УДК 519.677

А. Г. Бирюков[1], *А. И. Гриневич*[1,2]

[1]Московский физико-технический институт (государственный университет)
[2]ООО «ГринМарк»

Метод оценки погрешностей округления решений задач вычислительной математики в арифметике с плавающей запятой, основанный на сравнении решений с изменяемой длиной мантиссы машинного числа

Статья посвящена вопросам анализа погрешностей округления решений задач вычислительной математики на ЭВМ в арифметике с плавающей запятой и переменной длиной мантиссы машинного числа. Предложен метод оценки погрешностей округления, основанный на сравнении решений с различной длиной мантисс, сформулированы правила достижения требуемой точности.

Ключевые слова: погрешность округления, точность решения задач вычислительной математики, машинное число с переменной длиной мантиссы, бесконечношаговый и конечношаговый алгоритмы, K-решение задачи вычислительной математики, гарантированная точность решений задач.

1. Введение

В работе авторов [1] предложен метод анализа погрешностей округления решения задач вычислительной математики (ВМ) в арифметике с плавающей запятой и переменной длиной мантиссы машинного числа (МЧ).

Настоящая статья является продолжением работы [1]; в ней предлагаются численные оценки погрешностей округления решений задач ВМ, гарантирующие достижение заданной точности, и дается их теоретическое обоснование.

Проблема анализа влияния погрешностей округления на решение задач ВМ актуальна со времени появления ЭВМ и остается таковой по сей день. Научные исследования над указанной темой ведутся в разных направлениях. Отметим классические работы по исследованию погрешностей округления при решении задач линейной алгебры [2,3]; исследование погрешностей округления в рамках интервального анализа [9,10]; статистический анализ погрешностей округления [6, 11]; исследование новых моделей по выработке машинного числа [7]; алгоритмы с автоматической коррекцией ошибок округления первого порядка — метод CENA [8].

В настоящей работе предлагается метод оценки погрешностей округления решений задач, отличный от рассмотренных в приведенной выше литературе. В соответствии с [1] под решением задачи понимается значение некоторой вектор-функции $f(x) \in R^k$, $x \in R^n$. Для оценки погрешностей решений вычисляется несколько их значений $f_{m_i}(x)$ при различных длинах мантиссы МЧ: $m_1 < m_2 < ...m_i < ...m_L$. Эти решения имеют погрешности $\Delta_i = \|f_{m_i}(x) - f(x)\|$, $i \in [1, L]$. Значение погрешности решения Δ_i, $i \in [1, L-1]$ оценивается значением погрешности $\Delta_{iL} = \|f_{m_i}(x) - f_{m_L}(x)\|$, $i \in [1, L-1]$, при этом предполагается, что $\Delta_L << \Delta_i$. Здесь и далее под нормой $\|\bullet\|$ понимается евклидова норма $\|z\| = \sqrt{\sum_{i=1}^{s} z_i^2}$, $z \in R^s$. В разделе 2 рассматриваются свойства некоторого решения задачи $f_{m_L}(x)$ при значении длины мантиссы $m = m_L$, названного **контрольным** или **K-решением** (КР). В разделе 3 вводится понятие гарантированной точности решений и рассматриваются правила оценок погрешностей округлений. В разделе 4 рассматриваются оценки погрешностей округления по правилу совпадения первых десятичных знаков (СПЗ)

решений с различной длиной мантиссы. В разделе 5 рассматриваются правила оценки погрешностей решений для класса бесконечношаговых алгоритмов (БША). В разделе 6 кратко рассматриваются вопросы эффективности метода К-решений. В разделе 7 на примерах решений систем линейных уравнений и задачи численного дифференцирования приведены результаты численного эксперимента, иллюстрирующие основные свойства метода КР.

Пусть $f(x)$ – точное решение и $\tilde{f}(x)$ – приближенное решение некоторой задачи ВМ. Понятия абсолютной и относительной погрешностей $\Delta f = \left\| f(x) - \tilde{f}(x) \right\|$ и $\frac{\Delta f}{\|f(x)\|}$ рассматриваются по отношению к $f(x)$ – точному, но неизвестному значению решения. В ВМ используется также другое определение относительной погрешности: $\Delta f / \left\| \tilde{f} \right\|$. Не рассматривая специфики этих определений, в статье используется её первое значение. Незнание значения $f(x)$ является принципиальным ограничением для получения оценок погрешностей Δf и $\Delta f / \|f\|$. Использование машинных чисел с возможностью изменения длины мантиссы при решении задач ВМ позволяет предложить вариант устранения этого принципиального ограничения.

2. К-решения задачи ВМ и их свойства

Определение 1. Совокупность L решений задачи $f_{m_i}(x)$ при значениях длины мантиссы $m_i, i \in [1, L]$: $m_1 < m_2 < ... < m_L$ назовем **итерационной последовательностью с переменной мантиссой** (ИППМ) решения задачи. □

Определение 2. Числа η и η_0 называются малыми по сравнению с 1, если $0 < \eta \leqslant \eta_0 \leqslant 0,1$. Условие малости чисел по сравнению с 1 обозначается символом «<<» – много меньше: $\eta << 1$, $\eta_0 << 1$. □

Условие $\eta_0 << 1$ задается Вычислителем (лицом, решающим задачу ВМ). Пусть $a_1 > 0$, $a_2 > 0$, $a_1 = \eta a_2$, $\eta << 1$, тогда a_1 – малое число по сравнению с a_2, т.е. $a_1 << a_2$. В вычислительной практике, в зависимости от требований задачи, число $\eta_0 \leqslant 0,1$ может меняться в широком диапазоне значений. Например: $\eta_0 = 0,1$; $\eta_0 = 0,05$; $\eta_0 = 10^{-k}$, $k = 1, 2, ...$ и т.д.

Определение 3. Пусть значения погрешностей решений равны $\Delta_i = \|f_{m_i}(x) - f(x)\|$, $i \in [1, L]$; $\Delta_{ij} = \left\|f_{m_i}(x) - f_{m_j}(x)\right\|$, $j > i$, $i, j \in [1, L]$. **К-решением** (КР) задачи ВМ называется значение $f_{m_L}(x)$, если

$$\Delta_i = \Delta_{iL} + \xi_{iL}\Delta_{iL}, i \in [1, L-1], \tag{1}$$

где $|\xi_{iL}| << 1$, т.е. $|\xi_{iL}|$ малое число по сравнению с 1. □

По смыслу К-решение означает «Контрольное решение», т.е. решение, позволяющее оценить значение погрешности решения $f_{m_i}(x)$. В работе [1] (теорема 1) было получено значение погрешности функции $f(x)$ при вычислении её с длиной мантиссы m. Для упрощения изложения представим его для частного случая порядка погрешности $\alpha = 1$:

$$\bar{\Delta} = f_m(x) - f(x) = \bar{C}_m b^{-m}, \tag{2}$$

где b – основание МЧ. Введем следующее

Определение 4. Обозначим $g_i = \frac{\Delta_{i+1}}{\Delta_i} = \frac{\|f_{m_{i+1}}(x) - f(x)\|}{\|f_{m_i}(x) - f(x)\|}$, $i \in [1, L-1]$. Последовательность решений $f_{m_i}(x)$ назовем **g-устойчивой**, если $g_i \leqslant g_0 << 1$, $i \in [1, L-1]$. Число g_i, $i \in [1, L-1]$ назовем коэффициентом уменьшения (КУ) погрешности на i-м шаге. □

Обобщением КУ g_i является число $g_{ij} = \frac{\Delta_j}{\Delta_i}$, $i < j$, $i \in [1, L-1]$. Для него справедливо:

$$g_{ij} = g_{j-1}g_{j-2}...g_i << 1, g_{ij} \leqslant g_i, j \in [2, L], \tag{3}$$

если ИППМ g-устойчива. Приведем достаточное условие g-устойчивости.

Лемма 1. *Пусть для ИППМ выполнено условие (2) и погрешности:* $\Delta' = \|f_{m'}(x) - f(x)\| = \left\|\bar{C}_{m'}\right\| b^{-m'}$ *и* $\Delta'' = \|f_{m''}(x) - f(x)\| = \left\|\bar{C}_{m''}\right\| b^{-m''}$ *удовлетворяют условию:* $\frac{\left\|\bar{C}_{m''}\right\|}{\left\|\bar{C}_{m'}\right\|} \leqslant \xi_0 = const$, $\forall m', m'' : m_1 \leqslant m' < m'' \leqslant m_L$. *Тогда найдется*

такая $\Delta m = \min\limits_i (m_{i+1} - m_i)$, $i \in [1, L-1]]$, *что* $g_i \leqslant g_0$, $g_0 << 1$, *где* g_0– *заданное малое число, т.е. ИППМ g-устойчива.*

Доказательство

Доказательство очевидно: $\Delta m \geqslant \left\lfloor 1 - \log_b \frac{g_0}{\xi_0} \right\rfloor$, где $\lfloor A \rfloor$ – целая часть числа A. □

Приведем без доказательства следующее утверждение.

