На правах рукописи

ГРИНЕВИЧ Алексей Иванович

МЕТОД ОЦЕНКИ ПОГРЕШНОСТИ ОКРУГЛЕНИЙ ЗНАЧЕНИЙ ВЫЧИСЛЯЕМОЙ ФУНКЦИИ, ОСНОВАННЫЙ НА ВАРЬИРОВАНИИ ДЛИНЫ МАНТИССЫ В АРИФМЕТИКЕ С ПЛАВАЮЩЕЙ ЗАПЯТОЙ

Специальность 01.01.07 – вычислительная математика

АВТОРЕФЕРАТ

диссертации на соискание учётной степени кандидата физико-математических наук

МОСКВА – 2013

Работа выполнена на кафедре математических основ управления
Московского физико-технического института
(государственного университета)

Научный руководитель: кандидат физико-математических наук, доцент Бирюков Александр Гаврилович

Официальные оппоненты: доктор физико-математических наук, профессор Лобанов Алексей Иванович, кафедра вычислительной математики Московского физико-техинческого института (государственного университета), профессор

доктор физико-математических наук, доцент Рогов Борис Вадимович, Институт прикладной математики имени М.В. Келдыша РАН, ведущий научный сотрудник

Ведущая организация: Вычислительный центр им. А.А. Дородницына РАН

Защита состоится _____ 2013 г. в ___ часов на заседании диссертационного совета Д 212.156.05 при Московском физико-техническом институте (государственном университете) по адресу 141700, Московская обл., г. Долгопрудный, Институтский пер., д.9 ауд. 903 КПМ.

С диссертацией можно ознакомиться в библиотеке Московского физико-технического института (государственного университета).

Автореферат разослан _____ 2013г.

Ученый секретарь
диссертационного совета Федько О. С.

Общая характеристика работы

Актуальность темы

В настоящее время большое внимание научного сообщества уделяется вопросам «вычислений с высокой точностью», т.е. таким вычислениям, при которых возможны изменения длины мантиссы машинного числа (МЧ) в широком диапазоне значений. Как отмечено в обзоре D. Bailey за 2012 г., можно выделить целый ряд направлений исследований, где стандартной арифметики оказывается недостаточно. Среди них есть как давно известные проблемы, так и относительно новые, активное изучение которых началось вместе с появлением достаточных вычислительных мощностей. Это моделирование солнечной системы за период в миллионы лет, вычисление рекуррентных соотношений, определение экспоненциально малых явлений в динамических системах, компьютерное исследование новых математических соотношений, моделирование явлений в сверхновых звездах и др. В частности, А. Фролов утверждает, что «сейчас мы научились рассматривать и решать задачу нескольких тел с ограничениями, о чем нельзя было и мечтать всего несколько лет назад». Для решения задачи им использовалась арифметика высокой точности (120 знаков) для нахождения собственных векторов почти вырожденных матриц размерами 5000x5000 в рамках решения задачи n тел.

Как правило, вычислительные сложности возникают, когда решаемая проблема содержит, например, одну из следующих вычислительных задач.

1. Решение плохо обусловленных систем линейных уравнений (СЛУ).
2. Распараллеливание вычислений. Задачи, разрешимые на одном процессоре приобретают новую степень сложности при распараллеливании. В частности, вычисление суммы ряда в условиях параллелизма, когда точный порядок суммирования неконтролируем, приводит к появлению дополнительных вычислительных погрешностей.
3. Моделирование долговременных физических процессов. Как правило, в условиях большого количества итераций возникают искажения, вносимые накоплением ошибок округления промежуточных вычислений.
4. Проблемы экспериментальной математики. Такие, как вычисление большого количества знаков в числах π и e для установления возможной «нормальности», проверка гипотез о математических тождествах и т.п.

Время вычислений существенно возрастает (на 1-2 порядка) при переходе на «высокоточную» арифметику с варьируемой длиной мантиссы. Поэтому подобные вычислительные проблемы считались слишком трудоемкими до недавнего времени, и ученые старались изыскивать специфические методы решения для конкретных задач или упростить задачу до «разрешимого» варианта. В последние годы ситуация начала меняться с появлением библиотек программ с интерфейсами высокого уровня. Здесь следует отметить и специальные подпрограммы для нахождения значений функций (ARPREC, GMP, MFPR, MPFR++, MPFUN90, QD), пакеты компьютерной алгебры, позволяющие находить решение символически без потери точности (Maple, MathCad, Mathematica) и реализации языков программирования, позволяющие задавать длину мантиссы (LISP, Python, Perl, Haskell, Ruby). При переходе на высокоточную арифметику, как правило, оказывается возможным не переводить программу целиком, а заменять лишь часть ключевых алгоритмов на более «точные» варианты. Это позволяет локализовать вычисления, требующие высокой точности и минимизировать влияние возрастающего времени выполнения до приемлемых значений. Таким образом, безусловно, вычисления в арифметике высокой точности не могут рассматриваться отдельно от других подходов по оптимизации вычислений. Поскольку набор практических задач, где необходимы высокая или заданная точность, продолжает расти, возникает вопрос о построении метода (методики) оценок погрешностей округлений при варьировании точности арифметики для получения заданной точности решения (ЗТР). Приобретают актуальность вопросы выработки обобщенных подходов и методов, применимых к широкому спектру задач в условиях арифметики высокой точности.

Таким образом, построение методов (методики) оценок погрешности решений задач вычислительной математики (ВМ), обладающих вышеуказанными свойствами, является актуальной научной проблемой.

Цели диссертационной работы.

1. Построение алгоритма вычисления значения функции $f(x) \in R^k$, применимого для достаточно широкого класса задач ВМ. Для предложенного алгоритма – исследование оценок погрешностей округления приближённых значений $\tilde{f}(x)$, зависящих от длины m мантиссы машинного числа (МЧ) и получение такого значения m_0, которое обеспечивает достижение требуемой точности решения.
2. Разработка численных методов оценки погрешностей округления значений $f(x)$, гарантирующих достижение заданной точности (ЗТР), и исследование их свойств.

3. Численное исследование предложенных методов оценок погрешностей округления для некоторых классов задач ВМ.

Научная новизна.

1. Предложен численный алгоритм решения задачи ВМ, названный «Элементарный алгоритм вычислительной математики» (ЭАВМ), представляющий собой естественное объединение стандартных (базовых) вычислительных операций из какой-либо библиотеки программ. ЭАВМ обладает тем свойством, что при его реализации в точной арифметике будет получено точное значение вычисляемой функции.
2. Для ЭАВМ с конечным и бесконечным числом шагов (КША и БША) получены теоретические оценки погрешностей значений функции, зависящие от её аргументов и длины мантиссы m МЧ, гарантирующие достижение заданной точности решений.
3. Предложен метод К-решений (КР) – численной оценки погрешностей округления значения вычисляемой функции. Введены понятия итерационной последовательности значений функции с переменной мантиссой (ИППМ) и её g-устойчивости. Для g-устойчивой ИППМ доказана возможность достижения заданной точности решения.
4. Предложены алгоритмы, позволяющие оптимизировать процесс достижения заданной точности решения.

