УДК 537.322.2

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

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

О гарантированной точности решений задач вычислительной математики в арифметике с плавающей запятой и переменной длиной мантиссы

Статья посвящена вопросам анализа погрешностей округления решений задач вычислительной математики (ВМ) на ЭВМ в арифметике с плавающей запятой и переменной длиной мантиссы машинного числа (МЧ). Получены оценки погрешностей решения задач ВМ в зависимости от длины мантиссы МЧ и оценки длины мантиссы, гарантирующей достижение требуемой точности решения.

Ключевые слова: погрешность округления и точность решения задач вычислительной математики, машинное число с переменной длиной мантиссы, алгоритм типа Маркова вычисления значения функции, гарантированная точность решений задач ВМ.

1. Введение

В настоящее время в вычислительной практике при решении задач ВМ в арифметике с плавающей запятой преимущественно используются одинарная, двойная и четверная точности МЧ, что позволяет решать широкий круг практических задач. Однако существует множество задач, для решения которых четверной точности МЧ недостаточно по причине ошибок округления, и для их решения необходимо использование машинной арифметики с мантиссой числа большей длины. Классическим примером такой ситуации является задача решения систем линейных уравнений с плохо обусловленной матрицей. В настоящее время в бесплатном доступе получила распространение библиотека программ GNU GMP [4], реализующая стандарт IEEE 754 [1,9], в которой длина мантиссы в арифметике с плавающей запятой варьируется в широком диапазоне значений. Библиотека GNU GMP позволяет оперировать числами с длиной мантиссы от $m_{\min} = 24$ вплоть до $m_{\max} = 2^{31} = 2\,147\,483\,648$ двоичных знаков, чему соответствует 8 и 646 456 993 десятичных знаков. Верхнее значение длины мантиссы МЧ $m_{\max}$ невообразимо огромно. В указанной библиотеке также реализована возможность динамического изменения длины мантиссы m в различных сегментах программы от $m_{\min}$ до $m_{\max}$. Появление в свободном доступе программного обеспечения с такими возможностями расширяет границы для получения решений широкого круга задач ВМ с гарантированной точностью высокого порядка.

Проблема анализа влияния погрешностей округления (ВПО) 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 ВПО при решении задач линейной алгебры [2,3,7,11]; по исследованию ВПО в рамках интервального анализа [15,16,18]; по статистическому анализу ВПО [12,19]; исследованию новых моделей по выработке машинного числа [13]; алгоритмов с автоматической коррекцией ошибок округления первого порядка — метод CENA [14].

В настоящей работе предлагается новая схема анализа ВПО на решение задачи ВМ. Решение задач ВМ представлено как значение некоторой вектор-функции, определяемой методом решения задачи и его машинным алгоритмом. Последний представляется как алгоритм типа Маркова, алфавитом которого являются базовые (стандартные) математические функции библиотеки программ, на основе которой реализуется указанный алгоритм. Получена оценка решения задачи в зависимости от аргументов функции и длины мантиссы.

Для выделенного класса погрешностей получена оценка длины мантиссы МЧ, гарантирующей достижение требуемой точности решения.

Напомним определение машинного числа с плавающей запятой согласно стандарту IEEE 754 [1]:

Определение 1. *Машинным числом* будем называть число вида:

$$x_{m,p} = \pm 0, s_1 s_1 s_2 ... s_m \cdot b^e = \pm \left(\frac{s_1}{b^1} + \frac{s_2}{b^2} + ... + \frac{s_m}{b^m} \right) \cdot b^e = \pm \mu \cdot b^e,$$

где $b \in \{2, 3, ...\}$ — основание, μ — мантисса числа, m — количество знаков в мантиссе (число знаков, длина, размер), $s_i \in \{0, 1, ...b-1\}$, $i \in [1, m]$ — значащие цифры, $s_1 \neq 0$, $e \in [e_{\min}, e_{\max}]$ — порядок числа, p — размер порядка МЧ или количество знаков в представлении числа e. При заданных b, m, p машинные числа образуют конечное множество, обозначаемое далее $M_{b,m,p}$. Очевидно, $M_{b,m,p} \subset Q$, где Q — множество рациональных чисел. Для вещественного числа $x \in R$ ближайшее к нему машинное число $x_{m,p}$ есть *машинное представление* этого числа, т.е. число $x_{m,p}$ есть *результат округления* числа x. Число $\delta_\mu = b^{-m}$ назовем *погрешностью* (*точностью*) мантиссы или *точностью машинных чисел (машинной арифметики).* □

Машинное представление действительного числа x, полученное в результате его округления, имеет следующий вид [2,3]:

$$x_{m,p} = x\,(1 + u) + \nu, \tag{1}$$

где $u \cdot \nu = 0$.