Лемма 2. *Пусть* $V_1, V_2 \in R^n$ *и* $\|V_2\| < \|V_1\|$, *тогда имеет место неравенство* $\|V_1\| - \|V_2\| \leqslant \|V_1 - V_2\|$. □

Рассмотрим некоторые оценки погрешностей решений задач ВМ для g-устойчивой ИППМ. Ввиду ограниченности объема, приведем следующую лемму без доказательства.

Лемма 3. *Пусть ИППМ g-устойчива. Тогда для значений погрешностей* Δ_i, Δ_j, Δ_{ij} *имеют место двухсторонние оценки*:

$$\frac{\Delta_{ij}}{1+g_{ij}} \leqslant \Delta_i \leqslant \frac{\Delta_{ij}}{1-g_{ij}};\ \frac{g_{ij}\Delta_{ij}}{1+g_{ij}} \leqslant \Delta_j \leqslant \frac{g_{ij}\Delta_{ij}}{1-g_{ij}};\ (1-g_{ij})\,\Delta_i \leqslant \Delta_{ij} \leqslant (1+g_{ij})\,\Delta_i\,, \qquad (4)$$

где $i, j \in [1, L]$. □

Замечание. Из (4) при малых значениях g_{ij}: $g_{ij} \leqslant g_0 << 1$ можно получить важные практические оценки:

$$\Delta_{ij} \cong \Delta_i \text{ и } \Delta_j \cong g_{ij}\Delta_{ij}. \qquad (5)$$

Теорема 1. 1. *Пусть в ИППМ выполнено условие* $\frac{g_{ij}}{1-g_{ij}} \leqslant g_0 << 1$, $j > i$; $i, j \in [1, L]$. *Для того, чтобы значение функции* $f_{m_j}(x)$ *было K-решением, необходимо и достаточно, чтобы ИППМ была g-устойчивой. 2. Пусть ИППМ g-устойчива. Тогда для любого* $\varepsilon > 0$ *существует такое решение* $f_{m_i}(x)$, *что* $\Delta_i \leqslant \varepsilon$ *и* $\Delta_{ij} \leqslant (1+g_0)\,\varepsilon$, $j > i$.

Доказательство.

1. Доказательство п. 1 теоремы не приводится, ввиду ограниченности объема статьи.

2. Т.к.. ИППМ g-устойчива, т.е. $g_t \leqslant g_0$, $t = 1, 2, ..i, ..j$; $g_{ij} \leqslant g_0$ и $\Delta_i = g_{1i}\Delta_1 = (g_1 g_2 ... g_i)\,\Delta_1$, то $\Delta_i \leqslant g_0^i \Delta_1$. Потребуем, чтобы $\forall \varepsilon > 0$ выполнялось неравенство $\Delta_i \leqslant g_0^i \Delta_1 \leqslant \varepsilon$. Решая неравенство, найдем номер итерации, на которой гарантировано достижение требуемой точности решения: $i = \left\lfloor 1 + \frac{\ln(\varepsilon/\Delta_1)}{\ln g_0} \right\rfloor$. Из (4) следует: $\Delta_{ij} \leqslant (1+g_{ij})\,\Delta_i \leqslant (1+g_0)\,\varepsilon$. □

Замечание. Определение (3) называет К-решением задачи ВМ $f_{m_L}(x)$ – значение функции f при наибольшем заданном значении $m = m_L$. В доказанной теореме КР – значения $f_{m_j}(x)$, $j \in [2, L]$, оцениваемым решением является $f_{m_i}(x)$, $i \in [1, L-1]$. Из теоремы 1 следует важный практический вывод: если ИППМ g-устойчива, то достижима любая требуемая точность решения. На практике теорема 1 п.2 применима до максимального значения мантиссы, которую обеспечивает данная библиотека программ. В частности, для GNU MPFR $m_{\max} = 646\,456\,993$ десятичных знаков.

Введем понятие значения КУ, в котором точные значения $f(x)$ не используются.

Определение 5. Пусть $f_{m_L}(x)$ – КР задачи ВМ. Обозначим $g_i^L = \frac{\Delta_{i+1,L}}{\Delta_{iL}} = \frac{\|f_{m_{i+1}}(x) - f_{m_L}(x)\|}{\|f_{m_i}(x) - f_{m_L}(x)\|}$, $i \in [1, L-2]$. Последовательность решений задачи ВМ назовем **квазиустойчивой** (или g_i^L-устойчивой по отношению к КР), если $g_i^L << 1$. Число g_i^L назовем КУ значений Δ_{iL}. □

Обобщением параметра g_i^L является число $g_{ij}^L = \frac{\Delta_{jL}}{\Delta_{iL}} = \frac{\|f_{m_j}(x) - f_{m_L}(x)\|}{\|f_{m_i}(x) - f_{m_L}(x)\|}$, $j > i$ – КУ в ИППМ для любой пары $i, j \in [1, L-1]$. Очевидно, имеет место равенство

$$g_{ij}^L = g_{j-1}^L g_{j-2}^L ... g_i^L, \qquad (6)$$

Причем $g_{ij}^L << 1$, если $g_t^L << 1$, $i \leqslant t \leqslant j-1$.

Теперь необходимо сравнить значения g_{ij}^L и g_{ij}.

Лемма 4. *Пусть ИППМ g-устойчива. Тогда для чисел* g_{ij} *и* g_{ij}^L *справедливы оценки*

$$g_{ij}\frac{1-g_{jL}}{1+g_{iL}} \leqslant g_{ij}^{L} \leqslant g_{ij}\frac{1+g_{jL}}{1-g_{iL}}. \tag{7}$$

Если же $\frac{g_{iL}+g_{iL}}{1-g_{iL}} \leqslant g_0 << 1$, *то существует такое число* ξ_{ij}^{L}, *что* $|\xi_{ij}^{L}| \leqslant g_0$ *и*

$$g_{ij}^{L} = g_{ij}\left(1+\xi_{ij}^{L}\right). \tag{8}$$

Доказательство не приводится ввиду ограниченности объема статьи. □

Получим более простые приближенные двусторонние неравенства типа (7).

Лемма 5. *Пусть для мантисс выполнены условия:* $m_i < m_j < m_l$, $i,j,l \in [1, L-1]$, $m_l - m_j = m_j - m_i = \Delta m$ *и выполнено неравенство (7), в котором при* $L = l$ *значениями* g_{il} *можно пренебречь. Тогда имеют место оценки*

$$\frac{-1+\sqrt{1+4\omega g_{ij}^{l}}}{2\omega} \leqslant g_{ij} \leqslant \frac{1-\sqrt{1-4\omega g_{ij}^{l}}}{2\omega}, \tag{9}$$

где $\omega = \frac{\rho_i\rho_l}{\rho_j^2}$. *Если же* $4\omega g_{ij}^{l} << 1$, *то оценка (9) имеет вид*

$$g_{ij}^{l}\left(1-\omega g_{ij}^{l}\right) \leqslant g_{ij} \leqslant g_{ij}^{l}\left(1+\omega g_{ij}^{l}\right). \tag{10}$$

Доказательство не приводится ввиду ограничений на объем статьи. □

Рассмотрим теперь оценки погрешностей округлений решений $f_{m_i}(x)$.

Теорема 2. *Пусть решение* $f(x)$, $x \in R^n$, $f \in R^k$ *оценивается значением* $f_{m_i}(x)$, $i \in [1, L-1]$, *ИППМ g-устойчива,* $\|f_{m_i}(x)\| > \|\Delta_i\|$, $i \in [1, L]$. *Тогда имеет место оценка*

$$\frac{\sigma_1\Delta_{ij}}{\left\|f_{m_j}(x)\right\|} \leqslant \frac{\Delta_i}{\|f(x)\|} \leqslant \sigma_2\frac{\Delta_{ij}}{\left\|f_{m_j}(x)\right\|}, \tag{11}$$

где $j > i$, $j \in [i+1, L]$, $\alpha_j = \frac{\Delta_j}{\left\|f_{m_j}(x)\right\|}$; $\sigma_1 = \frac{1}{(1+g_{ij})(1+\alpha_i)}$, $\sigma_2 = \frac{1}{(1-g_{ij})(1-\alpha_j)}$ – *корректирующие множители погрешности решений. Машинное значение числа* $\frac{\Delta_{ij}}{\left\|f_{m_j}(x)\right\|}$ *имеет относительную погрешность* $\approx \frac{k+6}{2}b^{1-m_i}$, *которой можно пренебречь.*

Доказательство не приводится ввиду ограничений на объем статьи. □

3. Правила оценки погрешностей округления решений задач ВМ. Конечношаговый алгоритм (КША)

ИППМ можно рассматривать как метод решения задачи ВМ, позволяющий получать оценки погрешностей округления. Полезно следующее:

Определение 6. Пусть задано число $\varepsilon > 0$ – требуемая точность решения задачи ВМ, то есть если точность решения задачи достижима, то имеет место неравенство $\Delta \equiv \|f_m(x) - f(x)\| \leqslant \varepsilon$ или $\Delta/\|f(x)\| \leqslant \varepsilon$. Будем говорить, что имеет место гарантированная точность решения (ГТР) задачи ВМ, если задача решается на ЭВМ методом, для которого известны оценки погрешностей решения, гарантирующие достижение указанной точности и значения оценок погрешностей определяются вместе с искомым решением. □

В теореме 1 доказано, что в ИППМ существует такое решение $f_{m_i}(x)$, которое при определенных условиях обеспечивает ГТР задачи ВМ.

