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?

◁ 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. ▷

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 сами значения $f(x)$ вычисляются с абсолютной погрешностью Δ. Какую погрешность можно ожидать при вычислении производной, если $\left|f^{(k)}(x)\right|\leqslant M_k$, $\quad k=0,1,\ldots$?

◁ 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

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

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

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

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

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. ▷

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

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

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

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

Указание. Воспользоваться решением 1.7, учитывая незнакоопределенность a_i, и с точностью до слагаемых $O\left(\eta^2\right)$ получить

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

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.

Ответ: 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 $S = a_1 x_1 + \ldots + a_n x_n$, где коэффициенты a_i округлены с погрешностью η. Оценить погрешность вычисления S при условии, что $x_1^2 + \ldots + x_n^2 = 1$.

Ответ: 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?

Ответ: 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, $|f_n| \leqslant M$, $h \ll 1$. 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 $r_n = \eta \dfrac{n^2+n}{2}$. Сопоставляя точное решение y_n и погрешность, приходим к относительной погрешности порядка $h^{-2}\dfrac{\eta}{M}$. 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 $h^{-1}\dfrac{\eta}{M}$, 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 $R(y^*) = \dfrac{A(y^*)}{|y^*|}$.

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 $y^* \notin [y_1, y_2]$