При $\nu = 0$, $|u| \leqslant \delta_1$ и $x_{m,p} \neq 0$; при $u = 0$, $x_{m,p} = 0$ и $|x| \leqslant \delta_0$. Относительная погрешность представления числа x при $x_{m,p} \neq 0$ не более $\delta_1 = \frac{1}{2} b^{1-m}$, абсолютная погрешность $|x - x_{m,p}| \leqslant \delta_1 |x|$. Абсолютная погрешность при $x_{m,p} = 0$: $|x - x_{m,p}| = |x| \leqslant \delta_0$, $\delta_0 = 1 \cdot b^{e_{\min}} \approx b^{-b^p}$ — погрешность нуля, ближайшее к нулю число. На практике выбираются такие величины m и p, что можно всегда считать выполненным соотношение $\delta_0 << \delta_1$, а если учитывать что $p = 64$ [4], то $\delta_0 \cong 2^{-2^{64}}$ или $\delta_0 \cong 10^{-5{,}55 \cdot 10^{18}}$.

Число δ_1 также называют *машинным эпсилоном* или *погрешностью единицы.*

Учитывая, что величина p обычно подбирается так, чтобы перекрывать все необходимые значения порядков числа, точность машинной арифметики зависит прежде всего от длины мантиссы m, поэтому число $x_{m,p}$ можно считать не зависящим от p и представить как: $x_{m,p} \equiv x_m$. Введем определения.

Определение 2. Пусть $\varphi : R^n \to R^1$ — некоторая функция, $x \in R^n$ — вектор, x_m — его машинное покомпонентное представление. Тогда: $\varphi(x)$, $\varphi(x_m)$ — значения функции φ в точках x и x_m, $\varphi_m(x_m)$ — машинное представление значения $\varphi(x_m)$; $\varphi_m(x) \equiv \varphi_m(x_m)$. Значения функций $\varphi(x)$ и $\varphi(x_m)$ в общем случае представляются бесконечным числом знаков; в этом случае будем говорить, что их значения получены в *точной арифметике.* □

Определение 3. Математические функции и операции, реализуемые в библиотеках программ стандарта C99, назовем *базовыми или стандартными.* К базовым операциям относятся: округление чисел, арифметические операции, логические операции, операции вычисления математических функций, таких как $\sin x$, $\cos x$, $\tan x$, и их обратные, e^x, a^x, x^y, $\log_a x$ и т.д. □

Одной из библиотек, реализующих базовые функции стандарта C99, является GNU MPFR [4], в которой при заданной длине мантиссы m для значений базовых функций выполняется округление до последней значащей цифры, т.е.

$$\varphi_m(x_m) = \varphi(x_m)(1 + u), |u| \leqslant \delta_1,\ \varphi_m(x_m) \neq 0, \tag{2}$$

где $\varphi(x_m)$ и $\varphi_m(x_m)$ — значение одной из стандартных (базовых) функций, вычисленное в точке$x_m \in M_{b,m,p}$. Для случая, когда $\varphi_m(x_m) = 0$, $|\varphi(x_m) - \varphi_m(x_m)| = |\varphi(x_m)| \leqslant \delta_0$.

Для функций $\varphi(x)$, не являющихся базовыми, оценка (2), конечно, в общем случае не верна, и одной из целей настоящей работы является получение оценок погрешностей

для этих функций. Появление новых вычислительных возможностей, возникающих при варьировании m — длины мантиссы МЧ, позволяет взглянуть на многие известные задачи и методы их решения по-новому.

Во-первых, появляется возможность постановки задачи о достижении заданной точности решения (или вычисления с гарантированной точностью): для заданного $\varepsilon > 0$ найти решение задачи ВМ с погрешностью не более ε.

Во-вторых, появляется возможность рассмотреть вопрос о «реабилитации» численных методов решения задач ВМ, признанных в вычислительной практике неустойчивыми по причине влияния ошибок округления.

В-третьих, возникает также вопрос о влиянии повышения точности МЧ на эффективность применяемых на практике методов. Например, можно ли, увеличивая длину мантиссы МЧ, уменьшить общее время вычислений? И если да, то в каких случаях?

В настоящей работе дается вариант ответа на первый вопрос.

2. Задачи ВМ и алгоритмы их решения

Существует несколько классификаций задач вычислительной (прикладной) математики [8]. Общим для них является то, что процесс численного решения задачи ВМ на ЭВМ представляет собой упорядоченную последовательность конечного числа вычислительных базовых операций, которая определяется алгоритмом решения данной задачи. Этот процесс можно представить как вычисление значения некоторой функции (вектор-функции, отображения) $f \in R^k$ в точке $x \in G \subset R^n$, где G — область определения этой функции.

Определение 4. Задачей вычислительной математики будем называть совокупность понятий и условий $F(f, x, m, \varepsilon)$, определяющих возможность вычисления значений вектор-функции $f : R^n \to R^k$, где $f(x)$ — решение данной задачи в точке $x \in G$; ε — точность решения, m — длина мантиссы. Условие достижения точности означает: найти значение вектор-функции $f_m(x)$ при заданном $x \in G$ такое, что

$$\frac{\|f_m(x) - f(x)\|}{\|f(x)\|} \leqslant \varepsilon.$$

□

Возможны и другие определения задачи ВМ, но для целей настоящей работы определения 4 достаточно. Здесь и далее под нормой $\|\cdot\|$ понимается евклидова норма $\|z\| = \sqrt{\sum_{i=1}^{n} z_i^2}$, $z \in R^n$. Очевидно, что к вычислению значений вектор-функции сводятся, например, следующие задачи: решение системы линейных уравнений, вычисление производных функции разных порядков, задача интегрирования функции, задача Коши для дифференциальных уравнений, задачи поиска экстремума функции $\varphi(z)$, $z \in G \subset R^n$, и т.д.