В КША решение задачи ВМ получается за конечное число базовых (стандартных) вычислительных операций [1]. Оценки погрешности округления (ОПО) решения задачи ВМ проводятся путем сравнения значений $f_{m_i}(x)$ и $f_{m_j}(x)$, полученных при длине мантисс $m = m_i$ и $m = m_j$. При этом возможны различные варианты построения ОПО.

3.1. Алгоритм последовательной оценки погрешности округлений решений. ИППМ решения задач ВМ считаем g-устойчивой. Задается некоторое начальное значение длины мантиссы $m = m_1$. Следующие значения длин мантиссы задаем по правилу

$m_{i+1} = m_i + \Delta m_{i+1}$, $i = 1, 2, ...$ Особенностью настоящего алгоритма является то, что ОПО задачи ВМ проводится после каждого решения с мантиссой большей длины.

Пусть $L = 1$, т.е. найдено только одно решение задачи ВМ при длине мантиссы $m = m_1$, не исключающей одинарную, двойную и четверную точности или точности при большей длине мантиссы. Возможны ситуации, когда решаемая задача по мнению Вычислителя проста и нет необходимости находить её решение при других значениях длины мантиссы. В рассматриваемом случае за оценку погрешности решения ответственна интуиция Вычислителя, но указать величину погрешности решения в общем случае не представляется возможным.

Содержательные оценки точности решений можно получить для g-устойчивой ИППМ при числе решений $L \geqslant 2$. При достаточно большом Δm ИППМ g-устойчива (лемма 1) и по теореме 1 $f_{m_j}(x)$, $j > i$, – К-решение по отношению к $f_{m_i}(x)$. Тогда по теореме 2 абсолютная Δ_i и относительная $\Delta_i / \|f(x)\|$ погрешности приближенного решения $f_{m_i}(x)$ имеют вид

$$\begin{gathered} \Delta_i \equiv \|f_{m_i}(x) - f(x)\| \leqslant \frac{\left\|f_{m_i}(x) - f_{m_j}(x)\right\|}{1 - g_{ij}} \equiv \frac{\Delta_{ij}}{1 - g_{ij}} \leqslant \sigma_A \Delta_{ij}, \\ \frac{\Delta_i}{\|f(x)\|} \leqslant \frac{\Delta_{ij}}{(1 - g_{ij})(1 - \alpha_j)\left\|f_{m_j}(x)\right\|} \leqslant \sigma_0 \frac{\Delta_{ij}}{\left\|f_{m_j}(x)\right\|}, \end{gathered} \tag{12}$$

где Δ_{ij} и $\frac{\Delta_{ij}}{\left\|f_{m_j}(x)\right\|}$ – ОПО, $\sigma_A = \frac{1}{1-g_0}$, $\sigma_0 = \frac{\sigma_A}{1-\alpha_0}$ –коэффициенты коррекции ОПО для абсолютной и относительной погрешностей соответственно; $g_{ij} \leqslant g_0 \leqslant 0,1 << 1$, $\alpha_j \leqslant \alpha_0 \leqslant g_0 \leqslant 0,1$. Более точно, $\alpha_j << g_{ij}$ и $\alpha_0 << g_0$.

При решении задач ВМ возможны различные схемы получения ОПО. Выделим две основные схемы. Пусть $f_{m_i}(x)$ и $f_{m_j}(x)$ – решение и К-решение задачи ВМ.
1. Пусть задана требуемая точность решения ε_A и для некоторых решений $f_{m_i}(x)$, $f_{m_j}(x)$ выполнено неравенство: $\sigma_A \Delta_{ij} \equiv \Delta^1 \leqslant \varepsilon_A$. Тогда полученная относительная погрешность решения ε^1 удовлетворяет условию:

$$\frac{\sigma_A \Delta_{ij}}{\left\|f_{m_j}(x)\right\|(1 - \alpha_0)} \equiv \varepsilon^1 \leqslant \frac{\varepsilon_A}{\left\|f_{m_j}(x)\right\|(1 - \alpha_0)} \equiv \varepsilon_0. \tag{13}$$

2. Пусть задана требуемая точность решения ε_0 для относительной погрешности, т.е. выполняется неравенство: $\frac{\sigma_A \Delta_{ij}}{\left\|f_{m_j}(x)\right\|(1-\alpha_0)} \leqslant \varepsilon_0$. Тогда для абсолютной погрешности решения Δ^1 выполняется условие:

$$\sigma_A \Delta_{ij} \equiv \Delta^1 \leqslant \varepsilon_0 \left\|f_{m_j}(x)\right\|(1 - \alpha_0) \equiv \varepsilon_A. \tag{14}$$

Ниже рассматриваются два варианта алгоритма ОПО, в которых используются относительные погрешности решений (2-я схема); достижение требуемой точности решений гарантируется теоремой 1.

Пусть числа $g_0 \leqslant 0,1$ и $\alpha_0 \leqslant g_0$. Задавая различными способами Δm_{i+1}, $i = 1, 2, ...$, будем иметь варианты алгоритма ОПО.

3.1.1. Вариант 1. Пусть ИППМ g-устойчива и $\sigma_0 = \frac{1}{(1-g_0)(1-\alpha_0)}$, $\Delta m_{i+1} = \Delta m = \text{const}$, $i = 1, 2, ...$, и для некоторого $i \geqslant 1$ выполнено условие $\sigma_0 \frac{\Delta_{i,i+1}}{\left\|f_{m_{i+1}}(x)\right\|} \equiv \varepsilon_i \leqslant \varepsilon_0$, а условия $\varepsilon_t \leqslant \varepsilon_0$ не выполнены для $1 \leqslant t \leqslant i - 1$. Решением задачи является значение $f_{m_i}(x)$, абсолютная погрешность его (14) не более ε_A, относительная – ε_0; в (14) определено значение Δ^1. Гипотезу об g-устойчивости ИППМ можно подтвердить оценками, которые следуют из лемм 4, 5.

3.1.2. Вариант 2. Пусть определены m_1, $m_2 = m_1 + \Delta m$, решения $f_{m_1}(x)$ и $f_{m_2}(x)$ и условие $\sigma_0 \frac{\Delta_{12}}{\|f_{m_2}(x)\|} \leqslant \varepsilon_0$ не выполнено, число $\rho_1 \approx \Delta_{12} b^{m_1}$. Оценим значение m_3 из условия

$$\Delta_3 = \rho_3 b^{-m_3} = \varepsilon_0 \|f_{m_2}(x)\|. \tag{15}$$

Считаем, что $\rho_3 = \chi\rho_1$, где χ можно брать равным $\chi = 10^2$, $\chi = 10^3$ и т.д. Из (15) получим $\chi\Delta_{12} b^{m_1} b^{-m_3} = \varepsilon_0 \|f_{m_2}(x)\|$; откуда $b^{m_3} = \frac{\chi\Delta_{12} b^{m_1}}{\varepsilon_0\|f_{m_2}(x)\|}$. Если $\log_b \frac{\chi\Delta_{12}}{\varepsilon_0\|f_{m_2}(x)\|} \leqslant 2\Delta m$, то $m_3 = m_2 + \Delta m$.

Если $\log_b \frac{\chi\Delta_{12}}{\varepsilon_0\|f_{m_2}(x)\|} > 2\Delta m$, то $m_3 = m_2 + \left\lfloor \log_b \frac{\chi\Delta_{12}}{\varepsilon_0\|f_{m_2}(x)\|} \right\rfloor$. Находим $f_{m_3}(x)$ и проверяем условие $\sigma_0 \frac{\Delta_{23}}{\|f_{m_3}(x)\|} \leqslant \varepsilon_0$. Если это условие выполнено, то $f_{m_2}(x)$ — решение задачи ВМ и значение $\Delta^1 = \sigma_0\Delta_{23}$. Если не выполнено, то далее ИППМ реализуется по Варианту 1 при $m_{i+1} = m_i + i\Delta m$, $i \geqslant 3$.

