следует понимать как умножение компонент вектора $f$ последовательно на величины в круглых скобках. Значения $\delta_f$ указаны в первом столбце табл. 1, 3, а результаты $\alpha_T = \widetilde{T} - 1{,}0$, где $\widetilde{T}$ — решение системы

$$A\widetilde{T} = \widetilde{f},$$

приведены в следующих столбцах. При этом в качестве $\alpha_T$ выбиралось значение из той ячейки сетки, в которой находился $\max\limits_{i,j} \left|\widetilde{T}_{i+1/2,j+1/2} - 1{,}0\right|$.

**3.** Перейдем к анализу результатов. По данным табл. 1, 3 и двум расчетам табл. 2 можно вычислить значение параметра $10^K$:

$$10^K = \frac{|\alpha_T|}{\delta_f}.$$

Поясним кратко применяемую численную процедуру оценки параметров $10^K$ и $C_1$. Она заключается в отслеживании влияния на результат серии малых возмущений, вносимых в разряды десятичных мантисс характерных входных данных задачи, при последовательном сдвиге возмущаемого разряда в сторону младших разрядов, слева направо, навстречу возмущениям от ошибок округлений, распространяющимся по младшим разрядам рассчитываемых величин справа налево. Такой способ позволяет удовлетворительно оценить в результатах тот разряд десятичной мантиссы, до которого распространилась погрешность от округлений. При этом параметр $10^K$, рассчитываемый по формуле

$$10^K = \frac{|\delta_{\text{рез}}|}{\delta_{\text{вх}}},$$

где $\delta_{\text{рез}}$ и $\delta_{\text{вх}}$ есть относительные погрешности результата и входных данных, определяется с высокой точностью (доли процента). Точность же оценки параметра $C_1$ невысока из-за относительно невысокой точности фиксации самой величины $d$.

Продолжим анализ результатов.

На сетке $p \times q = 128 \times 128$ для обоих методов при надлежащем сведении итераций (см. последний столбец в табл. 1, 3) получается значение $10^K = 0{,}9817$.

Зависимость параметра $10^K$ от сетки для CG-метода приведена в табл. 2. На сетке $p \times q = 500 \times 500$ получилось $10^K = 0{,}9953$, а на сетке $p \times q = 1000 \times 1000$ — $10^K = 0{,}9976$.

Таким образом, оба метода при хорошем сведении итераций в каждом не увеличивают ($10^K \approx 1{,}0$) накапливаемую в последних разрядах результатов погрешность от округлений, которую в "чистом виде" можно наблюдать в таблицах в зависимости от варьируемых параметров.

**4.** При "недокрученных" итерациях результаты могут сильно отличаться от результатов при сведенных итерациях. Поэтому в расчетах на измельченных сетках количество итераций бралось таким, чтобы получаемые погрешности $\alpha_T$ уже не менялись при дальнейшем увеличении числа итераций.

**5.** Явление "установления" постоянного значения погрешности $\alpha_T$ при "перекрутке" итераций (см. табл. 2, 4) является особенностью взятых методов и использовано авторами для оценки необходимого числа итераций. Однако причины указанного явления еще не вполне ясны. В расчетах других задач при перекрутке итераций такого явления не наблюдалось — наблюдались ограниченные колебания содержимого младших разрядов результатов.

**6.** Оценку параметра $C_1$ формулы (1) сделаем из постулируемой для рассматриваемого теста зависимости

$$d = C_1\sqrt{N} \cdot \Delta_1. \quad (5)$$

Начнем с двух расчетов с одинарной точностью $\left(\Delta_1 = 10^{-7}\right)$ по CG-методу (см. левую часть табл. 2). В расчете на сетке 1 ($p \times q = 500 \times 500$) в качестве $d_1$ возьмем значение $|\alpha_T| = 0_5 56028$, достигнутое на итерации $\nu = 1\,000$. В расчете на сетке 2 ($p \times q = 1000 \times 1000$) в качестве $d_2$ возьмем значение $|\alpha_T| = 0_4 12338$, достигнутое на итерации $\nu = 2\,300$. В CG-методе количество арифметических операций на одну ячейку равно 20. В первом расчете полное количество операций равно $N_1 = 20 \cdot 500 \cdot 500 \cdot 1000 = 5 \cdot 10^9$, во втором расчете $N_2 = 20 \cdot 1000 \cdot 1000 \cdot 2300 = 4{,}6 \cdot 10^{10}$. Из зависимости (5) получаем: для первого расчета

$C_1^1 \approx 0{,}0008$, для второго расчета $C_1^2 \approx 0{,}0006$. Различие составляет $\approx 30\,\%$.