Для задач ВМ разработаны численные методы их решения, которые в форме, адаптированной к ЭВМ, представляют собой алгоритмы решения этих задач. По своей природе алгоритмы бывают конечношаговыми (КША) и бесконечношаговыми (БША). К первым, например, относится метод Гаусса решения систем линейных уравнений, ко вторым — метод Ньютона решения систем нелинейных уравнений. Алгоритмы КША и БША вычисления значений функции $f(x)$ в точной арифметике представим в следующем виде:

КША:

$$ALG(f(x)) = \varphi^1(a_1) \vee \varphi^2(a_2) \vee \cdots \vee \varphi^{N-1}(a_{N-1}) \vee \varphi^N(a_N), \tag{3}$$

БША:

$$ALG(f(x)) = \varphi^1(a_1) \vee \varphi^2(a_2) \vee \cdots \vee \varphi^{N-1}(a_{N-1}) \vee \varphi^N(a_N) \vee \cdots N \to \infty, \tag{4}$$

где символ $\vee$ означает объединение операций, $\varphi^i(a_i)$, $i \in [1, N]$ — базовые операции, $a_i \in R^s$ — аргумент i-й операции. Операции округления в (3) и (4) отсутствуют, а $ALG(f(x))$

представляет собой упорядоченную последовательность базовых операций, определенную алгоритмом решения задачи ВМ. Результатом реализации базовой операции (кроме логической) является действительное число. В (3) N — число шагов КША.

Пусть требуется вычислить значение функции $f(x), x \in R^n$, $f \in R^k$ при длине мантиссы m. Тогда некоторый КША вычисления значения $f_m(x)$ представляет собой упорядоченную последовательность N базовых операций $\varphi_m^i(a_i)$, $i \in [1, N]$, т.е.

$$ALG(f_m(x)) = \varphi_m^1(a_1) \vee \varphi_m^2(a_2) \vee \cdots \vee \varphi_m^{N-1}(a_{N-1}) \vee \varphi_m^N(a_N). \tag{5}$$

В отличие от (3), (4) в (5) присутствуют операции округления чисел.

БША для вычисления значения $f_m(x)$ имеет вид

$$ALG(f_m(x)) = \varphi_m^1(a_1) \vee \varphi_m^2(a_2) \vee \cdots \vee \varphi_m^{N_0-1}(a_{N_0-1}) \vee \varphi_m^{N_0}(a_{N_0}), \tag{6}$$

где аналогично (5) N_0 — число шагов алгоритма, которое в (6) определяется по некоторому правилу его окончания. Однако при решении конкретной задачи ВМ для вычисления значений $f_m(x)$ могут существовать различные алгоритмы. Введем следующее определение.

Определение 5. Алгоритм вычисления функции $f \in R^k$ (5) в точке $x \in R^n$ при длине мантиссы m будет называться *нормальным алгоритмом* для решения задач *вычислительной математики* (НАВМ), если вычисленное значение функции в точной арифметике по алгоритму (5), в котором логические операции не выполняются, дает точное значение функции $f(x)$, т.е. имеет место алгоритм

$$ALG(f(x)) = \varphi^1(a_1) \vee \varphi^2(a_2) \vee \cdots \vee \varphi^{N-1}(a_{N-1}) \vee \varphi^N(a_N). \tag{7}$$

□

Замечание. Будем считать, что при построении НАВМ использовались логические операции как из математической библиотеки, так и операции языка программирования, на котором реализуется алгоритм решения задачи ВМ. Структура $ALG\,(f_m\,(x))$ (5), т.е. упорядоченная последовательность выполнения базовых операций, определяется базовыми логическими операциями и для данного m становится фиксированной. При выполнении $ALG\,(f\,(x))$ базовые логические операции в (7) $\varphi^i\,(a_i) = \emptyset$, т.е. не выполняются, а остальные операции в точной арифметике выполняются в соответствии со структурой алгоритма (5). Отметим также, что в (7) операции округления из (5) представляют собой единичный оператор.

Аргументы $a_i \in R^s$ в (5) — это либо числа (векторы), являющиеся аргументами функции $f\,(x)$, либо числа a_j^t, $t \in [1, s]$ — результаты выполнения операции $\varphi_m^j(a_j)$, $j < i$, $j \in [1, N]$ на j-й итерации, полученные в процессе вычислений в (5).

Введенное понятие нормального алгоритма терминологически напоминает понятие *нормальный алгоритм Маркова* (НАМ) [6]. НАМ содержит понятие *алфавит*, на основании которого строятся дальнейшие выводы грамматики. Алфавит может быть произвольным. В случае НАВМ алфавитом можно считать базовые функции библиотеки программ. На этом основывается аналогия между НАВМ и НАМ. Таким образом, его можно считать сужением НАМ на класс задач вычислительной математики.

3. Погрешности решений задач ВМ

В настоящем разделе изучаются погрешности решений задач ВМ, возникающие в итерационном процессе НАВМ.