3.2. Табличный алгоритм оценки погрешностей округления решений. Пусть ИППМ решения задачи ВМ представлена в виде: $f_{m_i}(x)$, $i = [1, L]$, $m_{i+1} = m_1 + i\Delta m$, $i \in [1, L-1]$. При выполнении L решений $f_{m_i}(x)$, $i \in [1, L]$, становится возможным полученную информацию представить в виде таблиц. В частности, возможно построить таблицы значений Δ_{ij}, $\rho_i = \|\bar{C}_{m_i}\| \cong \Delta_{ij} b^{m_i}$, g_{ij}^L, ε_{ij} и т.д. Совокупность всех указанных таблиц дает информацию о величине погрешностей решений $f_{m_i}(x)$, $i \in [1, L-1]$, и о g-устойчивости ИППМ.

3.3. Округление решений задач ВМ. Часто бывает так, что решение $f_m(x)$ с требуемой точностью возможно представить при меньшем числе десятичных знаков. Такое представление реализуют различные процедуры округления чисел. Решением задачи ВМ в общем случае является вектор – набор чисел $f_{m_i} \in R^k$.

3.3.1. Рассмотрим сначала случай, когда $k = 1$, т.е. решением задачи является число. Далее, для упрощения записи обозначим $A = f(x)$ – точное решение, $a = f_{m_i}(x)$ – приближенное решение – m_i-значное число. Следуя [12], напомним известное понятие.

Определение 7. Говорят, что t первых значащих цифр (десятичных знаков) приближенного числа a являются **верными в узком смысле**, если абсолютная погрешность этого числа удовлетворяет условию

$$|A - a| \leqslant \frac{1}{2} 10^{e-t}; \tag{16}$$

являются **верными в широком смысле**, если выполнено условие

$$|A - a| \leqslant 1 \cdot 10^{e-t}, \tag{17}$$

где e – порядок числа. □

Округление решения проводится по известному правилу «по дополнению». Схема получения округленного решения следующая:

а) Пусть известна оценка решения $|A - a| \leqslant \Delta^1$ (см. пп. 3.1.1, 3.1.2).

б) Из решения неравенства $\Delta^1 \leqslant \frac{1}{2} 10^{e-t}$ определим число верных знаков решения:

$$t = e + \left\lfloor -\log_{10} 2\Delta^1 \right\rfloor \text{ и } e - t = -\left\lfloor -\log_{10} 2\Delta^1 \right\rfloor. \tag{18}$$

в) Округленное t-значное решение a_1 имеет t верных знаков в широком смысле и оценку погрешности

$$|A - a_1| \leqslant 1 \cdot 10^{-\lfloor -\log_{10} 2\Delta_1 \rfloor}. \tag{19}$$

Рассмотрим другой метод округления приближенного числа a, который, по нашему мнению, более гибок по сравнению с предыдущим. Пусть A – неизвестное точное конечно или бесконечнозначное число, a – его известное приближение с известной погрешностью $\Delta^1 : |A - a| \leqslant \Delta^1$, a_1 – другое приближение числа a, такое что $a = a_1 + \alpha$, где $|\alpha| < |a|$. Имеем очевидную оценку погрешности числа a_1:

$$|A - a_1| = |A - a + \alpha| \leqslant \Delta^1 + |\alpha| = \Delta^2. \tag{20}$$

Пусть мантисса числа a имеет m десятичных знаков. Представим числа a, a_1, α в виде

$$\begin{gathered} a = \pm\mu \cdot 10^e = \pm\left(\sum_{i=1}^{m} s_i \cdot 10^{-i}\right) \cdot 10^e = a_1 + \alpha, \\ a_1 = \pm\left(\sum_{i=1}^{t} s_i \cdot 10^{-i}\right) \cdot 10^e, \alpha = \pm\left(\sum_{i=t+1}^{m} s_i \cdot 10^{-i}\right) \cdot 10^e, \end{gathered} \tag{21}$$

где e – порядок числа, $1 \leqslant t \leqslant m$.

Определение 8. Разбиение (21) числа a на a_1 и α назовем **сечением числа по мантиссе**. В (21) t-значное число a_1 – округленное значение числа a, $m-t$-значное число, α – погрешность округления числа a_1, $|\alpha|$ – ошибка сечения числа a. Способ округления числа сечением по мантиссе назовем **методом отбрасывания** [17]. □

Для приближенного числа a оценка $\Delta^2 = \Delta^1 + |\alpha|$ состоит из численной m-значной оценки Δ^1, полученной в процессе решения задачи ВМ и $|\alpha|$ – ошибки сечения числа a. Число Δ^2 монотонно убывает при возрастании числа знаков сечения t. Числа a_1, α в (21) очевидно зависят от t, что можно представить как a_1^t, α^t. Значение числа t_0, гарантирующее достижение точности решения ε_A, рационально определять из следующего условия:

$$t_0 = \min\, t, \text{если} \Delta^2 = \Delta^1 + \left|\alpha^t\right| \leqslant \varepsilon_A \text{ и } 1 \leqslant t \leqslant m. \tag{22}$$

В свою очередь для удобства практического использования (экономичности записи) число $\Delta^2 = \Delta^1 + \left|\alpha^{t_0}\right|$ может быть округлено сверху и это значение $\overline{\Delta^2}$ не должно превышать значение ε_A – требуемой точности решения. □

3.3.2. В разделе 3.1 рассмотрены алгоритмы нахождения решения задачи ВМ, которое может быть как скаляром ($k = 1$), так и вектором ($k \geqslant 2$). Случай округления решения $f_{m_i}(x)$ при $k = 1$ рассмотрен в п. 3.3.1. Правила округления компонент решения $f^j_{m_i}(x)$, $j \in [1, k]$, $k \geqslant 2$ могут быть различными. Они зависят от тех требований, которые предъявляет к решению Вычислитель. Рассмотрим два варианта:

1. Задача ВМ решается в соответствии с алгоритмами пп. 3.1 и 3.2 и полученные значения решений $f_{m_i}(x)$ не округляются.

2. Задача ВМ решается в соответствии с алгоритмами п. 3.1 и, кроме выполненных требований точности, должны выполняться требования точности для компонент решения $f^j_{m_i}(x)$ после их округления по правилу «по дополнению» или по методу сечения. Требования точности для абсолютной погрешности для компонент ε_j, $j \in [1, k]$ должны удовлетворять условию $\sum_{j=1}^{k} \varepsilon_j^2 \leqslant \varepsilon_A^2$, где ε_A – требуемая точность решения из п. 3.1. Более подробного изложения этого правила, а также других возможных вариантов в рамках этой статьи приводить не будем.

4. Оценка погрешности округления по совпадению первых десятичных знаков (СПЗ) решений с различной длиной мантиссы

В работе [4] автор предлагает способ достижения требуемой точности решения (перевод наш): «В тех случаях, когда основным источником погрешностей является округление, общий подход к оценке точности вычисления таков: пересчитать результат с помощью более точной арифметики и сравнить количество совпавших знаков в первом и втором случае. Интуитивно мы предполагаем, что требуемая точность результата может быть достигнута при вычислениях с достаточно точной арифметикой». Обоснование предлагаемого способа оценки точности решения в работе [4] не приводится, идея высказана на уровне интуитивного предположения. Приведем описание способа оценки точности решения, основанного на учете числа первых совпадающих десятичных знаков решений $f_{m_i}(x)$ и $f_{m_j}(x)$.

