М.І. Жалдак
Ю.С. Рамський

ЧИСЕЛЬНІ МЕТОДИ МАТЕМАТИКИ

М.І. Жалдак
Ю.С. Рамський

ЧИСЕЛЬНІ МЕТОДИ МАТЕМАТИКИ

ПОСІБНИК
ДЛЯ САМООСВІТИ ВЧИТЕЛІВ

КИЇВ «РАДЯНСЬКА ШКОЛА» 1984

22.193
Ж24

Рекомендовано управлінням шкіл Міністерства освіти УРСР

ЖАЛДАК М. И., РАМСКИЙ Ю. С. Численные методы математики: Пособие для самообразования учителей.— К. : Рад. шк. 1984.— 206 с.— 50 к. 10 000 экз.

В пособии рассматриваются важные для практического применения методы численного анализа, в частности методы решения систем линейных уравнений и неравенств, нелинейных и трансцендентных уравнений, методы численного дифференцирования и интегрирования. Ко всем описанным методам дается геометрическая интерпретация. Изложение иллюстрируется значительным количеством задач из различных областей науки и производства, допускающих решение лишь с помощью численных методов.

Предназначается для учителей математики общеобразовательной школы.

Рукопис рецензували: доцент кафедри математики Кіровоградського педагогічного інституту, кандидат фізико-математичних наук *П. Д. Карпенко,* доцент кафедри математики Бердянського педагогічного інституту, кандидат фізико-математичних наук *М. В. Корчинський,* завідуючий кабінетом математики Закарпатського обласного ІУВ *Е. Є. Елек.*

Ж 4306010000—398 228 81 Видавництво «Радянська школа», 1984
М210(04)—84

ПЕРЕДМОВА

Однією з характерних особливостей нашого часу є широке застосування електронно-обчислювальних машин у найрізноманітніших галузях людської діяльності. Спілкування з ними неможливе без знання чисельних методів математики.

Цьому питанню присвячено чимало книжок, розрахованих лише на спеціалістів.

А вчитель математики й досі не має в своєму розпорядженні посібника, який би грунтовно в потрібних межах ознайомлював з чисельними методами розв'язування різноманітних практичних задач.

У пропонованій книжці і розглядаються саме ті питання цього розділу прикладної математики, які найтісніше пов'язані з шкільним її курсом.

У першому розділі розглядаються джерела похибок, способи їх врахування при обчисленнях, класифікація, а також правила оцінки при наближених обчисленнях. Вивчається поняття коректності задачі.

Другий розділ присвячено метричним та нормованим просторам, властивості яких широко використовуються далі.

У третьому розділі розглядаються методи розв'язування систем лінійних алгебраїчних рівнянь, а в четвертому — так званий симплекс-метод розв'язування систем лінійних нерівностей і задачі лінійного програмування.

П'ятий розділ присвячено наближеним методам розв'язування нелінійних рівнянь з одним невідомим.

У шостому розділі розглядаються питання наближення функцій, зокрема методи відшукання поліномів, які найкраще наближають дану функцію на певній множині точок. Подаються найпростіші інтерполяційні формули Лагранжа та Ньютона.

У сьомому розділі подаються деякі найпростіші методи наближеного обчислення похідних та визначених інтегралів, а також найпростіші чисельні методи побудови наближеного розв'язку звичайних диференціальних рівнянь.

Усі теоретичні положення ілюструються значною кількістю прикладів. Широко використовується також геометрична інтерпретація. Після кожного параграфа пропонуються вправи для самостійного розв'язування.

Розділ I
ЕЛЕМЕНТИ ТЕОРІЇ ПОХИБОК

§ 1. Розв'язування задачі. Джерела і класифікація похибок

Процес розв'язування будь-якої реальної фізичної задачі можна розбити на такі етапи:

1. Побудова математичної моделі (математичне формулювання задачі), що охоплює найважливіші для даної задачі сторони явища.

2. Вибір методу розв'язування. Для деяких найпростіших моделей вдається дістати аналітичні розв'язки задачі, а для складніших здебільшого не вдається. У цих випадках використовують наближені методи, зокрема чисельні.

3. Алгоритмізація процесу. При застосуванні чисельних методів потрібно записати алгоритм розв'язування задачі. Якщо задача розв'язуватиметься на ЕОМ, то треба також скласти програму.

4. Виконання обчислень (на ЕОМ чи вручну).

5. Аналіз результатів (осмислення математичного розв'язку і зіставлення його з експериментальними даними).

Важливо вміти оцінити точність розв'язку задачі, який здебільшого ми дістаємо з похибками. Похибки результатів зумовлюються такими причинами:

1) Математична модель лише наближено відображає реальні явища.

2) Вхідні дані, як правило,— це числа неточні (дані для обчислень часто дістають з експерименту, а кожний експеримент може дати результат лише з обмеженою точністю).

3) Метод розв'язування задачі часто є наближеним. У багатьох задачах точний результат можна дістати лише після нескінченної або досить великої кількості арифметичних операцій, які практично здійснити неможливо. Тому замість точного розв'язку здебільшого доводиться відшукувати наближений (наприклад, замість суми ряду беруть суму скінченної кількості його членів, нескінченний ітераційний процес обривають після скінченного числа ітерацій, інтеграл замінюють скінченною сумою і т. п.).

4) Округлення при обчисленнях. Усі обчислення (вручну і на ЕОМ) можна виконувати лише з обмеженою кількістю знача-

щих цифр. Тому при виконанні арифметичних дій потрібно вдаватися до округлень, які зумовлюють похибки, що нагромаджуються в процесі обчислень.

Похибки, що породжуються вищезгаданими причинами, відповідно називаються:

1) похибка математичної моделі;
2) неусувна похибка (вона не залежить від обчислювача);
3) похибка методу;
4) похибка округлень.

Повна похибка результату дорівнює сумі всіх перелічених похибок.

Зауважимо, що похибок, пов'язаних з особливостями побудови математичної моделі, не розглядатимемо. Похибки методів досліджуватимемо в кожному окремому випадку.

При виконанні обчислень потрібно дотримуватися такого ***правила:*** *у проміжних результатах похибка округлення має бути значно меншою від інших похибок.* При обчисленнях вручну це правило забезпечується збереженням запасних цифр. На сучасних ЕОМ числа записуються також з достатньою кількістю значущих цифр, тому похибкою окремого округлення здебільшого можна нехтувати на противагу похибці методу і неусувній похибці. У процесі виконання великої кількості операцій похибка округлень збільшується. До значного нагромадження таких похибок може призвести віднімання близьких за величиною чисел, знаходження коренів многочленів високих степенів тощо.

§ 2. Похибки наближених чисел

Часто на практиці з тих чи інших причин доводиться замість точного числа $\tilde{a}$ брати його наближене значення a. При цьому виникає похибка

$$\Delta a = \tilde{a} - a.$$

Абсолютною похибкою Δ наближеного числа a називають величину

$$\Delta = |\tilde{a} - a|.$$

Граничною абсолютною похибкою вважають усяке число Δ_a, що задовольняє умову

$$\Delta \leqslant \Delta_a.$$

Це означає, що $-\Delta_a \leqslant \tilde{a} - a \leqslant \Delta_a$.

Надалі ці нерівності скорочено записуватимемо так:

$$\tilde{a} = a \pm \Delta_a.$$

Зауважимо, що найчастіше точне значення $\tilde{a}$ буває невідомим, а тому невідома й похибка Δ наближеного числа a. Наприклад, неможливо точно визначити числа π, $\sqrt{2}$, $e = \lim\limits_{n\to\infty}\left(1+\frac{1}{n}\right)^n$ тощо, однак досить легко обчислити межі, у яких вони містяться. Можна знайти наближені значення їх та граничні абсолютні похибки. Зрозуміло, що чим менша гранична абсолютна похибка, тим точніше число a наближає число $\tilde{a}$.

Та гранична абсолютна похибка не завжди достатньо характеризує точність обчислень (чи вимірювань). Так, при вимірюванні величини, значення якої дорівнює $5 \cdot 10^7$, граничну абсолютну похибку 0,01 можна вважати досить малою, однак при вимірюванні величини, значення якої дорівнює $5 \cdot 10^{-5}$, похибка 0,01 завелика. Тому часто використовують *відносну похибку наближеного числа a*, яка означується рівністю:

$$\delta = \frac{\Delta}{|a|} \quad (a \neq 0).$$

Очевидно, похибку δ, як і Δ та саме число $\tilde{a}$, не завжди можна визначити. Тому на практиці здебільшого використовують граничну відносну похибку δ_a, що визначається умовою

$$\delta \leqslant \delta_a.$$

Отже,

$$\Delta = |a| \cdot \delta \leqslant |a| \cdot \delta_a = \Delta_a.$$

Якщо $a \neq 0$, то вважатимемо, що для Δ_a і δ_a справджуються співвідношення

$$\delta_a = \frac{\Delta_a}{|a|}, \quad \Delta_a = |a|\,\delta_a. \tag{2.1}$$

Запишемо додатне число a у вигляді скінченного десяткового дробу:

$$a = \alpha_1 10^m + \alpha_2 10^{m-1} + \cdots + \alpha_n 10^{m-n+1},$$

або

$$a = \alpha_1\alpha_2 \ldots \alpha_{m+1},\ \alpha_{m+2} \ldots \alpha_n$$

(де всі коефіцієнти $\alpha_i > 0$ і менші за число 10).

k-та цифра наближеного числа a називається правильною, якщо абсолютна похибка Δ цього числа не перевищує половини

одиниці k-го розряду, тобто коли

$$\Delta = |\tilde{a} - a| \leqslant \frac{1}{2}\,10^{m-k+1}.$$

У протилежному випадку цифру k-го розряду називають *сумнівною*.

Примітка. Інколи цифру k-го розряду називають правильною, якщо абсолютна похибка числа не перевищує одиниці цього розряду. Цифру за останнім означенням називають *правильною у широкому розумінні*, а за першим — *у вузькому*. Надалі ми вживатимемо поняття правильних цифр в розумінні першого означення.

Значущими цифрами наближеного числа a називатимемо всі його правильні цифри, починаючи з першої зліва, що не дорівнює нулю, до першої сумнівної цифри включно. Усі інші цифри називатимемо *незначущими*. Записуючи остаточні результати наближених обчислень, незначущі цифри числа відкидатимемо. При цьому кожне число записуватимемо як добуток деякого степеня числа 10 на число, всі цифри якого значущі. Якщо з наближеними числами виконуватимуться обчислення, то в них, крім значущих, треба зберігати ще одну або дві сумнівні цифри.

На практиці при виконанні обчислень часто виникає потреба округлювати числа.