Лемма 1. Пусть φ_i — базовые вычислительные операции (кроме логических) из некоторой библиотеки программ, $i \in [1, N_1]$ — номер базовой операции. Тогда значение $\varphi_m^i(a_i)$ можно представить в виде

$$\varphi_m^i(a_i) = \varphi^i(a_i) + \alpha_i\delta_1, \tag{8}$$

где N_1 — число базовых операций библиотеки, $|\alpha_i| \leqslant |\varphi^i(a_i)|$, $\delta_1 = \frac{1}{2}b^{1-m}$, $a_i \in R^s$, $a_i = a_{m,i}$ — вектор машинных чисел.

Доказательство.

Сначала рассмотрим операции по вычислению значений базовых функций, для которых справедливо равенство (2): $\varphi_m^i(a_i) = \varphi^i(a_i)(1+u_i)$, где $|u_i| \leqslant \delta_1$, $\delta_1 = \frac{1}{2}b^{1-m}$. Преобразуем его, и получим равенство (8):

$$\varphi_m^i(a_i) = \varphi^i(a_i) + \varphi^i(a_i)\cdot u_i = \varphi^i(a_i) + \varphi^i(a_i)\frac{u_i}{\delta_1}\delta_1 = \varphi^i(a_i) + \alpha_i\delta_1, \text{ где } \alpha_i = \frac{u_i}{\delta_1}\varphi^i(a_i),$$
$$\frac{u_i}{\delta_1} \in [-1,1] \text{ и } |\alpha_i| \leqslant |\varphi^i(a_i)|.$$

Для операции округления числа справедливо:

$$\varphi_m^i(a_i) = \varphi^i(a_i)(1+u_i) + v_i, \text{ где } u_iv_i = 0,\ |u_i| \leqslant \delta_1,\ |v_i| \leqslant \delta_0,\ \delta_0 = b^{-e_{\min}(p)}.$$

В случае, когда $v_i = 0$, имеем $\varphi^i(a_i) \equiv a_i$, т.к. φ^i — единичный оператор, $\varphi_m^i(a_i) = a_{m_i} = a_i(1+u_i) = a_i + \alpha_i\delta_1$, где $\alpha_i = \frac{u_i}{\delta_1}a_i$, $\frac{u_i}{\delta_1} \in [-1,1]$, т.е. справедливо (8). В случае, когда $u_i = 0$, $\varphi_m^i(a_i) = 0$ и $\varphi_m^i(a_i) = \varphi^i(a_i) + v_i = a_i + \nu_i$, где $|v| \leqslant \delta_0$, $\delta_0 = b^{e_{\min}(p)}$. Здесь значение α_i из (8) равно $\alpha_i = \frac{\nu_i}{\delta_1}$, где $|\alpha_i| \leqslant \frac{\delta_0}{\delta_1}$ — очень малое число. Для одинарной точности $\frac{\delta_0}{\delta_1} = 2\cdot 10^{-29}$, для двойной — $\frac{\delta_0}{\delta_1} = 2\cdot 10^{-308}$ [1]. Таким образом (8) также справедливо и для операции округления. □

Рассмотрим один частный, но важный случай оценивания погрешности операции деления МЧ на разность двух близких машинных чисел.

Лемма 2. Пусть для чисел d, y, z и их приближенных значений d_m, y_m, z_m выполнены условия $\Delta d = d_m - d = \xi d$, $\Delta y = y_m - y = \alpha y$, $\Delta z = z_m - z = \beta z$; $y_m = \mu_y b^t$, $z_m = \mu_z b^t$, μ_y и μ_z — мантиссы чисел y_m и z_m, t — порядок числа; $\mu_y - \mu_z = \eta_{m-q}b^{-q}$, $1 \leqslant q < m$, η_{m-q} — мантисса числа $\mu_y - \mu_z$, где $m-q$ её длина; $\xi = \tilde{\xi}\delta_1$, $\alpha = \tilde{\alpha}\delta_1$, $\beta = \tilde{\beta}\delta_1$. Тогда погрешность $\Delta = \frac{d_m}{y_m - z_m} - \frac{d}{y-z} = b^q\chi\delta_1$, где $\chi = \frac{d}{\eta_{m-q}b^t}\left(\tilde{\xi} - \frac{\tilde{\alpha}y - \tilde{\beta}z}{y-z}\right)$.

Доказательство.

Преобразуем значение погрешности Δ:

$$\Delta = \frac{d+\Delta d}{y-z+\Delta y-\Delta z} - \frac{d}{y-z} = \frac{\Delta d\,(y-z) - d\,(\Delta y - \Delta z)}{(y_m - z_m)\,(y-z)} =$$
$$= \frac{1}{y_m - z_m}\left(\xi d - \frac{d\,(\alpha y - \beta z)}{y-z}\right) = \frac{b^q d}{\eta_{m-q}b^t}\left(\tilde{\xi} - \frac{\tilde{\alpha}y - \tilde{\beta}z}{y-z}\right)\delta_1 = b^q\chi\delta_1. \tag{9}$$

Значение числа χ приведено в условии леммы. □

Отметим, что число χ может быть большим, если $y - z$ мало по сравнению с $\tilde{\alpha}y - \tilde{\beta}z$. Число q характеризует порядок потери точности вычислений. Желательно, чтобы оно было существенно меньше m, например $q \leqslant \frac{m}{4}$.

