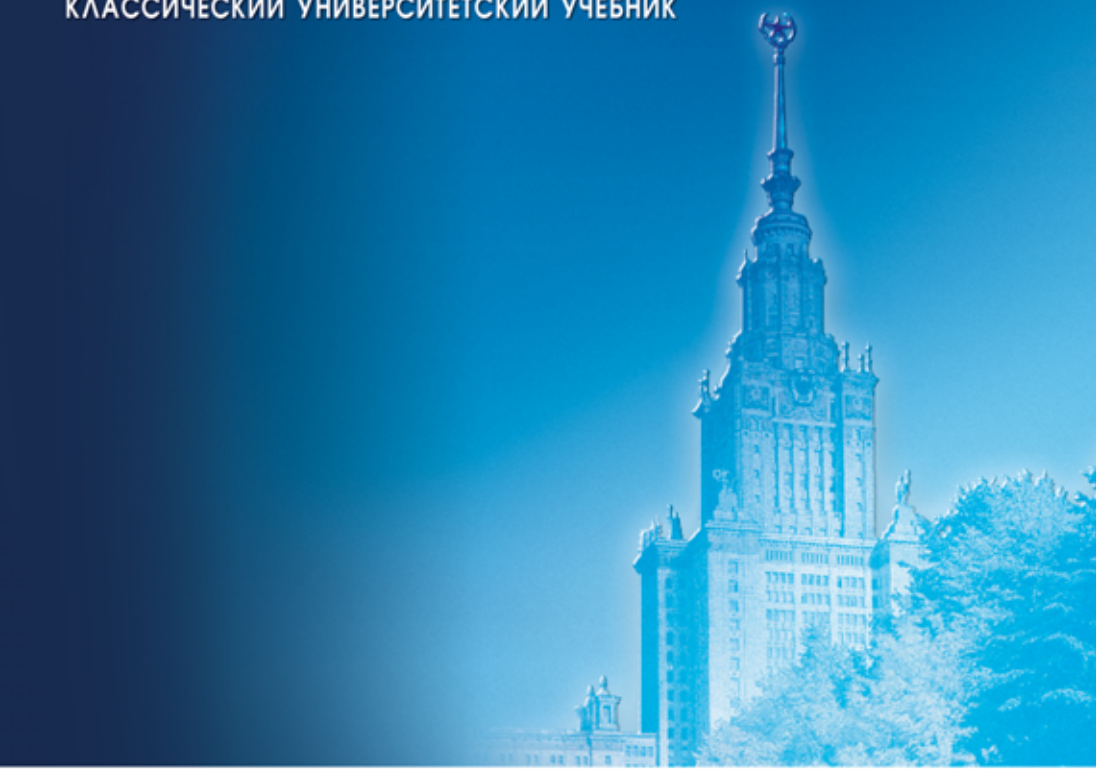
Московский государственный университет имени М.В. Ломоносова

КЛАССИЧЕСКИЙ УНИВЕРСИТЕТСКИЙ УЧЕБНИК



Н. С. Бахвалов, А. А. Корнев, Е. В. Чижонков

ЧИСЛЕННЫЕ МЕТОДЫ
РЕШЕНИЯ ЗАДАЧ И УПРАЖНЕНИЯ

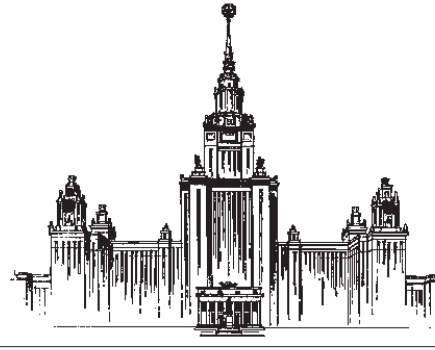
Серия

КЛАССИЧЕСКИЙ
УНИВЕРСИТЕТСКИЙ УЧЕБНИК

основана в 2002 году по инициативе ректора
МГУ им. М.В. Ломоносова
академика РАН В.А. Садовничего
и посвящена

250-летию
Московского университета

КЛАССИЧЕСКИЙ
УНИВЕРСИТЕТСКИЙ УЧЕБНИК



Уважаемый читатель!

Вы открыли одну из замечательных книг, изданных в серии «Классический университетский учебник», посвященной 250-летию Московского университета. Серия включает свыше 150 учебников и учебных пособий, рекомендованных к изданию Учеными советами факультетов, редакционным советом серии и издаваемых к юбилею по решению Ученого совета МГУ.

Московский университет всегда славился своими профессорами и преподавателями, воспитавшими не одно поколение студентов, впоследствии внесших заметный вклад в развитие нашей страны, составивших гордость отечественной и мировой науки, культуры и образования.

Высокий уровень образования, которое дает Московский университет, в первую очередь обеспечивается высоким уровнем написанных выдающимися учеными и педагогами учебников и учебных пособий, в которых сочетаются как глубина, так и доступность излагаемого материала. В этих книгах аккумулируется бесценный опыт методики и методологии преподавания, который становится достоянием не только Московского университета, но и других университетов России и всего мира.

Издание серии «Классический университетский учебник» наглядно демонстрирует тот вклад, который вносит Московский университет в классическое университетское образование в нашей стране и, несомненно, служит его развитию.

Решение этой благородной задачи было бы невозможным без активной помощи со стороны издательств, принявших участие в издании книг серии «Классический университетский учебник». Мы расцениваем это как поддержку ими позиции, которую занимает Московский университет в вопросах науки и образования. Это служит также свидетельством того, что 250-летний юбилей Московского университета — выдающееся событие в жизни всей нашей страны, мирового образовательного сообщества.

Ректор Московского университета
академик РАН, профессор *В. А. Садовничий*

Московский государственный университет имени М. В. Ломоносова

Н. С. Бахвалов, А. А. Корнев, Е. В. Чижонков

ЧИСЛЕННЫЕ МЕТОДЫ
РЕШЕНИЯ ЗАДАЧ И УПРАЖНЕНИЯ

2-е издание,
исправленное и дополненное

Допущено
УМО по классическому университетскому образованию
в качестве учебного пособия для студентов высших
учебных заведений, обучающихся по специальности
01.05.01 «Фундаментальные математика и механика»

Москва
Лаборатория знаний

УДК 519.6(075.8)
ББК 22.193я73
Б30

Печатается
по решению Ученого совета
Московского государственного университета
имени М. В. Ломоносова

Бахвалов Н. С.

