ЛЕКЦІЯ 1. НАБЛИЖЕНІ ЧИСЛА

Під час обчислень доводиться виконувати дії із двома видами чисел — точними та наближеними. Розгляньмо особливості дій з наближеними числами.

1.1. Точні та наближені числа

1. *Точними* називають числа, що дають справжнє значення досліджуваної величини. *Наближеними* називають числа, які близькі до справжнього значення, причому ступінь близькості визначається похибкою наближення.

Приміром, числа $20,31$ та 2 є точними у висловлюваннях: «в аудиторії 20 студентів», «у січні 31 день», $\sqrt{4} = 2$. Числа $10,200$ та $3,14$ є наближеними у висловлюваннях «на розв'язання задачі мені потрібно 10 хвилин», «200 грамів цукерок», $\pi \approx 3,14$.

2. Джерела похибок. Розв'язання більшості практичних задач складається із двох етапів:

1) створення математичної моделі задачі;

2) розв'язання сформульованої (в цій моделі) математичної задачі.

У цьому випадку джерелами похибок є:

1) недосконалість математичної моделі;

2) неточне задавання початкових даних, що є результатами експериментів;

3) похибка методу розв'язання математичної задачі;

4) похибки округлення під час виконання арифметичних дій з числами.

Похибки 1) та 2) називають *неусувними*. Вибраний чисельний метод є, як правило, наближеним у тому розумінні, що він не дає точного розв'язку задачі, а визначає його з деякою похибкою, яку називають *похибкою методу* (*похибкою апроксимації*).

Під час розв'язання задачі відбувається округлення початкових даних, проміжних та остаточних результатів і виникає *обчислювальна похибка* (*похибка округлення*).

Розв'язуючи наближено будь-яку задачу, потрібно вказувати найбільшу допустиму похибку результату.

Отже, якщо доводиться проводити обчислення з наближеними числами потрібно:

1) знаючи задану точність вихідних даних, уміти оцінювати точність результату;

2) вибирати початкові дані з такою точністю, яка б забезпечувала задану точність результату.

3. Округлення чисел. Нагадаймо, будь-яке десяткове число

$$x = \pm a_n a_{n-1} \ldots a_1 a_0, a_{-1} a_{-2} \ldots a_{-m}$$

можна записати у вигляді

$$x = \pm(a_n \cdot 10^n + a_{n-1} \cdot 10^{n-1} + \ldots + a_1 \cdot 10 + a_0 + \\ + a_{-1} \cdot 10^{-1} + \ldots + a_{-m} \cdot 10^{-m}).$$

Одним із джерел наближених чисел є округлення. Округлюють як наближені, так і точні числа.

Округленням заданого числа x до деякого розряду називають заміну його новим числом $\hat{x}$, яке одержують із заданого відкиданням усіх його цифр, справа від цифри цього розряду, або заміною їх нулями. Округлюють числа за таким правилом:

1) якщо перша з цифр (зліва), що відкидаються, менша за 5, то останню залишену цифру не змінюють;

2) якщо перша цифра, що відкидається, більша за 5 або дорівнює 5, то останню залишену цифру збільшують на одиницю.

Приміром,

$$3{,}1415 \approx 3{,}142;\ 3{,}1415 \approx 3{,}14;\ 3{,}1415 \approx 3.$$

Похибка, що виникає при округленні, не перевищує половини одиниці молодшого розряду. Повторні округлення зазвичай не роблять, оскільки це може призвести до збільшення похибки.

1.2. Абсолютна та відносна похибки

1. Нехай x — точне значення, $\hat{x}$ — його наближене значення (його наближення). Якщо $\hat{x} < x$, то кажуть, що $\hat{x}$ є наближеним значенням числа x *з недостачею*; якщо ж $\hat{x} > x$, — наближеним значенням *з надлишком*.