Замечание. Вычисления в НАВМ, в которых будут использованы в качестве промежуточных значений величины $\frac{d_m}{y_m - z_m}$, будут иметь оценку погрешности, в которой сомножитель b^q может сохраниться, возможны случаи, когда значение q возрастет, уменьшится или станет равным нулю. Следовательно, при анализе погрешности вычислений необходимо каким-то образом учитывать этот эффект потери точности вычислений. На практике данный эффект возникает при численном дифференцировании функций, при решении плохо обусловленных систем линейных уравнений и т.д.

Лемма 3. Пусть функция $\varphi : R^n \to R^1$ удовлетворяет условию Липшица:

$$|\varphi(x + \Delta x) - \varphi(x)| \leqslant L\,\|\Delta x\|;\ x, x+\Delta x \in G, \tag{10}$$

где $G \subset R^n$ — компакт. Тогда существуют такие числа $l_i \in [-L, L]$, что $\varphi(x+\Delta x) - \varphi(x) = \sum_{i=1}^n l_i \Delta x_i$.

Доказательство.
Из (10) следует, что для каждой пары точек $x+\Delta x$, x существует число $L_1 \in [-L, L]$ такое, что

$$\varphi(x+\Delta x) - \varphi(x) = L_1 \|\Delta x\|. \tag{11}$$

Представим (11) в виде

$$\varphi(x+\Delta x) - \varphi(x) = \frac{L_1}{\|\Delta x\|}\|\Delta x\|^2 = \sum_{i=1}^{n} \frac{L_1 \Delta x_i}{\|\Delta x\|}\Delta x_i = \sum_{i=1}^{n} L_1 \alpha_i \Delta x_i = \sum_{i=1}^{n} l_i \Delta x_i,$$

где $\alpha_i \in [-1, 1]$, т.е. $l_i \in [-L, L]$. □

Рассмотрим свойства погрешностей конечношаговых алгоритмов.

Теорема 1. Пусть функция $f(x) \in R^k$, $x \in G \subset R^n$; базовые функции $\varphi^i(a_i)$ (кроме логических) либо являются функциями округления числа, либо непрерывны по Липшицу, т.е. в некоторой окрестности $\Omega_i(a_i)$ точки $a_i \in R^s$удовлетворяют условию: $\varphi^i(a_i) - \varphi^i(b_i) \leqslant L_i|a_i - b_i|$, $i \in [1, N]$, а алгоритм (5) вычисления функции $f(x)$ является нормальным для $x \in G$. Тогда существует такой вектор $\tilde{C} \in R^k$, что

$$f_m(x) - f(x) = \tilde{C}\delta_1, \tag{12}$$

где $\delta_1 = \frac{1}{2}b^{1-m}$, m – длина мантиссы.

Доказательство.
Для доказательства утверждения достаточно показать, что

$$a_{m,i} = a_i + \alpha_i\delta_1 \text{ и } \varphi^i_m(a_{m,i}) = \varphi^i(a_i) + \beta_i\delta_1, \forall\ i \in [1, N], \tag{13}$$

кроме логических операций φ_i, где α_i и β_i — некоторые константы, a_i — точное значение аргумента.

Доказательство проводится индукцией по номеру $i \in [1, N]$ вычислительной операции алгоритма (5) (кроме логической). В качестве базы индукции можно взять операцию округления числа a_1 — первого числа, с которого начинаются вычисления. По Лемме 1: $a_{m,1} = \varphi^1_m(a_1) = \varphi^1(a_1) + \alpha_1\delta_1 = a_1 + \alpha_1\delta_1$, $|\alpha_1| \leqslant |a_1|$. Предположим, что для некоторого $i \in [1, N]$ выполнено (13). Тогда необходимо доказать, что (13) выполнено и для операции с номером $i+1$. Компонентами аргумента $a_{m,i+1}$ могут быть либо числа, входящие в условие задачи (аргументы функции $f(x)$), либо числа $a^j_{m,i+1} \equiv \varphi^t_m(a_t)$, $j \in [1, s]$, $t \leqslant i$, для которых условие (13) выполнено по условию индукции. Если операция $i+1$ есть разность двух близких чисел одинакового порядка, то по Лемме 2 справедливо (9). Если операция $i+1$ является округлением, то справедливо (8). Теперь надо показать, что (13) выполнено для функции φ^{i+1}, удовлетворяющей условию Липшица. По Лемме 1 $\varphi^{i+1}_m(a_{m,i+1}) = \varphi^{i+1}(a_{m,i+1})(1+u_{i+1})$, где $|u_{i+1}| \leqslant \delta_1$. Учитывая, что $a_{i+1} \in R^s$ и Лемму 3, получим

$$\varphi^{i+1}_m(a_{m,i+1}) = \varphi^{i+1}(a_{i+1} + \alpha_{i+1}\delta_1)(1+u_{i+1}) = (\varphi^{i+1}(a_{i+1}) + \sum_{j=1}^{s}\alpha^j_{i+1}l^j_{i+1}\delta_1)(1+u_{i+1}) =$$
$$= \varphi^{i+1}(a_{i+1}) + \left[\frac{u_{i+1}}{\delta_1}\varphi^{i+1}(a_{i+1}) + (1+u_{i+1})\sum_{j=1}^{s}\alpha^j_{i+1}l^j_{i+1}\right]\delta_1 = \varphi^{i+1}(a_{i+1}) + \beta_{i+1}\delta_1, \tag{14}$$