Практическая ценность. 1. Предложенный метод оценок погрешностей округления применим для достаточно широкого класса вычисляемых функций. 2. Метод не требует изменения используемых вычислительных алгоритмов. Для получения значений погрешностей необходимо вычислить значения функции при разных длинах мантисс МЧ и сравнить их. 3. Предложен вариант метода варьирования длины мантиссы для получения заданной точности решения при решении задачи безусловной минимизации методом Ньютона.

Апробация. Основные положения диссертационной работы докладывались, обсуждались и получили одобрение специалистов на научных семинарах кафедр высшей математики и математических основ управления Московского физико-технического института (государственного университета) (2007-2013), научном семинаре отдела прикладных проблем оптимизации в Вычислительном центре им. А.А. Дородницына РАН (2013).

Публикации. По теме диссертации опубликованы 6 печатных работ, из них четыре [1, 2, 5, 6] из списка, рекомендованного ВАК РФ.

Личный вклад автора. Как содержание диссертации, так и основные положения, выносимые на защиту, отражают личный вклад автора в

опубликованные по теме диссертации работы. Все представленные в диссертации результаты получены лично автором.

Структура. Диссертация состоит из введения, четырех глав, заключения, списка использованной литературы, включающего 72 наименования, и двух приложений. Общий объем работы составляет 157 страниц.

Содержание работы

Во введении обоснована актуальность, сформулированы цели диссертационной работы, изложены основные положения, выносимые на защиту, характеристика их научной новизны, практической значимости и апробации полученных результатов.

Для достижения требуемой точности решения рассматриваются оценки вида:

$$\Delta = \left\| \tilde{f} - f \right\| \le \Phi_1\left(\tilde{f}_1,..., \tilde{f}_l\right) \le \varepsilon, \qquad (1)$$

где $\tilde{f}_i, i = \overline{1,l}$ – приближенные значения решений. Предложен метод К-решений (Глава 2) для построения функций оценок погрешности $\Phi_1\left(\tilde{f}_1,..., \tilde{f}_l\right)$. Для того, чтобы оценка (1) выполнялась для ε, изменяющемся в некотором интервале значений, функции $\tilde{f}_1,..., \tilde{f}_l$, определяются при их вычислениях с различной длиной мантиссы МЧ. Факт достижения требуемой точности характеризует следующее определение.

Определение 1. Пусть задано $\varepsilon > 0$ – требуемая точность вычисления приближённого значения $\tilde{f}$ функции f, т.е. если точность решения задачи достижима, то имеет место неравенство $\Delta = \left\| \tilde{f} - f \right\| \le \varepsilon$ или $\Delta / \left\| f \right\| \le \varepsilon$. Будем говорить, что имеет место **заданная точность решения (ЗТР)**, если задача решается на ЭВМ методом, для которого известна функция оценки погрешностей решения $\Phi_1\left(\tilde{f}_1,.., \tilde{f}_r\right)$, значения оценок погрешностей определяются вместе с искомым решением и для них выполнено условие достижимости точности (1). □

Глава 1 диссертационной работы является вводной, носящей вспомогательный характер. Рассматриваются свойства машинного числа (МЧ) в формате IEEE 754; рассмотрены понятия **машинного числа** (МЧ), **точности представления** МЧ, **функции округления** МЧ и **арифметических операций**; рассмотрены ситуациии **потери значимости** и **переполнения,** возникающие при выполнении арифметических операций; рассматриваются характеристики библиотек програм GNU GMP и GNU MPFR, в которых реализуется вычисление в арифметике с плавающей запятой базовых (стандартных)

математических функций. Важнейшим свойством этих программ является то, что пользователь может задавать длину мантиссы МЧ в очень широком диапазоне от $\mathrm{m}_{\min}=8$ до $m_{\max}=646456993$ десятичных знаков. Появление в свободном доступе программного обеспечения с такими характеристиками расширяет границы возможностей для получения решений широкого круга задач ВМ с гарантированной точностью высокого порядка. Диссертационная работа посвящена изучению погрешностей округления решений, значения которых получены для варьируемых значений длины мантиссы.

Также рассмотрены другие подходы для достижения требуемой точности решений. В частности, рассмотрены свойства модели рациональных чисел и модель чисел с многоуровневой экспонентой. Отмечается, что в широкой практике они пока не нашли своего применения.

В **главе 2** диссертации предложен алгоритм решения задачи ВМ, названный «элементарным» и исследованы некоторые его свойства.

В §2.1 вводятся определения понятий, на основе которых строится «элементарный алгоритм ВМ» (ЭАВМ).

Определение 2. Пусть $\varphi: R^n \to R^1$ – некоторая функция, $x \in R^n$ – вектор, x_{m} – его машинное представление. Тогда: $\varphi(x)$, $\varphi(x_{\mathrm{m}})$ – значения функции φ в точках x и x_{m}, $\varphi_{\mathrm{m}}(x_{\mathrm{m}})$ – **машинное представление значения** $\varphi(x_{\mathrm{m}})$; $\varphi_{\mathrm{m}}(x) \equiv \varphi_{\mathrm{m}}(x_{\mathrm{m}})$. 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, реализующих математическую библиотеку стандарта ANSI 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, в которой при заданной длине мантиссы m для значений базовых функции выполняется округление до последней значащей цифры, т.е.:

$$\varphi_{\mathrm{m}}(x_{\mathrm{m}}) = \varphi(x_{\mathrm{m}})(1+u),\ |u| \le \delta_1,\ \varphi_{\mathrm{m}}(x_{\mathrm{m}}) \ne 0, \text{ где} \qquad (2)$$

$\varphi(x_{\mathrm{m}})$ и $\varphi_{\mathrm{m}}(x_{\mathrm{m}})$ – значение одной из стандартных (базовых) функций, вычисленное в точке $x_{\mathrm{m}} \in M_{\mathrm{b,m,p}}$. Для случая, когда $\varphi_{\mathrm{m}}(x_{\mathrm{m}}) = 0$,

$|\varphi(x_{\mathrm{m}})-\varphi_{\mathrm{m}}(x_{\mathrm{m}})|=|\varphi(x_{\mathrm{m}})|\le\delta_0$. В (2) $\delta_1=\frac{1}{2}b^{1-m}$, b-основание системы исчисления, δ_0-погрешность нуля; она на много порядков меньше δ_1 (см. Главу 1).

<u>Определение 4</u>. **Задачей вычислительной математики** будем называть совокупность понятий и условий $F(f,x,\mathrm{m},\varepsilon)$, 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 $f_{\mathrm{m}}(x)$ для которой выполнено одно из условий:

$$\Phi(f_{m_1},...,f_{m_l})\le\varepsilon \text{ или } \frac{\|f_{\mathrm{m}}-f\|}{\|f\|}\le\varepsilon \text{ или } \|f_m-f\|\le\varepsilon. \qquad \square$$

Функция $\Phi(\bullet)$ определяется методом решения задачи; f_m, f_{m_1}, …, f_{m_l} – приближенные значения решений вычисляемые в точке $x\in G$ для длин мантисс МЧ $m,m_1,...,m_l$.

