Семинар 2

Варламов Антоний Михайлович

15 сентября 2021 г.

1 Погрешности округления. Погрешности арифметики с плавающей точкой

Существуют различные типы данных:

1. Integer(целые числа) – используется 4 байта/8 байт
2. Real (действительные числа) – используется 4 байта (single precision) или 8 байт (double precision) (Альтернативные названия – float и double). Число хранится в виде:

$$a = \pm 2^{\pm e} (1 + f) \tag{1}$$

   single precision: $\mathrm{OFL} = 10^{38}$

   double precision: $\mathrm{OFL} = 10^{324}$

   Представление чисел в компьютере – дискретное. Плотность распределения чисел непостоянна.

Характерные особенности компьютерной арифметики:

1. $a := 0.1$ но $a \neq 0.1$
2. $10^{20} + 1 = 10^{20}$
   $\left(10^{20} + 1\right) - 10^{20} = 0$
   $\left(10^{20} - 10^{20}\right) + 1 = 0$
   Данные следствия говорят о том, что ноль в компьютерном представлении не единственен.
3. Все числа в компьютере неточны. Степень неточности:

$$\delta a = \frac{\Delta a}{a} \leqslant \varepsilon_{\text{маш}} \tag{2}$$

4. Все операции неточны. Для обозначения "компьютерных"операций используются символы операции в кружке.

$$a \oplus b = (a + b)(1 \pm \delta) \tag{3}$$

   Для анализа обычно используются два метода:

   а Все операции точны, но числа имеют погрешность.
   б Все операции неточны, но числа точны

   Рассмотрим пример применения разных методов:

$$a_1 + a_2 + a_3 \tag{4}$$

$$\begin{aligned}
(a_1 + a_2)(1 + \delta_1) \oplus a_3 &= \qquad (5)\\
((a_1 + a_2)(1 + \delta_1) + a_3)(1 + \delta_2) &= \qquad (6)\\
a_1 + a_2 + a_3 + \delta_1 (a_1 + a_2) + \delta_1\delta_2 (a_1 + a_2) + \delta_2 (a_1 + a_2 + a_3) &\approx \qquad (7)\\
a_1 + a_2 + a_3 + a_1 (\delta_1 + \delta_2) + a_2 (\delta_1 + \delta_2) + a_3\delta_2& \qquad (8)
\end{aligned}$$

Вспомним прошлые результаты:

$$\left| f_i' - \frac{f_{i+1} - f_i}{h} \right| \leq \frac{h}{2} M_2 \tag{9}$$

$$\left| f_i' - \frac{f_{i+1} - f_{i-1}}{2h} \right| \leq \frac{h^2}{6} M_3 \tag{10}$$

Проведем некоторый анализ:

$$\tilde{f}_i = f_i \left(1 + \delta\right) \tag{11}$$

$$\frac{\tilde{f_{i+1}} - \tilde{f}_i}{h} - \frac{f_{i+1} - f_i}{h} = \frac{\delta_1 f_{i+1} - \delta_2 f_i}{h} \tag{12}$$

$$\left| \frac{\delta_1 f_{i+1} - \delta_2 f_i}{h} \right| = \frac{\left|\delta_1\right| f_{i+1} + \left|\delta_1\right| f_{i+1}}{h} \leq \frac{2\varepsilon_{\text{маш}} M}{h} \tag{13}$$

$$err_1 = \frac{h}{2} M_2 + \frac{2\varepsilon_{\text{маш}} M}{h} \tag{14}$$

$$\left| \frac{\tilde{f_{i+1}} - \tilde{f_{i-1}}}{2h} - \frac{f_{i+1} - f_{i-1}}{2h} \right| = \frac{\varepsilon_{\text{маш}} M}{h} \tag{15}$$

$$err_2 = \frac{h^2}{6} M_3 + \frac{\varepsilon_{\text{маш}} M}{h} \tag{16}$$

Нарисуем картинку:

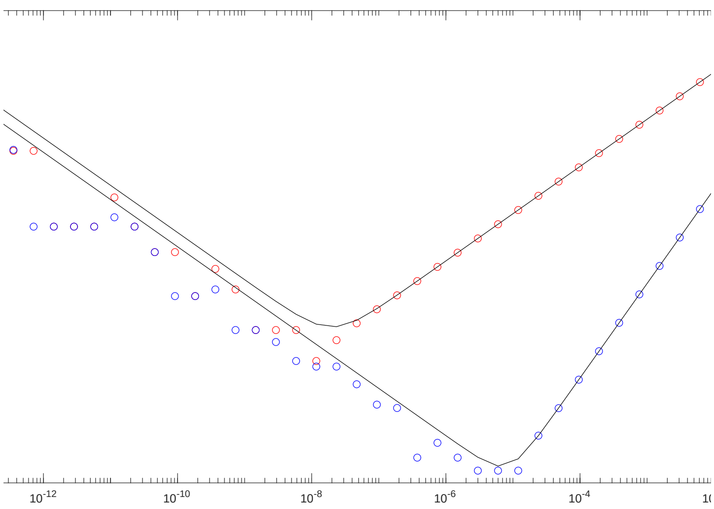

Рис. 1: Визуализация погрешностей вычисления на компьютере

2 Основы линейной алгебры

Вспомним основные понятия:

2.1 Норма

1. $\parallel x \parallel \geq 0, \parallel x \parallel = 0 \Leftrightarrow x = 0$
2. $\parallel \alpha x \parallel = |\alpha| \parallel x \parallel$
3. $\parallel x + z \parallel \leq \parallel x + y \parallel + \parallel y + z \parallel$

$$l_p = \sqrt[p]{\sum_i x_i^p} \tag{17}$$

Матричная норма:

1. $\parallel A \parallel \geq 0, \parallel A \parallel = 0 \Leftrightarrow A = 0$
2. $\parallel \alpha A \parallel = |\alpha| \parallel A \parallel$
3. $\parallel A + B \parallel \leq \parallel A + C \parallel + \parallel C + B \parallel$
4. $\parallel AB \parallel \leq \parallel A \parallel \parallel B \parallel$

Матричная норма Фробениуса:

$$\parallel A \parallel_F = \sqrt{\sum a_{ij}^2} \tag{18}$$

Понятие подчиненной нормы:

$$\parallel A \parallel = \sup_{x \neq 0} \frac{\parallel Ax \parallel}{\parallel x \parallel} = \max_{\parallel y \parallel = 1} \parallel Ay \parallel \tag{19}$$

Рассмотрим некоторые виды норм:

1. $\parallel A \parallel_1 = \max\limits_j \sum\limits_i |a_{ij}|$
2. $\parallel A \parallel_2 = \sqrt{\max\limits_i \lambda_i A^T A}$
3. $\parallel A \parallel_\infty = \max\limits_i \sum\limits_j |a_{ij}|$