**7.** Рассмотрим расчеты по CG-методу с двойной точностью ($\Delta_1 = 10^{-16}$ — см. правую часть табл. 2). В расчете на сетке 1 (см. п. 6) в качестве $d_1$ возьмем значение $|\alpha_T| = 0_{14}9992$, достигнутое на итерации $\nu = 1500$ за $N_1 = 7{,}5 \cdot 10^9$ арифметических операций. В расчете на сетке 2 (также см. п. 6) в качестве $d_2$ возьмем значение $|\alpha_T| = 0_{13}3819$, достигнутое на итерации $\nu = 3000$ за $N_2 = 6 \cdot 10^{10}$ арифметических операций. Из зависимости (5) находим: для первого расчета $C_1^1 \approx 0{,}0011$, для второго — $C_1^2 \approx 0{,}0015$, т. е. также получаем расхождение $\approx 30\,\%$.

**8.** Рассмотрим теперь два расчета по R-методу с одинарной точностью (см. левую часть табл. 4). В расчете на сетке 1 ($p \times q = 400 \times 400$) в качестве $d_1$ возьмем значение $|\alpha_T| = 0_3 7054$, достигнутое на итерации $\nu = 250\,000$. В расчете на сетке 2 ($p \times q = 600 \times 600$) в качестве $d_2$ возьмем значение $|\alpha_T| = 0_2 1583$, достигнутое на итерации $\nu = 500\,000$. В R-методе количество арифметических операций на одну ячейку равно 12. В первом расчете полное количество операций $N_1 = 12 \cdot 400 \cdot 400 \cdot 250000 = 0{,}48 \cdot 10^{12}$, во втором расчете $N_2 = 12 \cdot 600 \cdot 600 \cdot 500000 = 2{,}16 \cdot 10^{12}$. Из зависимости (5) для первого расчета получаем $C_1^1 \approx 0{,}0108$, для второго — $C_1^2 \approx 0{,}0102$. Различие составляет $\approx 6\,\%$ .

**9.** Рассмотрим расчеты по R-методу с двойной точностью (см. правую часть табл. 4). На сетке 1 из п. 8 в качестве $d_1$ возьмем значение $|\alpha_T| = 0_{11}1346$, достигнутое на итерации $\nu = 900\,000$ за $N_1 = 1{,}728 \cdot 10^{12}$ операций. На сетке 2 (также см. п. 8) в качестве $d_2$ возьмем значение $|\alpha_T| = 0_{11}2952$, достигнутое на итерации $\nu = 2\,100\,000$ за $N_2 = 9{,}072 \times 10^{12}$ арифметических операций. Из зависимости (5) находим: для первого расчета $C_1^1 \approx 0{,}0098$, для второго — $C_1^2 \approx 0{,}0102$. Полученное различие ($\approx 4\,\%$) следует считать весьма небольшим. Отметим формулу для отношения $C_1^2/C_1^1$:

$$C_1^2/C_1^1 = \frac{d_2}{d_1} \bigg/ \frac{\sqrt{N_2}}{\sqrt{N_1}},$$

следующую из соотношения (5).

## 2. Примеры расчета предельных относительных погрешностей одной операции округления

Предельные относительные погрешности одной операции округления $\Delta_1$ определяют суммарную погрешность $d$ результатов, и их знание необходимо при оценке параметров зависимости (1). В табл. 5 приведены значения этих величин для ПЭВМ PC, в которых используется двоичная система для представления чисел [4, 5], и для сравнения — для ЕС ЭВМ [6], в которых использовалась шестнадцатеричная система. В табл. 5 $m$ — количество двоичных разрядов слова, используемое для записи мантиссы числа; $m_1 - m$ — количество разрядов для записи порядка числа и его знака.