Алгоритм решения задачи ВМ представляет собой композицию (объединение) базовых операций, т.е.

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

$$ALG(f(x))=\varphi^1(a_1)\vee\varphi^2(a_2)\vee\cdots\vee\varphi^{N_0-1}(a_{N_0-1})\vee\varphi^N(a_N),\ N\to\infty \qquad (4)$$

где символ $\vee$ означает объединение операций. Первый алгоритм вычисления $f(x)$ является конечношаговым (КША), второй – бесконечношаговым (БША). Оба алгоритма (3) и (4) рассматриваются в точной арифметике. Соответствующие КША и БША алгоритмы для вычисления $f_m(x)$ представлены следующими формулами:

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

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

N_0 – число шагов БША, которое определяется по некоторому правилу его окончания.

<u>Определение 5</u>. Алгоритм вычисления функции $f\in R^k$ (5) в точке $x\in R^n$ при длине мантиссы m будет называться **элементарным алгоритмом** для решения задач **вычислительной математики (ЭАВМ)**, 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:

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

В §2.2 изучаются погрешности округлений, возникающие в итерационном процессе ЭАВМ.

Лемма 1. Пусть φ_i – базовые вычислительные операции (кроме логических) из некоторой библиотеки программ, $i \in [1, N_б]$ – номер базовой операции. Тогда значение $\varphi^i_{\mathrm{m}}(a_i)$ можно представить в виде

$$\varphi^i_{\mathrm{m}}(a_i) = \varphi^i(a_i) + \alpha_i \delta_1, \text{ где} \qquad (8)$$

$N_б$ – число базовых операций библиотеки, $|\alpha_i| \le |\varphi^i(a_i)|$, $\delta_1 = \frac{1}{2} b^{1-\mathrm{m}}$, $a_i \in R^s$, $a_i = a_{\mathrm{m},i}$ – вектор машинных чисел. □

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

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

$$|\varphi(x + \Delta x) - \varphi(x)| \le L \|\Delta x\|, \; x, x + \Delta x \in G, \text{ где} \qquad (9)$$

$G \subset R^n$ – компакт. Тогда существуют такие числа $l_i \in [-L, L]$, что

$$\varphi(x + \Delta x) - \varphi(x) = \sum_{i=1}^{n} l_i \Delta x_i \,. \qquad \square$$

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: $\varphi^i(a_i) - \varphi^i(b_i) \le L_i |a_i - b_i|$, $i \in [1, N]$, а алгоритм (3) вычисления функции $f(x)$ является элементарным для $x \in G$. Тогда существует такой вектор $\tilde{C} \in R^k$, что

$$f_{\mathrm{m}}(x) - f(x) = \tilde{C}\delta_1, \text{ где } \delta_1 = \frac{1}{2} \mathrm{b}^{1-\mathrm{m}}, \; \mathrm{m} - \text{длина мантиссы. } \square \qquad (10)$$

Определение 6. Пусть в бесконечношаговом алгоритме выполнено N первых базовых операций и для $f(x)$ получено приближение значения вычисляемой функции $f^{\mathrm{N}}_{\mathrm{m}}(x)$. БША вычисления функции f называется **элементарным**, если для всех указанных N алгоритм вычисления значения $f^N_{\mathrm{m}}(x)$ будет элементарным. □

Определение 7. Бесконечношаговый алгоритм называется **сходящимся**, если $f(x)=\lim\limits_{N\to\infty} f^N(x)$, где $f^N(x)$ – точное решение задачи после N операций БША. □

Из доказательства теоремы 1 следует, что для ЭАВМ решение задачи представляется как:

$$f_m(x)= f(x)+\tilde{C}\delta_1, \tag{11}$$

где $\tilde{C}$ – вектор параметров погрешностей округления, $\|\tilde{C}\|<\infty$, $\delta_1=\frac{1}{2}b^{1-m}$, m – длина мантиссы МЧ. Анализ формулы, приведенный в Следствии к теореме 1 показывает, что решение (11) возможно представить в другом виде:

$$f_{\mathrm{m}}(x)= f(x)+\overline{C}b^{-\alpha m}, \text{ где } 0<\alpha\le 1 \text{ и } \overline{C}=\frac{\tilde{C}}{2}b^{1-m(1-\alpha)}. \tag{12}$$

Если БША решения задачи является сходящимся т.е. $f(x)=\lim\limits_{N\to\infty} f^N(x)$, где $f^N(x)$ – точное решение задачи после выполнения N операций его определяющих, то результат теоремы 1 уточняется.

Теорема 2. Пусть для функции $f_{\mathrm{m}}^{\mathrm{N}}(x)$ БША является элементарным, сходящимся и выполнены условия теоремы 1. Тогда существуют векторы $\tilde{C}\in R^k$, $\gamma\in R^k$, такие, что $\forall\varepsilon>0:\|\gamma\|\le\varepsilon$ имеет место представление

$$f_{\mathrm{m}}^{\mathrm{N}}(x)= f(x)+\tilde{C}\delta_1+\gamma \qquad \square \tag{13}$$

В §2.3 получены оценки длины мантиссы, гарантирующей достижение требуемой точности ЭАВМ решения задачи ВМ для КША и БША.

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:

$$\Delta=\|f(x)-f_{\mathrm{m}}(x)\|\le\varepsilon, \text{ или } \frac{\Delta}{\|f(x)\|}\le\varepsilon, \text{ при } \|f(x)\|\ne 0. \tag{14}$$

Величину ε в (14) будем называть **требуемой точностью**. □

Определение 9. Будем говорить, что значение функции $f_{\mathrm{m}}(x)$ **имеет погрешность порядка** α, $0<\alpha\le 1$, относительно погрешности мантиссы δ_μ, если существуют константы C и C_0 такие, что $\forall x\in G\subset R^n$, $\|\overline{C}\|\le C$, $C/\|f(x)\|\le C_0$, $\forall m\ge m_{\min}$, где $m_{\min}$ – минимальное значение длины мантиссы при которой могут проводиться вычисления и

$$\Delta=\|f(x)-f_{\mathrm{m}}(x)\|\le C\delta_\mu^\alpha \text{ , для абсолютной погрешности;} \tag{15}$$

$\frac{\|f(x)-f_{\mathrm{m}}(x)\|}{\|f(x)\|}=\frac{\Delta}{\|f(x)\|}\le C_0\delta_\mu^\alpha$ при $\|f(x)\|\neq 0$, для относительной погрешности,

где $\delta_\mu=\mathrm{b}^{-\mathrm{m}}$. □

Векторы $\tilde{C}$ в формуле (10) (или $\overline{C}$ в (12)) назовем параметром погрешности (ПП) значения функции

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

Таким образом теорема 3 (и 4) дают условия, при которых имеет место заданная точность решения (ЗТР). Конечно, верхнее значение длины мантиссы, необходимое для обеспечения ЗТР, не может превышать максимальных значений, которые имеет используемая библиотека программ. В частности для пакета Maple $m_{\max}=65535$ десятичных знаков, а для GNU GMP $m_{\max}=646456993$ десятичных знаков.