Якщо округлюється наближене чи точне число a до n значущих цифр, то за правилом доповнення число a замінюють числом a_1 з n значущими цифрами так, щоб похибка округлення не перевищувала половини одиниці розряду, що зберігається, тобто щоб справджувалась умова

$$|a_1 - a| \leqslant \frac{1}{2} \cdot 10^{m-n+1}.$$

Отже, якщо при округленні числа a до n значущих цифр відкидається менше ніж $\frac{1}{2} \cdot 10^{m-n+1}$, то в числі a_1 зберігаються без зміни всі перші n цифр числа a. А коли відкидається не менше як $\frac{1}{2} \cdot 10^{m-n+1}$, то в числі a_1 остання цифра береться на одиницю більшою, ніж у числі a.

Примітка. Внаслідок округлення наближеного числа a до n значущих цифр похибка округлення Δ_0 може досягти величини $0{,}5 \cdot 10^{m-n+1}$. Якщо врахувати ще й похибку Δ_a числа a, то n-а цифра округленого числа може вже бути неправильною. Та коли похибка Δ_a досить мала порівняно з максимально можливим значенням похибки округлення Δ_0 (наприклад, $\Delta_a = 0{,}5 \cdot 10^{m-n}$), то похибкою Δ_a можна знехтувати і n-у цифру числа a після округлення вважати правильною, навіть якщо похибка округлення досягає

$$0{,}5 \cdot 10^{m-n+1}.$$

Отже, щоб знайти результат з точністю до $0{,}5 \cdot 10^{m-n+1}$, треба виконувати обчислення так, щоб дістати результат з точністю принаймні до $0{,}5 \times \times 10^{m-n}$, тобто з $n+1$ правильними цифрами. Тоді після округлення можна буде вважати, що результат має n правильних цифр.

Легко встановити зв'язок відносної похибки наближеного числа a з кількістю правильних значущих цифр.

Якщо додатне наближене число a має n правильних значущих цифр, то його відносна похибка δ задовольняє співвідношення

$$\delta \leqslant \frac{1}{\alpha_1}\left(\frac{1}{10}\right)^{n-1},$$

де α_1 — перша значуща цифра наближеного числа a.

Справді,

$$\Delta \leqslant \frac{1}{2} \cdot 10^{m-n+1},$$

тому

$$\delta = \frac{\Delta}{a} \leqslant \frac{\frac{1}{2} \cdot 10^{m-n+1}}{\alpha_1 10^m + \alpha_2 10^{m-1} + \ldots + \alpha_n 10^{m-n+1} - \frac{1}{2} 10^{m-n+1}} \leqslant$$

$$\leqslant \frac{10^{m-n+1}}{\alpha_1 10^m} = \frac{1}{\alpha_1}\left(\frac{1}{10}\right)^{n-1}.$$

Таким чином, за граничну відносну похибку можна взяти

$$\delta_a = \frac{1}{\alpha_1}\left(\frac{1}{10}\right)^{n-1}.$$

Якщо число a має більше ніж дві правильні значущі цифри, то на практиці можна користуватись оцінкою

$$\delta_a = \frac{1}{2\alpha_1}\left(\frac{1}{10}\right)^{n-1}.$$

Навпаки, за відносною похибкою δ можна визначити кількість правильних цифр числа a. Для цього знаходимо граничну абсолютну похибку Δ_a із співвідношення (2.1), після чого кількість правильних цифр числа a легко встановлюється.

Приклади.

1. Якою буде гранична відносна похибка, якщо замість числа $\sqrt{2}$ взяти 1,41?

Маємо: $n = 3$, $\alpha_1 = 1$. Отже $\delta_a = \frac{1}{1}\left(\frac{1}{10}\right)^{3-1} = 0{,}01$.

2. Знайти наближено $\sqrt{50}$ так, щоб відносна похибка не перевищувала 0,001 (або 0,1 %).

Маємо: $\alpha_1 = 7$. Отже, повинно справджуватись співвідношення $\delta_a = \frac{1}{7}\left(\frac{1}{10}\right)^{n-1} \leqslant 0{,}001$. Звідки $n - 1 \geqslant 3$ і $n \geqslant 4$. Отже, щоб відносна похибка не перевищувала 0,001, $\sqrt{50}$ треба взяти не менше як з чотирма правильними цифрами.

3. Число 7432 має граничну відносну похибку $\delta_a = = 0{,}01$ (1 %). Скільки в ньому правильних цифр?

Знайдемо граничну абсолютну похибку:

$$\Delta_a = |a|\,\delta_a = 7432 \cdot 0{,}01 < 0{,}75 \cdot 10^2.$$

Отже, число 7432 має лише одну правильну цифру (друга цифра сумнівна). Його треба записати так: $0{,}74 \cdot 10^4$, або $74 \cdot 10^2$, або $7{,}4 \cdot 10^3$ і т. п.

Вправи.

1. Знайти граничну абсолютну та граничну відносну похибки, якщо замість точного числа $\tilde{a}$ взято наближене число a: а) $\tilde{a} = e$, $a = 2{,}7$; б) $\tilde{a} = = \lg 2$, $a = 0{,}30$; в) $\tilde{a} = \sqrt{5}$, $a = 2{,}236$; г) $\tilde{a} = \sqrt{99}$, $a = 9{,}95$.

2. Знайти граничну відносну похибку числа a:

а) $\tilde{a} = 273{,}4 \pm 0{,}1$; б) $\tilde{a} = 82\,900 \pm 500$; в) $\tilde{a} = 0{,}00126 \pm 0{,}00003$.

3. Знайти межі точного числа $\tilde{a}$, якщо відома гранична відносна похибка δ_a числа a: $a = 1427{,}394$; $\delta_a = 0{,}00005$; $a = 97{,}45$; $\delta_a = 0{,}015$; $a = 389 \times \times 10^5$; $\delta_a = 3$ %.

4. Записати число:

а) π з п'ятьма правильними значущими цифрами;

б) $\sqrt{2}$ з трьома правильними значущими цифрами;

в) 99,995 з чотирма правильними значущими цифрами.

5. Скільки правильних цифр треба взяти в наближенні числа $\tilde{a}$, щоб відносна похибка не перевищувала δ:

$\tilde{a} = \sqrt[3]{69}$, $\delta = 0{,}0001$; $\tilde{a} = \sqrt[4]{100}$, $\delta = 0{,}01$ %; $\tilde{a} = \lg 127$, $\delta = 1$ %?

6. Число a має відносну похибку δ. Скільки в ньому правильних цифр, якщо

а) $a = 97{,}432$, $\delta = 0{,}05$ %; б) $a = 11{,}4814$, $\delta = 0{,}01$;

в) $a = 547{,}92$, $\delta = 0{,}001$?

§ 3. Загальна формула для обчислення похибки

Нехай потрібно обчислити значення функції $\tilde{y} = f(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$ при заданих значеннях незалежних змінних. Якщо при цьому замість точних значень $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$, які нам не відомі, підставляються їх наближені значення $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ (причому $\tilde{x}_i = x_i + \Delta x_i$, $i = 1, 2, \ldots, n$), то при обчисленні значення функції виникає похибка

$$\Delta y = f(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_n + \Delta x_n) - f(x_1, x_2, \ldots, x_n).$$

Припустивши, що функція $f(x_1, x_2, \ldots, x_n)$ диференційовна (достатню кількість разів), Δy можна подати у вигляді

$$\Delta y = \frac{\partial f(x_1, x_2, \ldots, x_n)}{\partial x_1}\Delta x_1 + \frac{\partial f(x_1, x_2, \ldots, x_n)}{\partial x_2} \times$$

$$\times \Delta x_2 + \cdots + \frac{\partial f(x_1, x_2, \ldots, x_n)}{\partial x_n}\Delta x_n$$ плюс члени другого і вищих порядків відносно Δx_i. (3.1)

Якщо похибки Δx_i за абсолютною величиною досить малі, то членами другого і вищих порядків у виразі (3.1) можна знехтувати. Тоді для абсолютної похибки наближеного значення y матимемо:

$$|\Delta y| \leqslant \sum_{i=1}^{n} \left|\frac{\partial f(x_1, x_2, \ldots, x_n)}{\partial x_i}\right| \Delta_{x_i},$$

де через Δ_{x_i} позначені граничні абсолютні похибки аргументів. Таким чином,

$$\Delta_y = \sum_{i=1}^{n} \left|\frac{\partial f(x_1, x_2, \ldots, x_n)}{\partial x_i}\right| \Delta_{x_i}. \quad (3.2)$$

Для граничної відносної похибки δ_y маємо:

$$\delta_y = \frac{\Delta y}{|y|} = \frac{\sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}\right| \Delta_{x_i}}{|y|} = \sum_{i=1}^{n} \left|\frac{\frac{\partial f}{\partial x_i}}{f}\right| \Delta_{x_i},$$

або

$$\delta_y = \sum_{i=1}^{n} \left|\frac{\partial}{\partial x_i} \ln f(x_1, x_2, \ldots, x_n)\right| \Delta_{x_i}. \quad (3.3)$$

З формули (3.2) випливає, що ***гранична абсолютна похибка суми наближених чисел*** $a_1 + a_2 + \ldots + a_n$ ***... дорівнює сумі їх граничних абсолютних похибок.*** Для різниці двох чисел $a = a_1 - a_2$ маємо: $\Delta_a = \Delta_{a_1} + \Delta_{a_2}$. При цьому відносна похибка різниці $\delta_a = \frac{\Delta_{a_1} + \Delta_{a_2}}{|a|}$ може виявитись досить великою, якщо числа a_1 та a_2 між собою мало відрізняються і їх різниця близька до нуля. У цьому випадку для забезпечення потрібної точності треба мати в зменшуваному і від'ємнику достатню кількість цифр, щоб гранична абсолютна похибка різниці a була меншою від самої різниці. Тому по можливості слід уникати віднімання близьких чисел.

Для похибки добутку додатних чисел $a = a_1 a_2 \ldots a_n$ з (3.3) маємо:

$$\delta_a = \frac{\Delta_{a_1}}{a_1} + \frac{\Delta_{a_2}}{a_2} + \cdots + \frac{\Delta_{a_n}}{a_n},$$

тобто ***гранична відносна похибка добутку кількох наближених додатних чисел, що відрізняються від нуля, дорівнює сумі граничних відносних похибок цих чисел.***

Легко переконатись, що ***гранична відносна похибка частки дорівнює сумі граничних відносних похибок діленого і дільника.***

Справді, нехай $a = \frac{a_1}{a_2}(a_1, a_2 > 0)$. Тоді з (3.3) маємо:

$$\delta_a = \frac{\Delta_{a_1}}{a_1} + \frac{\Delta_{a_2}}{a_2} = \delta_{a_1} + \delta_{a_2}.$$

Можна переконатися також, що відносна похибка k-го степеня числа a у k раз більша, ніж відносна похибка числа a. Аналогічно відносна похибка кореня k-го порядку з числа a у k раз менша, ніж відносна похибка числа a.

