Ассоциативный закон сложения на компьютере?

$$A + B + C + D = \left(A + C\right) + (B + D) \text{ ???}$$

Пусть мантиссы чисел с плавающей точкой представляются в компьютере пятью десятичными разрядами. Каждая операция сопровождается округлением до пяти разрядов.

$A = 28.654 = 0.28654 \cdot 10^{2}$; $B = -28.653 = -0.28653 \cdot 10^{2}$;

$C = 0.0015178 = 0.15178 \cdot 10^{-2}$; $D = -0.0014963 = -0.14963 \cdot 10^{-2}$.

Вариант 1. Складываем последовательно слева направо:

$$A + B + C + D = 28.654 - 28.653 + 0.0015178 - 0.0014963 =$$
$$= 0.001 + 0.0015178 - 0.0014963 = 0.0025178 - 0.0014963 = 0.0010215$$

Вариант 2. Расставим скобки:

$$\left(A + C\right) + (B + D) = \left(28.654 + 0.0015178\right) + \left(-28.653 - 0.0014963\right) =$$
$$= 28.6555178 - 28.6544963 \approx 28.656 - 28.654 = 0.002$$

0.0010215 и 0.0020000 !!!!!

Задача 1.

Пусть требуется определить значение E_9, где E_n представляет собой следующий интеграл

$$E_n = \int_0^1 x^n e^{x-1} dx\,.$$

Интегрируя по частям

$$E_n = \int_0^1 x^n e^{x-1} dx = x^n e^{x-1}\Big|_0^1 - \int_0^1 n x^{n-1} e^{x-1} dx = 1 - n\int_0^1 x^{n-1} e^{x-1} dx = 1 - nE_{n-1}\,,$$

получаем рекуррентную формулу для определения E_9

$$E_n = 1 - nE_{n-1}\,, \qquad n=2, 3,\ldots,9. \qquad (*)$$

Задавая E_1 с точностью в шесть десятичных разрядов и выполняя в дальнейшем восемь раз все вычисления по формуле (*) без дополнительных округлений, получаем

$E_1 = 0.367879$; $E_2 = 0.264242$; ...

$E_7 = 0.110160$; $E_8 = 0.118720$; $E_9 = -0.068480$

\

Очевидно, что значение E_9 определено не верно. В пользу этого вывода говорят следующие два замечания.

Замечание 1. Подынтегральная функция на промежутке [0, 1] всегда положительная, и вычисляемый интеграл положителен для любого n.

Замечание 2. График подынтегральной функции с увеличением n располагается все ближе и ближе к оси абсцисс, а функция E_n является монотонно убывающей от n. У нас же $E_8 > E_7$!!!

Кто виноват и что делать?

Как можно было бы исправить ситуацию и возможно ли это вообще?

Поступило предложение переписать формулу (*) в ином виде:

$$E_{n-1} = \frac{1 - E_n}{n} \qquad (**)$$

и начать пользоваться ею с E_{20}.

Так Куликовская битва была в 1380 году, выберем $E_{20} = 1380$. Правда, при этом значение E_{19} будет отрицательным. Но ведь нам нужно не E_{19}, а E_9! Применяя формулу (**) одиннадцать раз, получаем $E_9 \approx 0.0916123$. Очень правдоподобно!? Попробуем еще раз.

Первая мировая война началась в 1914 году. Выберем $E_{20} = 1914$. Применяя формулу (**) одиннадцать раз, получаем несколько иной результат, но первые разряды такие же: $E_9 \approx 0.0916123$. Попробуем еще раз.

Значение E_{20} ожидаемо мало. Выберем $E_{20} = 0$. Применяя формулу (**) одиннадцать раз, получаем опять чуть иной результат, но первые разряды такие же: $E_9 \approx 0.0916123$.

Попробуем разобраться с тем, что произошло. Начнем с формулы (*).

Погрешность задания E_1 не превышает величины $\Delta \le 0.5 \cdot 10^{-6}$. Так как дальнейшие вычисления выполнялись точно, то при вычислении E_2 эта погрешность составила 2Δ, E_3 – $3 \cdot 2\Delta$, E_4 – $4 \cdot 3 \cdot 2\Delta$, … , E_9 - $9!\Delta \approx 0.36 \cdot 10^6 \cdot 0.5 \cdot 10^{-6} = 0.18$. Таким образом, полученное нами значение $E_9 = -0.068480$ вполне укладывается в интервал (точное значение ± 0.18)!!

Теперь перейдем к формуле (**), выполняя все промежуточные действия с высокой точностью. Считая погрешность E_{20} , равной Δ, уменьшим эту погрешность для E_{19} до величины $\Delta / 20$. Для E_{18} она уменьшается уже до величины $\Delta / (20 \cdot 19)$. Наконец, для E_9 она сократится до

$$\frac{\Delta}{\left(20 \cdot 19 \cdot 18 \cdot 17 \cdot \ \dots 10\right)} < 10^{-11} \cdot \Delta .$$

Учитывая этот факт, можно позволять себе выбирать E_{20} = 1380 и получать для E_9 более восьми верных разрядов.