Теорема 4. Пусть БША является сходящимся и элементарным, для каждого N выполнены условия теоремы 2, погрешность округления для каждого N в (13) имеет порядок α; т.е. $\tilde{C}\delta_1=\overline{C}b^{-\alpha m}$, $0<\alpha\le 1$ и выполнено условие $\|\overline{C}\|\le C\ \ \forall x\in G$, $G\subset R^n$, $\forall m\ge m_{\min}$, где $m_{\min}$ – минимальная длина мантиссы при которой могут проводиться вычисления. Тогда существуют такое число базовых операций N и такая длина мантиссы m, при которых достигается требуемая точность решения ε, т.е. $\|f_m^N(x)-f(x)\|\le\varepsilon$. □

Глава 3 диссертации посвящена построению вычислимых оценок погрешностей округлений значений $f_m(x)$.

В §3.1 третьей главы вводится определение **итерационной последовательности с переменной мантиссой** (ИППМ).

Определение 10. Совокупность L решений задачи $f_{m_i}(x)$ при значениях длины мантиссы $\mathrm{m}_i, i\in[1,L]$: $\mathrm{m}_1<\mathrm{m}_2<\ldots<\mathrm{m}_L$ назовем **итерационной последовательностью с переменной мантиссой** (ИППМ) решения задачи ВМ. □

Определение 11. Числа η и η_0 называются малыми по сравнению с 1, если $0<\eta\le\eta_0\le 0{,}1$. Условие малости чисел по сравнению с 1 обозначается символом «<<» – много меньше: $\eta<<1$, $\eta_0<<1$. □

Определение 12. Пусть значения погрешностей решений равны $\Delta_i=\left\|f_{m_i}(x)-f(x)\right\|$, $i\in[1,L]$; $\Delta_{ij}=\left\|f_{m_i}(x)-f_{m_j}(x)\right\|$, $j>i$, $i,j\in[1,L]$. **К-решением** (Контрольным решением, КР) задачи ВМ называется значение $f_{m_L}(x)$, если

$$\Delta_i=\Delta_{iL}+\xi_{iL}\Delta_{iL},\ i\in[1,L-1], \qquad (16)$$

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

Определение 13. Обозначим $g_i=\dfrac{\Delta_{i+1}}{\Delta_i}=\dfrac{\left\|f_{m_{i+1}}(x)-f(x)\right\|}{\left\|f_{m_i}(x)-f(x)\right\|}$, $i\in[1,L-1]$.

Последовательность решений задачи назовем g-**устойчивой**, если $g_i\le g_0<<1$, $i\in[1,L-1]$. Число g_i, $i\in[1,L-1]$ назовем **коэффициентом уменьшения** (КУ) погрешности на i-м шаге. □

Для упрощения изложения представим значение погрешности для частного случая его порядка $\alpha=1$:

$$\overline{\Delta}=f_m(x)-f(x)=\overline{C}_m b^{-m}, \qquad (17)$$

где b – основание МЧ. Тогда очевидно, что при достаточно большом значении m, и $\left\|\overline{C}_m\right\|\le C$, где C не зависит от m, $\left\|\overline{\Delta}\right\|$ может быть малым числом.

Лемма 4. Пусть для ИППМ выполнено условие (17) и погрешности: $\Delta'=\left\|f_{m'}(x)-f(x)\right\|=\left\|\overline{C}_{m'}\right\|b^{-m'}$ и $\Delta''=\left\|f_{m''}(x)-f(x)\right\|=\left\|\overline{C}_{m''}\right\|b^{-m''}$ удовлетворяют условию: $\dfrac{\left\|\overline{C}_{m''}\right\|}{\left\|\overline{C}_{m'}\right\|}\le\xi_0=const$, $\forall m',m'':m_1\le m'<m''\le m_L$. Тогда найдется такая $\Delta m=\min_i(m_{i+1}-m_i)$, $i\in[1,L-1]$, что $g_i\le g_0$, $g_0<<1$, где g_0 – заданное малое число, т.е. ИППМ g-устойчива. □

Приведем достаточное условие g-устойчивости ИППМ. Рассмотрим некоторые оценки погрешностей решений задач ВМ для g-устойчивой ИППМ.

Лемма 5. Пусть ИППМ g- устойчива. Тогда для значений погрешностей Δ_i, Δ_j, Δ_{ij} имеют место двухсторонние оценки:

$$\frac{\Delta_{ij}}{1+g_{ij}} \le \Delta_i \le \frac{\Delta_{ij}}{1-g_{ij}};\ \frac{g_{ij}\Delta_{ij}}{1+g_{ij}} \le \Delta_j \le \frac{g_{ij}\Delta_{ij}}{1-g_{ij}};\ (1-g_{ij})\Delta_i \le \Delta_{ij} \le (1+g_{ij})\Delta_i, \qquad (18)$$

где $i, j \in [1, L]$. □

Теорема 5. 1. Пусть в ИППМ выполнено условие $\frac{g_{ij}}{1-g_{ij}} \le g_0 << 1,\ j > i;\ i, j \in [1, L]$. Для того, чтобы значение функции $f_{m_j}(x)$ было К-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 $\Delta_i \le \varepsilon$ и $\Delta_{ij} \le (1+g_0)\varepsilon,\ j > i$. □

Из П.2 теоремы следует, что существует такой номер i ИППМ, для которого имеет место ЗТР $\forall \varepsilon > 0$. Номеру i соответствует длина мантиссы m_i, которая, конечно, не должна превышать значения $m_{\max}$ библиотеки программ.

Препятствием к практическому применению полученных оценок является значение g_{ij}, в котором присутствует «истинное» решение $f(x)$, в общем случае неизвестное при вычислениях.

Для практического применения указанных оценок вводятся новые понятия.

Определение 14. Пусть $f_{m_L}(x)$ – КР задачи ВМ. Обозначим $g_i^L = \frac{\Delta_{i+1,L}}{\Delta_{iL}} = \frac{\left\| f_{m_{i+1}}(x) - f_{m_L}(x) \right\|}{\left\| f_{m_i}(x) - f_{m_L}(x) \right\|},\ i \in [1, L-2]$. 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}. □

В леммах 6 и 7 приводятся оценки, связывающие теоретические значения для коэффициента уменьшения g_{ij} и вычисляемые на практике величины g_{ij}^L.

В теореме 6 рассматриваются оценки погрешностей округления для g-устойчивой ИППМ, которые можно применять при численном решении.