Вправи.

1. Знайти суму наближених чисел так, щоб усі цифри результату були значущими (усі цифри доданків значущі): а) $a_1 = 28{,}85$; $a_2 = 491{,}000285$; $a_3 = 1{,}18294$; $a_4 = 0{,}18498932$; $a_5 = 9544{,}2$; б) $a_1 = 25{,}4$; $a_2 = 2900{,}3241$; $a_3 = 9{,}7 \cdot 10^2$; $a_4 = 0{,}00097$; $a_5 = 469{,}99999$.

2. Знайти абсолютну і відносну похибки компонентів дії віднімання, якщо цифри даних правильні: а) 127,434 — 48,815; б) 281,428 — 125,341; в) 824, 481 — 73,5889.

3. Знайти: $\sqrt{105} - \sqrt{103}$ з трьома правильними значущими цифрами; $\sqrt{127} - \sqrt{124}$ з чотирма правильними значущими цифрами.

4. Знайти граничні абсолютну і відносну похибки добутку наближених чисел, добуток і кількість правильних цифр у ньому, якщо цифри-множників правильні: 144,8 · 145,12; 19,875 · 99,9996; 21,545 · 49,876; 87,894 × × 39,878.

5. Знайти граничні абсолютну і відносну похибки частки, частку і кількість правильних цифр у частці, якщо всі цифри діленого і дільника правильні: 25,343 : 34,97; 43,82 : 8442; 39,8412 : 481,7; 32,872 : 99,435.

6. Знайти степінь числа, граничні відносну і абсолютну похибки і кількість правильних цифр, якщо всі цифри основи правильні: $(393{,}4)^2$; $(48{,}15243)^3$; $(97{,}4312)^3$.

§ 4. Поняття про ймовірнісну оцінку похибки. Обчислення без точного врахування похибок

Наближене число a може відхилятися від точного числа $\tilde{a}$ в той чи інший бік на певну величину. Тому наближене число можна розглядати як випадкову величину з математичним сподіванням $M[a] = \tilde{a}$. Сама похибка $\Delta a = \tilde{a} - a$ також

є випадковою величиною з математичним сподіванням, що дорівнює нулю, $M[\Delta_a] = 0$.

Розглянемо суму наближених чисел $a = a_1 + a_2 + \dots + a_n$. Гранична абсолютна похибка суми є $\Delta_a = \Delta_{a_1} + \Delta_{a_2} + \dots + \Delta_{a_n}$.

Вважатимемо, що випадкові величини Δ_{a_1}, Δ_{a_2}, ..., Δ_{a_n} мають один і той самий нормальний розподіл імовірностей з математичним сподіванням, що дорівнює нулю, і середнім квадратичним відхиленням σ. Тоді, як відомо з курсу теорії імовірностей, математичне сподівання суми випадкових величин дорівнює сумі їх математичних сподівань, тобто $M[\Delta_a] = 0$. Дисперсія суми незалежних випадкових величин дорівнює сумі дисперсій цих величин, тобто $D[\Delta_a] = D[\Delta_{a_1}] + D[\Delta_{a_2}] + \dots + D[\Delta_{a_n}] = n\sigma^2$ (нагадаємо, що за означенням середнє квадратичне відхилення $\sigma[X]$ випадкової величини X є $\sqrt{D[X]}$, де $D[X]$ — дисперсія випадкової величини X), а тому $\sigma[\Delta_a] = \sigma\sqrt{n}$.

Сума незалежних випадкових величин з нормальним розподілом імовірностей також має нормальний розподіл імовірностей. Як відомо, випадкова величина з нормальним розподілом імовірностей з імовірністю 0,997 відхиляється в той чи інший бік від свого математичного сподівання не більш як на три середніх квадратичних відхилення. Отже, якщо практично $|\Delta_{a_i}| < 3\sigma$ ($P(|\Delta_{a_i}| < 3\sigma) \approx 0{,}997 \approx 1$), то $|\Delta_a| < 3\sigma\sqrt{n}$.

Таким чином, абсолютна похибка суми наближених чисел $a_1 + a_2 + \dots + a_n$ пропорційна числу $\sqrt{n}$, а не числу n, як це випливало з формули (3.2). Це пояснюється тим, що практично похибки різних чисел можуть частково «гасити» одна одну, в той час як формула (3.2) передбачає, так би мовити, крайній випадок.

Аналогічно приходимо до висновку, що й відносна похибка добутку наближених чисел $a_1, a_2, \dots, a_n$ зростає пропорційно числу $\sqrt{n}$, а не числу n.

Точний підрахунок похибок результатів обчислень наближених чисел досить громіздкий. Тому здебільшого на практиці точно не враховують похибок, а користуються ***правилами підрахунку цифр за В. М. Брадісом:***

1. При додаванні і відніманні наближених чисел молодший збережений розряд результату має бути найбільшим серед розрядів, що виражаються останніми значущими цифрами вихідних даних.

2. При множенні і діленні наближених чисел у результаті потрібно зберегти стільки значущих цифр, скільки їх

має наближене дане з найменшою кількістю значущих цифр.

3. При піднесенні до квадрата або до куба в результаті треба зберегти стільки значущих цифр, скільки їх має число, яке підносять до степеня.

4. При добуванні кореня в результаті слід брати стільки значущих цифр, скільки їх у підкореневому виразі.

5. При обчисленні проміжних результатів потрібно брати на одну — дві цифри більше, ніж рекомендують попередні правила.

6. Якщо дані можна брати з довільною точністю, то, щоб знайти результат з k правильними цифрами, дані треба брати з такою кількістю цифр, яка забезпечує k + 1 правильну цифру в результаті, відповідно до правил 1—4.

7. При обчисленнях одночленних виразів за допомогою логарифмів слід підрахувати кількість значущих цифр у тому наближеному даному, яке має їх найменше, і взяти таблицю логарифмів з кількістю знаків, на одиницю більшою. У кінцевому результаті остання цифра відкидається.

Примітка. В основу цих правил покладено *принцип академіка О. М. Крилова* — основний принцип звичайних обчислень, тобто обчислень без строгого врахування похибок: *Наближене число слід писати так, щоб у ньому всі цифри, крім останньої, були правильними і лише остання була сумнівною, притому не більш як на одну одиницю.*

§ 5. Обернена задача теорії похибок

Досить часто доводиться розв'язувати таку задачу: визначити допустимі похибки аргументів $x_1, x_2, \dots, x_n$ так, щоб гранична абсолютна похибка функції y не перевищувала заданої величини Δ_y.

Користуючись формулою (3.2), помічаємо, що ця задача не визначена, оскільки, по-різному вибравши значення Δ_{x_i}, можна дістати одне й те саме значення Δ_y.

Отже, задачу можна розв'язувати кількома способами.

1. Доберемо Δ_{x_i} так, щоб в (3.2) всі величини $\left|\frac{\partial f}{\partial x_i}\right|\Delta_{x_i}$ були рівними між собою (принцип рівних впливів).

Тоді, очевидно, $\left|\frac{\partial f}{\partial x_i}\right|\Delta_{x_i} = \frac{\Delta_y}{n}$ і $\Delta_{x_i} = \frac{\Delta_y}{n\left|\frac{\partial f}{\partial x_i}\right|}$.

2. Поклавши, що всі Δ_{x_i} рівні між собою, дістанемо:

$$\Delta_{x_i} = \frac{\Delta_y}{\sum_{i=1}^{n}\left|\frac{\partial f}{\partial x_i}\right|}.$$

3. Вважаючи всі відносні похибки аргументів рівними між собою

$$\frac{\Delta_{x_1}}{|x_1|} = \frac{\Delta_{x_2}}{|x_2|} = \cdots = \frac{\Delta_{x_n}}{|x_n|} = \delta$$

та підставляючи в (3.2) вирази для Δ_{x_i}, дістанемо:

$$\Delta_y = \delta \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| |x_i|.$$

Визначивши δ, для Δ_{x_i} матимемо:

$$\Delta_{x_i} = |x_i| \delta = \frac{|x_i| \Delta_y}{\sum_{i=1}^{n} |x_i| \left| \frac{\partial f}{\partial x_i} \right|}.$$

При розв'язуванні задачі можуть бути використані і деякі інші варіанти.

Аналогічно можна розв'язувати і другу обернену задачу теорії похибок, коли потрібно підібрати граничні абсолютні чи відносні похибки аргументів так, щоб відносна похибка функції не перевищувала наперед заданої величини.

Вправи.

1. Визначити точність, з якою треба виконувати обчислення, щоб знайти результат з точністю $0{,}5 \cdot 10^{-3}$:

а) $y = \dfrac{0{,}0327x + \sqrt{\pi}}{0{,}4231x + 0{,}9248}$, $x \in [0{,}7;\ 2{,}3]$;

б) $y = \dfrac{4}{3}\pi R^3$, $R \in [1;\ 5]$.

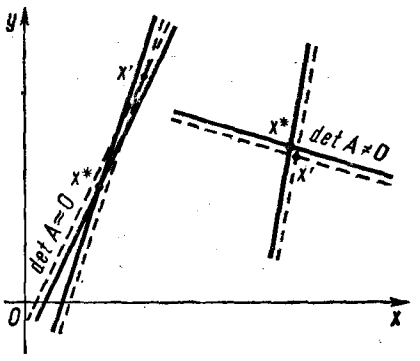
2. Корені рівняння $x^2 - 2x + \lg 3 = 0$ треба знайти з чотирма правильними значущими цифрами. З якою точністю треба взяти вільний член рівняння?

§ 6. Коректність задач

Розглянемо тепер питання так званої коректності. Більшість задач, які доводиться розв'язувати, можна записати у вигляді $y = A(x)$, де x — деяка відома величина, y — шукана величина, $A(x)$ — задана функція (оператор). Зауважимо, що тут y і x можуть бути числами, масивами чисел, функціями однієї чи багатьох змінних тощо.

Задача $y = A(x)$ називається коректно поставленою, якщо для будь-яких вхідних даних x з деякого класу розв'язок y існує, єдиний і стійкий за вхідними даними.

Розглянемо поняття стійкості. *Стійка за вхідними даними* та задача, розв'язок якої неперервно залежить від вхідних даних, тобто для таких задач $|\Delta y| \to 0$, коли $|\Delta x| \to 0$. Якщо ця умова не виконується, то задача називається *нестійкою за вхідними даними*. У цьому випадку навіть незначна похибка у вхідних даних може викликати як завгодно великі похибки в розв'язку, тобто розв'язок може бути зовсім спотворений. Прикладом некоректної задачі є задача диференціювання (див. розділ VII).