Б30 Численные методы. Решения задач и упражнения : учебное пособие для вузов / Н. С. Бахвалов, А. А. Корнев, Е. В. Чижонков. — 2-е изд., испр. и доп. — М. : Лаборатория знаний, 2016. — 352 с. : ил. — (Классический университетский учебник).

ISBN 978-5-906828-04-0

Материал пособия соответствует программе курса «Численные методы», рекомендованной Министерством образования и науки РФ. Содержатся основные положения теории, большое количество подробно разобранных примеров, которые являются основой для компьютерного решения практических и учебных задач различного уровня сложности — от домашних упражнений до курсовых и дипломных работ. Включены упражнения для самостоятельной работы.

Книга такого типа по численным методам не имеет аналогов как в нашей стране, так и за рубежом.

Для студентов университетов, педагогических вузов, вузов с углубленным изучением математики, а также для студентов технических вузов, аспирантов и преподавателей, инженеров и научных работников, использующих в практической деятельности численные методы.

УДК 519.6(075.8)
ББК 22.193я73

Учебное издание

Серия: «Классический университетский учебник»

Бахвалов Николай Сергеевич, **Корнев** Андрей Алексеевич,
Чижонков Евгений Владимирович

ЧИСЛЕННЫЕ МЕТОДЫ.
РЕШЕНИЯ ЗАДАЧ И УПРАЖНЕНИЯ

Учебное пособие для вузов

Ведущий редактор *М. С. Стригунова*. Художественный редактор *В. Е. Шкерин*
Оригинал-макет подготовлен *О. Г. Лапко* в пакете LaTeX 2_ε

Подписано в печать 29.09.15. Формат $70\times100/16$.
Усл. печ. л. 28,60. Тираж 500 экз. Заказ

Издательство «Лаборатория знаний»
125167, Москва, проезд Аэропорта, д. 3
Телефон: (499) 157-5272, e-mail: info@pilotLZ.ru, http://www.pilotLZ.ru

---

ISBN 978-5-906828-04-0

Оглавление

Предисловие

Учебное пособие написано на основе многолетнего опыта преподавания численных методов студентам механико-математического факультета и факультета вычислительной математики и кибернетики МГУ им. М. В. Ломоносова и полностью соответствует требованиям Государственного образовательного стандарта по математике, рекомендованного Министерством образования Российской Федерации.

Как правило, классический университетский курс, ориентированный на приближенное решение задач, состоит из теоретической (лекции) и практической (семинары) частей и сопровождается лабораторными работами. Поэтому учебная литература традиционно представлена теоретическими учебниками, сборниками задач и вычислительными практикумами. Предлагаемая вниманию читателя книга содержит в форме задач и упражнений наиболее ценные, по мнению авторов, сведения по численным методам из пособий всех указанных типов, и ее можно использовать не только в учебных, но и в справочных целях.

Пособие охватывает материал по разностным уравнениям, приближению функций, численному интегрированию и дифференцированию, интегральным уравнениям, задачам алгебры и решению нелинейных уравнений, приближенным методам решения дифференциальных уравнений как обыкновенных, так и с частными производными, а также по влиянию вычислительной погрешности в различных алгоритмах.

Главная цель пособия — помочь читателю глубоко и последовательно освоить предмет. Для этого материал разбит на крупные теоретические части — главы и, кроме того, на темы — параграфы, содержание которых структурировано специальным образом. Изучение каждой новой темы начинается со знакомства с основными определениями, формулировками фундаментальных теоретических результатов (теорем), полезными вспомогательными фактами и т. п., затем разбираются и анализируются типичные упражнения, отражающие специфику постановок задач и методы их решений. Первые задачи каждого параграфа решены подробно и сопровождаются комментариями. Сложность задач постепенно возрастает, поэтому нередки ссылки на уже разобранные примеры. Далее приводятся упражнения для самостоятельных занятий. Они, как правило, достаточно разнообразны и могут удовлетворить запросы большинства читателей. Затем содержатся наборы из нескольких упражнений, которые при одинаковом задании имеют различные условия. Это — образцы для контрольных работ по изучаемой теме, они сопровождены только ответами. В конце каждого параграфа имеются упражнения повышенной сложности, как правило, снабженные только указаниями и/или ответами. Их целесообразно использовать в качестве зачетных задач или как основу для небольших курсовых проектов.

Важная методическая особенность пособия — расположение решений, указаний и ответов непосредственно за условиями задач и упражнений, а не в конце книги, как это принято в задачниках. Тщательный отбор и подача материала в такой форме способствуют эффективному усвоению численных методов даже при самостоятельной работе. Поэтому данное учебное пособие рекомендуется студентам, аспирантам и преподавателям высших учебных заведений с углубленным изучением математики и всем, кто по роду своей деятельности сталкивается с приближенным решением задач, допускающих математическую формулировку. Даже специалист в области вычислительной математики может найти сформулированные в виде упражнений необычные формулы, факты, утверждения, неизвестные ему ранее. Например, различные численные аспекты решения уравнения Шрёдингера и задачи Стокса.

Критические замечания и предложения по совершенствованию книги просьба сообщать авторам на кафедру вычислительной математики механико-математического факультета МГУ им. М. В. Ломоносова.

Авторы.

Глава 1

Погрешность решения задачи

Если a — точное значение некоторой величины, a^* — известное приближение к нему, то *абсолютной погрешностью* приближенного значения a^* обычно называют некоторую величину $\Delta(a^*)$, про которую известно, что

$$|a^* - a| \leqslant \Delta(a^*)\,.$$

Относительной погрешностью приближенного значения называют некоторую величину $\delta(a^*)$, про которую известно, что

$$\left|\frac{a^* - a}{a^*}\right| \leqslant \delta(a^*)\,.$$

Относительную погрешность часто выражают в процентах.

В этой главе на модельных упражнениях показано принципиальное отличие между математически точными вычислениями и вычислениями с произвольно высокой, но конечной точностью. Приведены примеры *катастрофического* накопления вычислительной погрешности в стандартных алгоритмах, рассмотрены методы возможного улучшения исследуемых алгоритмов.

