1. Аналіз похибок заокруглення

1.1. Види похибок

Нехай необхідно розв'язати рівняння

$$Au = f. \tag{1}$$

За рахунок неточно заданих вхідних даних насправді ми маємо рівняння

$$\tilde{A}\tilde{u} = \tilde{f}. \tag{2}$$

Означення: Назвемо $\delta_1 = u - \tilde{u}$ *неусувною похибкою.*

Застосування методу розв'язання (2) приводить до рівняння

$$\tilde{A}_h\tilde{u}_h = \tilde{f}_h, \tag{3}$$

де $h > 0$ — малий параметр.

Означення: Назвемо $\delta_2 = \tilde{u} - \tilde{u}_h$ *похибкою методу.*

Реалізація методу на ЕОМ приводить до рівняння

$$\tilde{A}_h^\star\tilde{u}_h^\star = \tilde{f}_h^\star. \tag{4}$$

Означення: Назвемо $\delta_3 = \tilde{u}_h - \tilde{u}_h^\star$ *похибкою заокруглення.*

Означення: Тоді *повна похибка* $\delta = u - \tilde{u}_h^\star = \delta_1 + \delta_2 + \delta_3$.

Означення: кажуть, що задача (1) *коректна*, якщо

- $\forall f \in F$: $\exists! u \in U$;
- Задача (1) *стійка*, тобто $\forall\varepsilon > 0$: $\exists\delta > 0$:

$$|A - \tilde{A}| < \delta, |f - \tilde{f}| < \delta \implies |u - \tilde{u}| < \varepsilon. \tag{5}$$

Якщо задача (1) некоректна, то або розв'язок її не існує, або він неєдиний, або він нестійкий, тобто $\exists\varepsilon > 0$: $\forall\delta > 0$:

$$|A - \tilde{A}| < \delta, |f - \tilde{f}| < \delta \implies |u - \tilde{u}| > \varepsilon. \tag{6}$$

Означення: *Абсолютна похибка* $\Delta x \le |x - x^\star|$.

Означення: *Відносна похибка* $\delta x \le \Delta x / |x|$, або $\Delta x / |x^\star|$.

Означення: *Значущими цифрами* називаються всі цифри, починаючи з першої ненульової зліва.

Означення: *Вірна цифра* — це значуща, якщо абсолютна похибка за рахунок відкидання всіх молодших розрядів не перевищує одиниці розряду цієї цифри.

Тобто, якщо $x^\star = \alpha_n \ldots \alpha_0 . \alpha_{-1} \ldots \alpha_{-p} \ldots$, то α_{-p} вірна, якщо $\Delta x \le 10^{-p}$ (інколи $\Delta x \le w \cdot 10^{-p}$, де $1/2 \le w < 1$ наприклад, $w = 0.55$).

1.2. Підрахунок похибок в ЕОМ

Підрахуємо відносну похибку заокруглення числа x на ЕОМ з плаваючою комою. В β-ічній системі числення число представляється у вигляді

$$x = \pm(\alpha_1\beta^{-1} + \alpha_2\beta^{-2} + \ldots + \alpha_t\beta^{-t} + \ldots) \cdot \beta^p, \tag{7}$$

де $0 \le \alpha_k < \beta$, $\alpha_1 \ne 0$, $k = 1, 2, \ldots$

Якщо в ЕОМ t розрядів, то при відкиданні молодших розрядів ми оперуємо з наближеним значенням

$$x^\star = \pm(\alpha_1\beta^{-1} + \alpha_2\beta^{-2} + \ldots + \alpha_t\beta^{-t}) \cdot \beta^p \tag{8}$$

і відповідно похибка заокруглення

$$x - x^\star = \pm\beta^p \cdot (\alpha_{t+1}\beta^{-t-1} + \ldots). \tag{9}$$

Тоді її можна оцінити так

$$\begin{aligned} |x - x^\star| &\le \beta^{p-t-1} \cdot (\beta - 1) \cdot (1 + \beta^{-1} + \ldots) \le \\ &\le \beta^{p-t-1} \cdot (\beta - 1) \cdot \frac{1}{1 - \beta^{-1}} = \beta^{p-t}. \end{aligned} \tag{10}$$

Якщо в представлені (7) взяти $\alpha_1 = 1$, то $|x| \ge \beta^p \cdot \beta^{-1}$. Звідси остаточно

$$\delta x \le \frac{\beta^{p-t}}{\beta^{p-1}} = \beta^{-t+1}. \tag{11}$$

При більш точних способах заокруглення можна отримати оцінку $\delta x \le \frac{1}{2} \cdot \beta^{-t+1} = \varepsilon$. Число ε називається «машинним іпсилон». Наприклад, для $\beta = 2$, $t = 24$, $\varepsilon = 2^{-24} \approx 10^{-7}$.