Поясним получение значений $\Delta_1$.

**Значения $\Delta_1$ для ЭВМ двух типов**

*Таблица 5*

| Количество разрядов $m_1$ | ПЭВМ PC (двоичная система) | | ЭВМ ЕС (шестнадцатеричная система) | |
|---|---|---|---|---|
| | $m$ | $\Delta_1$ | $m$ | $\Delta_1$ |
| 32 | 23 | $\frac{1}{2}2^{-23} \approx 10^{-7,2}$ | 24 | $\left(\frac{1}{16}\right)^{-1}\frac{1}{2}16^{-6} \approx 10^{-6,3}$ |
| 64 | 52 | $\frac{1}{2}2^{-52} \approx 10^{-15,9}$ | 56 | $\left(\frac{1}{16}\right)^{-1}\frac{1}{2}16^{-14} \approx 10^{-15,9}$ |
| 80 | 64 | $\frac{1}{2}2^{-64} \approx 10^{-19,5}$ | 72 | $\left(\frac{1}{16}\right)^{-1}\frac{1}{2}16^{-18} \approx 10^{-20,7}$ |
| 96 | — | — | 88 | $\left(\frac{1}{16}\right)^{-1}\frac{1}{2}16^{-22} \approx 10^{-25,5}$ |
| 128 | — | — | 120 | $\left(\frac{1}{16}\right)^{-1}\frac{1}{2}16^{-30} \approx 10^{-35,1}$ |

При общепринятой записи чисел в позиционных системах счисления с основанием $\beta$, используемой и в ЭВМ, и стандартном способе округления "длинной" нормализованной мантиссы $m_x$ числа $x > 0$

$$m_x = 0{,}n_1 n_2 \ldots n_{k-1} n_k n_{k+1} \ldots; \quad (6)$$
$$1 \le n_1 \le \beta - 1; \quad 0 \le n_i \le \beta - 1 \text{ для } 2 \le i;$$
$$\beta^{-1} \equiv 0{,}1 \le m_x < 1$$

до "короткой" мантиссы $\widetilde{m}_x$, помещаемой в $K$ разрядах,

$$\widetilde{m}_x = 0{,}n_1 n_2 \ldots n_{k-1} n_k, \ 0 \le n_{k+1} \le \left[\frac{\beta}{2}\right] - 1; \quad (7)$$
$$\widetilde{m}_x = 0{,}n_1 n_2 \ldots n_{k-1} n_k$$
$$+ 0{,}0\ 0 \ldots\ 0\ \ 1, \quad \left[\frac{\beta}{2}\right] \le n_{k+1} \le \beta - 1; \quad (8)$$
$$\beta^{-1} \equiv 0{,}1 \le \widetilde{m}_x \le 1 \quad (9)$$

нетрудно вывести следующие формулы [2]:

$$|\widetilde{m}_x - m_x| < \frac{1}{2}\,\beta^{-k}; \quad (10)$$
$$|\delta_{\widetilde{x}}| = \frac{|x - \widetilde{x}|}{\widetilde{x}} \equiv \frac{|\widetilde{m}_x - m_x|}{\widetilde{m}_x} \le \frac{|\widetilde{m}_x - m_x|}{\beta^{-1}} <$$
$$< \frac{1}{\beta^{-1}}\,\frac{1}{2}\,\beta^{-k} = \Delta_1. \quad (11)$$

Для $\beta = 16$, полагая $K = m/4$, по формуле (11) получаем значения $\Delta_1$ в правой части табл. 5.