1.1. Вычислительная погрешность

Наиболее распространенная форма представления действительных чисел в компьютерах — *числа с плавающей точкой*. Множество F чисел с плавающей точкой характеризуется четырьмя параметрами: основанием системы счисления p, разрядностью t и интервалом показателей $[L, U]$. Каждое число x, принадлежащее F, представимо в виде

$$x = \pm\left(\frac{d_1}{p} + \frac{d_2}{p^2} + \cdots + \frac{d_t}{p^t}\right) p^{\alpha}\,,$$

где целые числа $p, \alpha, d_1, \ldots, d_t$ удовлетворяют неравенствам $0 \leqslant d_i \leqslant p-1$, $i = 1, \ldots, t$; $L \leqslant \alpha \leqslant U$. Часто d_i называют *разрядами*, t — *длиной мантиссы*, α — *порядком числа*. *Мантиссой* (дробной частью) x называют число в скобках. Множество F называют *нормализованным*, если для каждого $x \neq 0$ справедливо условие $d_1 \neq 0$.

Удобно определить, что округление с точностью ε — это некоторое отображение fl действительных чисел $\mathbf{R}$ на множество F чисел с плавающей точкой, удовлетворяющее следующим аксиомам.

1) Для произвольного $y \in \mathbf{R}$ такого, что результат отображения $fl(y) \in F$, имеет место равенство при $fl(y) \neq 0$

$$fl(y) = y\,(1+\eta), \quad |\eta| \leqslant \varepsilon\,.$$

2) Обозначим результат арифметической операции $*$ с числами $a, b \in F$ через $fl(a * b)$. Если $fl(a * b) \neq 0$, то

$$fl(a * b) = (a * b)(1+\eta), \quad |\eta| \leqslant \varepsilon\,.$$

Приведенные соотношения позволяют изучать влияние ошибок округления в различных алгоритмах.

Если результат округления не принадлежит F, то его обычно называют *переполнением* и обозначают ∞.

Будем считать, что ε — точная верхняя грань для $|\eta|$. При традиционном способе округления чисел имеем $\varepsilon = \frac{1}{2}\,p^{1-t}$, при округлении отбрасыванием разрядов $\varepsilon = p^{1-t}$. Величину ε часто называют *машинной точностью*.

1.1. Построить нормализованное множество F с параметрами $p = 2$, $t = 3$, $L = -1$, $U = 2$.

$\lhd$ Каждый элемент $x \in F$ имеет вид

$$x = \pm\left(\frac{d_1}{2} + \frac{d_2}{4} + \frac{d_3}{8}\right) 2^{\alpha}\,, \text{ где } \alpha \in \{-1, 0, 1, 2\},\ d_i \in \{0, 1\}$$

и $d_1 \neq 0$ для $x \neq 0$.

Зафиксируем различные значения мантисс m_i для ненулевых элементов множества:

$$\frac{1}{2}\,, \quad \frac{1}{2} + \frac{1}{8} = \frac{5}{8}\,, \quad \frac{1}{2} + \frac{1}{4} = \frac{3}{4}\,, \quad \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = \frac{7}{8}\,,$$

или $m_i \in \left\{\frac{1}{2}\,,\ \frac{5}{8}\,,\ \frac{3}{4}\,,\ \frac{7}{8}\right\}$. Далее, умножая m_i на 2^{α} с $\alpha \in \{-1, 0, 1, 2\}$ и добавляя знаки $\pm$, получим все ненулевые элементы множества F: $\pm\frac{1}{4}$, $\pm\frac{5}{16}$, $\pm\frac{3}{8}$, $\pm\frac{7}{16}$, $\pm\frac{1}{2}$, $\pm\frac{5}{8}$, $\pm\frac{3}{4}$, $\pm\frac{7}{8}$, ± 1, $\pm\frac{5}{4}$, $\pm\frac{3}{2}$, $\pm\frac{7}{4}$, ± 2, $\pm\frac{5}{2}$, ± 3, $\pm\frac{7}{2}$. После добавления к ним числа *нуль* имеем искомую модель системы действительных чисел с плавающей точкой. $\rhd$

1.2. Сколько элементов содержит нормализованное множество F с параметрами p, t, L, U?

Ответ: $2\,(p-1)\,p^{t-1}(U - L + 1) + 1$.

1.3. Каков результат операций $fl(x)$ при использовании модельной системы из 1.1 для следующих значений x:

$$\frac{23}{32}\ ,\ \frac{1}{8}\ ,\ 4\,,\ \frac{1}{2} + \frac{3}{4}\ ,\ \frac{3}{8} + \frac{5}{4}\ ,\ 3 + \frac{7}{2}\ ,\ \frac{7}{16} - \frac{3}{8}\ ,\ \frac{1}{4} \cdot \frac{5}{16}\ .$$

Ответ: $\frac{3}{4}$, 0, ∞ $\left(x > \frac{7}{2}\right)$, $\frac{5}{4}$, $\frac{3}{2}$ или $\frac{7}{4}$, ∞, 0, 0.

1.4. Верно ли, что всегда $fl\left(\frac{a+b}{2}\right) \in [a, b]$?

Ответ: нет (см. 1.3).

1.5. Пусть отыскивается наименьший корень уравнения $y^2-140y+1=0$. Вычисления производятся в десятичной системе счисления, причем в мантиссе числа после округления удерживается четыре разряда. Какая из формул $y=70-\sqrt{4899}$ или $y=\dfrac{1}{70+\sqrt{4899}}$ дает более точный результат?

$\lhd$ Воспользуемся первой формулой. Так как $\sqrt{4899}=69,992...$, то после округления получаем $\sqrt{4899}\approx 69,99$, $\quad y_1\approx 70-69,99=0,\underline{01}$.

Вторая формула представляет собой результат «избавления от иррациональности в числителе» первой формулы. Последовательно вычисляя, получаем $70+69,99=139,99\approx 140,0$, $\frac{1}{140}=0,00714285\ldots$. Наконец, после последнего округления имеем $y_2=0,00\underline{7143}$.

Если произвести вычисления с большим количеством разрядов, то можно проверить, что в y_1 и y_2 все подчеркнутые цифры результата верные; однако во втором случае точность результата значительно выше. В первом случае пришлось вычитать близкие числа, что привело к эффекту *пропадания значащих цифр*, часто существенно искажающему конечный результат вычислений. Увеличение абсолютной погрешности также может происходить в результате деления на малое (умножение на большое) число. Еще одна опасность — выход за диапазон допустимых значений в промежуточных вычислениях, например после умножения исходного уравнения на достаточно большое число. $\rhd$

1.6. Пусть приближенное значение производной функции $f(x)$ определяется при $h\ll 1$ по формуле $f'(x)\approx\dfrac{f(x+h)-f(x-h)}{2h}$, а самими значения $f(x)$ вычисляются с абсолютной погрешностью Δ. Какую погрешность можно ожидать при вычислении производной, если $\left|f^{(k)}(x)\right|\leqslant M_k$, $\quad k=0,1,\ldots$?