1.3. Підрахунок похибок обчислення значення функції

Нехай задана функція $y = f(x_1, \ldots, x_n) \in C^1(\Omega)$. Необхідно обчислити її значення при наближеному значенні аргументів $\vec{x}^\star = (x_1^\star, \ldots, x_n^\star)$, де $|x_i - x_i^\star| \le \Delta x_i$ та оцінити похибку обчислення значення функції $y^\star = f(x_1^\star, \ldots, x_n^\star)$. Маємо

$$\begin{aligned} |y - y^\star| &= \left|f(\vec{x}) - f(\vec{x}^\star)\right| = \\ &= \left|\sum_{i=1}^n \frac{\partial f}{\partial x_i}\left(\vec{\xi}\right) \cdot (x_i - x_i^\star)\right| \le \\ &\le \sum_{i=1}^n B_i \cdot \Delta x_i, \end{aligned} \tag{12}$$

де $B_u = \max\limits_{\vec{x} \in U} \left|\frac{\partial f}{\partial x_i}(\vec{x})\right|$.

Тут

$$U = \left\{\vec{x} = (x_1, \ldots, x_n) : |x_i - x_i^\star| \le \Delta x_i\right\} \subset \Omega, \tag{13}$$

для $i = \overline{1, n}$. Отже з точністю до величин першого порядку малості по

$$\Delta x = \max_i \Delta x_i, \tag{14}$$

$$\Delta y = |y - y^\star| \prec \sum_{i=1}^n b_i \cdot \Delta x_i, \tag{15}$$

де $b_i = \left|\frac{\partial f}{\partial x_i}(\vec{x}^\star)\right|$ та «$\prec$» означає приблизно менше.

Розглянемо похибки арифметичних операцій.

- Сума: $y = x_1 + x_2$, $x_1, x_2 > 0$:

$$\Delta y \le \Delta x_1 + \Delta x_2, \tag{16}$$

$$\delta y \le \frac{\Delta x_1 + \Delta x_2}{x_1 + x_2} \le \max(\delta x_1, \delta x_2). \tag{17}$$

- Різниця: $y = x_1 - x_2$, $x_1 > x_2 > 0$:

$$\Delta y \le \Delta x_1 + \Delta x_2, \tag{18}$$

$$\delta y \le \frac{x_2\delta x_1 + x_1\delta x_2}{x_1 - x_2}. \tag{19}$$

  При близьких x_1, x_2 зростає відносна похибка (за рахунок втрати вірних цифр).

- Добуток: $y = x_1 \cdot x_2$, $x_1, x_2 > 0$:

$$\Delta y \prec x_2\Delta x_1 + x_1\Delta x_2, \tag{20}$$

$$\delta y \le \delta x_1 + \delta x_2. \tag{21}$$

- Частка $y = x_1 / x_2$, $x_1, x_2 > 0$:

$$\Delta y \prec \frac{x_2\Delta x_1 + x_1\Delta x_2}{x_2^2}, \tag{22}$$

$$\delta y \le \delta x_1 + \delta x_2. \tag{23}$$

  При малих x_2 зростає абсолютна похибка (за рахунок зростання результату ділення).

Означення: *Пряма задача* аналізу похибок: обчислення Δy, δy по заданих Δx_i, $i = \overline{1, n}$.

Означення: *Обернена задача*: знаходження Δx_i, $i = \overline{1, n}$ по заданих Δy, δy. Якщо $n > 1$, маємо одну умову

$$\sum_{i=1}^n b_i \cdot \Delta x_i < \varepsilon \tag{24}$$

для багатьох невідомих Δx_i.

Вибирають їх із однієї з умов:

$$\forall i : b_i \cdot \Delta x_i < \frac{\varepsilon}{n} \tag{25}$$

або

$$\Delta x_i < \frac{\varepsilon}{\sum\limits_{i=1}^n b_i}. \tag{26}$$