Теорема 6. 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, $\left\| f_{m_i}(x) \right\| > \left\| \Delta_i \right\|,\ i \in [1, L]$. Тогда имеет место оценка

$$\frac{\sigma_1 \Delta_{ij}}{\left\| f_{m_j}(x) \right\|} \le \frac{\Delta_i}{\left\| f(x) \right\|} \le \sigma_2 \frac{\Delta_{ij}}{\left\| f_{m_j}(x) \right\|}, \qquad (19)$$

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. □

В §3.2 рассматриваются подходы к практической оценке погрешностей, получаемых в процессе построения ИППМ. Исследованы алгоритмы построения оценок погрешностей округлений.

В разделе 3.2.1 рассматривается алгоритм последовательной оценки погрешности округлений решений. Задается некоторое начальное значение длины мантиссы $m = m_1$. Следующие значения длин мантиссы задаем по правилу $m_{i+1} = m_i + \Delta m_{i+1}$, $i = 1,2,\ldots$. Особенностью настоящего алгоритма является то, что оценка погрешности округления (ОПО) задачи ВМ проводится после каждого решения с мантиссой большей длины. 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.

Введем числа $\sigma_A = \frac{1}{1-g_0}$ и $\sigma_0 = \frac{\sigma_A}{1-\alpha_0}$ – коэффициент коррекции погрешности, где $g_{ij} \le g_0$ и $\alpha_0 << g_0$.

1. Пусть задана требуемая точность решения ε_A и для некоторых решений $f_{m_i}(x)$, $f_{m_j}(x)$ выполнено неравенство: $\Delta_i \le \sigma_A \Delta_{ij} \equiv \Delta^1 \le \varepsilon_A$. Тогда полученная относительная погрешность решения ε^1 удовлетворяет условию:

$$\frac{\sigma_A \Delta_{ij}}{\left\| f_{m_j}(x) \right\| (1-\alpha_0)} \equiv \varepsilon^1 \le \frac{\varepsilon_A}{\left\| f_{m_j}(x) \right\| (1-\alpha_0)} \equiv \varepsilon_0 . \qquad (20)$$

2. Пусть задана требуемая точность решения ε_0 для относительной погрешности, т.е. выполняется неравенство: $\frac{\Delta_i}{\left\| f(x) \right\|} \le \frac{\sigma_0 \Delta_{ij}}{\left\| f_{m_j}(x) \right\|} \le \varepsilon_0$. Тогда для абсолютной погрешности решения Δ^1 выполняется условие:

$$\sigma_A \Delta_{ij} \equiv \Delta^1 \le \varepsilon_0 \left\| f_{m_j}(x) \right\| (1-\alpha_0) \equiv \varepsilon_A . \qquad (21)$$

В оценках (20) и (21) значению функций $\Phi_1(\cdot,\cdot)$ из (1) соответствует функция $\Phi_1(f_{m_i}, f_{m_j}) \equiv \sigma_A \Delta_{ij}$ для абсолютной погрешности и функция $\Phi_2(f_{m_i}, f_{m_j}) \equiv \frac{\sigma_0 \Delta_{ij}}{\left\| f_{m_j}(x) \right\|}$ – для относительной погрешности.

П1 (Вариант 1) В этом варианте $\sigma_0 = \frac{1}{(1-g_0)(1-\alpha_0)}$, $\Delta m_{i+1} = \Delta m = const$, $i = 1,2,...$, и для некоторого $i \geq 1$ выполнено условие $\sigma_0 \frac{\Delta_{i,i+1}}{\left\| f_{m_{i+1}}(x) \right\|} \equiv \varepsilon_i \leq \varepsilon_0$, а условия $\varepsilon_t \leq \varepsilon_0$ не выполнены для $1 \leq t \leq i-1$. Решением задачи является значение $f_{m_i}(x)$, абсолютная погрешность его не более ε_A: $\Delta_i \leq \Delta_{i,i+1}/(1-g_0)$ относительная $-\varepsilon_0$. Функции оценки ЗТР (1) $\Phi_1(\cdot,\cdot)$ здесь равны: $\Phi_1(\cdot,\cdot) = \frac{\varepsilon_0 \Delta_{i,i+1}}{\left\| f_{m_{i+1}}(x) \right\|}$ для относительной погрешности и $\Phi_1(\cdot,\cdot) = \frac{\Delta_{i,i+1}}{1-g_0}$ – для абсолютной.

П2 (Вариант 2) Пусть определены m_1, $m_2 = m_1 + \Delta m$, решения $f_{m_1}(x)$ и $f_{m_2}(x)$ и условие $\sigma_0 \frac{\Delta_{12}}{\left\| f_{m_2}(x) \right\|} \leq \varepsilon_0$ не выполнено, число $\rho_1 \approx \Delta_{12} b^{m_1}$. Оценим значение m_3 из условия $\Delta_3 = \rho_3 b^{-m_3} = \varepsilon_0 \left\| f_{m_2}(x) \right\|$. Считаем, что $\rho_3 \cong \chi \rho_1$, где χ можно брать равным $\chi = 10^2$, $\chi = 10^3$ и т.д. Если $\log_b \frac{\chi \Delta_{12}}{\varepsilon_0 \left\| f_{m_2}(x) \right\|} \leq 2\Delta m$, то $m_3 = m_2 + \Delta m$.

Если $\log_b \frac{\chi \Delta_{12}}{\varepsilon_0 \left\| f_{m_2}(x) \right\|} > 2\Delta m$, то $m_3 = m_2 + \left\lfloor \log_b \frac{\chi \Delta_{12}}{\varepsilon_0 \left\| f_{m_2}(x) \right\|} \right\rfloor$. Находим $f_{m_3}(x)$ и проверяем условие $\sigma_0 \frac{\Delta_{23}}{\left\| f_{m_3}(x) \right\|} \leq \varepsilon_0$. Если это условие выполнено, то $f_{m_2}(x)$ - решение задачи ВМ и значение $\Delta^1 \equiv \Delta_{23}/(1-g_0)$. Если не выполнено, то далее ИППМ реализуется по Варианту 1 при $m_{i+1} = m_i + i\Delta m$, $i \geq 3$.

В разделе 3.2.2 (П2) рассматривается табличный алгоритм оценки погрешностей округления решений. Пусть ИППМ решений представлена в виде: $f_{m_i}(x)$, $i = [1, L]$, $m_{i+1} = m_1 + i\Delta m$, $i \in [1, L-1]$. Значения m_1, Δm, число решений L задает Вычислитель (лицо, решающее задачу ВМ). Построим таблицу погрешностей Δ_{ij}, $j > i$, $i \in [1, L-1]$, $j \in [2, L]$. Таблица значений $\rho_i \cong \Delta_{ij} b^{m_i}$ – треугольник значений ρ_i, на i строке которого находятся приближенные значения ρ_i, $j > i$. Строится также таблица чисел g_{ij}^L и таблица оценок относительной погрешности решения: $\varepsilon_{ij} = \sigma_0 \frac{\Delta_{ij}}{\left\| f_{m_j}(x) \right\|}$, $j > i$, $i, j \in [1, L]$, где $\sigma_0 = 1 + t_0$, $t_0 \leq 0{,}1$ и т.п. Совокупность всех указанных таблиц дает информацию о величине погрешностей решений $f_{m_i}(x)$, $i \in [1, L-1]$, и об g-устойчивости ИППМ.