Введем соответствующее

Определение 9. Пусть дано l значений функции $\varphi_j(x)$, $j \in [1,l]$, $l \geqslant 2$ одинакового порядка, представление которых в десятичной системе имеет вид: $\varphi_j(x) = \pm\left(\sum_{i=1}^{r} s_i^j \cdot 10^{-i}\right)\cdot 10^e$, $j \in [1,l]$, r –натуральное число или $r = \infty$. Будем говорить, что у l функций одинакового знака **совпадают t первых** десятичных **знаков** (СПЗ), если $s_i^1 = s_i^2 = ... = s_i^l$, $i \in [1,t]$, $t \geqslant 1$, $l \geqslant 2$. □

Пусть $f(x) \in R^1$, т.е. f – число, и получены два решения $f_{m_i}(x)$ и $f_{m_j}(x)$ при длинах мантисс m_i, m_j, $m_i < m_j$. Используя (2) при $b = 10$, запишем:

$$f_{m_i}(x) = f(x) + \bar{C}_{m_i}10^{-m_i}, f_{m_j}(x) = f(x) + \bar{C}_{m_j}10^{-m_j}. \tag{23}$$

В общем случае числа $f(x), C_{m_i}10^{-m_i}, C_{m_j}10^{-m_j}$ – бесконечнозначны, а $f_{m_i}(x)$, $f_{m_j}(x)$, $f^t(x)$, $h^t(x)$ – конечнозначны. Представим их в виде:

$$\begin{gathered}
f(x) = \pm\left(\sum_{i=1}^{\infty} s_i 10^{-i}\right)10^e; f(x) = f^t(x) + h^t(x);\\
f^t(x) = \pm\left(\sum_{i=1}^{t} s_i\cdot 10^{-i}\right)\cdot 10^e; h^t(x) = \pm\left(\sum_{i=t+1}^{\infty} s_i\cdot 10^{-i}\right)\cdot 10^e;\\
\overline{\Delta}_i = \bar{C}_{m_i}10^{-m_i} = \pm\left(\sum_{\alpha=1}^{\infty} r_\alpha 10^{-\alpha}\right)10^{e-t_i};\\
\overline{\Delta}_j = \bar{C}_{m_j}10^{-m_j} = \pm\left(\sum_{i=1}^{\infty} \xi_i 10^{-i}\right)10^{e-t_j};\\
f_{m_i}(x) = \pm\left(\sum_{\alpha=1}^{m_i} \eta_\alpha 10^{-\alpha}\right)10^e; f_{m_j}(x) = \pm\left(\sum_{i=1}^{m_j} \psi_i 10^{-i}\right)10^e,
\end{gathered} \tag{24}$$

где $-t_i$ и $-t_j$ – порядки погрешностей чисел $f_{m_i}(x)$ и $f_{m_j}(x)$. Знаки + или – у чисел f, f_{m_i}, f_{m_j} одинаковы. Знаки у чисел $\overline{\Delta}_i$, $\overline{\Delta}_j$ могут быть различными. Порядки погрешностей решений $f_{m_i}(x)$ и $f_{m_j}(x)$ при g-устойчивой ИППМ удовлетворяют условию: $t_j \geqslant t_i \geqslant 1$. Очевидно, что число первых совпадающих десятичных знаков может изменяться от 0 до t_i. Например, у чисел $f_{m_i}(x) = 0,4001111$ и $f_{m_j}(x) = 0,3999989$, $\Delta_i = 0,1111\cdot 10^{-3}$, $\Delta_j = -0,11\cdot 10^{-5}$, $f(x) = 0,4$, нет ни одного совпадающего десятичного знака. Однако на практике при g-устойчивой ИППМ **совпадение первых десятичных знаков встречается часто** и этот вариант полезно использовать при оценке погрешностей решений. Итак, пусть у решений $f_{m_i}(x)$ и $f_{m_j}(x)$ совпало t первых десятичных знаков. Тогда в качестве решения берется число $a_1 = f^t(x) = \pm\left(\sum_{\alpha=1}^{t}\eta_\alpha 10^{-\alpha}\right)10^e$. Погрешность числа $f^t(x)$ оценивается по методу сечений, рассмотренном в п. 3.3.1 (теорема 3).

Совпадение t первых десятичных знаков у $f_{m_i}(x)$ и $f_{m_j}(x)$ ещё не означает, что совпадают t первых знаков у решений $f(x)$ и $f_{m_i}(x)$, $f(x)$ и $f_{m_j}(x)$. Например, пусть $f(x) = 0,4$, тогда у решений $f_{m_i}(x) = 0,3999888$ и $f_{m_j}(x) = 0,3999998$ совпадают 4 первых знака и нет ни одного совпадающего знака с решением $f(x)$.

Рассмотрим условия, при которых у решений $f(x)$, $f_{m_i}(x)$, $i \leqslant 1 < L$ совпадают t первых десятичных знака, $1 \leqslant t < m_i$.

Теорема 4. *Пусть $h_m^t \equiv h_m^t(x) = f_m(x) - f^t(x)$, h_m^t – $(m-t)$-значное число, $\left|h_m^t\right| < 1\cdot 10^{e-t}$, $\left|\bar{\Delta}\right| < 1\cdot 10^{e-t}$, e – порядок числа. Для того чтобы решения $f(x)$ и $f_m(x)$ имели t СПЗ, необходимо и достаточно, чтобы*

$$0 \leqslant \left|h_m^t - \bar{\Delta}\right| < 1\cdot 10^{e-t}, \tag{25}$$

причем если h_m^t и $\bar{\Delta}$ имеют разные знаки, то должно выполняться неравенство $\left|h_m^t - \bar{\Delta}\right| < 1\cdot 10^{e-t}$, если одинаковые, то $\left|h_m^t\right| \geqslant \left|\bar{\Delta}\right|$.

Погрешность решения $f^t(x)$ удовлетворяет условию $\left|f^t(x) - f(x)\right| < 1 \cdot 10^{e-t}$.

Доказательство. Представим решение $f(x)$ в виде

$$f(x) = f^t(x) + h^t(x) \equiv f^t + h^t = f_m(x) - \bar{\Delta} \equiv f^t + h_m^t - \bar{\Delta}. \tag{26}$$

Из (26) следует, что $h^t = h_m^t - \bar{\Delta}$. Так как $0 \leqslant \left|h^t\right| < 1 \cdot 10^{e-t}$, то условие (25), как следует из (26), эквивалентно тому, что функции $f(x)$ и $f_m(x)$ имеют t СПЗ. Из неравенств $\left|h_m^t\right| < 10^{e-t}$ и $\left|\bar{\Delta}\right| < 10^{e-t}$ следует, что условию $\left|h_m^t - \bar{\Delta}\right| < 1 \cdot 10^{e-t}$ соответствует случай, когда h_m^t и $-\bar{\Delta}$ имеют одинаковые знаки (т.е. h_m^t и $\bar{\Delta}$ имеют разные знаки), а условию $0 \leqslant \left|h_m^t - \bar{\Delta}\right|$, при одинаковых знаках чисел h_m^t и $\bar{\Delta}$, эквивалентно неравенство $\left|h_m^t\right| \geqslant \left|\bar{\Delta}\right|$. Неравенство $\left|f(x) - f^t(x)\right| < 1 \cdot 10^{e-t}$ следует из (25) и (26). □

Следствие 1. Пусть у функций $f(x)$ и $f_m(x)$ найдено t СПЗ. Тогда для любого s, $1 \leqslant s \leqslant t$ имеет место неравенство $|f(x) - f^s(x)| < 1 \cdot 10^{e-s}$. □

Для практического применения неравенств $\left|h_m^t - \bar{\Delta}\right| < 1 \cdot 10^{e-t}$ и $\left|h_m^t\right| \geqslant \left|\bar{\Delta}\right|$ необходимо знать оценки погрешностей $\bar{\Delta}$, т.к. значение $\bar{\Delta}$ в общем случае неизвестно.

Но оценку погрешности $\bar{\Delta}_i$ некоторого решения $f_{m_i}(x)$ возможно получить, только зная К-решение $f_{m_j}(x)$, $j > i$. Введем обозначения: $f_{m_i}(x) = f^t(x) + h_i^t(x) \equiv f^t + h_i^t$, $f_{m_i}(x) = f(x) + \bar{\Delta}_i$, $f_{m_j}(x) = f^t(x) + h_j^t(x) \equiv f^t + h_j^t$, $f_{m_j}(x) = f(x) + \bar{\Delta}_j$, $\bar{\Delta}_{ij} = f_{m_i}(x) - f_{m_j}(x) = \bar{\Delta}_i - \bar{\Delta}_j$. Так как ИППМ g-устойчива, то $f_{m_j}(x)$ – К-решение и выполнены равенства $\bar{\Delta}_i = \bar{\Delta}_{ij} + \bar{\Delta}_j = \bar{\Delta}_{ij}(1 + \xi_{ij})$ (1), $\bar{\Delta}_j = \xi_{ij}\bar{\Delta}_{ij}$, $|\xi_{ij}| \leqslant \frac{g_{ij}}{1-g_{ij}} \leqslant g_0 \leqslant 0,1$ при $g_{ij} \leqslant 1/11$ (теорема 1). Представим далее: пусть $1 \leqslant t \leqslant t_i$ и

$$h^t = \nu^t 10^{e-t}; h_i^t = \nu_i^t 10^{e-t}; h_j^t = \nu_j^t 10^{e-t};$$

$$\bar{\Delta}_{ij} = \pm\mu_{ij} 10^{e-t_i} = \nu_{ij} 10^{e-t}; j > i, t \leqslant t_i;$$