Абсолютною похибкою наближеного числа $\hat{x}$ називають число

$$\Delta(\hat{x}) = \left| x - \hat{x} \right|.$$

Відносною похибкою наближеного числа $\hat{x} \neq 0$ називають число

$$\delta(\hat{x}) = \frac{\Delta(\hat{x})}{|\hat{x}|} = \left|\frac{x-\hat{x}}{\hat{x}}\right|.$$

Відносну похибку часто виражають у відсотках:

$$\delta(\hat{x}) = \frac{\Delta(\hat{x})}{|\hat{x}|} \cdot 100\%.$$

Оскільки x невідоме, то $\Delta(\hat{x})$ та $\delta(\hat{x})$ також невідомі. Однак для цих величин бувають відомі оцінки зверху

$$\Delta(\hat{x}) \le \Delta \text{ та } \delta(\hat{x}) \le \delta\,.$$

У цьому випадку кажуть, що абсолютна та відносна похибки наближеного числа $\hat{x}$ не перевищують відповідно *граничної абсолютної похибки* Δ та *граничної відносної похибки* δ.

2. Точні й сумнівні значущі цифри. Першу ліворуч відмінну від нуля цифру числа x і всі розташовані справа від неї цифри називають *значущими*. Приміром, числа $0{,}0201$ та $1{,}310$ мають 3 та 4 значущі цифри.

Цифру a_i числа x називають *точною*, якщо $\Delta(\hat{x}) \le 10^i$, тобто абсолютна похибка числа $\hat{x}$ не перевищує однієї одиниці відповідного розряду десяткового числа. Якщо точна цифра a_i значуща, то її називають *точною значущою* цифрою.

Якщо цифра наближення є точною значущою цифрою, то це не означає, що вона збігається з відповідним десятковим знаком точного числа.

Приміром, $x = 3{,}000$; $\hat{x} = 2{,}999$; $\Delta(\hat{x}) = 10^{-3}$. У наближення $\hat{x}$ чотири точних значущих цифри, однак жодна з них не збігається з відповідними десятковим знаками числа x.

Наближене число записують таким чином, щоб вигляд запису числа показував його абсолютну похибку, яка не перевищує одиниці останнього розряду, який зберігають при записі. Тобто виписують тільки точні цифри числа, при цьому точні нулі справа не відкидають. Цифри, що не є точними називають *сумнівними*.

Так у наближенні числа $x = 2{,}0302$ числом $\hat{x}_1 = 2{,}03$ усі цифри значущі, а числом $\hat{x}_2 = 2{,}0300$ — усі цифри, окрім останнього нуля.

3. Показникова нормалізована форма числа. Щоб уникнути непорозумінь із записом наближених чисел використовують *показникову нормалізовану форму* запису чисел, у якій число записують у вигляді

$$x = \pm M \cdot 10^p,$$

де $M = \overline{0,a_{-1}a_{-2}...a_{-m}...}$ — додатне число із проміжку $(0;1]$, усі цифри, якого після коми значущі, та $a_{-1} \neq 0$, $p \in \mathbb{Z}$.

Число M називають *мантисою*, а p — *порядком* числа x.

4. Зв'язок між числом точних знаків і похибкою числа. Нехай число $\hat{x} > 0$ є наближеним значенням точного числа x і має вигляд

$$x = a_n \cdot 10^n + a_{n-1} \cdot 10^{n-1} + ... + a_1 \cdot 10 + a_0 + \\ + a_{-1} \cdot 10^{-1} + ... + a_{-m} \cdot 10^{-m},$$

де цифри $a_n,...,a_i$ — точні. За означенням число точних знаків числа $\hat{x}$ визначають з нерівності

$$\Delta(\hat{x}) = \left|x - \hat{x}\right| \leq 10^i.$$

Ділячи обидві частини цієї нерівності на $\left|\hat{x}\right|$, дістаємо

$$\delta(\hat{x}) = \frac{\left|x - \hat{x}\right|}{\left|\hat{x}\right|} \leq \frac{10^i}{\left|a_n \cdot 10^n + a_{n-1} \cdot 10^{n-1} + ... + a_{-m} \cdot 10^{-m}\right|} \leq \\ \leq \frac{10^i}{a_n \cdot 10^n} = \frac{1}{a_n} \cdot 10^{i-n},$$

де a_n — перша значуща цифра числа $\hat{x}$; а a_i — остання значуща цифра цього числа.