$\lhd$ В данном случае имеется два источника погрешности: *погрешность метода* и *вычислительная погрешность*. Первая связана с неточностью формулы в правой части при отсутствии ошибок округления. Разложим функцию $f(x\pm h)$ в ряд Тейлора в точке x:

$$f(x\pm h)=f(x)\pm h\,f'(x)+\frac{h^2}{2}\,f''(x)\pm\frac{h^3}{6}\,f'''(x_\pm).$$

Подставляя полученные разложения в правую часть приближенного равенства, получим

$$\frac{f(x+h)-f(x-h)}{2h}=f'(x)+\frac{h^2}{6}\left[\frac{f'''(x_+)+f'''(x_-)}{2}\right].$$

Ограничиваясь главным членом в разложении по степеням h, имеем оценку для погрешности метода

$$\left|\frac{f(x+h)-f(x-h)}{2h}-f'(x)\right|\leqslant\frac{h^2}{6}\,M_3\,.$$

С другой стороны, в силу наличия ошибок округления в вычислениях участвуют не точные значения $f(x\pm h)$, а их приближения $f^*(x\pm h)$ с заданной абсолютной погрешностью. Поэтому полная погрешность выглядит так:

$$Err = \left| \frac{f^*(x+h) - f^*(x-h)}{2h} - f'(x) \right| .$$

Добавляя в числитель дроби $\pm f(x+h)$ и $\pm f(x-h)$, после перегруппировки слагаемых получим

$$Err \leqslant \left| \frac{f^*(x+h)-f(x+h)}{2h} - \frac{f^*(x-h)-f(x-h)}{2h} \right| + \left| \frac{f(x+h)-f(x-h)}{2h} - f'(x) \right| .$$

Оценка вычислительной погрешности для каждого из двух первых слагаемых имеет вид $\frac{\Delta}{2\,h}$, а погрешность метода в предположении ограниченности третьей производной получена выше. Окончательно имеем $Err \leqslant \frac{\Delta}{h} + \frac{h^2}{6}\, M_3$.

Зависимость такого рода при малых h наблюдается при численных экспериментах: при уменьшении h сначала погрешность квадратично убывает, а затем линейно растет; начиная с некоторого h ошибка может стать больше, чем сама производная $f'(x)$. Здесь эффект пропадания значащих цифр (см. 1.5) усиливается за счет деления на малую величину. $\triangleright$

Ответ: $Err \leqslant \frac{\Delta}{h} + \frac{h^2}{6}\, M_3$.

1.7. Найти абсолютную погрешность вычисления суммы $S = \sum\limits_{j=1}^{n} x_j$, где все x_j — числа одного знака.

$\triangleleft$ Используя аксиому

$$fl(a+b) = (a+b)(1+\eta), \quad |\eta| \leqslant \tfrac{1}{2}\, p^{1-t},$$

имеем

$$fl\,(S) = (\ldots((x_1+x_2)(1+\eta_2)+x_3)(1+\eta_3)+\ldots+x_n)(1+\eta_n) =$$
$$= (x_1+x_2)\prod_{j=1}^{n-1}(1+\eta_{j+1}) + x_3\prod_{j=2}^{n-1}(1+\eta_{j+1}) + \ldots + x_n\prod_{j=n-1}^{n-1}(1+\eta_{j+1}) .$$

Перепишем полученное выражение в виде

$$fl\,(S) = \sum_{j=1}^{n} x_j\,(1+E_j)\ ,$$

где для модулей E_j справедливы равенства

$$|E_1| = \tfrac{n-1}{2}\, p^{1-t} + O(p^{2(1-t)})\,,$$

$$|E_i| = \left| \prod_{j=i-1}^{n-1}(1+\eta_{j+1}) \right| = \tfrac{n+1-i}{2}\, p^{1-t} + O(p^{2(1-t)})$$

при $2 \leqslant i \leqslant n$.

Найденное представление означает, что суммирование чисел на компьютере в режиме с плавающей точкой эквивалентно точному суммированию с относительным возмущением E_j в слагаемом x_j. При этом относительные возмущения неодинаковы: они максимальны в первых слагаемых и минимальны в последних. Абсолютная погрешность Δ вычисления суммы равна $\Delta = \sum\limits_{j=1}^{n} |x_j|\,|E_j|$. Оценки E_j не зависят от x_j, поэтому в общем случае погрешность Δ будет наименьшей, если числа суммировать в порядке возрастания их абсолютных значений начиная с наименьшего. $\triangleright$

Ответ: $\Delta = \sum\limits_{j=1}^{n} |x_j|\,|E_j|$.

1.8. Пусть вычисляется сумма $\sum\limits_{j=1}^{10^6} \frac{1}{j^2}$. Какой алгоритм $S_0 = 0$, $S_n = S_{n-1} + \frac{1}{n^2}$, $n = 1, \ldots, 10^6$, или $R_{10^6+1} = 0$, $R_{n-1} = R_n + \frac{1}{n^2}$, $n = 10^6, \ldots, 1$, $\tilde{S}_{10^6} = R_0$, следует использовать, чтобы суммарная вычислительная погрешность была меньше?

Ответ: следует воспользоваться вторым алгоритмом (см. решение 1.7).

1.9. Можно ли непосредственными вычислениями проверить, что ряд $\sum\limits_{j=1}^{\infty} \frac{1}{j}$ расходится?

1.10. Предложить способ вычисления суммы, состоящей из слагаемых одного знака, минимизирующий влияние вычислительной погрешности.

$\triangleleft$ Рассмотрим оценки величин E_j из 1.7. Имеем

$$|E_1| = \frac{n-1}{2}\, p^{1-t} + O(p^{2(1-t)})\,,$$

$$|E_i| = \frac{n+1-i}{2}\, p^{1-t} + O(p^{2(1-t)})\,,\ 2 \leqslant i \leqslant n\,.$$

Из этих оценок следует, что $\left|\frac{E_1}{E_n}\right| \approx n$, т. е. первое слагаемое вносит возмущение примерно в n раз большее, чем последнее. Неравноправие слагаемых объясняется тем, что в образовании погрешностей каждое слагаемое участвует столько раз, сколько суммируются зависящие от него частичные суммы.