$\left|\nu^t\right| < 1$, $\left|\nu_i^t\right| \leqslant 1 - 10^{-m_i}$, $\left|\nu_j^t\right| \leqslant 1 - 10^{-m_j}$, $|\nu_{ij}| = \mu_{ij} 10^{t-t_i}$, μ_{ij} – мантисса числа $|\Delta_{ij}|$, округленная до m_i знаков, т.е. $\mu_{ij} \leqslant 1 - 10^{-m_i}$. Теперь критерий для t СПЗ (теорема 4) можно переформулировать в новых обозначениях. □

Следствие 2. Для того чтобы решения $f(x)$, $f_{m_i}(x)$, $f_{m_j}(x)$ имели t СПЗ, необходимо и достаточно, чтобы

$$0 \leqslant \left|\nu_i^t - (1 + \xi_{ij})\nu_{ij}\right| \equiv \left|\nu_j^t - \xi_{ij}\nu_{ij}\right| < 1 - 10^{-m_i}. \tag{27}$$

Имеет место тождество: $\nu_i^t - (1 + \xi_{ij})\nu_{ij} \equiv \nu_j^t - \xi_{ij}\nu_{ij}$. □

Учитывая неравенство $|\xi_{ij}| \leqslant g_0$, из (27) получим достаточные условия t СПЗ, которые можно использовать на практике:

$$\left|\nu_i^t\right| + (1 + g_0)\mu_{ij} 10^{t-t_i} < 1 - 10^{-m_i} \text{или} \left|\nu_j^t\right| + g_0\mu_{ij} 10^{t-t_i} < 1 - 10^{-m_i}, \tag{28}$$

при ν_i^t и ν_{ij}; ν_j^t и ν_{ij} имеющих разные знаки и

$$\left|\nu_i^t\right| \geqslant (1 + g_0)\mu_{ij} 10^{t-t_i} \text{или} \left|\nu_j^t\right| \geqslant g_0\mu_{ij} 10^{t-t_i}, \tag{29}$$

при ν_i^t и ν_{ij}, ν_j^t и ν_{ij}, имеющих одинаковые знаки. Для определения знака погрешности $\bar{\Delta}_j = \xi_{ij}\bar{\Delta}_{ij}$ необходимо найти решение $f_{m_l}(x)$, $l > j$. Если решение $f_{m_l}(x)$ не найдено, то достаточно проверить условия (28) и (29) при некотором значении g_0 только для решения $f_{m_i}(x)$.

5. Правило оценки погрешности решений задач ВМ. Бесконечношаговый алгоритм

В работе [1] введено понятие сходящегося и нормального бесконечношагового алгоритма решения задач ВМ (БША) и доказано, что решение задачи $f_m^N(x)$ после выполнения некоторых N базовых вычислительных операций представляется в виде

$$f_m^N(x) = f(x) + \tilde{C}\delta_1 + \gamma, \quad \tilde{C}\delta_1 = f_m^N(x) - f^N(x), \quad \gamma = f^N(x) - f(x), \tag{30}$$

где $f^N(x)$ – точное значение решения после выполнения N операций, m – длина мантиссы МЧ, $\delta_1 = \frac{1}{2}b^{1-m}$. По своей структуре БША являются итерационными алгоритмами, т.е. очередное приближение решения определяется после выполнения дополнительных N_s базовых вычислительных операций, где s – номер итерации, $N = \sum_{s=1}^{L} N_s$, L – число решений (итераций) ИППМ, N – число базовых (стандартных) вычислительных операций.

Теорема 5. Пусть БША является сходящимся и нормальным, для каждого N выполнены условия теоремы 1 из [1], погрешность округления в (30) имеет порядок $\alpha : \tilde{C}\delta_1 = \bar{C}b^{-\alpha m}$, $0 < \alpha \leqslant 1$ и выполнено условие $\left\|\bar{C}\right\| \leqslant C \; \forall x \in G, G \subset R^n, \forall m \geqslant m_{\min}$, где $m_{\min}$ – минимальная длина мантиссы, при которой могут проводиться вычисления. Тогда существуют такой номер N – число базовых операций — и такая длина мантиссы m, при которых достигается требуемая точность решения ε, т.е. $\left\|f_m^N(x) - f(x)\right\| \leqslant \varepsilon$.

Доказательство

Из уравнения (30) имеем оценки:

$$\left\|f_m^N(x) - f(x)\right\| = \left\|\gamma + \tilde{C}\delta_1\right\| = \left\|\gamma + \bar{C}b^{-\alpha m}\right\| \leqslant \|\gamma\| + \left\|\bar{C}\right\| b^{-\alpha m} \leqslant \|\gamma\| + Cb^{-\alpha m},$$

где $\bar{C} = \frac{\tilde{C}}{2}b^{1+m(\alpha-1)}$. Так как БША сходящийся, то для любого $\varepsilon_1 > 0$ существует такой N, что $\|\gamma\| = \left\|f^N(x) - f(x)\right\| \leqslant \varepsilon_1$. Константа C не зависит от m, поэтому для любого $\varepsilon_2 > 0$ существует такая m, что $Cb^{-\alpha m} \leqslant \varepsilon_2$. Последнее неравенство будет верно при $m \geqslant \left\lfloor 1 - \frac{1}{\alpha}\log_b \frac{\varepsilon_2}{C} \right\rfloor$. Выберем $\varepsilon_1 = \alpha_1\varepsilon$, $\varepsilon_2 = \alpha_2\varepsilon$, где $0 < \alpha_1 \leqslant \frac{1}{2}$, $0 < \alpha_2 \leqslant \frac{1}{2}$. Для погрешности имеем оценку $\left\|f_m^N(x) - f(x)\right\| \leqslant \|\gamma\| + Cb^{-\alpha m} \leqslant \varepsilon_1 + \varepsilon_2 = \alpha_1\varepsilon + \alpha_2\varepsilon \leqslant \varepsilon$. □

Замечание 1. Существенными условиями для достижимости требуемой точности решений в БША являются его сходимость в точной арифметике и ограниченность нормы параметра погрешности: $\left\|\bar{C}\right\| \leqslant C, \forall m \geqslant m_{\min}$ аналогично требованию $\left\|\bar{C}\right\| \leqslant C, \forall m \geqslant m_{\min}$ для КША в [1] (определение 9, теорема 3). Именно ограниченность $\left\|\bar{C}\right\|$ позволяет получить в БША и КША требуемую точность решения задачи ВМ.

2. Как показано в теореме 1, требуемое значение точности решения в КША достижимо, когда ИППМ g-устойчива. Очевидно, в БША требуемое значение точности решения будет также достижимо, если он будет g-устойчив, т.е. $g_i = \frac{\Delta_{i+1}}{\Delta_i} = \frac{\left\|f_{m_{i+1}}^{i+1}(x) - f(x)\right\|}{\left\|f_{m_i}^{i}(x) - f(x)\right\|} = \frac{\left\|\gamma_{i+1} + \bar{C}_{i+1}b^{-m_{i+1}}\right\|}{\left\|\gamma_i + \bar{C}_i b^{-m_i}\right\|}$, $g_i << 1$, $m_{i+1} > m_i$, где $f_{m_i}^i(x) \equiv f_{m_i}^{N(i)}(x)$, $N(i) = \sum_{s=1}^{i} N_s$ – число базовых операций, выполненных за i шагов (итераций) ИППМ. Очевидно, для g-устойчивости БША достаточно, чтобы $\beta_i = \frac{\|\gamma_{i+1}\|}{\|\Delta_i\|}$ и $q_i = \frac{\left\|\bar{C}_{i+1}\right\|}{\left\|\bar{\Delta}_i\right\|}b^{-(m_{i+1}-m_i)}$ были достаточно малы, т.е. $\beta_i << 1$ и $q_i << 1$.

Таким образом, имеет место важный вывод: при g-устойчивости БША для них справедливы все результаты теории, сформулированные в разделах 2 и 3, а потому методика получения гарантированной точности решений для БША будет той же, что и для КША.

Представляет интерес изучение свойств погрешности $\Delta = \left\|\gamma + \bar{C}b^{-\alpha m}\right\|$ для конкретных классов задач ВМ, решаемых БША. Учет специфики погрешности Δ для отдельных классов задач позволит повысить эффективность решения задач ВМ. К этим задачам относятся: нахождение суммы числового ряда; нахождение численного значения производной k-го порядка; задача приближенного вычисления определенного интеграла; задача численного решения дифференциальных уравнений методом конечных разностей; численное решение систем нелинейных уравнений; численное решение экстремальных задач и т.д.

6. Об эффективности метода K-решений

Некоторый метод решения задачи ВМ назовем эффективным, если на данной ЭВМ решение задачи получено с заданной точностью за приемлемое время. Точность решения задается по-разному для разных классов решаемых задач и определяется некоторым признаком (условием) окончания решения задачи. Заданная точность решений часто отличается от гарантированной точности решений (ГТР), получаемой, например, в методах линейной алгебры [2]; методах решения задач, использующих интервальный анализ [10]; в ИППМ, использующей K-решения для оценки погрешностей округления.