1.3. Похибки арифметичних дій

Оцінімо тепер похибки, які виникають при арифметичних діях над числами.

Нехай x та y — точні числа, $\hat{x}$ та $\hat{y}$ — їх наближення, $\Delta(\hat{x})$ та $\Delta(\hat{y})$ — їх абсолютні, а $\delta(\hat{x}), \delta(\hat{y})$ — їх відносні похибки відповідно.

1. Похибка суми та різниці. Правдиві формули:

$$\boxed{\begin{aligned} &\Delta(\hat{x} \pm \hat{y}) \leq \Delta(\hat{x}) + \Delta(\hat{y}); \\ &\delta(\hat{x} \pm \hat{y}) \leq \frac{\Delta(\hat{x}) + \Delta(\hat{y})}{\left|\hat{x} \pm \hat{y}\right|}. \end{aligned}}$$

► Справді,

$$\Delta(\hat{x} \pm \hat{y}) = \left|(x \pm y) - (\hat{x} \pm \hat{y})\right| = \left|(x - \hat{x}) \pm (y - \hat{y})\right| \leq \\ \leq \left|x - \hat{x}\right| + \left|y - \hat{y}\right| = \Delta(\hat{x}) + \Delta(\hat{y}). \blacktriangleleft$$

Отже, абсолютна похибка суми або різниці двох наближених чисел не перевищує їх абсолютних похибок.

Звідси випливає **правило** додавання (віднімання) наближених чисел різної абсолютної точності:

1) виділяють числа, які мають найбільшу абсолютну похибку (числа найменшої точності);

2) точніші числа округлюють так, щоб зберегти в них на одну значущу цифру більше, ніж у виділеному числі;

3) додають (віднімають) усі числа з урахуванням збережених цифр;

4) одержаний результат округлюють на одну цифру.

2. Віднімання двох близьких чисел. Нехай числа x_1 та x_2 близькі. Із співвідношення

$$\delta(\hat{x}_1 - \hat{x}_2) \le \frac{\Delta(\hat{x}_1) + \Delta(\hat{x}_2)}{\left|\hat{x}_1 - \hat{x}_2\right|}$$

випливає, що при будь-якій, навіть малій абсолютній похибці наближень, відносна похибка результату $\delta(\hat{x}_1 - \hat{x}_2)$ значно погіршується порівняно з відносними похибками $\delta(\hat{x}_1)$ та $\delta(\hat{x}_2)$.

Це означає, що для таких обчислень треба брати наближення із значно більшою кількістю точних значущих цифр, ніж вимагає результат. Якщо такої можливості немає, то процедуру треба модифікувати.

Приміром,

$$\sqrt{x} - \sqrt{y} = \frac{x - y}{\sqrt{x} + \sqrt{y}};$$
$$x^2 - y^2 = (x - y)(x + y).$$

3. Похибки добутку. Правдиві формули

$$\Delta(\hat{x}\hat{y}) \le \left|\hat{x}\right|\Delta(\hat{y}) + \left|\hat{y}\right|\Delta(\hat{x}) + \Delta(\hat{x})\Delta(\hat{y});$$
$$\delta(\hat{x}\hat{y}) \le \delta(\hat{x}) + \delta(\hat{y}) + \delta(\hat{x})\delta(\hat{y}).$$

►Справді, нехай $x = \hat{x} + \Delta(\hat{x}), y = \hat{y} + \Delta(\hat{y})$. Тоді

$$\Delta(\hat{x}\hat{y}) = \left|xy - \hat{x}\hat{y}\right| = \left|(\hat{x} + \Delta(\hat{x}))(\hat{y} + \Delta(y)) - \hat{x}\hat{y}\right| =$$
$$\left|\hat{x}\Delta(\hat{y}) + \hat{y}\Delta(\hat{x}) + \Delta(\hat{x})\Delta(\hat{y})\right| \le \left|\hat{x}\right|\Delta(\hat{y}) + \left|\hat{y}\right|\Delta(\hat{x}) + \Delta(\hat{x})\Delta(\hat{y})$$

Ділячи обидві частини нерівності на $\left|\hat{x}\right|\left|\hat{y}\right| \ne 0$, одержуємо нерівність для відносної похибки добутку.◄