где $|l^j_{i+1}| \leqslant L_{i+1}$, $\frac{u_{i+1}}{\delta_1} \in [-1, 1]$, т.е. для операции $i+1$ выполнено (13). Таким образом, равенства (13) выполнены $\forall\ i \in [1, N]$. Так как компоненты вектор-функции f^t_m — это некоторые числа $\varphi^{i_t}_m$, $i_t \in [1, N]$, $t \in [1, k]$, значения которых по доказанному представимы в виде $\varphi^{i_t}_m = \varphi^{i_t}(a_{i_t}) + \beta_{i_t}\delta_1$, то, переобозначая β_{i_t} в c_t, получим $f^t_m(x) = f^t(x) + c_t\delta_1$ и $f_m(x) = f(x) + \tilde{C}\delta_1$, где $\tilde{C} = (c_1, c_2, ..., c_k)$. □

Таким образом, система уравнений (15) для $i \in [1, N]$ представляет собой систему условий для рекуррентного оценивания погрешностей вычисления значений функции $f(x)$, $x \in G$.

Приведем качественные оценки (выводы) полученных результатов. По необходимости при неизвестной природе решаемой задачи, их можно принять за гипотезы.

Следствие.

1) Вектор $\tilde{C}$ в формуле (12) назовем параметром погрешности (ПП) значения функции. Он является функцией точки x и размера мантиссы m. Структура его сложна и может включать элементы погрешности вида (9). Тогда погрешность $\tilde{C}\delta_1$ можно представить в виде $\tilde{C}\delta_1 = \left(\tilde{A}_1 b^{q_1}, \tilde{A}_2 b^{q_2}, ..., \tilde{A}_k b^{q_k}\right) \delta_1$. Возьмем компоненту вектора $\tilde{A}_{i_0}$, для которой $\left|\tilde{A}_{i_0} b^{q_{i_0}}\right| = \max\limits_i \left|\tilde{A}_i b^{q_i}\right|$. Обозначим $\mathrm{q}_{i_0} = q$. Тогда существуют $\bar{C}_i$ такие, что $\tilde{C}\delta_1 = \left(\bar{C}_1, ..., \bar{C}_k\right) b^q \delta_\mu = \left(\bar{C}_1, ..., \bar{C}_k\right) (\delta_\mu)^\alpha = \bar{C} (\delta_\mu)^\alpha$, где $\alpha = 1 - \frac{q}{m}$, $\delta_1 = \frac{b}{2}\delta_\mu$.
Таким образом, теперь формулу погрешности (12) можно представить в виде

$$\Delta f_m(x) = f_m(x) - f(x) = \bar{C} (\delta_\mu)^\alpha, 0 < \alpha \leqslant 1. \tag{15}$$

Когда $q = 0$, то $\alpha = 1$ и $\Delta f_m(x) = f_m(x) - f(x) = \bar{C}\delta_\mu$, где $\bar{C} = \frac{b}{2}\tilde{C}$. Оценка погрешности (15) дана для одной точки. Её обобщением является определение 9.

2) Из Теоремы 1 можно получить следующие выводы.

Во-первых, $\left\|\tilde{C}\right\| < \infty$, т.е. компоненты c_i, $i \in [1, k]$, для фиксированного достаточно большого m, ограниченные сверху и снизу числа. Этот вывод следует из того, что каждая компонента c_i получена по рекуррентной формуле (14) за конечное число шагов.

Во-вторых, значение параметра $\tilde{C}$ может быть таким, что $\left\|\tilde{C}\right\| \to \infty$ при $m \to \infty$. Например, если $q = (1 - \alpha) m$ при фиксированном значении α, $m \to \infty$. Известно [7], что погрешность значения функции f величина случайная, зависящая от длины мантиссы m. Но так как при увеличении длины мантиссы m модуль погрешности значений базовых функций уменьшается, то и величина колебаний значений параметра $\left\|\bar{C}\right\|$, при определенных условиях, также может уменьшаться и следует ожидать, что его значения будут иметь колебания около некоторой средней его величины. В определении (9) приведено понятие, удовлетворяющее этим условиям.

Рассмотрим теперь свойства бесконечношаговых алгоритмов вычисления $f(x)$. Для БША имеет место следующее определение:

Определение 6. Пусть в бесконечношаговом алгоритме выполнено N первых базовых операций и для $f(x)$ получено приближение значения вычисляемой функции $f_m^N(x)$. БША вычисления функции f называется *нормальным*, если для всех N алгоритм вычисления значения $f_m^N(x)$ будет нормальным. □

Определение 7. Бесконечношаговый алгоритм называется *сходящимся*, если $f(x) = \lim\limits_{N \to \infty} f^N(x)$. □

Теорема 1 может быть уточнена на случай сходящегося бесконечношагового алгоритма следующим образом:

Теорема 2. Пусть для вычисления значения функции $f_m^N(x)$ БША является нормальным и выполнены условия Теоремы 1. Тогда существуют векторы $\tilde{C} \in R^k$, $\gamma \in R^k$ такие, что $\forall\, \varepsilon > 0 :\ \|\gamma\| \leqslant \varepsilon$, имеет место представление

$$f_m^N(x) = f(x) + \tilde{C}\delta_1 + \gamma. \tag{16}$$

Доказательство

Из сходимости бесконечношагового алгоритма следует, что $\forall\, \varepsilon > 0\ \exists\, N : \left\|f^N(x) - f(x)\right\| \leqslant \varepsilon$ и $f^N(x) - f(x) = \gamma$, $\|\gamma\| \leqslant \varepsilon$. Из Теоремы 1 для данного N имеем равенство $f_m^N(x) = f^N(x) + \tilde{C}\delta_1$, где $\tilde{C} \in R^k$. Тогда для значения $f_m^N(x)$ получим

$$f_m^N(x) = f(x) - f(x) + f^N(x) + \tilde{C}\delta_1 = f(x) + \tilde{C}\delta_1 + \gamma.$$

Теорема доказана. □

4. О гарантированной точности решений задач ВМ

В этом разделе для случая, когда значение параметра погрешности $\|\bar{C}\|$ ограничено, получена оценка длины мантиссы МЧ, гарантирующая достижение требуемой точности решения задачи ВМ.

Определение 8. Будем говорить, что метод (алгоритм) вычисления значения функции $f \in R^k$ называется *корректным* (КМ), если для любого $\varepsilon > 0$ найдется такой размер мантиссы m, что

$$\Delta = \|f(x) - f_m(x)\| \leqslant \varepsilon, \text{ или } \frac{\Delta}{\|f_m(x)\|} \leqslant \varepsilon, \text{ при } \|f_m(x)\| \neq 0. \tag{17}$$

Величину ε в (17) будем называть *требуемой точностью*. □

Обратимся к вопросу определения достаточной (гарантированной) точности, которую должно обеспечить ВУ — вычислительное устройство, т.е. ЭВМ, имеющая необходимое программное обеспечение.

Пусть ВУ имеет переменную длину мантиссы m, т.е. пользователь может выбрать размер мантиссы, необходимый для проведения вычислений. Пусть $f : G \subset R^n \to R^k$, а $f_m(x)$ — значение функции f, вычисленное в точке x_m с помощью данного ВУ.

Определение 9. Будем говорить, что значение функции $f_m(x)$ *имеет погрешность порядка* α, $0 < \alpha \leqslant 1$, относительно погрешности мантиссы δ_μ, если существуют константы C и C_0 такие, что $\forall\, x \in G$:

$$\begin{gathered}\Delta = \|f(x) - f_m(x)\| \leqslant C\,(\delta_\mu)^\alpha\,, \text{ для абсолютной погрешности;} \\ \frac{\|f(x) - f_m(x)\|}{\|f_m(x)\|} = \frac{\Delta}{\|f_m(x)\|} \leqslant C_0\,(\delta_\mu)^\alpha \text{ при } \|f_m(x)\| \neq 0, \text{ для относительной погрешности,}\end{gathered} \tag{18}$$

где $\delta_\mu = b^{-m}$. □

Приведем без доказательства простое, но полезное утверждение, в котором гарантируется погрешность значения $f_m(x)$ порядка a.

Лемма 4. Пусть для функции $f(x)$ в некоторой точке $x \in \Omega \subset G$, где Ω — компакт, выполнены условия Теоремы 1, а в равенстве (15) α удовлетворяет условию $0 < \alpha \leqslant 1$. Тогда значение функции $f_m(x)$ имеет погрешность порядка α_0, $0 < \alpha_0 \leqslant 1$, $\forall\, x \in \Omega$. □

Докажем теперь теорему о длине мантиссы МЧ, гарантирующей требуемую точность решения задачи ВМ.

Теорема 3. Пусть погрешности Δ значения функции $f_m(x)$ или $\frac{\Delta}{\|f_m(x)\|}$ имеют порядок α. Тогда для любого $\varepsilon > 0$ данный метод вычисления функции будет корректным при $m \geqslant \left\lfloor 1 - \frac{1}{\alpha}\log_b \frac{\varepsilon}{C} \right\rfloor$ или $m \geqslant \left\lfloor 1 - \frac{1}{\alpha}\log_b \frac{\varepsilon}{C_0} \right\rfloor$, где $\lfloor A \rfloor$ — целая часть числа A.

Доказательство.

Докажем теорему для абсолютной погрешности Δ. Доказательство для относительной погрешности аналогично.