Влияние всех слагаемых можно уравнять с помощью следующего приема. Пусть для простоты количество слагаемых равно $n = 2^k$. На первом этапе разобьем близкие слагаемые x_j на пары и сложим каждую из них. При этом в каждое слагаемое вносится относительное возмущение одного порядка. Далее будем складывать уже полученные суммы. Для этого повторяем процесс разбиения и попарного суммирования до тех пор, пока получающиеся суммы не превратятся в одно число (степень двойки 2^k

нужна только здесь). Абсолютная погрешность по-прежнему имеет вид $\Delta = \sum\limits_{j=1}^{n} |x_j|\,|\tilde{E}_j|$, но теперь для всех $\tilde{E}_j$ справедлива оценка

$$\left|\tilde{E}_j\right| = \frac{1+\log_2 n}{2}\; p^{1-t} + O(p^{2(1-t)})\,,\; 1 \leqslant j \leqslant n\,.$$

Таким образом, меняя только порядок суммирования можно уменьшить оценку погрешности примерно в $\frac{n}{\log_2 n}$ раз. Значения $\tilde{E}_j$ отличаются от E_j в силу другого порядка суммирования. $\triangleright$

1.11. Предложить способ вычисления знакопеременной суммы, минимизирующий влияние вычислительной погрешности.

1.12. Пусть значение многочлена $P_n(x) = a_0 + a_1 x + ... + a_n x^n$ вычисляется в точке $x = 1$ по схеме Горнера:

$$P_n(x) = a_0 + x(a_1 + x(...(a_{n-1} + a_n x)...))\,.$$

Какую погрешность можно ожидать в результате, если коэффициенты округлены с погрешностью η?

У к а з а н и е. Воспользоваться решением 1.7, учитывая незнакоопределенность a_i, и с точностью до слагаемых $O\left(\eta^2\right)$ получить

$$|P_n(1) - P_n^*(1)| \leqslant n\,\eta\,(|a_0| + |a_1| + ... + |a_n|)\,.$$

1.13. Оценить погрешность вычисления скалярного произведения двух векторов $S = \sum\limits_{j=1}^{n} x_j y_j$, если их компоненты округлены с погрешностью η.

О т в е т: с точностью до слагаемых $O\left(\eta^2\right)$ имеем $|S - S^*| \leqslant n\,\eta\,\|x\|_2\|y\|_2$, где $\|z\|_2^2 = \sum\limits_{j=1}^{n} z_j^2$.

1.14. Пусть вычисляется величина $S = a_1 x_1 + ... + a_n x_n$, где коэффициенты a_i округлены с погрешностью η. Оценить погрешность вычисления S при условии, что $x_1^2 + ... + x_n^2 = 1$.

О т в е т: с точностью до слагаемых $O\left(\eta^2\right)$ имеем $|S - S^*| \leqslant n\,\eta\,\|a\|_2$, где $\|a\|_2^2 = \sum\limits_{j=1}^{n} a_j^2$.

1.15. Для элементов последовательности

$$I_n = \int\limits_0^1 x^n \mathrm{e}^{x-1} dx$$

справедливо точное рекуррентное соотношение $I_n = 1 - n\,I_{n-1}$, $I_1 = \frac{1}{\mathrm{e}}$. Можно ли его использовать для приближенного вычисления интегралов, считая, что ошибка округления допускается только при вычислении I_1?

$\lhd$ Пусть в результате округления значения I_1 получено значение I_1^*, использование которого приводит к величинам $I_n^* = 1 - n\,I_{n-1}^*$. Для погрешности $\Delta_n = I_n - I_n^*$ имеем соотношение $\Delta_n = -n\,\Delta_{n-1}$, откуда следует $\Delta_n = (-1)^{n+1} n!\,\Delta_1$. Полученная формула гарантирует факториальный рост погрешности и ее знакопеременность. Учитывая, что точные значения удовлетворяют неравенству

$$0 < I_n < \int\limits_0^1 x^n dx = \frac{1}{n+1}\,,$$

получим, что начиная с некоторого n величина погрешности существенно больше искомого результата. Алгоритмы такого рода называются *неустойчивыми*. $\rhd$

1.16. Можно ли использовать для приближенного вычисления интегралов

$$I_n = \int\limits_0^1 x^n \mathrm{e}^{x-1} dx$$

точное рекуррентное соотношение $I_{n-1} = \dfrac{1 - I_n}{n}$ (в обратную сторону по сравнению с 1.15), считая, что ошибка округления допускается только при вычислении стартового значения I_N? Как выбрать это значение?

О т в е т: да (см. решение 1.15), $I_N \approx 0$ при достаточно больших N.

1.17. Пусть вычисления ведутся по формуле

$$y_{n+1} = 2y_n - y_{n-1} + h^2 f_n\,,$$

где $n = 1, 2, \ldots$; y_0, y_1 заданы точно, $|f_n| \leqslant M$, $\quad h \ll 1$. Какую вычислительную погрешность можно ожидать при вычислении y_n для больших значений n? Улучшится ли ситуация, если вычисления вести по формулам $\dfrac{z_{n+1} - z_n}{h} = f_n\,,\ \dfrac{y_n - y_{n-1}}{h} = z_n\,$?

$\lhd$ Формулы, приведенные в условии, являются численными алгоритмами решения задачи Коши для уравнения $y'' = f(x)$. Рассмотрим модельную задачу $y'' = M\,, y(0) = y'(0) = 0$, имеющую точное решение $y(x) = x^2 \dfrac{M}{2}$. Введем сетку с шагом h: $x_n = n\,h$ и будем искать приближенное решение по формуле

$$y_{n+1} = 2y_n - y_{n-1} + h^2 M\,, \quad n = 1, 2, \ldots\,; \quad y_0 = 0\,, y_1 = h^2 \frac{M}{2}\,.$$

При отсутствии ошибок округлений получим $y_n = (n\,h)^2 \dfrac{M}{2}$, т. е. проекцию точного решения на сетку.

Вычисления приводят к соотношениям

$$y_0^* = 0\,, y_1^* = h^2 \frac{M}{2} + \eta_1,$$

$$y_{n+1}^* = 2y_n^* - y_{n-1}^* + h^2 M + \eta_{n+1}\,, \quad n = 1, 2, \ldots\,.$$

Отсюда для погрешности $r_n = y_n^* - y_n$ получим

$$r_{n+1} = 2r_n - r_{n-1} + \eta_{n+1}\,, \quad n = 1, 2, \ldots\,; \quad r_0 = 0\,, r_1 = \eta_1\,.$$

Для простоты вычислений предположим, что все η_n постоянны и равны η, тогда для погрешности справедлива формула $r_n = \eta\frac{n^2+n}{2}$. Сопоставляя точное решение y_n и погрешность, приходим к относительной погрешности порядка $h^{-2}\frac{\eta}{M}$. Требование малости этой величины накладывает ограничение на шаг интегрирования h снизу, так как обычно $\eta \sim p^{1-t}$.