Якщо похибки наближень множників достатньо малі, то можна вважати, що

$$\Delta(\hat{x})\Delta(\hat{y}) \approx 0,\ \delta(\hat{x})\delta(\hat{y}) \approx 0.$$

Отже,

$$\boxed{\begin{aligned}&\Delta(\hat{x}\hat{y}) \le \left|\hat{x}\right|\Delta(\hat{y}) + \left|\hat{y}\right|\Delta(\hat{x});\\ &\delta(\hat{x}\hat{y}) \le \delta(\hat{x}) + \delta(\hat{y}).\end{aligned}}$$

Звідси випливає **правило** множення наближених чисел різної абсолютної точності:

1) виділяють число з найменшою кількістю точних значущих цифр;

2) решту множників округлюють так, щоб вони зберегли на одну значущу цифру більше, ніж кількість точних значущих цифр у виділеному числі;

3) у добутку зберігають стільки значущих цифр, скільки точних значущих цифр має виділене число.

3. Похибка частки. Для наближення частки при $y \neq 0$ та $\hat{y} \neq 0$ маємо точні оцінки:

$$\Delta\left(\frac{\hat{x}}{\hat{y}}\right) \le \frac{\left|\hat{x}\right|\Delta(\hat{y}) + \left|\hat{y}\right|\Delta(\hat{x})}{\hat{y}^2} \cdot \frac{1}{1-\delta(\hat{y})};$$

$$\delta\left(\frac{\hat{x}}{\hat{y}}\right) \le \frac{\delta(\hat{x}) + \delta(\hat{y})}{1-\delta(\hat{y})}$$

і спрощені:

$$\boxed{\begin{aligned}&\Delta\left(\frac{\hat{x}}{\hat{y}}\right) \le \frac{\left|\hat{x}\right|\Delta(\hat{y}) + \left|\hat{y}\right|\Delta(\hat{x})}{\hat{y}^2};\\ &\delta\left(\frac{\hat{x}}{\hat{y}}\right) \le \delta(\hat{x}) + \delta(\hat{y}).\end{aligned}}$$

Правило ділення чисел різної абсолютної точності таке саме як і добутку.

1.4. Типи обчислювальних методів

1. Обчислювальні методи, які використовують, щоб перетворити задачу до вигляду, зручному для реалізації на комп'ютері, можна поділити на такі класи:

1) методи еквівалентних перетворень;

2) методи апроксимації;

3) прямі (точні) методи;

4) ітераційні методи;

5) методи статистичних випробувань.

2. Методи еквівалентних перетворень. Ці методи дозволяють замінити початкову задачу іншою, яка має той самий розв'язок. Еквівалентні перетворення є корисними, якщо нова задача простіше початкової або має кращі властивості, або для неї існує відомий метод розв'язання.

Приміром, еквівалентне перетворення квадратного рівняння

$$x^2 + bx + c = 0$$

за допомогою виділення повного квадрату до вигляду

$$\left(x + \frac{b}{2}\right)^2 = \frac{b^2 - 4c}{4}$$

зводить задачу до проблеми обчислення квадратного кореня і приводить до відомих формул коренів квадратного рівняння.

Або, приміром задачу відшукання кореня нелінійного рівняння

$$f(x) = 0$$

можна звести до еквівалентної задачі пошуку точки глобального мінімуму функції

$$\Phi(x) = f^2(x).$$

3. Методи апроксимації. Ці методи дозволяють наблизити (апроксимувати) початкову задачу іншою, розв'язок якої в певному сенсі близький до розв'язку початкової задачі. Похибку, яка виникає при такій заміні, називають *похибкою апроксимації*. Зазвичай апроксимувальна задача містить деякі параметри, що дозволяють регулювати значення похибки апроксимації або впливати на інші властивості задачі. Кажуть, що метод апроксимації *збігається*, якщо похибка апроксимації прямує до нуля, коли параметри методу прямують до деякого граничного значення.

Приміром, ураховуючи означення похідної функції

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h},$$

для її наближеного значення можна використовувати формулу

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}.$$

Похибка апроксимації цієї формули чисельного диференціювання прямує до нуля, коли $h \to 0$.