ГТР – новое качество решений в отличие от многих методов решений, не обеспечивающих выполнение этого требования. Метод ИППМ обладает значительной универсальностью в решении задач ВМ, т.к. он не ориентирован на какие-либо классы решаемых задач. Таким образом, если метод ИППМ решает задачу за приемлемое время, а традиционный метод (ТМ), использующий «стандартное» программное обеспечение, решает, но не дает ГТР, то метод ИППМ можно считать высокоэффективным по сравнению с ТМ.

В методе ИППМ используется программное обеспечение (ПО), реализующее стандарт машинной арифметики IEEE 745 [14–16]. При выходе за диапазон длин мантисс «стандартной» арифметики (одинарной, двойной, четверной точности), наблюдается скачок увеличения времени вычислений от 10 до 100 раз в зависимости от сложности задачи и количества операций в ней. К примеру, решение СЛУ размерности $K = 30$ в «стандартной» арифметике с двойной точностью ($m = 15$, $b = 10$) находится, в среднем, за $5 \cdot 10^{-4}$с, а решение той же системы при $m = 16$, $b = 10$, т.е. при выходе за пределы стандартной арифметики и «подключении» специального ПО, уже требует $3,3 \cdot 10^{-2}$ с. Рост времени решения при увеличении длины мантиссы – это плата за новое качество – ГТР.

7. Численный эксперимент

Решение системы линейных уравнений методом Гаусса

Рассмотрим задачу нахождения решения системы линейных уравнений (СЛУ):

$$Hz = c, \tag{31}$$

где H – матрица Гильберта порядка K, т.е.

$$H = \{h_{ij}\}, \; i, j \in [1, K], h_{ij} = \frac{1}{i + j - 1}. \tag{32}$$

Алгоритм решения задачи 1 конечношаговый (КША), где функции $f(x)$ соответствует $f \equiv z$, а аргументам x соответствует матрица H и вектор c, т.е. $x \equiv \{H, c\}$, $z \in R^k$, $x \in R^n$, где $n = k + \frac{k(k+1)}{2}$, т.к. матрица H симметричная.

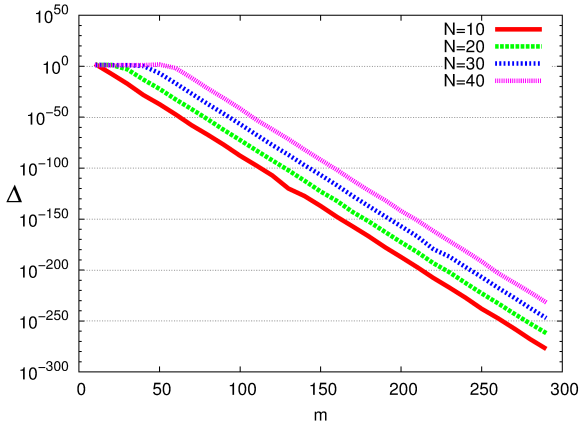
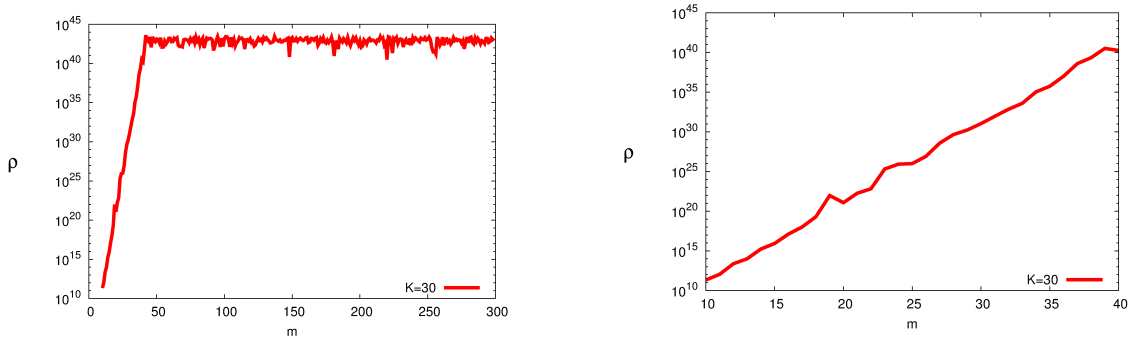


Рис. 1. Зависимость абсолютной погрешности решения Δ_m от длины мантиссы m

Рис. 1 представляет зависимость погрешности Δ_m решения СЛУ (31) от величины m при различных значениях K, $\Delta_m = \|z_m - z\|$, z — точное решение системы (31). Из данного графика видно, что, начиная с некоторого значения длины мантиссы m, погрешность решения монотонно уменьшается при достаточно большом локальном увеличении длины мантиссы Δm.



а) б)

Рис. 2. Зависимость ρ в области стабильности ($m > 45$) и области роста ($m < 45$) для СЛУ вида (31) от длины мантиссы m при $K = 30$ и шаге изменения длины мантиссы $\Delta m = 1$

Как указывалось выше [11], погрешность округления носит случайный характер. На графике (рис. 2) представлено значение параметра погрешности $\rho = \|z_m - z\| b^m$ при $b = 10$ и известном точном решении z. Из графиков видно, что зависимость параметра погрешности ρ от m при $m > 45$ содержит элемент «случайности».

Приближенные значения мантисс чисел g_{ij}^L для тех же значений $i, j \in [1, 10]$ и $m_L = 110$ отличались от значений мантисс g_{ij} в 10-м или 11-м знаках, поэтому таблица для g_{ij}^L не приводится.

Таблица 1

Значение g_{ij}^{j+1} для СЛУ вида (31) от длин мантиссы m_i, m_j при $K = 40$

	10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
10		1.33e+00	1.51e+00	3.09e+00	1.02e+00	2.18e-03	4.52e-13	1.99e-23	4.91e-33	4.55e-43
20			1.16e+00	2.47e+00	9.38e-01	1.63e-03	3.37e-13	1.49e-23	3.67e-33	3.40e-43
30				2.35e+00	9.73e-01	1.12e-03	2.32e-13	1.02e-23	2.52e-33	2.34e-43
40					1.19e+00	1.15e-03	2.38e-13	1.05e-23	2.59e-33	2.39e-43
50						3.23e-04	6.70e-14	2.95e-24	7.28e-34	6.75e-44
60							2.07e-10	9.14e-21	2.25e-30	2.09e-40
70								4.41e-11	1.09e-20	1.01e-30
80									2.47e-10	2.28e-20
90										9.26e-11

Таблица 1 представляет значения g_{ij}^{j+1}, которые близки со значениями g_{ij} при $m_i \geqslant 60$ (зона устойчивости) и отличаются от них при $m_i \leqslant 50$.

Таблица 2

Зависимость значений ε_i и ε_{ij} от длины мантиссы m_i, m_j при $k = 40$

	ε_i	**20**	**30**	**50**	**60**	**70**	**80**	**90**	**100**
10	1.10945	1.75e+00	1.78e+00	7.36e+00	1.11e+00	1.11e+00	1.11e+00	1.11e+00	1.11e+00
20	1.48567		2.32e+00	7.98e+00	1.49e+00	1.49e+00	1.49e+00	1.49e+00	1.49e+00
30	2.1614			7.69e+00	2.16e+00	2.16e+00	2.16e+00	2.16e+00	2.16e+00
40	2.10893			6.31e+00	2.11e+00	2.11e+00	2.11e+00	2.11e+00	2.11e+00
50	7.48466				7.48e+00	7.48e+00	7.48e+00	7.48e+00	7.48e+00
60	0.00241895					2.42e-03	2.42e-03	2.42e-03	2.42e-03
70	5.01304e-13						5.01e-13	5.01e-13	5.01e-13
80	2.21032e-23							2.21e-23	2.21e-23
90	5.45175e-33								5.45e-33

Таблица 2 представляет значения относительной погрешности решений $\varepsilon_i = \frac{\Delta_i}{\|f(x)\|}$, $i \in [1, 10]$ и $\varepsilon_{ij} = \frac{\Delta_{ij}}{\|f_j(x)\|}$, $j > i$. При $m_i \geqslant 60$ имеет место хорошее совпадение их значений.