Мал. 1

Якщо для похибок розв'язку і вхідних даних існує співвідношення $|\Delta y| \leqslant c\,|\Delta x|$, де c — досить велика константа, то задача формально стійка, але неусувна похибка в цьому випадку може бути значною. Це випадок так званої слабкої стійкості *(слабкої обумовленості)*. Для задовільної практичної стійкості константа c, очевидно, повинна бути не досить великою.

Із слабкою стійкістю ми зустрічаємося, наприклад, при розв'язуванні систем лінійних алгебраїчних рівнянь, визначник яких близький до нуля. Для таких систем похибки в коефіцієнтах системи або похибки округлення при розрахунках можуть призвести до результату, далекого від шуканого розв'язку. У випадку системи двох рівнянь з двома невідомими слабкій стійкості можна дати просту геометричну ілюстрацію. Слабкій стійкості такої системи геометрично відповідає пара майже паралельних прямих, причому невелика зміна нахилу чи зсув однієї з прямих істотно змінює положення точки перетину (мал. 1). Зауважимо, що близькість до нуля визначника системи є лише необхідною умовою слабкої стійкості, але не достатньою.

Примітка. Слід розрізняти стійкість задачі і *стійкість алгоритму* розв'язування задачі. Задача може бути стійка, а алгоритм її розв'язування нестійкий.

Якщо при виконанні алгоритму, наприклад, доводиться обчислювати різницю близьких за величиною чисел (може, й не один раз), то це може призвести до великої похибки результату.

Для ілюстрації наведемо досить простий приклад на знаходження коренів квадратного рівняння $x^2 + px + q = 0$, де $p = -120$, $q = 1$.

Корені шукатимемо за відомою формулою $x_{1,2} = -\frac{p}{2} \mp \sqrt{\frac{p^2}{4} - q}$, причому, всі обчислення виконуватимемо з чотирма значущими цифрами. Тоді

$$x_{1,2} = 60 \mp \sqrt{3599} = 60 \mp 59{,}99 \text{ і } x_1 = 0{,}01, \quad x_2 = 120{,}0.$$

У другому результаті правильними є чотири значущі цифри, а в першому — лише одна. Для порівняння наводимо значення першого кореня з чотирма значущими цифрами: $x_1 = 0{,}008334$. Перший корінь, таким чином, обчислили з відносною похибкою 20 %, а другий — 0,07 %. Втрата точності при обчисленні першого кореня сталася за рахунок віднімання близьких чисел.

А якщо значення першого кореня обчислити за формулою $x_1 = \frac{q}{-\frac{p}{2} + \sqrt{\frac{p^2}{4} - q}}$, яку дістанемо з попередньої, помноживши і поділивши вираз $-\frac{p}{2} - \sqrt{\frac{p^2}{4} - q}$ на спряжений, то дістанемо: $x_1 = \frac{1}{60 + 59{,}99} = 0{,}008334$.

Так нескладними перетвореннями вдалося підвищити точність результату. Отже, при відшукуванні коренів квадратних рівнянь треба передбачити можливу втрату точності і вибирати щоразу потрібну формулу.

Коли використовуються стійкі алгоритми, то завжди дістають результат, достатньо близький до шуканого.

Вправи.

1. Задано систему:

$$\begin{cases} 5x - 331y = 3{,}5, \\ 6x - 397y = 5{,}2. \end{cases}$$

а) розв'яжіть її;

б) розв'яжіть систему, в якій права частина другого рівняння дорівнює не 5,2, а 5,1. Порівняйте результати;

в) підставте в систему $x = 358{,}173$, $y = 5{,}4$ і з'ясуйте, як вони задовольняють систему. Поясніть усі ці факти.

Розділ II
МЕТРИЧНІ І НОРМОВАНІ ПРОСТОРИ

§ 7. Метричні простори. Принцип стискуючих відображень

Означення 1. *Метричним простором називається довільна множина X деяких елементів (їх називають ще точками), в якій для будь-яких двох елементів x, $y \in X$ означено дійсне число $\rho(x, y)$ — відстань між точками x і y, яке задовольняє такі умови (аксіоми метрики):*

1) $\rho(x, y) \geqslant 0$, причому $\rho(x, y) = 0$ тоді і тільки тоді, коли $x = y$;

2) $\rho(x, y) = \rho(y, x)$ (аксіома симетрії);

3) $\rho(x, z) \leqslant \rho(x, y) + \rho(y, z)$ для довільних x, y, z з X (аксіома трикутника).

Очевидно, відстань між елементами однієї й тієї самої множини можна вводити по-різному. При цьому діставатимемо різні простори.

Наведемо кілька прикладів.

1. Множина дійсних чисел з відстанню

$$\rho(x, y) = |x - y|$$

є метричним простором, який позначається $\mathbf{R}$ або $\mathbf{R}^1$.

2. Множина впорядкованих сукупностей з n дійсних чисел, тобто множина n-вимірних векторів.

$$x = (x_1;\ x_2;\ \ldots;\ x_n)$$

з дійсними координатами x_i $(i = 1, 2, \ldots, n)$ і відстанню

$$\rho(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \tag{7.1}$$

є метричним простором. Цей простір називається *n-вимірним арифметичним евклідовим простором* і позначається $\mathbf{R}^n$.

Покажемо, що тут справджується аксіома 3) (виконання аксіом 1) і 2) очевидне). Аксіому 3) можна записати у вигляді:

$$\sqrt{\sum_{i=1}^{n} (z_i - x_i)^2} \leqslant \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} + \sqrt{\sum_{i=1}^{n} (z_i - y_i)^2}. \tag{7.2}$$

Для доведення нерівності (7.2) скористаємося відомою нерівністю Коші*

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leqslant \sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2.$$

Тоді

$$\sum_{i=1}^{n} (a_i + b_i)^2 = \sum_{i=1}^{n} a_i^2 + 2\sum_{i=1}^{n} a_i b_i + \sum_{i=1}^{n} b_i^2 \leqslant \sum_{i=1}^{n} a_i^2 +$$

$$+ 2\sqrt{\sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2} + \sum_{i=1}^{n} b_i^2 = \left(\sqrt{\sum_{i=1}^{n} a_i^2} + \sqrt{\sum_{i=1}^{n} b_i^2}\right)^2.$$

Звідси

$$\sqrt{\sum_{i=1}^{n} (a_i + b_i)^2} \leqslant \sqrt{\sum_{i=1}^{n} a_i^2} + \sqrt{\sum_{i=1}^{n} b_i^2}.$$

Якщо в останній нерівності покласти $a_i = y_i - x_i$; $b_i = z_i - y_i$, то дістанемо потрібну нерівність (7.2).

3. Множина n-вимірних векторів з відстанню

$$\rho_1(x, y) = \max_{1 \leqslant i \leqslant n} |y_i - x_i| \qquad (7.3)$$

є метричним простором і позначається $\mathbf{R}_0^n$.

4. Множина n-вимірних векторів з відстанню

$$\rho_2(x, y) = \sum_{i=1}^{n} |y_i - x_i| \qquad (7.4)$$

є метричним простором і позначається $\mathbf{R}_1^n$.

У прикладах 2, 3 і 4 розглядалася одна й та сама множина. Вводячи по-різному відстань між елементами множини, ми дістали три різні метричні простори.

5. Розглянемо множину всіх неперервних дійсних функцій, визначених на відрізку $[a; b]$. Відстань введемо так:

$$\rho(f, q) = \max_{a \leqslant x \leqslant b} |g(x) - f(x)|. \qquad (7.5)$$

Покажемо, що аксіоми метрики справджуються. Очевидно, що $\rho(f, g) \geqslant 0$ і $\rho(f, g) = 0$ лише тоді, коли $f(x) \equiv g(x)$, а $\rho(f, g) = \rho(g, f)$.

---

* Нерівність Коші випливає з тотожності

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 = \sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2 - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} (a_i b_j - b_i a_j)^2,$$

яку можна перевірити безпосередньо.

Перевіримо аксіому 3. Для будь-якого $x \in [a; b]$ і будь-яких $g(x)$, $f(x)$, $h(x)$ маємо:

$$|g(x) - f(x)| = |g(x) - h(x) + h(x) - f(x)| \leqslant |g(x) - h(x)| +$$

$$+ |h(x) - f(x)| \leqslant \max_{x \in [a;b]} |g(x) - h(x)| + \max_{x \in [a;b]} |h(x) - f(x)|.$$

Звідси випливає, що

$$\max_{x \in [a;b]} |g(x) - f(x)| \leqslant \max_{x \in [a;b]} |g(x) - h(x)| + \max_{x \in [a;b]} |h(x) - f(x)|,$$

тобто

$$\rho(g, f) \leqslant \rho(g, h) + \rho(h, f).$$

Отже, множина неперервних функцій, заданих на відрізку $[a; b]$ з відстанню (7.5), утворює метричний простір. Його позначають $C[a; b]$. Цей простір називають ще *простором неперервних функцій з чебишовською метрикою.*

6. Розглянемо, як і в прикладі 5, множину всіх неперервних функцій, заданих на відрізку $[a; b]$, але відстань між елементами тепер введемо інакше, а саме:

$$\rho(f, g) = \sqrt{\int_a^b (f(x) - g(x))^2 \, dx}.$$

Аксіоми метрики справджуються. Зокрема, справедливість аксіоми трикутника випливає з нерівності Буняковського для інтегралів

$$\left(\int_a^b f(x) g(x) \, dx\right)^2 \leqslant \int_a^b f^2(x) \, dx \int_a^b g^2(x) \, dx,$$

аналогічної нерівності Коші для сум.

Цей метричний простір позначається $C_2[a; b]$ і називається *простором неперервних функцій з квадратичною метрикою.*

Означення 2. *Точку x^* метричного простору X називають границею послідовності $x^{(1)}, x^{(2)}, \ldots, x^{(n)} \ldots$ точок метричного простору X і записують $x^{(n)} \to x^*$ $(n \to \infty)$ або* $\lim\limits_{n \to \infty} x^n = x^*$, *якщо* $\lim\limits_{n \to \infty} \rho(x^{(n)}, x^*) = 0$.