4. Прямі методи. Метод розв'язання задачі називають *прямим*, якщо він дозволяє одержати розв'язок після виконання скінченної кількості елементарних операцій.

Приміром, метод розв'язання лінійного рівняння

$$ax + b = 0$$

за формулою

$$x = -\frac{b}{a}$$

є прямим методом.

Елементарна операція прямого методу може бути досить складною (обчислення значень елементарної або спеціальної функції, розв'язання системи лінійних алгебричних рівнянь, обчислення визначеного інтеграла тощо). Те, що її вважають елементарною, означає, що її виконання істотно простіше знаходження розв'язку всієї задачі.

Для побудови прямих методів істотну увагу приділяють мінімізації кількості елементарних операцій.

Іноді прямі методи називають *точними*, розуміючи під цим, що при відсутності похибок у початкових даних і при точному виконанні елементарних операцій одержаний результат також буде точним.

5. Приклад прямого методу (схема Горнера). Нехай задача полягає в обчисленні значення многочлена

$$P_n(x) = a_0x^n + a_1x^{n-1} + ... + a_{n-1}x + a_n,$$

у точці $x = c$.

Якщо обчислювати значення многочлена безпосередньо, причому $c^2, c^3, ..., c^n$ знаходити послідовним домноженням на c, то буде потрібно виконати $2n - 1$ операцій множення та n операцій додавання.

Значно економнішим є метод обчислення, який називають *схемою Горнера*, що полягає в записі многочлена у вигляді

$$P_n(x) = a_0x^n + a_1x^{n-1} + ... + a_{n-1}x + a_n = \\ = a_n + x(a_{n-1} + ... + x(a_1 + a_0x)...).$$

Обчислення за схемою Горнера проводять так:

$$\begin{array}{ll} & b_0 = a_0 \\ s_1 = b_0c & b_1 = a_1 + s_1 \\ s_2 = b_1c & b_2 = a_2 + s_2 \\ s_3 = b_2c & b_3 = a_3 + s_3 \\ ... & ... \\ s_n = b_{n-1}c & b_n = a_n + s_n, \end{array}$$

де $s_n = P_n(c)$, і потребують n операцій множення і n операцій додавання. Можна довести, що значення многочлена в точці не можна обчислити меншою кількістю операцій.

Схему Горнера можна задати рекуретним співвідношенням:

$$b_i = b_{i-1}c + a_i,\ i = \overline{1,n},$$

де $b_0 = a_0$, а b_i — значення виразу, який містить i-та дужка.

6. Ітераційні методи. Це спеціальні методи побудови послідовних наближень до розв'язку задачі. Застосування методу починають з вибору одного чи кількох початкових наближень. Щоб одержати кожне наступне наближення виконують однотипний набір дій з використанням знайдених раніше наближень — *ітерацію*. Необмежене продовження цього ітераційного процесу дозволяє будувати нескінченну послідовність наближень до розв'язку — ітераційну послідовність. Якщо ця послідовність збігається до розв'язку задачі, то кажуть, що ітераційний метод *збігається*. Множину початкових наближень, для яких метод збігається, називають *областю збіжності* методу.

7. Приклад ітераційного процесу (обчислення квадратного кореня). Розгляньмо добування квадратного кореня $x = \sqrt{a}$ за ітераційний методом Герона, відповідно до якого наближене значення кореня визначають рівністю

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) = x_n + \frac{1}{2}\left(\frac{a}{x_n} - x_n\right),\quad n = 0,1,2,\ldots$$

Тут x_0 — деяке грубе значення кореня. Точність ε обчислення визначається нерівністю

$$\left|x_{n+1} - x_n\right| < \varepsilon \Leftrightarrow \frac{1}{2}\left|\frac{a}{x_n} - x_n\right| < \varepsilon.$$

Відомо, що цей метод збігається при будь-якому початковому наближенні $x_0 > 0$, так що його область збіжності — множина всіх додатних чисел.

Ітераційний метод за своєю суттю є наближеним; жодне з одержуваних не є точним значенням розв'язку. Однак збіжний ітераційний процес дає принципову можливість знайти розв'язок за будь-якою заданою точністю $\varepsilon > 0$, після досягнення якої процес переривають.