Однако для $\beta = 2$, полагая $K = m$, по формуле (11) получаются значения, в 2 раза бо́льшие, чем приведенные в левой части табл. 5. Дело заключается в следующем. В двоичной системе счисления для всех чисел $x > 0$ у их нормализованных мантисс $m_x$ (6) всегда первый разряд $n_1 = 1$. Это обстоятельство использовано конструкторами ПЭВМ PC: первый разряд не задается, он аппаратно полагается равным 1, а в разрядах $1, \ldots, K, \ldots$ ячейки в действительности располагаются значения разрядов $n_2, \ldots, n_{k+1}$ мантиссы $m_x$ (6). Таким образом, в ПЭВМ PC установлена следующая форма записи нормализованной мантиссы $m'_x$ числа $x > 0$ (см. [5], с. 30—32):

$$\begin{array}{ccccccc} m'_x & = & 1, n'_1\, n'_2 \ldots & n'_k & n'_{k+1} \ldots \\ & & \ \ \| \quad \| \quad \| & \| & \| \\ & & n_1\, n_2\, n_3 \ldots & n_{k+1} & n_{k+2} \ldots \end{array} \quad (12)$$

и $n'_0 = n_1 = 1$ не хранится.

Порядок числа $x$ при способе хранения (12) на 1 меньше, чем порядок числа с записью мантиссы в форме (6). Для (12) имеем

$$1 \le m'_x < 2.$$

Округляя (12) до $K$ разрядов (после запятой) по содержимому $n'_{k+1}$ разряда $K + 1$ точно так же, как в (7) и (8), вместо (9) получаем

$$1 \le \widetilde{m'_x} \le 2, \quad (13)$$

но соотношение (10) при $\beta = 2$ сохраняется:

$$\left|\widetilde{m'_x} - m'_x\right| < \frac{1}{2}\beta^{-k}. \quad (14)$$

В силу (13) и (14) вместо (11) при $\beta = 2$ имеем

$$\delta_{\widetilde{x}} \equiv \frac{|\widetilde{m}_x - m_x|}{\widetilde{m}_x} \le |\widetilde{m}_x - m_x| < \frac{1}{2}\beta^{-k} = \Delta_1. \quad (15)$$

Именно значения $\Delta_1$, полученные при $\beta = 2$ из (15), приведены в левой части табл. 5. Они в 2 раза меньше значений $\Delta_1$, полученных из (11). Достигнутое преимущество, однако, теряется с увеличением $m_1$, и авторы склонны отдать предпочтение представлению чисел в испытанной форме ЕС ЭВМ.

## Заключение

Подведем некоторые итоги выполненных исследований.

1. На нескольких примерах с помощью предложенных процедур удается подобрать значения параметров $10^K$ и $C_1$ формулы (1), с удовлетворительной точностью описывающей поведение относительной погрешности $d$ результата вычислений с округлениями на ЭВМ.

   Желательно дальнейшее продолжение численных экспериментов со значениями $N > > 10^{12}$ для проверки гипотезы о возрастании параметра $C_1$ с ростом $N$.

2. Формула (1) позволяет получить априорную оценку количества $m_1$ двоичных разрядов основного слова ЭВМ, необходимого для сохранения требуемой точности в расчетах с большим числом $N$. Так, при $N = 10^{20}$, $k = 6$, $C_1 = 0{,}01$ для размещения чисел целесообразно отвести 96-разрядные слова ($m_1 = 96$, $\Delta_1 = 10^{-25{,}5}$) . Указанное значение $N$ вполне достижимо для быстро развивающихся современных вычислительных систем.

Авторы выражают благодарность П. А. Авдееву и Б. Н. Шамраеву — авторам программ, по которым выполнялись расчеты.

## Список литературы


1. *Бахвалов Н. С. , Жидков Н. П. , Кобельков Г. М.* Численные методы. М.: Наука, 1987.
2. *Березин И. С., Жидков Н. П.* Методы вычислений. Т. 1. М.: Физматгиз, 1959.
3. *Чудов Л. А., Кудрявцев В. П.* Об ошибках округления при решении разностными методами задач с начальными условиями для эллиптических уравнений и систем // Сб. работ ВЦ МГУ, II / Под ред. Г. С. Рослякова, Л. А. Чудова. М.: МГУ, 1963. С. 20—47.
4. *Брэдли Д.* Программирование на языке ассемблера для персональной ЭВМ фирмы IBM: Пер. с англ. М.: Радио и связь, 1988.
5. *Маргулев А. И.* Язык С для PC. М.: АГАР, 1997.
6. *Айнберг В. Д., Геронимус Ю. В.* Основы программирования для Единой системы ЭВМ. М.: Машиностроение, 1980.