Задача нахождения производной 1-го порядка

Рассмотрим метод численного дифференцирования первого порядка:

$$\varphi'(x) \cong \varphi'_m(x) = \frac{\varphi_m(x+h) - \varphi_m(x)}{h}, \quad \varphi(x) \in R^1. \tag{33}$$

Метод численного дифференцирования относится к классу бесконечношаговых алгоритмов (БША) в том смысле, что для нахождения $\varphi'(x)$ с требуемой точностью надо решить последовательность задач (33) для последовательности значений шага дифференцирования $h_i \to 0$, $i \to \infty$, а значение $\varphi'(x) = \lim\limits_{h_i \to 0} \frac{\varphi_{m_i}(x+h_i) - \varphi_{m_i}(x)}{h_i}$ – при одновременном увеличении длины мантиссы $m_i \to \infty$. Метод (33) превращается в конечношаговый при установлении зависимости между h и m. В [13] значение оптимального шага h рекомендуется брать пропорциональным $\sqrt{\mathrm{E}}$, где E – ошибка вычисления значения функции φ. Далее будем брать шаг $h = 2\sqrt{b^{-m}} = 2 \cdot 10^{-\frac{m}{2}}$. Задача (32) исследуется на примере функции

$$\varphi(x) = (\sin x)^{\cos x}, \quad x = {}^{2}/_{3}\,. \tag{34}$$

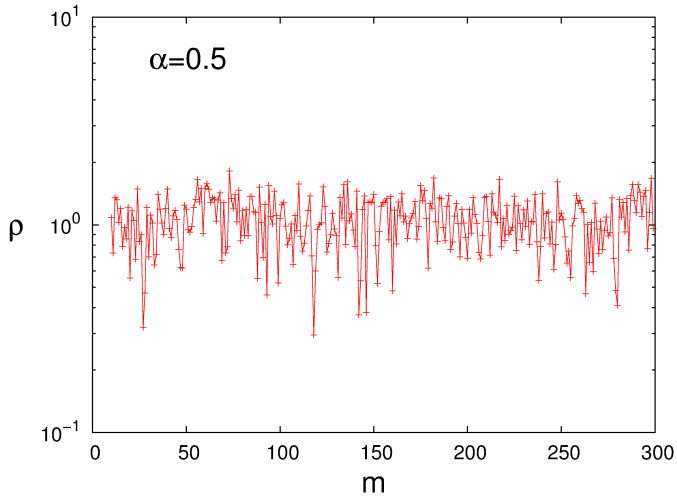


Рис. 3. Зависимость параметра $\rho(x)$ от m в предположении, что порядок погрешности $\alpha = 1/2$

Рис. 3 представляет зависимость параметра погрешности ρ от m для значения порядка погрешности $\alpha = 0,5$, т.е. $\rho = \Delta m 10^{0,5m}$. Из этого графика видно, что значение $\rho \leqslant 2$ при $10 \leqslant m \leqslant 300$.

Таблица 3

Значения решений φ'_{m_i}, $i \in [1, 10]$

m_i	$\varphi'_{m_i}(x)$
10	**8.882**8000000e-01
20	**8.8829085280**0000000000e-01
30	**8.882908528554890**000000000000000e-01
40	**8.8829085285548970217**5000000000000000000000e-01
50	**8.882908528554897021898675**0000000000000000000000000000e-01
60	**8.88290852855489702189867619213**500000000000000000000000000000000e-01
70	**8.8829085285548970218986761921499853**0000000000000000000. . . 0e-01
80	**8.882908528554897021898676192149985427833**5000000000000. . . 00000000000e-01
90	**8.88290852855489702189867619214998542783966863**450000000. . . 0000000000000000000000e-01
100	**8.88290852855489702189867619214998542783966863**552550000. . . 0000000000000000000000000000000e-01

Таблица 3 представляет значения решений $f_{m_i}(x) \equiv \varphi'_{m_i}(x)$ при $m_{i+1} = m_1 + i\Delta m$, $m_1 = 10$, $\Delta m = 10$, $i \in [1, 9]$. Как видно из таблицы, числа t_i СПЗ для решений $f_{m_i}(x)$ равны: $t_1 = 4$, $t_2 = 10$, $t_3 = 15$, $t_4 = 19$, $t_5 = 24$, $t_6 = 29$, $t_7 = 34$, $t_8 = 39$, $t_9 = 44$. Для оценки числа t_{10} необходимо вычислить значение $\varphi'_{m_{11}}(x)$. Значения t_i СПЗ выделены жирным шрифтом. Эта таблица иллюстрирует большую практическую важность правила СПЗ как

метода нахождения решения задачи ВМ (решение $f^{t_i}(x)$ указано в явном виде), так и его погрешности, равной $\Delta_i = \left\| f(x) - f^{t_i}(x) \right\| \leqslant 1 \cdot 10^{e-t_i}$.

8. Заключение

В настоящей работе предложен метод численного анализа погрешностей округления решения задач ВМ. Результаты, полученные в статье, сводятся к следующим положениям:

1) Введено понятие КР задачи, которое в оценках погрешностей решений с некоторой точностью заменяет истинное решение задачи $f(x)$. Исследованы свойства g-устойчивости КР, в том числе доказана теорема о достижимости решения с требуемой гарантированной точностью и получены оценки погрешности, далее численно реализуемые в ИППМ.

2) Предложены алгоритмы, позволяющие оптимизировать процесс решения задачи в ИППМ. Рассмотрены методы округления полученных решений, причем округленное решение имеет гарантированную точность.

3) Доказана теорема об оценках погрешности метода (правила) округления решения по совпадению t первых десятичных знаков (СПЗ); погрешность метода СПЗ не превышает значения $\varepsilon = 10^{e-t}$, где e – порядок числа.

4) Для бесконечношаговых алгоритмов (БША) решения задач ВМ доказана теорема о достижимости требуемой точности решения.

5) Предложенный метод КР оценки погрешностей округления обладает следующими свойствами:

   а) В g-устойчивой ИППМ обеспечивается ГТР задач ВМ.

   б) Метод КР обладает универсальностью в том смысле, что он не ориентирован на решение конкретных классов задач ВМ.

6) Приведены результаты численного эксперимента, иллюстрирующие основные свойства метода КР оценки погрешности решений.

Актуальность предложенного метода КР в первую очередь заключается в возможности получения ГТР для различных классов задач ВМ. Метод имеет перспективы развития в том смысле, что определение границ его применимости и численной эффективности для различных классов задач открывает новую область исследований в вычислительной математике.

Литература

1. *Бирюков А.Г., Гриневич А.И.* О гарантированной точности решений задач вычислительной математики в арифметике с плавающей запятой и переменной длиной мантиссы // Труды МФТИ. – 2012. – Т. 4, № 3. – С. 171–180.
2. *Годунов С.К., Антонов А.Г., Кирилюк О.П., Костин В.И.* Гарантированная точность решения систем линейных уравнений в евклидовых пространствах. – Новосибирск: Наука. Сиб. Отд-ние, 1988. – 456 с. ISBN 5-02-028593-5.
3. *Wilkinson J.H.* Rounding Errors in algebraic processes. – Englewood Cliffs, N.J.: Prentice-Hall, 1963. ISBN 0-486-67999-3.
4. *Higham N. J.* Accuracy and stability of numerical algorithms. – Philadelphia : Society for Industrial and Applied Mathematics, 1996.
5. *Воеводин В.В.* Вычислительные основы линейной алгебры. – М.: Наука, 1977. – 304 с.

6. *Henrici P.* Elements of Numerical Analysis. – New York. – John Wiley & Sons Inc., 1964.

7. *Clenshaw C. W. and Olver F. W. J.* Beyond floating point // J. Assoc. Comput. Mach. – 1984. – V 31. – P. 319–328.

8. *Langlois P.* A Revised Presentation of the CENA Method. – ARENAIRE – INRIA Grenoble Rhфne-Alpes / LIP Laboratoire de l'Informatique du Parallelisme.

9. *Шокин Ю. И.* Интервальный анализ. – Новосибирск: Сибирское отд. изд-ва «Наука», 1981.

10. *Алефельд Г., Херцбергер Ю.*. Введение в интервальные вычисления. – М.: Мир, 1987.

11. *Воеводин В.В.* Ошибки округления и устойчивость в прямых методах линейной алгебры. – М.: Изд-во МГУ, 1969. – 140 с.

12. *Демидович Б.П., Марон И.А.* Основы вычислительной математики. – СПб.: Лань, 2009. ISBN 978-5-8114-0695-1.

13. *Бахвалов Н.С., Жидков Н.П., Кобельков Г.М.* Численные методы. – М.: Наука, 1987.

14. IEEE 754-2008: 754-2008 IEEE Standard for Floating-Point Arithmetic. – ISBN: 978-0-7381-5753-5.

15. GNU GMP: Multiple precision arithmetic library / http://gmplib.org/

16. GNU MPFR, http://www.mpfr.org/

17. Математическая энциклопедия Т. 4 – М.: Советская энциклопедия, 1984.

Поступила в редакцию 13.01.2013.