Відомо, що ***збіжна послідовність точок метричного простору може мати тільки одну границю.***

Неважко довести, що відстань $\rho(x, y)$ є неперервною функцією своїх аргументів, тобто якщо $x^{(n)} \to x$ і $y^{(n)} \to y$ при $n \to \infty$, то

$$\rho(x^{(n)}, y^{(n)}) \to \rho(x, y).$$

Означення 3. *Послідовність $\{x^{(n)}\}$ точок метричного простору X називається фундаментальною, якщо для будь-якого $\varepsilon > 0$ існує таке натуральне число $N = N(\varepsilon)$, що $f(x^{(m)}, x^{(n)}) < \varepsilon$ для всіх $m > N$, $n > N$.*

Очевидно, що всяка збіжна послідовність фундаментальна. Справді, якщо $\lim\limits_{n\to\infty} x^{(n)} = a \in X$, то для будь-якого $\varepsilon > 0$ можна знайти таке число N, що $\rho(x^{(n)}, a) < \frac{\varepsilon}{2}$ для $n > N$. Тоді для $n > N$ і $m > N$, скориставшись аксіомою трикутника, маємо:

$$\rho(x^{(m)}, x^{(n)}) \leqslant \rho(x^{(m)}, a) + \rho(a, x^{(n)}) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Обернене твердження не завжди правильне, оскільки існують метричні простори, в яких є фундаментальні послідовності, що не збігаються.

Означення 4. *Метричний простір називається повним, якщо кожна фундаментальна послідовність має в ньому границю.*

Усі метричні простори, наведені в прикладах 1—5, є повними. Повнота простору $\mathbf{R}$ випливає з критерію Коші існування границі числової послідовності: для того щоб послідовність збігалася, необхідно і достатньо, щоб для будь-якого $\varepsilon > 0$ існувало таке N, що для всіх $n > N$, $m > N$ справджувалася нерівність $|x_n - x_m| < \varepsilon$.

Доведемо повноту простору $\mathbf{R}^n$.

Нехай $x^{(m)} = (x_1^{(m)}, x_2^{(m)}, \ldots, x_n^{(m)}) \in \mathbf{R}^n$ $(m = 1, 2, \ldots)$ — фундаментальна послідовність. Тоді для будь-якого $\varepsilon > 0$ знайдеться $N = N(\varepsilon)$, таке, що

$$\rho(x^{(m)}, x^{(l)}) = \sqrt{\sum_{i=1}^{n} (x_i^{(l)} - x_i^{(m)})^2} < \varepsilon$$

при всіх $m > N$, $l > N$.

Звідси для всіх $l > N$, $m > N$ і кожного $i = 1, 2, \ldots, n$ маємо:

$$|x_i^{(l)} - x_i^{(m)}| \leqslant \rho(x^{(m)}, x^{(l)}) < \varepsilon.$$

Отже, числова послідовність $\{x_i^{(m)}\}$ є фундаментальною, тобто збіжною. Нехай

$$\lim_{m\to\infty} x_i^{(m)} = x_i^*.$$

Тоді, очевидно,

$$\lim_{m\to\infty} x^{(m)} = x^*,$$

де

$$x^* = (x_1^*; x_2^*; \ldots; x_n^*) \in \mathbf{R}^n.$$

Отже, повноту простору $\mathbf{R}^n$ доведено.

Аналогічно можна довести повноту просторів $\mathbf{R}_0^n$ і $\mathbf{R}_1^n$.

Зауважимо, що в просторах $\mathbf{R}^n$, $\mathbf{R}_0^n$, $\mathbf{R}_1^n$ збіжність послідовності елементів

$$x^{(m)} = (x_1^{(m)}; x_2^{(m)}; \ldots; x_n^{(m)}) \to x^* = (x_1^*; x_2^*; \ldots; x_n^*) \ (m \to \infty)$$

рівносильна збіжності по координатах.

Прикладом неповного простору є множина раціональних чисел з відстанню

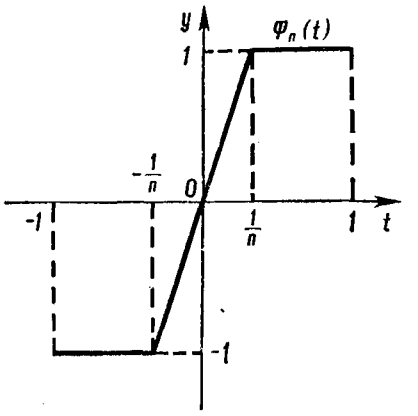
$$\rho(x, y) = |x - y|.$$

Так, послідовність

$$x_n = \left(1 + \frac{1}{n}\right)^n$$

$$(n = 1, 2, \ldots)$$

в цьому просторі є фундаментальною, але в просторі раціональних чисел границі не має. Границею цієї послідовності при $n \to \infty$, як відомо, є число e, яке не є раціональним.



Мал. 2

Неповним простором є також простір $C_2[a; b]$. Наведемо приклад фундаментальної послідовності з $C_2[a; b]$, яка не має в цьому просторі границі. Для простоти за $[a; b]$ візьмемо відрізок $[-1; 1]$ і розглянемо послідовність неперервних функцій (мал. 2):

$$\varphi_n(t) = \begin{cases} -1, \text{ якщо } -1 \leqslant t \leqslant -\frac{1}{n}, \\ nt, \text{ якщо } -\frac{1}{n} \leqslant t \leqslant \frac{1}{n}, \\ 1, \text{ якщо } \frac{1}{n} \leqslant t \leqslant 1. \end{cases}$$

Послідовність $\{\varphi_n(t)\}$ фундаментальна в $C_2[a; b]$, але границі в цьому просторі не має. Вона збігається до розривної

функції $\varphi(t)$

$$\varphi(t)=\begin{cases}-1, & \text{якщо } t<0,\\ 0, & \text{якщо } t=0,\\ 1, & \text{якщо } t>0.\end{cases}$$

Розглянемо тепер так званий принцип стискуючих відображень. Нехай задано дві довільні множини X і Y. Якщо кожному елементу $x\in X$ певним чином ставиться у відповідність один елемент $y\in Y$, то вважають, що задано відображення (оператор) T множини X в множину Y (при цьому пишуть: $y=Tx$).

Означення 5. *Відображення T метричного простору X в себе називається стискуючим, якщо існує таке додатне число $q<1$, що для будь-яких двох точок x, $y\in X$ виконується нерівність*

$$\rho(Tx, Ty)\leqslant q\rho(x, y). \tag{7.6}$$

Зауважимо, що тут q не залежить від x і y.

Означення 6. *Точка x називається нерухомою для відображення T, якщо $Tx=x$.*

Очевидно, нерухомі точки відображення T є розв'язками рівняння

$$Tx=x.$$

Для стискуючих відображень справджується дуже важлива теорема, яка називається теоремою С. Банаха про нерухому точку або принципом стискуючих відображень.

Теорема. ***Стискуюче відображення T повного метричного простору X в себе має нерухому точку і притому лише одну. Цю нерухому точку можна визначити як границю послідовності $x^{(k)}=Tx^{(k-1)}$ $(k=1, 2, \ldots)$, де $x^{(0)}$ — довільний елемент з X.***

Доведення. Нехай $x^{(0)}$ — довільний фіксований елемент простору X і

$$x^{(1)}=Tx^{(0)},\quad x^{(2)}=Tx^{(1)},\quad \ldots,\quad x^{(n)}=Tx^{(n-1)}\quad \ldots$$

З умови теореми випливає, що

$$x^{(i)}\in X\quad (i=0, 1, 2, \ldots).$$

Доведемо, що послідовність $\{x^{(n)}\}$ — фундаментальна.

Справді, нехай $m>n$. Скориставшись аксіомою трикутника, маємо:

$$\rho(x^{(n)}, x^{(m)})\leqslant\rho(x^{(n)}, x^{(n+1)})+\rho(x^{(n+1)}, x^{(n+2)})+\cdots$$
$$\cdots+\rho(x^{(m-1)}, x^{(m)}). \tag{7.7}$$

Врахувавши, що

$$\rho(x^{(1)}, x^{(2)})=\rho(Tx^{(0)}, Tx^{(1)})\leqslant q\rho(x^{(0)}, x^{(1)})=q\rho(x^{(0)}, Tx^{(0)}),$$
$$\rho(x^{(2)}, x^{(3)})=\rho(Tx^{(1)}, Tx^{(2)})\leqslant q\rho(x^{(1)}, x^{(2)})\leqslant q^2\rho(x^{(0)}, Tx^{(0)}),$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\rho(x^{(n)}, x^{(n+1)})\leqslant q^n\rho(x^{(0)}, Tx^{(0)}),$$

з останньої нерівності дістанемо:

$$\rho(x^{(n)}, x^{(m)})\leqslant(q^n+q^{n+1}+\cdots+q^{m-1})\rho(x^{(0)}, Tx^{(0)})=$$
$$=\frac{q^n(1-q^{m-n})}{1-q}\rho(x^{(0)}, Tx^{(0)})$$

Оскільки $q<1$, то

$$\rho(x^{(n)}, x^{(m)})<\frac{q^n}{1-q}\rho(x^{(0)}, Tx^{(0)}). \tag{7.8}$$

Отже, для досить великих n і m число $\rho(x^{(n)}, x^{(m)})$ — як завгодно мале. Іншими словами, для будь-якого $\varepsilon>0$ існує таке $N=N(\varepsilon)$, що для всіх $n>N$, $m>N$ $\rho(x^{(n)}, x^{(m)})<\varepsilon$, тобто послідовність $\{x^{(n)}\}$ фундаментальна.

З повноти простору X випливає, що границя послідовності $\{x^{(n)}\}$ існує і належить X. Припустимо, що

$$\lim_{n\to\infty}x^{(n)}=x^*.$$

Доведемо, що x^* є нерухомою точкою відображення T, тобто що

$$Tx^*=x^*.$$

Справді,

$$\rho(x^*, Tx^*)\leqslant\rho(x^*, x^{(n)})+\rho(x^{(n)}, Tx^*)=\rho(x^*, x^{(n)})+$$
$$+\rho(Tx^{(n-1)}, Tx^*)\leqslant\rho(x^*, x^{(n)})+q\rho(x^{(n-1)}, x^*).$$

Але при будь-якому $\varepsilon>0$ і досить великому n

$$\rho(x^*, x^{(n)})<\frac{\varepsilon}{2},\quad \rho(x^*, x^{(n-1)})<\frac{\varepsilon}{2}.$$

Отже,

$$\rho(x^*, Tx^*)<\varepsilon.$$

Оскільки $\varepsilon>0$ довільне, то $\rho(x^*, Tx^*)=0$, тобто $Tx^*=x^*$. Доведемо тепер, що існує тільки одна нерухома точка.

Припустимо, що існують дві нерухомі точки x^*, $y^*\in X$, тобто $Tx^*=x^*$ і $Ty^*=y^*$.

Тоді

$$\rho(x^*, y^*)=\rho(Tx^*, Ty^*)\leqslant q\rho(x^*, y^*).$$

Якщо припустити, що $\rho(x^*, y^*) > 0$, то з останнього випливає, що $q \geqslant 1$. А це суперечить умові теореми. Теорему доведено.