Хоча сам факт збіжності безумовно важливий, він недостатній для практичного використання методу, адже метод може збігатися повільно.

Практична реалізація ітераційних методів завжди потребує критерія завершення ітераційного процесу. Обчислення не можуть тривати нескінченно довго і повинні бути перервані відповідно до деякого критерію, зв'язаного з досягненням заданої точності.

Використання з цією метою *апріорних оцінок* (теоретичних, одержаних до початку процесу) частіше всього неможливо або неефективно.

Для формування критерію закінчення при досягненні заданої точності, як правило, використовують *апостеріорні оцінки* похибки — нерівності, у яких значення похибки оцінюють через відомі або одержані під час обчислювального процесу величини.

Для методу Герона правдива така апостеріорна оцінка:

$$\left|x_{n+1} - \sqrt{a}\right| \le \left|x_{n+1} - x_n\right|, \ n = 0,1,2,\ldots,$$

що дозволяє оцінювати абсолютну похибку наближення через модуль різниці двох послідовних наближень.

Вона дозоляє сформулювати при заданій точністю $\varepsilon > 0$ дуже простий критерій завершення. Як тільки буде виконано нерівність

$$\left|x_{n+1} - x_n\right| < \varepsilon,$$

обчислення треба припинити і взяти x_{n+1} за наближення до $\sqrt{a}$ з точністю ε.

8. Методи статистичних випробувань. Це чисельні методи, що ґрунтуються на моделюванні випадкових величин і побудові статистичних оцінок розв'язання задач.

1.5. Коректність обчислювальних алгоритмів

1. Обчислювальний алгоритм. Обчислювальний метод доведений до ступеня деталізації, що дозволяє реалізувати його на комп'ютері, набуває форму *обчислювального алгоритму* — точного переліку дій над вхідними даними, що задають обчислювальний процес, скерований на перетворення довільних вхідних даних x (із множини допустимих для цього алгоритму вхідних даних X) у повністю визначений цими вхідними даними результат.

Реальний обчислювальний алгоритм складається із двох частин: абстрактного обчислювального алгоритму, сформулюваного в загальноприйнятних математичних термінах, і програми, записаної на одній з алгоритмічних мов.

Приміром, обчислення значення многочлена за схемою Горнера можна задати таким алгоритмом.

Крок 0. Покласти $b_0 = a_0$.

Крок 1. Збільшити значення i на одиницю.

Крок 2. Якщо $i > n$, то перейти до кроку 4, якщо ні — кроку 3.

Крок 3. Обчислити $b_i = b_{i-1}c + a_i$ і перейти до кроку 1.

Крок 4. Покласти $P_n(c) = b_n$ і припинити процес обчислення.

Кроки 1—3 цього алгоритму повторюють багатократно циклічно, причому число цих повторень (циклів) визначається степенем n многочлена.

2. Коректність алгоритму. Обчислювальний алгоритм називають коректним, якщо виконано такі умови:

1) він дозволяє після виконання скінченної кількості елементарних для комп'ютера операцій перетворити будь-яке вхідне дане $x \in X$ у результат y;

2) результат y стійкий відносно до малих збурень вхідних даних;

3) результат y має обчислювальну стійкість.

Стійкість результату до малих збурень вхідних даних означає, що результат неперервно залежить від вхідних даних за умови, що відсутня обчислювальна похибка. Звісно, що потрібно припускати, що разом із вхідними даними x у множині допустимих вхідних даних X належать і всі близькі до x наближені вхідні дані x^*.

Через наявність похибок округлення при вводі даних і при виконанні арифметичних операцій обов'язково з'являється обчислювальна похибка. Для фіксованого алгоритму значення похибки визначається машинною точністю ε_m. Алгоритм називають *обчислювально стійким*, якщо обчислювальна похибка результату прямує до нуля, коли $\varepsilon_m \to 0$.

Обчислювальний алгоритм називають *стійким*, якщо він стійкий до вхідних даних і обчислювально стійкий, і — *нестійким*, якщо хоча б одну з умов не виконано.