Зададим число $\varepsilon > 0$. Так как для достаточно больших m число $(\delta_\mu)^\alpha = b^{-\alpha m}$ может быть достаточно малым, то выберем m такое, что $(\delta_\mu)^\alpha = b^{-\alpha m} \leqslant \frac{\varepsilon}{C}$. Подставляя в (18) значение $\frac{\varepsilon}{C}$ вместо $(\delta_\mu)^\alpha$, получим, что $\Delta = \|f(x) - f_m(x)\| \leqslant \varepsilon$, т.е. выполнено (17) и метод вычисления будет корректным при выбранном m. Теперь из неравенства $b^{-\alpha m} \leqslant \frac{\varepsilon}{C}$ найдем значение m: $-\alpha m \leqslant \log_b \frac{\varepsilon}{C}$, откуда $m \geqslant -\frac{1}{\alpha}\log_b \frac{\varepsilon}{C}$. Так как m — целое число, то очевидно: $m \geqslant \left\lfloor 1 - \frac{1}{\alpha}\log_b \frac{\varepsilon}{C} \right\rfloor$. Значение мантиссы m для относительной погрешности определяется по формуле $m \geqslant \left\lfloor 1 - \frac{1}{\alpha}\log_b \frac{\varepsilon}{C_0} \right\rfloor$. □

Замечание.
1. В важном частном случае, когда $\alpha = 1$, оценки для m из Теоремы 3 имеют вид $m \geqslant \lfloor 1 - \log_b \frac{\varepsilon}{C} \rfloor$ и $m \geqslant \left\lfloor 1 - \log_b \frac{\varepsilon}{C_0} \right\rfloor$.
2. Константы C и C_0 связаны следующими условиями:

$$\frac{\Delta}{\|f_m(x)\|} = \frac{\|f(x) - f_m(x)\|}{\|f_m(x)\|} \leqslant \frac{C\,(\delta_\mu)^\alpha}{\|f_m(x)\|} \leqslant C_0\,(\delta_\mu)^\alpha, \tag{19}$$

где $C_0 = \frac{C}{\|f_m(x)\|}$ или $C_0 = \frac{C}{\underline{f}}$, т.е. $\frac{C}{\|f_m(x)\|} \leqslant \frac{C}{\underline{f}} = C_0$, $\underline{f} = \min\limits_z \|f_m(z)\|$, $z \in \Omega$, $\Omega \subset G$ – компакт. В случае, когда $f_m(x) = 0$, относительная погрешность не рассматривается.

Ввиду особой важности Теоремы 3 для прикладных исследований переформулируем её в следующем виде:

Теорема 4. (**правило гарантированной точности** решений задач ВМ). Пусть погрешность (абсолютная или относительная) вычисленного значения функции f имеет порядок $\alpha, 0 < \alpha \leqslant 1$. Тогда для любой требуемой точности $\varepsilon > 0$ существует такой размер мантиссы m, при котором достигается заданная точность ε решения задачи. □

На практике Теоремы 3, 4 применимы до максимального значения мантиссы $m_{\max}$, которую обеспечивает данная библиотека программ. В частности, для библиотеки GNU GMP $m_{\max} = 646456993$ десятичных знаков.

5. Вывод

В работе введено понятие нормального алгоритма решения задач вычислительной математики (НАВМ). Для НАВМ получена оценка погрешности решения задачи ВМ, зависящая от значений её аргументов и длины мантиссы числа. Для задач ВМ, в которых значение параметра погрешности ограничено, получена оценка длины мантиссы, гарантирующая достижение требуемой точности решения. В продолжении настоящей работы будут предложены и обоснованы эмпирические правила оценки точности решений задач ВМ.

Литература

1. IEEE 754-2008: 754-2008 IEEE Standard for Floating-Point Arithmetic. – ISBN: 978-0-7381-5753-5.
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. *Higham N. J.* Accuracy and stability of numerical algorithms. — Philadelphia: Society for Industrial and Applied Mathematics, 1996.
4. *Torbjorn Granlund* [*et al.*] GNU Multiple Precision Arithmetic Library 4.1.2 / http://swox.com/gmp/, 2002.
5. *Fousse, Laurent and Hanrot, Guillaume and Lefèvre, Vincent and Pélissier, Patrick and Zimmermann, Paul* MPFR: A multiple-precision binary floating-point library with correct rounding. — ACM Trans. Math. Softw., 2007.
6. *Марков А. А., Нагорный Н. М.* Теория алгорифмов. — М.: Наука, 1984.
7. *Воеводин В. В.* Вычислительные основы линейной алгебры. — М.: Наука, 1977. — 304 с.
8. *Иванов В. В.* Методы вычислений на ЭВМ: Справочное пособие. — Киев: Наукова думка, 1986. — 584 с.
9. IEEE standard for radix-independent floating-point arithmetic: ANSI/IEEE Std 854-1987, 1987. http://grouper.ieee.org/groups/754/
10. ISO/IEC 9899:1999 Standard for the C programming language (C99) — 1999.

11. *Wilkinson J. H.* Rounding Errors in algebraic processes. — Englewood Cliffs, N.J.: Prentice-Hall, 1963. ISBN 0-486-67999-3.

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

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

14. *Langlois P.* A Revised Presentation of the CENA Method. — ARENAIRE — Inria Grenoble Rhône-Alpes / LIP Laboratoire de l'Informatique du Parallélisme

15. *Шокин Ю. И.* Интервальный анализ. — Новосибирск: Наука. Сиб. отд-ние, 1981.

16. *Шарый С. П.* Конечномерный интервальный анализ. — М., 2007.

17. *Francoise Chaitin-Chatelin and Valdrie Fraysse.* Lectures on Finite Precision Computations. — Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996.

18. *Алефельд Г., Херцбергер Ю.* Введение в интервальные вычисления. — М.: Мир, 1987. — 356 с.

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

Поступила в редакцию 05.06.2012.