Важливість доведеної теореми про нерухому точку полягає не тільки в тому, що вона вказує на існування нерухомої точки відображення, а й у тому, що вона дає метод її знаходження. Це метод послідовних наближень. Важливою властивістю його є те, що він дає можливість діставати з довільною точністю наближення нерухомої точки.

Примітка. Якщо виконуються умови теореми, то починати відшукування послідовних наближень $x^{(n)}$, які збігаються до нерухомої точки x^*, можна з будь-якого елемента $x^{(0)} \in X$. Вибір початкового наближення $x^{(0)}$ впливатиме лише на кількість ітерацій при знаходженні границі із заданим ступенем точності. Коли перейти в формулі (7.8) до границі при $m \to \infty$, то дістанемо таку оцінку n-го наближення:

$$\rho(x^{(n)}, x^*) \leqslant \frac{q^n}{1-q} \rho(x^{(0)}, Tx^{(0)}). \tag{7.9}$$

З цієї оцінки випливає, що метод послідовних наближень тим швидше збігається, чим менше q.

Виведемо ще одну оцінку для похибки n-го наближення. Якщо $m > n$,

$$\rho(x^{(n)}, x^{(m)}) \leqslant \rho(x^{(n)}, x^{(n+1)}) + \rho(x^{(n+1)}, x^{(n+2)}) + \cdots$$

$$\cdots + \rho(x^{(m-1)}, x^{(m)}) \leqslant \rho(x^{(n)}, x^{(n+1)}) + q\rho(x^{(n)}, x^{(n+1)}) +$$

$$+ q^2\rho(x^{(n)}, x^{(n+1)}) + \cdots + q^{m-n-1}\rho(x^{(n)}, x^{(n+1)}) =$$

$$= (1 + q + q^2 + \cdots + q^{m-n-1})\rho(x^{(n)}, x^{(n+1)}) =$$

$$= \frac{1-q^{m-n}}{1-q} \rho(x^{(n)}, x^{(n+1)}) \leqslant \frac{q(1-q^{m-n})}{1-q} \rho(x^{(n-1)}, x^{(n)}),$$

тобто

$$\rho(x^{(n)}, x^{(m)}) \leqslant \frac{q(1-q^{m-n})}{1-q} \rho(x^{(n)}, x^{(n-1)}).$$

Звідки при $m \to \infty$ дістанемо:

$$\rho(x^{(n)}, x^*) \leqslant \frac{q}{1-q} \rho(x^{(n)}, x^{(n-1)}), \tag{7.10}$$

де x^* — нерухома точка ($\lim\limits_{m\to\infty} x^{(m)} = x^*$).

Часто на практиці доводиться відшукувати нерухому точку x^* відображення T з точністю до ε.

У цьому випадку за наближення x^* треба взяти таке $x^{(n)}$, що задовольняє умову

$$\rho(x^{(n)}, x^*) \leqslant \varepsilon.$$

Якщо x^* шукати методом послідовних наближень (звичайно, при умові, що він збігається), то з нерівності (7.10) випливає, що обчислення за формулами $x^{(n)} = Tx^{(n-1)}$ ($n = 1, 2, \ldots$) треба продовжувати доти, поки не дістанемо нерівність:

$$\frac{q}{1-q} \rho(x^{(n)}, x^{(n-1)}) \leqslant \varepsilon,$$

або, що те саме,

$$\rho(x^{(n)}, x^{(n-1)}) \leqslant \frac{1-q}{q} \varepsilon = \varepsilon_1. \tag{7.11}$$

§ 8. Лінійні нормовані простори

Поняття лінійного простору широко вживається не тільки в математиці, а й в інших науках, зокрема у фізиці. У шкільному курсі математики з лінійними просторами зустрічаємося, наприклад, розглядаючи множини векторів на прямій, на площині і в просторі.

Означення 1. *Непорожня множина L елементів x, y, z, ... довільної природи називається лінійним, або векторним, простором, якщо вона задовольняє такі умови (аксіоми лінійного простору):*

1. *Для будь-яких двох елементів x, $y \in L$ однозначно задається третій елемент з множини L, який називається їх сумою (позначається $x + y$), причому операція додавання задовольняє умови:*

а) $x + y = y + x$ *(комутативність);*

б) $x + (y + z) = (x + y) + z$ *(асоціативність);*

в) *в L існує нульовий елемент 0, такий, що $x + 0 = x$ для всіх $x \in L$;*

г) *для кожного $x \in L$ в L існує такий елемент — x (його називають протилежним елементу x), що $x + (-x) = 0$.*

Замість $x + (-y)$ пишуть $x - y$.

2. *Для будь-якого числа α і будь-якого $x \in L$ визначений елемент $\alpha x \in L$, який називається добутком елемента x на число α, причому добуток задовольняє умови:*

а) $\alpha(\beta x) = (\alpha\beta)x$, *де β — довільне число (асоціативність множення);*

б) $1x = x$;

в) $(\alpha + \beta)x = \alpha x + \beta x$; $\alpha(x + y) = \alpha x + \alpha y$ *(закони дистрибутивності).*

Примітка. Якщо при цьому використовується множина дійсних чисел, то лінійний простір називається дійсним, а коли використовується множина комплексних чисел, то — комплексним. Ми розглядатимемо лише дійсні простори.

Зауважимо, що правила виконання операцій над геометричними векторами підпорядковані загальним аксіомам лінійного простору.

Приклади лінійних просторів.

1. Простір дійсних чисел **R** із звичайними операціями додавання і множення.

2. Множина всіх алгебраїчних многочленів степеня, який не перевищує n.

3. Множина n-вимірних векторів $x = (x_1; x_2; \ldots; x_n)$, додавання яких і множення на число визначається формулами

$$(x_1; x_2; \ldots; x_n) + (y_1; y_2; \ldots; y_n) = \\ = (x_1 + y_1; x_2 + y_2; \ldots; x_n + y_n),$$

$$\alpha (x_1; x_2; \ldots; x_n) = (\alpha x_1; \alpha x_2; \ldots; \alpha x_n).$$

4. Множина неперервних функцій, заданих на $[a; b]$ із звичайними операціями додавання двох функцій і множення функцій на число.

5. Нехай маємо множину всіх функцій, заданих на відрізку $[a; b]$. Дві функції $f(x)$ і $g(x)$ вважатимемо тотожними, якщо $f(x_i) = g(x_i)$ (x_i ($i = 0, 1, 2, \ldots, n$) — деякі точки з відрізка $[a; b]$). Дві тотожні функції не розрізняються між собою і вважаються одним і тим самим елементом простору. Таким чином, елементами простору тут є цілі класи функцій. *Нульовим елементом* у цьому просторі є клас функцій, які в точках $x_0, x_1, \ldots, x_n$ дорівнюють нулю. Кожна конкретна функція, що задана в точках x_i ($i = 0, 1, \ldots, n$) на відрізку $[a; b]$, належить, очевидно, одному з класів еквівалентних функцій і є його представником. *Сумою* двох елементів (двох класів) простору називатимемо клас, якому належить сума представників класів доданків. *Добутком* елемента (класу) на число називатимемо клас, якому належить добуток представника класу на це число. Суму представників і добуток представника на число вводимо звичайним способом.

Неважко довести, що результати введених в просторі операцій не залежать від вибору згаданих представників.

Очевидно, що така множина утворює лінійний простір. (У даному випадку інформація про функцію задається її значеннями на скінченній множині точок $x_0, x_1, \ldots, x_n$ з відрізка $[a; b]$; таку ситуацію маємо, коли функції задаються таблично).

У лінійному просторі природно розглядати вирази $\sum_{k=}^{n} \alpha_k x_k$ (α_k — числа, x_k — елементи цього лінійного простору), так звані лінійні комбінації елементів $x_1, x_2, \ldots, x_n$.

О з н а ч е н н я 2. *Елементи $x_1, x_2, \ldots, x_n$ лінійного простору L називаються лінійно залежними, якщо існують такі*

числа $\alpha_1, \alpha_2, \ldots, \alpha_n$ (серед них не всі дорівнюють нулю), що

$$\sum_{k=1}^{n} \alpha_k x_k = \theta.$$

Якщо остання рівність справджується лише при $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$, то ці елементи називаються лінійно незалежними.

Якщо в просторі L існує $n > 0$ лінійно незалежних елементів, а будь-які $n + 1$ елементи цього простору лінійно залежні, то кажуть, що простір L має *розмірність n*.

Якщо ж у просторі L можна вибрати систему з довільного числа лінійно незалежних елементів, то простір називається *нескінченновимірним*. Розмірність простору L позначається через $\dim L$.

Неважко встановити, що $\dim R = 1$; $\dim R^n = n$, $\dim C[a; b] = \infty$. Наприклад, те, що простір $C[a; b]$ нескінченновимірний випливає вже з того, що функції з цього простору $1, x, x^2, \ldots, x^n$ при будь-якому n є лінійно незалежними. Справді, $\alpha_0 1 + \alpha_1 x + \alpha_2 x^2 + \ldots + \alpha_n x^n \equiv 0$ лише тоді, коли $\alpha_i = 0$ ($i = 0, 1, 2, \ldots, n$).

Якщо лінійний простір є метричним, то він називається *лінійним метричним простором*. Важливим класом лінійних метричних просторів є клас *нормованих* просторів.

О з н а ч е н н я 3. *Лінійний простір L називається нормованим, якщо кожному елементу $x \in L$ поставлено у відповідність дійсне число $\| x \|$, яке називається його нормою і задовольняє такі умови (аксіоми норми):*

1. $\| x \| \geqslant 0$, *причому* $\| x \| = 0$ *тоді і тільки тоді, коли* $x = 0$ *(додатна визначеність);*

2. $\| x + y \| \leqslant \| x \| + \| y \|$, $x, y \in L$ *(нерівність трикутника);*

3. $\| \alpha x \| = |\alpha| \cdot \| x \|$, α — *число (однорідність).*

Кожний лінійний нормований простір можна зробити метричним, ввівши таку метрику: $\rho(x, y) = \| x - y \|$.

Тоді норма будь-якого елемента дорівнюватиме відстані його від нульового елемента.

Неважко перевірити, що всі аксіоми метрики при цьому виконуються. Отже, всі поняття, введені для метричних просторів, переносяться на нормовані простори. Збіжність у цьому випадку називають збіжністю за нормою.

Якщо лінійний нормований простір є повним, то він називається *банаховим*, або B-*простором*.

Приклади нормованих просторів.

1. Множина дійсних чисел **R** буде нормованим простором,

якщо для всякого числа $x \in \mathbf{R}$ норму ввести так: $\|x\| = |x|$.