Аналогичные рассуждения для второго способа расчетов приводят к относительной погрешности порядка $h^{-1}\frac{\eta}{M}$, что, в свою очередь, приводит к более слабым ограничениям на h при одном и том же η. Другими словами, используя формулы

$$\frac{z_{n+1} - z_n}{h} = f_n\,, \quad \frac{y_n - y_{n-1}}{h} = z_n\,,$$

как правило, получаем меньшую вычислительную погрешность. $\triangleright$

1.2. Погрешность функции

Пусть искомая величина y является функцией параметров x_j, $j = 1, 2, \ldots, n$: $y = y(x_1, x_2, \ldots, x_n)$. Область G допустимого изменения параметров x_j известна, требуется получить приближение к y и оценить его погрешность. Если y^* — приближенное значение величины y, то *предельной абсолютной погрешностью* называют величину

$$A(y^*) = \sup_{(x_1, x_2, \ldots, x_n) \in G} |y(x_1, x_2, \ldots, x_n) - y^*|\,;$$

при этом *предельной относительной погрешностью* называют величину $R(y^*) = \frac{A(y^*)}{|y^*|}$.

1.18. Доказать, что предельная абсолютная погрешность $A(y^*)$ минимальна при

$$y^* = \frac{y_1 + y_2}{2},$$

где $y_1 = \inf\limits_G y(x_1, x_2, \ldots, x_n)$, $y_2 = \sup\limits_G y(x_1, x_2, \ldots, x_n)$.

$\triangleleft$ Используя определения величин y_1 и y_2, выражение для $A(y^*)$ перепишем в виде

$$A(y^*) = \sup_{y(x_1, x_2, \ldots, x_n) \in [y_1, y_2]} |y(x_1, x_2, \ldots, x_n) - y^*|\,,$$

при этом $A(y_1) = A(y_2) = y_2 - y_1$. Обозначим $A = y_2 - y_1$. Так как нас интересует минимальное значение величины $A(y^*)$, то достаточно проанализировать только $y^* \in [y_1, y_2]$. Это следует из того, что для $y^* \notin [y_1, y_2]$

справедливо неравенство $A(y^*) > A$. Введем для y^* параметризацию $y^* = \alpha\, y_1 + (1-\alpha)\, y_2$ с $\alpha \in [0,1]$ и рассмотрим предельную абсолютную погрешность

$$A(y^*) = \sup_{y\in[y_1,y_2]} |y - [\alpha\, y_1 + (1-\alpha)\, y_2]| = \\ = \max\{\alpha\, A(y_1),\ (1-\alpha)\, A(y_2)\} = A\, \max\{\alpha, 1-\alpha\}\,.$$

Минимум величины $\max\{\alpha, 1-\alpha\}$ равен $\frac{1}{2}$ и достигается при $\alpha = \frac{1}{2}$, т. е. минимум $A(y^*)$ имеет место при $y^* = \frac{y_1+y_2}{2}$. $\triangleright$

1.19. Показать, что предельная абсолютная погрешность суммы или разности чисел равна сумме их предельных абсолютных погрешностей.

$\triangleleft$ Если известны оценки $|x_j - x_j^*| \leqslant \Delta(x_j^*),\ j = 1,2$, то можно определить область G:

$$G = \left\{(x_1, x_2)\ :\ x_j^* - \Delta(x_j^*) \leqslant x_j \leqslant x_j^* + \Delta(x_j^*)\,,\ j = 1,2\right\}\,.$$

Рассмотрим в этой области функции $y_\pm = x_1 \pm x_2$ и их предельные абсолютные погрешности. Имеем

$$A(y^*) = \sup_{(x_1,x_2)\in G} |y_\pm - y_\pm^*| = \sup_{(x_1,x_2)\in G} |(x_1 \pm x_2) - (x_1^* \pm x_2^*)| \leqslant \\ \leqslant \sum_{j=1}^{2} \sup_{x_j} |x_j - x_j^*| = \Delta(x_1^*) + \Delta(x_2^*)\,. \qquad \triangleright$$

1.20. Показать, что предельная относительная погрешность произведения или частного с точностью до членов второго порядка малости равна сумме предельных относительных погрешностей.

$\triangleleft$ Если известны оценки $\dfrac{|x_j - x_j^*|}{|x_j^*|} \leqslant \delta(x_j^*),\ j = 1,2$, то можно определить область G:

$$G = \left\{(x_1, x_2)\,:\ x_j^* - \Delta(x_j^*) \leqslant x_j \leqslant x_j^* + \Delta(x_j^*)\,,\ j = 1,2\right\},$$

где $\Delta(x_j^*) = |x_j^*|\, \delta(x_j^*)$. Рассмотрим в этой области функцию $y = x_1 x_2$ и ее предельную относительную погрешность

$$R(y^*) = \frac{A(y^*)}{|y^*|} = \frac{1}{|x_1^* x_2^*|} \sup_{(x_1,x_2)\in G} |x_1 x_2 - x_1^* x_2^*| \leqslant \\ \leqslant \frac{1}{|x_1^* x_2^*|}\left(\Delta(x_1^*)x_2^* + \Delta(x_2^*)x_1^* + \Delta(x_1^*)\Delta(x_2^*)\right)\,.$$

Отбрасывая члены второго порядка малости, получим

$$R(y^*) \leqslant \frac{\Delta(x_1^*)}{|x_1^*|} + \frac{\Delta(x_2^*)}{|x_2^*|} = \delta(x_1^*) + \delta(x_2^*)\,.$$

Аналогично рассматривается случай функции $y = \dfrac{x_1}{x_2}$. $\triangleright$

1.21. Пусть $y = y(x_1, x_2, \ldots, x_n)$ — непрерывно дифференцируемая функция. Положим

$$A_{\sup}(y^*) = \sum_{j=1}^{n} B_j \Delta(x_j^*)\,, \quad \text{где } B_j = \sup_G \left| \frac{\partial y(x_1, x_2, \ldots, x_n)}{\partial x_j} \right| ;$$

$$A_{\mathrm{lin}}(y^*) = \sum_{j=1}^{n} b_j \Delta(x_j^*)\,, \quad \text{где } b_j = \left| \frac{\partial y(x_1, x_2, \ldots, x_n)}{\partial x_j} \right|_{\mathbf{x}=(x_1^*, x_2^*, \ldots, x_n^*)}$$

Доказать, что $A(y^*) \leqslant A_{\sup}(y^*)$, и если величина $\rho = \left(\sum\limits_{j=1}^{n} \Delta^2(x_j^*) \right)^{1/2}$ мала, то справедливо равенство $A_{\sup}(y^*) = A_{\mathrm{lin}}(y^*) + o(\rho)$.