В разделе 3.2.3 рассматриваются методы округления решений. Полученное значение функции представляется как $f_{m_i}(x)$ m_i-значное, т.е. представление решения в 10-ичной системе имеет m_i 10-ичных знаков. Значение $f_{m_i}(x)$ округляется до t знаков, $t < m_i$.

В §3.2 рассматривается естественное обобщение сформулированных выше правил округления на случай $k \geq 2$, т.е. на случай вектор-функций.

В §3.3 исследуется метод оценки погрешности округления скалярных решений по правилу «**совпадения t первых десятичных знаков**».

В §3.4 показано, что при g-устойчивости БША для них справедливы все результаты теории, сформулированные в разделах 2 и 3, а потому методика получения заданной точности решений для БША будет той же, что и для КША.

В §3.5 рассматривается вопрос эффективности предложенного подхода. Метод КР обладает значительной универсальностью в определении погрешностей округления значений функции, т.к. он не ориентирован на какие-либо классы решаемых задач. Таким образом, если метод КР решает задачу за приемлемое время, а традиционный метод (ТМ), использующий «стандартное» программное обеспечение решает, но не дает ЗТР, то метод КР можно считать высокоэффективным по сравнению с ТМ.

В **главе 4** анализируются результаты численных экспериментов, проделанных для проверки предложенных теоретических результатов. Для практических целей использовался математический пакет Maple 13, в котором регулировка точности вычислений производится в соответствии со стандартом IEEE 854, который является расширением стандарта IEEE 754 на случай $b = 10$. Ввиду ограничений на размер автореферата, результаты приводятся для наиболее значимых экспериментов.

В §4.1 рассматривается задача вычисления полинома. Рассмотрена зависимость погрешности от m и точки x.

В §4.2 рассматриваются различные варианты методов решения СЛУ: метод Гаусса с выбором главного элемента. Рассмотрим задачу нахождения решения системы линейных уравнений (СЛУ):

$$Hz = c \tag{22}$$

где H - матрица Гильберта порядка K т.е.:

$$H = \{h_{ij}\},\ i, j \in [1, K], h_{ij} = \frac{1}{i + j - 1}. \tag{23}$$

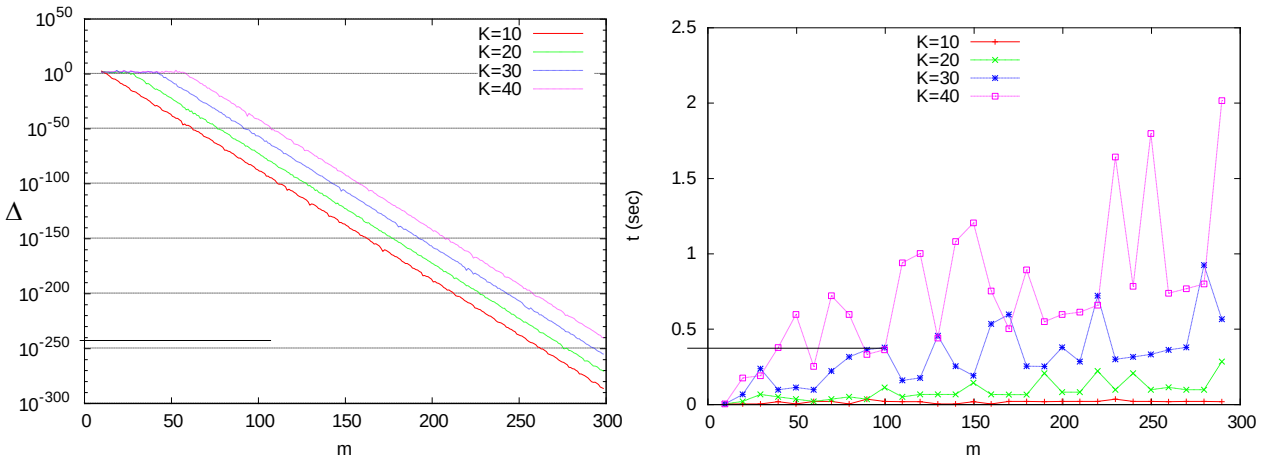

Рис. 1. Левый график: Зависимость абсолютной погрешности решения СЛУ (23) Δ_m от длины мантиссы m. Правый график: Зависимость t – времени решения от длины мантиссы m и размерности СЛУ ($K = 10,20,30,40$)

Из первого графика (Рис. 1) видно, что зависимость погрешности для плохо обусловленной матрицы имеет зону неопределенности, после чего выходит на участок экспоненциального убывания вместе с ростом длины мантиссы. На этом участке проявляется свойство g-устойчивости.

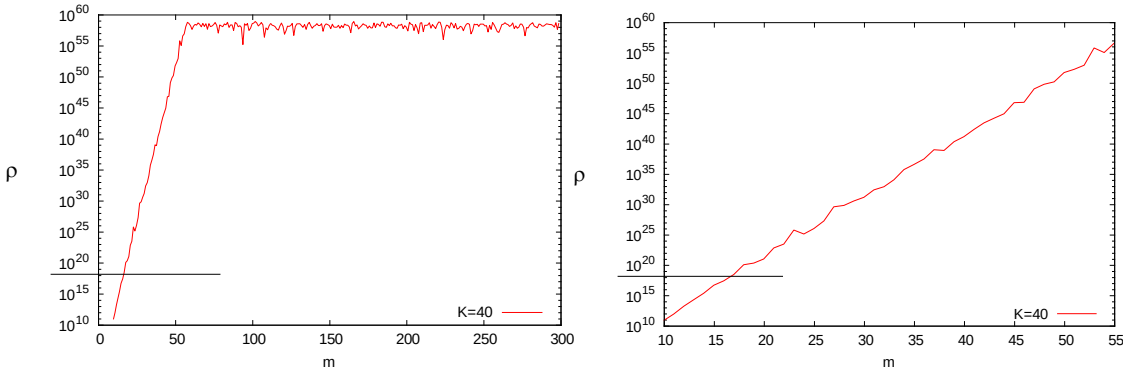

Рис. 2. Зависимость ρ в области стабильности ($m > 60$) и области роста ($m < 60$) для СЛУ вида ((23) от длины мантиссы m при $K = 40$ и шаге изменения длины мантиссы $\Delta m = 1$.

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 $m > 45$ можно трактовать как постоянное значение, на которое наложено «случайное» возмущение.