2. Простір n-вимірних векторів $x = (x_1; \ldots; x_n)$ $\mathbf{R}^n$ стає нормованим, якщо покласти:

$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Неважко перевірити, що всі аксіоми норми виконуються. Ця норма породжує метрику

$$\rho(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Таким чином, $\mathbf{R}^n$ є B-простором, причому породжена нормою метрика збігається з уведеною раніше. Зауважимо, що в цьому просторі норму можна ввести й інакше:

$$\|x\| = \begin{cases} \|x\|_1 = \max\limits_{1 \leqslant i \leqslant n} |x_i| \text{ або} \\ \|x\|_2 = \sum\limits_{i=1}^{n} |x_i|. \end{cases}$$

Ці норми породжують метрики, які розглядалися в прикладах 3 і 4 § 7.

3. У просторі $C[a; b]$ неперервних функцій, заданих на відрізку $[a; b]$, норму введемо так: $\|f\| = \max\limits_{x \in [a;b]} |f(x)|$. Очевидно, всі аксіоми норми виконуються. Ця норма дає відстань, яка вже нами була розглянута (приклад 5, § 7).

Легко довести, що всі розглянуті простори є повними, тобто банаховими.

Зауважимо, що не всі лінійні нормовані простори є повними. Неповними є, наприклад, простір всіх раціональних чисел з нормою $\|x\| = |x|$; простір усіх неперервних функцій, заданих на $[a; b]$, з нормою $\|f\| = \sqrt{\int_a^b f^2(x)\,dx}$.

Розглянемо тепер так звані евклідові простори. Відомо, яку важливу роль у звичайному векторному просторі (на площині, у просторі) відіграє скалярний добуток. Поняття скалярного добутку можна узагальнити на довільні лінійні простори. Простори, в яких введено скалярний добуток, займають особливе місце серед усіх просторів. Виявляється, що норму в таких просторах можна вводити через скалярний добуток.

О з н а ч е н н я 4. *Скалярним добутком двох довільних елементів x і y з дійсного лінійного простору L називається дійсне число, яке позначається через (x, y) і задовольняє такі умови (аксіоми):*

1. $(x, x) \geqslant 0$, *причому* $(x, x) = 0$ *тоді і тільки тоді, коли* $x = \theta$.
2. $(x, y) = (y, x)$.
3. $(x + y, z) = (x, z) + (y, z)$, $x, y, z \in L$.
4. $(\alpha x, y) = \alpha (x, y)$, α — *довільне дійсне число.*

Очевидно, скалярний добуток векторів (на площині, у просторі), який відомий з шкільного курсу математики, задовольняє аксіоми 1—4.

Лінійний простір з фіксованим в ньому скалярним добутком називають *евклідовим* * (його часто позначають через E).

З аксіом 1—4 можна вивести ряд властивостей скалярного добутку. Зокрема, з аксіом 2—4 випливає:

$$(x, \lambda y) = \lambda (x, y) \text{ і } (x, y + z) = (x, y) + (x, z).$$

В евклідовому просторі справджується важлива нерівність:

$$|(x, y)| \leqslant \sqrt{(x, x)}\,\sqrt{(y, y)}. \qquad (8.1)$$

Рівність виконується тоді і тільки тоді, коли x і y лінійно залежні.

Для доведення нерівності розглянемо вираз $\varphi(\alpha) = (\alpha x + y, \alpha x + y) \geqslant 0$, де α — довільне дійсне число. Розкривши скалярний добуток, матимемо:

$$\varphi(\alpha) = (x, x)\alpha^2 + 2(x, y)\alpha + (y, y).$$

Отже, $\varphi(\alpha)$ — квадратний тричлен від дійсного аргументу α. Оскільки $\varphi(\alpha) \geqslant 0$, то дискримінант цього тричлена недодатний:

$$(x, y)^2 - (x, x)(y, y) \leqslant 0.$$

Звідси і випливає шукана нерівність.

В евклідовому просторі норму можна ввести за формулою

$$\|x\| = \sqrt{(x, x)}. \qquad (8.2)$$

У цьому випадку кажуть, що *норма породжується скалярним добутком.*

Як бачимо, для звичайного векторного простору (на прямій, на площині, в тривимірному просторі) норма вектора буде не що інше, як довжина вектора.

Норма елемента, введена формулою (8.2), задовольняє всі аксіоми норми. Аксіоми 1 і 3, очевидно, справджуються, а

* Інколи цей простір називають *передгільбертовим*. А евклідовим називають тільки простір $\mathbf{R}^n$ із скалярним добутком $(x, y) = \sum\limits_{i=1}^{n} x_i y_i$.

аксіома 2 (аксіома трикутника) випливає з нерівності (8.1):

$$\|x+y\|^2=(x+y,\ x+y)=(x,\ x)+2(x,\ y)+(y,\ y)\leqslant$$
$$\leqslant(x,\ x)+2\sqrt{(x,\ x)}\sqrt{(y,\ y)}+(y,\ y)=(\|x\|+\|y\|)^2.$$

Користуючись означенням норми, нерівність (8.1) можна записати так:

$$|(x,\ y)|\leqslant\|x\|\|y\|. \tag{8.3}$$

Нерівність (8.3), як і нерівність (8.1), називається *нерівністю Коші — Буняковського*.

В евклідовому просторі справджуються й інші властивості, відомі для звичайного векторного простору. Зокрема, легко переконатися, що для будь-яких елементів x і y евклідового простору виконується рівність

$$\|x+y\|^2+\|x-y\|^2=2(\|x\|^2+\|y\|^2),$$

яка називається *рівністю паралелограма*. Для векторів на площині остання рівність має такий геометричний зміст: сума квадратів діагоналей паралелограма дорівнює сумі квадратів його сторін.

Оскільки кожний евклідів простір можна зробити метричним, ввівши відстань за формулою

$$\rho(x,\ y)=\|x-y\|=\sqrt{(x-y,\ x-y)},$$

то всі поняття і твердження, доведені для метричних просторів, можна перенести на евклідові простори.

В евклідових просторах за аналогією до звичайних векторних просторів можна ввести поняття ортогональності елементів.

Два елементи x і y з простору E називаються *ортогональними*, якщо $(x,\ y)=0$. Система ненульових векторів $\{x_i\}$ називається *ортогональною*, якщо $(x_i,\ x_j)=0$ при $i\neq j$. Якщо, крім того, $(x_i,\ x_i)=1$, то система називається *ортонормованою*.

Виявляється, що будь-яку скінченну чи зчисленну множину лінійно незалежних елементів $\{x_i\}$ можна ортонормувати, тобто скласти з лінійних комбінацій елементів x_i нову систему елементів, яка буде ортонормованою.

За аналогією до звичайних векторних просторів у довільному евклідовому просторі можна ввести поняття *базису*.

Ортонормовану систему називають *повною*, якщо в просторі не існує жодного елемента (відмінного від нульового), ортогонального до всіх елементів системи, тобто повнота системи означає, що її не можна розширити приєднанням нових елементів простору.

Повна ортонормована система елементів евклідового простору називається ортонормованим базисом евклідового простору.

Повні евклідові простори називають *гільбертовими просторами*.

Приклади.

1. Простір n-вимірних векторів $\mathbf{R}^n$ є гільбертовим простором, якщо скалярний добуток ввести за формулою:

$$(x,\ y)=\sum_{i=1}^{n}x_iy_i,\ \text{де}\ x=(x_1;\ x_2;\ \ldots;\ x_n),\ y=(y_1;\ y_2;\ \ldots;\ y_n).$$

Очевидно, система векторів

$$e_1=(1;\ 0;\ 0;\ \ldots;\ 0),\quad e_2=(0;\ 1;\ 0;\ \ldots;\ 0),\ \ldots$$
$$\ldots,\ e_n=(0;\ 0;\ 0;\ \ldots;\ 1)$$

буде тут ортонормованим базисом. Кожний вектор $x\in R^n$ можна однозначно подати у вигляді $x=\sum_{i=1}^{n}\alpha_ie_i$.

2. В просторі $C_2[a;\ b]$, що складається з неперервних функцій, скалярний добуток можна ввести так:

$$(f,\ g)=\int_a^b f(x)\,g(x)\,dx.$$

Неважко довести, що в цьому разі всі аксіоми скалярного добутку виконуються. Доведемо, наприклад, аксіому 1. Для будь-якої функції $f\in C_2[a;\ b]$ $(f,\ f)\geqslant 0$. Якщо $f=0$, тобто $f(x)\equiv 0$, то очевидно, $(f,\ f)=0$. Нехай тепер $(f,\ f)=0$, тобто

$$\int_a^b f^2(x)\,dx=0. \tag{8.4}$$

Доведемо, що тоді $f(x)\equiv 0$.

Припустимо, що $f(x)\not\equiv 0$ на $[a;\ b]$. Тоді в силу неперервності функції $f^2(x)$ знайдеться відрізок $[a';\ b']\subset[a;\ b]$, на якому $f^2(x)>0$, і оскільки неперервна функція досягає на відрізку свого мінімуму, то $m=\min\limits_{x\in[a';b']}f^2(x)>0$. Звідси

$$\int_a^b f^2(x)\,dx\geqslant\int_{a'}^{b'}f^2(x)\,dx\geqslant\int_{a'}^{b'}m\,dx=m(b'-a')>0.$$

Отже, дістали протиріччя з (8.4). Тому $f(x)=0$.

Аксіома 1, таким чином, виконується. Перевірку аксіом 3 і 4 легко здійснити, використовуючи відомі лінійні властивості визначеного інтеграла, а виконання аксіоми 2 очевидне.

Простір $C_2[a; b]$ є евклідовим, але гільбертовим він не буде. Ортогональною системою в $C_2[a; b]$ є система функцій

$$1,\ \cos k\frac{2\pi x}{b-a},\quad \sin k\frac{2\pi x}{b-a}\quad (k=1,\ 2,\ \ldots).$$

3. Розглянемо простір функцій, що задані на скінченній (дискретній) множині точок $x_0, x_1, \ldots, x_n$ (див. лінійні простори, приклад 5). Скалярний добуток введемо так:

$$(f, g) = \sum_{i=1}^{n} f(x_i)\, g(x_i).$$

Будь-яку функцію f з цього простору можна вважати $n+1$-вимірним вектором $f = (f_0, f_1, \ldots, f_n)$, де $f_i = f(x_i)$. Введений скалярний добуток, по суті, нічим не відрізнятиметься від скалярного добутку в $(n+1)$-вимірному евклідовому векторному просторі $\mathbf{R}^{n+1}$. Тоді для норми елемента f матимемо:

$$\|f\| = \sqrt{(f, f)} = \sqrt{\sum_{i=1}^{n} f^2(x_i)}.$$

Отже, простір таких функцій є гільбертовим простором розмірності $n+1$.