Этот пример – первое знакомство с *устойчивыми* и *неустойчивыми* алгоритмами. Перенос известного метода интегрирования по частям на компьютер в условиях ограниченной точности исходных данных дал негативный результат. Важно отметить, что замена алгоритма (**) на любой устойчивый метод позволяет решить все возникшие проблемы.

Задача 2.

Требуется решить следующую систему линейных алгебраических уравнений **Ax** = **b**, где элементы матрицы **A** и вектора **b** являются экспериментальными данными, заданы с предельной абсолютной погрешностью $\varepsilon \approx 0.005$ и принимают следующие значения

$$\mathbf{A} = \begin{pmatrix} 1.00 & 0.99 \\ 0.99 & 0.98 \end{pmatrix}; \qquad \mathbf{b} = \begin{pmatrix} 1.99 \\ 1.97 \end{pmatrix}.$$

$$x_1 + 0.99x_2 = 1.99;$$
$$0.99x_1 + 0.98x_2 = 1.97$$

Точное решение имеет вид:

Решение №1. $\mathbf{x} = (1.00, 1.00)^T$.

Попробуем решать эту систему. Так как элементы матрицы и вектора имеют по два верных знака после точки, воспользуемся часто применяющимся на практике «инженерным» подходом и будем выполнять все промежуточные действия с тремя разрядами после точки, а в конце результат округлим до двух разрядов. Выражаем x_1 из первого уравнения, подставляем во второе и приводим подобные слагаемые.

$$\left(0.98 - 0.99^2\right) x_2 = 1.97 - 0.99 * 1.99;$$

$0.98 - 0.99^2 = -0.0001 \approx 0$ (округлили до трех разрядов и делим на ноль!!)

Если провести вычисления не с тремя разрядами, а с четырьмя, то получается ожидаемый ответ $\mathbf{x} = (1.00, 1.00)^T$. Однако, тревожный звонок прозвучал…

Поступило неожиданное предложение считать решением следующий вектор:

Решение №2. $\mathbf{x} = \left(3.0000, -1.0203\right)^T$.

Зная **решение №1**, такое предложение выглядит невероятным.

Сделаем проверку, подставив **решение №2** в исходную систему. Вектор правых частей должен быть $\left(1.99, 1.97\right)^T$, а получается следующим: $\left(1.9899, 1.9701\right)^T$. Так какое же решение из двух является настоящим? Погрешность эксперимента лежит в третьем разряде после точки, а правые части для обоих решений различаются в четвертом разряде! *Кто победил?*

Поступила новая информация от постановщика задачи. Все элементы матрицы и вектора были измерены значительно более точно с погрешностью в пятом-шестом разряде после точки. При этом пять чисел сохранили свой внешний вид, а элемент матрицы a_{22} несколько уточнился (вместо 0.98 он стал 0.98015):

$$1.0000x_1 + 0.9900x_2 = 1.9900;$$

$$0.9900x_1 + 0.98015x_2 = 1.9700$$

Решаем систему, как и прежде:

$$\left(0.98015 - 0.9900^2\right)x_2 = 1.9700 - 0.9900 * 1.9900;$$

$$0.98015 - 0.9900^2 = 0.00005; \quad 1.9700 - 0.9900 * 1.9900 = -0.0001;$$

$$x_2 = -0.0001 / 0.00005 = -2; \quad x_1 = 3.98$$

Решение №3. $\mathbf{x} = \left(3.98, -2.00\right)^T$

Сравнивая все три решения, попробуем ответить на вопрос:

«***Кто виноват и что делать?***»

Виновата задача в своей постановке или алгоритм ее решения?

В исходной системе умножим первое уравнение на 0.99, второе оставим без изменения и получаем

$$0.99x_1 + 0.9801x_2 = 1.9701;$$
$$0.99x_1 + 0.98x_2 = 1.97$$

Оба уравнения ***почти*** совпали. Графически им отвечают две прямые на плоскости почти параллельные друг другу. Очень малое изменение параметров любой из этих прямых приводит к тому, что точка их пересечения (*решение системы*) резко изменяется.

Исходная система обладает очень «вредным» свойством: малое изменение исходной информации приводит к сильному изменению решения.

Этим **задача 2** заметно отличается в худшую сторону от **задачи 1**.

Что возможно сделать в таких условиях?

Может помочь:

- повышение точности исходных экспериментальных данных,
- использование любой дополнительной априорной информации о решении,
- внесение изменений в постановку задачи.