Рассмотрим различные правила варьирования длины мантиссы МЧ (ПВМ) для для нахождения решения задачи ((23) с заданной точностью $\varepsilon_0 = 10^{-20}$ в условиях ИППМ:

Для всех правил $m_1 = 100$. Критерий остановки варьирования: $\Delta_{i+1,i+2} < 0,1$ и $\Delta_{i,i+1} < 0,1$ - это грубые условия, характеризующие выход из «зоны

неопределенности». $g_{ii+1}^{i+2} = \frac{\Delta_{i+1,i+2}}{\Delta_{i,i+1}} \le g_0 = 0{,}01$ - это оценка g-устойчивости данного ИППМ; $\frac{\Delta_{i,i+1}}{\|x_i\|} \le \varepsilon_0$.

ПВМ №1 clog: Правило, на основании **П1.2 Вариант 2**, описанного в п. 3.2.

$m_{i+1} = m_i + \Delta m,\ m_{i+1} = m_i + \min\left(10, -\left\lfloor \log_b \frac{\chi \Delta_{i-1,i}}{\varepsilon_0 \| f_{m_i}(x) \|} \right\rfloor\right)$, $\chi = 10^3$ – выбран на основе анализа графиков для ρ, рассмотренных выше.

ПВМ №2 lin10: ПВМ №2. $m_{i+1} = m_i + \Delta m, \Delta m = 10$

ПВМ №3 lin20: $m_{i+1} = m_i + \Delta m, \Delta m = 20$

ПВМ №4 2x: ПВМ №4. $m_{i+1} = 2m_i$

ПВМ №5 optm: $m_i \equiv m^*$ - единственное вычисление при «необходимой» длине мантиссы.

ПВМ №6 m1000: $m_i = 1000$ – правило с достаточно большой длиной мантиссы и единственным вычислением.

Эксперимент проведён для матриц Гильберта порядка N=100,200,300.

Подход	N	t, сек	m_i	m_L	ε	Δ_{iL}	i
clog	100	30.563	175	254	7.48402e-103	2.53773e-25	6
lin10		55.234	180	200	9.86545e-49	6.83909e-29	11
lin20		36	180	220	1.85175e-68	6.83909e-29	7
2x		32.954	200	800	1.34404e-648	9.86545e-49	4
optm		4.922	175	175	2.53773e-25		1
m1000		18.203	1000	1000	1.47395e-848		1
clog	200	542.266	350	499	2.25154e-194	1.46094e-46	9
lin10		908.375	330	350	1.46094e-46	2.87443e-27	16
lin20		579.719	340	380	2.96797e-75	1.53653e-36	10
2x		589.953	400	1600	2.33753e-1296	4.24806e-96	4
optm		57.063	350	350	1.46094e-46		1
m1000		168.859	1000	1000	4.28177e-696		1
clog	300	3651.16	500	638	5.88136e-181	8.38046e-42	11
lin10		6545.81	480	500	8.38046e-42	1.19533e-21	21
lin20		3819.86	480	520	1.94937e-61	1.19533e-21	12
2x		3349.88	600	2400	1.76402e-1941	9.43380e-142	4
optm		248.844	480	480	8.38046e-42		1

m1000		517.891	1000	1000	8.60282e-542		1

Таблица 1. Зависимость времени решения t для СЛУ вида ((23) от ПВМ при $N=100,200,300$, где m_i – достаточная длина мантиссы, ε - истинная погрешность полученного решения, Δ_{iL} – практическая оценка погрешности, i – число шагов и ИППМ.

Из результатов эксперимента (Таблица 1) видно, что подход с варьированием мантиссы позволяет получать решение с требуемой точностью вне зависимости от способа варьирования. Далее в работе для решения СЛУ с матрицей Гильберта рассмотрен метод сопряженных градиентов. Произведены оценки коэффициентов g_{ij}^{L} и их сравнение с теоретическим результатом. На основании этого сделан вывод о g-устойчивости данных методов. Рассмотрено поведение параметра погрешности $C(x,m)$.

В §4.3 рассматривается задача численного нахождения производной. Показано, что параметр погрешности $\alpha=1/2$ при соответствующем выборе шага дифференцирования $h\sim\sqrt{\delta_\mu}$. Изучен вопрос g-устойчивости ИППМ для этой задачи.

В §4.4 задача нахождения собственных чисел методом Данилевского. На эксперименте показано, что ИППМ для метода Данилевского является g-устойчивой начиная с некоторой длины мантиссы m.

В §4.5 рассмотрены задачи безусловной минимизации и решения систем нелинейных уравнений. Для метода Ньютона приведены формулы для предсказания Δm при варьировании мантиссы на основании правил изложенных в главе 3. Рассмотрены другие правила к варьированию мантиссы и показана эффективность предложенного метода предсказания. В частности, произведен следующий численный эксперимент.

Задача БМ Б (Функция Розенброка 1)

$$f(x)=\sum_{i=1}^{n-1}(1-x_i)^2+\alpha\left[0{,}5\cdot n\cdot(n-1)\cdot x_n-\sum_{t=1}^{n-1}t\cdot x_t^p\right]^2, \text{ где} \qquad (24)$$

$\alpha>0$, $p=1,2,3\ldots$. Для функции (24) известно точное решение – $x_i^*=1, i=\overline{1,n}$, а значение $f(x^*)=0$, причем решение (точка x^*) у этой функции – единственное. При численном эксперименте принимались следующие *общие условия*:

а) Алгоритм заканчивал работу, если $\Delta x_k\equiv\|\lambda_k p_k\|\le\varepsilon_1$ и $\|\nabla f(x_{k+1})\|\le\varepsilon_2$ где $\varepsilon_1=\varepsilon_2=10^{-20}$ – требуемая точность. При такой величине требуемой точности, стандартной арифметики двойной точности уже недостаточно и варьирование мантиссы становится необходимым.

б) Шаг λ_k находится из условия $x^* = \arg\min_{x \in R^n} f(x_{k-1} + \lambda p_k)$. Одномерная минимизация происходит с помощью метода золотого сечения. Причем критерий остановки одномерной минимизации – погрешность $\Delta\lambda \le b^{-\left\lfloor \frac{\mathrm{m}}{3} \right\rfloor}$.

Для нахождения требуемой длины мантиссы использовались различные ПВМ. Общий подход к реализации ИППМ состоит в использовании последовательности длин мантисс $\mathrm{m}_g = \mathrm{m}_1 < \mathrm{m}_2 < \mathrm{m}_3 \ldots$, где $g = 1,2,\ldots$ – группы вычислений с указанным значением мантиссы. Для каждой группы вычислений определены следующие параметры: 1. Требуемая точность для данной группы: $\varepsilon_i = \max\left(10^{-\left\lfloor \frac{m}{3} \right\rfloor}, \varepsilon_1\right)$. 2. Параметр перехода. Способ определения точности для каждой группы: $\mathrm{m}_{i+1} = T(\mathrm{m}_i, x_k, F)$. Для первой группы $m_1 = T_1(x_0, F, \varepsilon_1)$. 3. Шаг численного дифференцирования в зависимости от m_i: $h_i = 10^{-\left\lfloor \frac{\mathrm{m}_i}{3} \right\rfloor}$.