$\lhd$ Используя формулу конечных приращений Лагранжа, получим

$$y(x_1, x_2, \ldots, x_n) - y^* = \sum_{j=1}^{n} b_j(\theta)(x_j - x_j^*)\,,$$

где

$$b_j(\theta) = \left. \frac{\partial y(x_1, x_2, \ldots, x_n)}{\partial x_j} \right|_{\mathbf{x}=\mathbf{x}(\theta)}\,,$$

$$\mathbf{x}(\theta) = (x_1^* + \theta_1(x_1 - x_1^*), \ldots, x_n^* + \theta_n(x_n - x_n^*))\,,\ \theta_j \in [0,1]\,.$$

Отсюда следует $A(y^*) \leqslant A_{\sup}(y^*)$, так как $|b_j(\theta)| \leqslant B_j$.

В силу непрерывности производных $\frac{\partial y}{\partial x_j}$ справедливо представление $B_j = |b_j(0)| + o(1)$ при $\rho \to 0$. Поэтому величину $A_{\sup}(y^*)$ можно записать в виде $A_{\sup}(y^*) = A_{\mathrm{lin}}(y^*) + o(\rho)$, так как $b_j = |b_j(0)|$.

На практике часто используют, вообще говоря, неверную «оценку» $|y(x_1, x_2, \ldots, x_n) - y^*| \leqslant A_{\mathrm{lin}}(y^*)$, называемую *линейной оценкой погрешности*. Величина $A_{\mathrm{lin}}(y^*)$ вычисляется значительно проще, чем $A_{\sup}(y^*)$ или $A(y^*)$, но не следует забывать о требуемой малости ρ. $\rhd$

1.22. Пусть $y = x^{10}$, $x^* = 1$ и задано: 1) $\Delta(x^*) = 0{,}001$; 2) $\Delta(x^*) = 0{,}1$. Вычислить величины $A_{\sup}(y^*)$, $A_{\mathrm{lin}}(y^*)$, $A(y^*)$.

$\lhd$ 1) Здесь $y^* = 1$, $\frac{\partial y}{\partial x} = 10 \cdot x^9$, $b(0) = 10$. Пусть $\Delta(x^*) = 0{,}001$, тогда

$$B = \sup_{|x-1| \leqslant 0{,}001} |10 \cdot x^9| = 10{,}09\ldots\,,$$

$$A_{\sup}(y^*) = B\,\Delta(x^*) = 0{,}01009\ldots\,,$$

$$A_{\mathrm{lin}}(y^*) = |b(0)|\,\Delta(x^*) = 0{,}01\,,$$

$$A(y^*) = \sup_{|x-1| \leqslant 0{,}001} |x^{10} - 1| = 1{,}001^{10} - 1 = 0{,}010045\ldots\,.$$

В этом случае верхняя оценка, предельно точная оценка и линейная оценка отличаются несущественно.

2) Здесь

$$B = \sup_{|x-1|\leqslant 0,1} |10\cdot x^9| = 10\cdot(1,1)^{10} = 23,\dots,$$

$$A_{\sup}(y^*) = B\,\Delta(x^*) = 2,3\dots,$$

$$A_{\rm lin}(y^*) = |b(0)|\,\Delta(x^*) = 1\,,$$

$$A(y^*) = \sup_{|x-1|\leqslant 0,1} |x^{10}-1| = (1,1)^{10}-1 = 1,5\dots.$$

Различие между рассматриваемыми величинами в этом случае более заметно. $\rhd$

1.23. Получить линейную оценку погрешности функции, заданной неявно уравнением $F(y,x_1,\dots,x_n)=0$.

$\lhd$ Дифференцируя по x_j, имеем $\dfrac{\partial F}{\partial y}\,\dfrac{\partial y}{\partial x_j} + \dfrac{\partial F}{\partial x_j} = 0$, откуда $\dfrac{\partial y}{\partial x_j} = -\dfrac{\partial F}{\partial x_j}\left(\dfrac{\partial F}{\partial y}\right)^{-1}$. При фиксированных $x_1^*,\dots,x_n^*$ можно найти y^* как решение нелинейного уравнения $F(y,x_1^*,\dots,x_n^*)=0$ с одним неизвестным y. Далее вычисляем значения $b_j = -\dfrac{\partial F}{\partial x_j}\left(\dfrac{\partial F}{\partial y}\right)^{-1}\Big|_{(y^*,x_1^*,\dots,x_n^*)}$, приводящие к искомой величине $A_{\rm lin}(y^*) = \sum\limits_{j=1}^{n} |b_j|\Delta(x_j^*)$. $\rhd$

1.24. Пусть y^* — простой (не кратный!) корень уравнения $y^2+b\,y+c=0$, вычисленный при заданных приближенных значениях коэффициентов b^*, c^*, и известны погрешности $\Delta(b^*), \Delta(c^*)$. Доказать, что

$$A_{\rm lin}(y^*) = \frac{|y^*|\Delta(b^*)+\Delta(c^*)}{|2y^*+b^*|}\ .$$

У к а з а н и е. Воспользоваться решением 1.23, где $F(y,b,c)\equiv y^2+b\,y+c = 0$ — неявная функция, и вычислить следующие величины:

$$b_1 = -\ \frac{\partial F}{\partial b}\left(\frac{\partial F}{\partial y}\right)^{-1}\Big|_{(y^*,b^*,c^*)} = -\ \frac{y^*}{2\,y^*+b^*}\ ,\quad b_2 = -\ \frac{1}{2\,y^*+b^*}\ .$$

1.25. Показать, что если уравнение из 1.24 имеет кратный корень, то погрешность приближенного значения корня имеет порядок $O(\sqrt{\rho})$, где $\rho = \left(\Delta^2(b^*)+\Delta^2(c^*)\right)^{1/2} \ll 1$.

$\lhd$ Пусть уравнение $F(y,b,c)\equiv y^2+b\,y+c=0$ имеет y^* — двухкратный корень при $b=b^*$, $c=c^*$. Разложим F в ряд Тейлора в окрестности точки (y^*,b^*,c^*):

$$F(y,b,c) = F(y^*,b^*,c^*) + F_y(y^*,b^*,c^*)(y-y^*) + F_b(y^*,b^*,c^*)(b-b^*) +$$
$$+F_c(y^*,b^*,c^*)(c-c^*) + \tfrac{1}{2}\,F_{yy}(y^*,b^*,c^*)(y-y^*)^2 + o(\rho) = 0\,.$$

Из условия имеем

$$F(y^*,b^*,c^*) = F_y(y^*,b^*,c^*) = 0\,,\quad \tfrac{1}{2}\,F_{yy}(y^*,b^*,c^*) = 1\,,$$

что приводит к неравенству

$$(y-y^*)^2 \leqslant |F_b(y^*,b^*,c^*)|\,|b-b^*| + |F_c(y^*,b^*,c^*)|\,|c-c^*| + o(\rho)\,,$$

т. е. $|y-y^*| = O(\sqrt{\rho})$. ▷

1.26. Показать, что в случае, когда алгебраическое уравнение $\sum\limits_{i=0}^{N} a_i y^i = 0$ имеет корень кратности n, погрешность значения корня, вычисленного при заданных приближенных значениях коэффициентов a_i^* с известными погрешностями $\Delta(a_i^*)$, имеет порядок $O(\rho^{1/n})$, где $\rho = \left(\sum\limits_{i=0}^{N} \Delta^2(a_i^*)\right)^{1/2}$.

У к а з а н и е. Воспользоваться решением 1.25.

1.27. Имеется приближение y^* к простому корню уравнения $f(y) = 0$. Вывести приближенное равенство $y - y^* \approx -\dfrac{f(y^*)}{f'(y^*)}$.

◁ Рассмотрим более общее уравнение $f(y) = a$ и вычислим $a^* = f(y^*)$. При малых $|y^* - y|$ из равенства $f(y) - f(y^*) = a - a^*$ следует, что $f'(y^*)(y - y^*) \approx a - a^*$, откуда получаем

$$y - y^* \approx \frac{a-a^*}{f'(y^*)} = \frac{a-f(y^*)}{f'(y^*)}\ .$$

Заметим, что $f'(y^*) \neq 0$ в силу того, что y^* — простой корень. Полагая $a = 0$ (по условию), приходим к искомой формуле. ▷

1.28. С каким минимальным числом верных знаков надо взять $\lg 2$ для того, чтобы вычислить корни уравнения $y^2 - 2\,y + \lg 2 = 0$ с четырьмя верными знаками?

◁ Уточним условие. Если $\lg 2 = 0,30102999566\ldots$, то корни принимают значения $y_1 = 1,83604425979\ldots$ и $y_2 = 0,16395574020\ldots$. Требуется найти приближение к числу $\lg 2$, обеспечивающее значения корней $y_1^* = 1,836$ и $y_2^* = 0,164$. Теперь воспользуемся решением 1.24 при $b = -2$, $\Delta(b^*) = 0$ и $c = \lg 2$. После подстановки в $A_{\lim}(y^*) = \dfrac{|y^*|\Delta(b^*) + \Delta(c^*)}{|2y^* + b^*|}$ имеем

$$A_{\lim}(y_{1,2}^*) = \frac{\Delta(c^*)}{2\sqrt{1-c^*}} = \Delta(c^*)\cdot 0,5980544\ldots\,.$$

Из этой формулы следует: если требуется в решении получить n верных знаков, то достаточно в c^* взять также n верных знаков, так как постоянная, связывающая величины погрешностей, не превосходит единицы. Таким образом, требуется взять $\lg 2$ с четырьмя верными знаками, т. е. $\lg 2 \approx 0,301$.

Если вычисления провести аккуратно, то при $\lg 2 \approx 0,301$ получим $y_1^* = 1,83606\cdots \approx 1,836$ и $y_2 = 0,16393\cdots \approx 0,164$. Меньшее количество верных знаков брать нельзя: при $\lg 2 \approx 0,30$ имеем $y_1^* = 1,83666\cdots \approx 1,837$ и $y_2 = 0,16333\cdots \approx 0,163$. ▷

1.29. Пусть ограниченные по модулю величиной M коэффициенты уравнения $a\,y^2 + b\,y + c = 0$ заданы с одинаковой относительной погрешностью δ. Найти максимальную абсолютную (относительную) погрешность, с которой могут вычисляться их корни.

У к а з а н и е. Воспользоваться решениями 1.24 и 1.25.

1.30. Найти приближенное значение интеграла $I_{100} = \int\limits_0^{2\pi} \sin^{100} x dx$ с относительной погрешностью не более 10%.

У к а з а н и е. Вывести по индукции с помощью интегрирования по частям формулу для точного значения интеграла $I_{100} = \frac{100!\,2\pi}{2^{100}(50!)^2}$, затем применить формулу Стирлинга (см. указание к 5.69) с учетом 1.20.

О т в е т: с заданной погрешностью $I_{100} \approx \frac{1}{2}$.

[. . .]

Московский государственный университет имени М.В. Ломоносова

КЛАССИЧЕСКИЙ УНИВЕРСИТЕТСКИЙ УЧЕБНИК

Основная парадигма вычислительной математики гласит: «Цель расчетов – понимание, а не числа». Это означает, что ни гигантские хранилища информации, ни вычислительная мощь современных суперкомпьютеров не в состоянии заменить интеллектуальные возможности математика-исследователя. Развитие цивилизации ставит перед обществом проблемы, решение которых вынуждает не только использовать все уже накопленные знания, но и интенсивно раздвигать научные горизонты. Изучение нелинейностей окружающего мира приводит к естественной математизации других, в том числе гуманитарных наук, что проявляется в построении и анализе математических моделей численными и аналитическими методами.

Данное пособие написано на основе многолетнего опыта преподавания численных методов в МГУ им. М. В. Ломоносова. Содержание пособия тесно связано с классическим учебником Н. С. Бахвалова, Н. П. Жидкова, Г. М. Кобелькова «Численные методы» и книгой Н. С. Бахвалова, А. В. Лапина, Е. В. Чижонкова «Численные методы в задачах и упражнениях».