При вычислениях в каждой группе используются следующие условия окончания: Условие окончания 1: Если $\|x_k - x_{k-1}\| \le \varepsilon_i$, то происходит увеличение мантиссы и вычисления продолжаются для следующей группы начиная с достигнутого x_k. Условие окончания 2: Если количество шагов в данной группе превосходит $n = N_i$, то происходит переход к следующей группе. Условие окончания 3: Если на любом шаге группы срабатывает более сильное условие полного окончания а) то алгоритм считается завершенным.

Рассмотрены ПВМ, аналогичные приведенным ранее для СЛУ:

ПВМ №1: (clog): $m_{i+1} = m_{i-1} + \max\left\lfloor 0, \frac{1}{\alpha}\log_b \frac{\varepsilon_1 \|x_k\|}{\chi \|\lambda_k p_k\|} \right\rfloor$, $m_1 = 15$.

ПВМ №2: (lin10): $m_{i+1} = m_i + 10$, $m_1 = 15$

ПВМ №3: (lin20): $m_{i+1} = m_i + 20$, $m_1 = 15$

ПВМ №5: (2x): $m_{i+1} = 2 \cdot m_i$, $m_1 = 15$

ПВМ №5: (m*): использует m^* – наименьшую из достаточных точностей полученную на одном из предыдущих шагов. $m_{i+1} = m_i, m_1 = m^*$.

ПВМ №6 (m100): В этом подходе фиксированная достаточно большая длина мантиссы 100, т.е. $m_{i+1} = m_i, m = 100$.

Подход	N	m^*	t, сек	$\|\nabla f(x_k)\|$	$\|x - x_k\|$	$f_k - f^*$	k	g
clog	5	52	2.628	9.28103e-26	4.64229e-21	9.80325e-55	55	31
lin10		75	5.016	4.68859e-36	3.41326e-28	4.23552e-76	33	7
lin20		135	5.19	4.68869e-36	3.41303e-28	4.23540e-76	33	7
2x		960	5.72	4.68877e-36	3.41264e-28	4.23510e-76	33	7

optm		52	6.781	4.55641e-29	1.04900e-33	1.02147e-61	33	7
m100		100	8.327	6.98179e-54	1.34388e-33	1.35510e-111	33	7
clog	10	43	13.727	4.17797e-23	8.68153e-21	1.56645e-51	94	50
lin10		105	26.046	2.52116e-34	2.61099e-23	9.24818e-74	55	10
lin20		195	28.048	2.52112e-34	2.61112e-23	9.24846e-74	55	10
2x		7680	78.735	2.52108e-34	2.61126e-23	9.24875e-74	55	10
optm		43	28.563	9.12954e-22	2.66973e-23	4.99117e-48	54	10
m100		100	33.469	3.96826e-41	2.69291e-23	2.72010e-87	55	10

Таблица 2. Время решения задачи (24) с ЗТР $\varepsilon_1 = \varepsilon_2 = 10^{-20}$ с помощью различных ПВМ, градиент вычисляется численно. $\|x - x_k\|$ – погрешность полученного решения.

Сравнительные результаты применения различных подходов к варьированию мантиссы показывают, что предложенный адаптивный подход (ПВМ №1) для предсказания Δm, в целом дает эффективный метод для варьирования длин мантисс в ИППМ.

В **заключении** приведены основные результаты диссертации.

В **приложении 1** рассмотрены различные подходы к варьированию мантисс применительно к методу штрафных функций.

В **приложении 2** рассмотрены сеточные методы для решения задачи Коши и уравнения теплопроводности. В частности, исследован вопрос о возможности связывания шага сетки h с длиной мантиссы m по аналогии с задачей о численном дифференцировании.

Основные результаты диссертации

1. Введены понятия элементарного алгоритма решений задач вычислительной математики, конечношагового (КША) и бесконечношагового (БША) алгоритмов. Для элементарного алгоритма получены оценки погрешности решений как для КША, так и для БША, зависящие от длины мантиссы и аргументов алгоритмов. Для КША получены оценки длины мантиссы, обеспечивающие достижение требуемой точности значения функции.
2. Введены понятие контрольного решения (КР) задачи, которое в оценках погрешностей решений с некоторой точностью заменяет истинное значение функции и понятие g-устойчивой последовательности решений. Исследованы свойства g-устойчивости, в том числе доказана теорема о достижимости заданной точности значения функции; получены оценки погрешности, далее численно реализуемые в итерационной последовательности с переменной мантиссой.

3. Предложены алгоритмы, позволяющие оптимизировать процесс решения задачи в итерационной последовательности с переменной мантиссой. Рассмотрены методы округления полученных решений, причем округленное решение имеет заданную точность.

Автор выражает глубокую благодарность своему научному руководителю доценту Бирюкову А.Г. за постановку задачи, ценные советы, терпение и постоянное внимание к работе.

Список публикаций автора по теме диссертации

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. *Бирюков А.Г., Гриневич А.И.* Метод оценки погрешностей округления решений задач вычислительной математики в арифметике с плавающей запятой, основанный на сравнении решений с изменяемой длиной мантиссы машинного числа // Труды МФТИ. – 2013 – Том 5, № 2(18), С.160-174.
3. *Гриневич А.*И. Математическая модель погрешностей округления при вычислениях в арифметике с плавающей запятой и переменной длиной мантиссы. // Труды 55-й научной конференции / Управление и прикладная математика. Том 1. – М.: МФТИ, 2012, С.15-16.
4. *Бирюков А.Г., Гриневич А.И.* Итерационный процесс с переменной длиной мантиссы для решения задач вычислительной математики с заданной точностью // Труды 55-й научной конференции / Управление и прикладная математика. Том 1. – М.: МФТИ, 2012, С. 86-87.
5. *Бирюков А.Г., Гриневич А.И.* Анализ погрешностей некоторых алгоритмов безусловной минимизации. Труды Института системного анализа РАН. Динамика неоднородных систем. Том 42(1), 2009, С. 106-111.
6. *Бирюков А.Г., Гриневич А.И.* Об эффективности и устойчивости численных методов решения систем нелинейных уравнений и задач безусловной минимизации // Труды Института системного анализа РАН. Динамика линейных и нелинейных систем. Том 25(1), 2006, С. 111-123.

Гриневич Алексей Иванович

МЕТОД ОЦЕНКИ ПОГРЕШНОСТИ ОКРУГЛЕНИЙ ЗНАЧЕНИЙ ВЫЧИСЛЯЕМОЙ ФУНКЦИИ, ОСНОВАННЫЙ НА ВАРЬИРОВАНИИ ДЛИНЫ МАНТИССЫ В АРИФМЕТИКЕ С ПЛАВАЮЩЕЙ ЗАПЯТОЙ

Автореферат

Подписано в печать 12.04.2013. Формат 60 × 84 1/16.
Усл. печ. л. 1,0. Тираж 100 экз. Заказ № 275
Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Московский физико-технический институт (государственный университет)»
Отдел оперативной полиграфии «Физтех-полиграф»
141700, Московская обл., г. Долгопрудный, Институтский пер., 9
E-mail: rio@mail.mipt.ru