4. Простір $C[a; b]$ неперервних на відрізку $[a; b]$ функцій з нормою $\|f\| = \max\limits_{a\leqslant x\leqslant b} |f(x)|$ є нормованим, але не евклідовим.

Розділ III
МЕТОДИ РОЗВ'ЯЗУВАННЯ СИСТЕМ ЛІНІЙНИХ АЛГЕБРАЇЧНИХ РІВНЯНЬ

Методи розв'язування систем лінійних рівнянь можна поділити на два типи: *прямі*, або *точні*, та *ітераційні*. Прямі методи дають можливість дістати розв'язок, виконавши скінченне число дій. Якщо всі проміжні обчислення виконувати точно (без округлень), то дістанемо точний розв'язок. Ітераційні методи дають нескінченну послідовність наближених розв'язків, границя якої є розв'язком системи.

§ 9. Жорданові виключення

Розглянемо рівності:

$$\begin{aligned}
\delta_1(x_1, x_2, \ldots, x_n) &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + b_1,\\
\delta_2(x_1, x_2, \ldots, x_n) &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + b_2,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
\delta_n(x_1, x_2, \ldots, x_n) &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n + b_n.
\end{aligned} \tag{9.1}$$

Нехай потрібно знайти значення змінних x_j^0 $(j = 1, 2, \ldots, n)$, при яких змінні $\delta_i(x_1, x_2, \ldots, x_n)$ набувають наперед заданих значень δ_i^0 $(i = 1, 2, \ldots, n)$. Геометрично кожна рівність $a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n + b_i = 0$ $(i = 1, 2, \ldots, n)$ (або, що те саме, $\delta_i(x_1, x_2, \ldots, x_n) = 0$ $(i = 1, 2, \ldots, n)$ (визначає деяку *гіперплощину* (у тривимірному просторі — площину, в двовимірному — пряму). Якщо в рівняння деякої гіперплощини

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n + b_i = 0$$

підставити координати довільної точки $M(x_1; x_2; \ldots; x_n)$, то дістанемо так зване *відхилення* $\delta_i(x_1, x_2, \ldots, x_n)$ *точки* $M(x_1; x_2; \ldots; x_n)$ *від гіперплощини* $a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n + b_i = 0$. Якщо при цьому виявиться, що відхилення $\delta_i(x_1, x_2, \ldots, x_n)$ дорівнює нулю, тобто координати точки M задовольняють рівняння $a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n + b_i = 0$, то говорять, що *точка* M *належить гіперплощині*

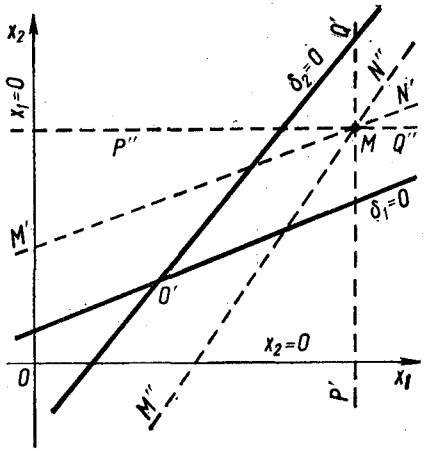
$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n + b_i = 0,$$

або, іншими словами, гіперплощині

$$\delta_i(x_1, x_2, \ldots, x_n) = 0.$$

Зауважимо, що самі змінні $x_1, x_2, \ldots, x_n$ є відхиленнями точки M від координатних гіперплощин $x_1 = 0$, $x_2 = 0$, ..., $x_n = 0$ (рівняння $x_i = 0$, як відомо, є рівнянням деякої гіперплощини). Такі гіперплощини $x_1 = 0$, $x_2 = 0$, ..., $x_n = 0$, як правило, вибираються взаємно ортогональними.

Однак координатні гіперплощини не обов'язково повинні бути взаємно ортогональні. Якщо задати відхилення точки від гіперплощин $\delta_1 = 0$, $\delta_2 = 0$, ..., $\delta_n = 0$, які мають єдину спільну точку, то положення такої точки в просторі визначається однозначно. На мал. 3 множина точок прямої $M'N'$ є множиною точок, рівновідхилених від прямої $\delta_1 = 0$ на одну й ту саму величину δ_1^0, $M''N''$ — множина точок, рівновідхилених від прямої $\delta_2 = 0$ на одну й ту саму величину δ_2^0. Точка M належить обом цим множинам. Аналогічно $P'Q'$ є множиною точок, рівновідхилених від прямої $x_1 = 0$ на одну й ту саму величину x_1^0, $P''Q''$ — множина точок, рівновідхилених від прямої $x_2 = 0$ на одну й ту саму величину x_2^0. Положення точки M цілком визначається парою чисел (δ_1^0, δ_2^0). Отже, пару чисел (δ_1^0, δ_2^0) можна розглядати як координати точки M, якщо за координатні прямі взяти $\delta_1 = 0$, $\delta_2 = 0$, аналогічно тому, як пару чисел (x_1^0, x_2^0) вважають координатами точки M, якщо за координатні прямі взяти $x_1 = 0$, $x_2 = 0$. Очевидно, відповідні пари чисел (x_1^0, x_2^0) та (δ_1^0, δ_2^0) визначають одну й ту саму точку M.



Мал. 3

У n-вимірному випадку відхилення $\delta_i(x_1, x_2, \ldots, x_n)$ можна тлумачити як координати точки M, якщо за координатні взято площини $\delta_i(x_1, x_2, \ldots, x_n) = 0$ (які, за припущенням, мають єдину спільну точку).

Очевидно (мал. 3), у двовимірному просторі за координатні можна взяти будь-які дві прямі, що мають одну спільну точку, наприклад прямі $\delta_1 = 0$ та $x_2 = 0$, $\delta_1 = 0$ та $x_1 = 0$, $\delta_2 = 0$ та $x_1 = 0$, $\delta_2 = 0$ та $x_2 = 0$ і т. д. Якщо задати відхилення деякої точки від будь-якої пари таких прямих, то положення точки в просторі (двовимірному) повністю визначається. Аналогічно міркують щодо n-вимірного простору.

Надалі замість слова «гіперплощина» вживатимемо просто «площина» (в двовимірному випадку, маємо на увазі пряму). Про що саме йтиметься, залежатиме від того, якої розмірності простір розглядатиметься.

Повернемось тепер до нашої задачі. Щоб розв'язати задачу, діятимемо так. Одну за однією вилучатимемо змінні $x_1, x_2, \ldots, x_n$ з незалежних змінних, вводячи щоразу до незалежних одну із змінних $\delta_1, \delta_2, \ldots, \delta_n$.

Нехай, наприклад, $a_{11} \neq 0$, тоді з першої рівності системи (9.1) матимемо:

$$x_1 = \frac{1}{a_{11}}\delta_1 - \frac{a_{12}}{a_{11}}x_2 - \cdots - \frac{a_{1n}}{a_{11}}x_n - \frac{b_1}{a_{11}}. \qquad (9.2)$$

Підставивши вираз для x_1 у вирази для $\delta_2, \delta_3, \ldots, \delta_n$, дістанемо рівності, в яких до незалежних змінних вже не входитиме змінна x_1. Незалежними змінними тепер будуть $\delta_1, x_2, x_3, \ldots, x_n$, а змінні $x_1, \delta_2, \delta_3, \ldots, \delta_n$ виражатимуться через $\delta_1, x_2, \ldots, x_n$ і будуть, таким чином, залежними.

Для δ_2, наприклад, маємо:

$$\delta_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + b_2 =$$

$$= a_{21}\left(\frac{1}{a_{11}}\delta_1 - \frac{a_{12}}{a_{11}}x_2 - \cdots - \frac{a_{1n}}{a_{11}}x_n - \frac{b_1}{a_{11}}\right) +$$

$$+ a_{22}x_2 + \cdots + a_{2n}x_n + b_2 = \frac{a_{21}}{a_{11}}\delta_1 + \left(a_{22} - a_{21}\frac{a_{12}}{a_{11}}\right) \times$$

$$\times x_2 + \cdots + \left(a_{2n} - a_{21}\cdot\frac{a_{1n}}{a_{11}}\right)x_n + \left(b_2 - a_{21}\cdot\frac{b_1}{a_{11}}\right) =$$

$$= \frac{a_{21}}{a_{11}}\delta_1 + \left(\frac{a_{22}a_{11} - a_{21}a_{12}}{a_{11}}\right)x_2 + \cdots + \left(\frac{a_{2n}a_{11} - a_{21}a_{1n}}{a_{11}}\right)x_n +$$

$$+ \frac{b_2a_{11} - a_{21}b_1}{a_{11}}. \qquad (9.3)$$

Аналогічно дістанемо вирази для $\delta_3, \ldots, \delta_n$.

При цьому відбувається перехід до системи координат $(\delta_1, x_2, \ldots, x_n)$, в якій положення точки в просторі визначається відхиленнями від площини $\delta_1 = 0$, $x_2 = 0$, ..., $x_n = 0$.

Аналогічно з будь-якого виразу, наприклад з виразу для δ_2, визначимо змінну x_2 і підставимо результат у всі інші вирази для $x_1, \delta_3, \delta_4, \ldots, \delta_n$.

Таким чином, ми виключимо змінну x_2 з незалежних, ввівши замість неї до незалежних змінну δ_2.

Кожний такий крок вилучення якоїсь змінної з незалежних змінних і введення замість неї до незалежних нової змінної називають *кроком жорданового виключення*.

Обчислення при цьому зручно подати у вигляді послідовності таблиць. Вихідною є таблиця:

Таблиця 1

	x_1	x_2	...	x_n	1
δ_1	a_{11}	a_{12}	...	a_{1n}	b_1
δ_2	a_{21}	a_{22}	...	a_{2n}	b_2
...	...	...	...	...	...
δ_n	a_{n1}	a_{n2}	...	a_{nn}	b_n

де в скороченій формі записано рівності (9.1).

Позначивши через $a_{ij}^{(1)}$ коефіцієнти при незалежних змінних, знайдені після першого кроку жорданового виключення, матимемо таблицю:

Таблиця 2

	δ_1	x_2	...	x_n	1
x_1	$a_{11}^{(1)}$	$a_{12}^{(1)}$	...	$a_{1n}^{(1)}$	$b_1^{(1)}$
δ_2	$a_{21}^{(1)}$	$a_{22}^{(1)}$	...	$a_{2n}^{(1)}$	$b_2^{(1)}$
...	...	...	...	...	...
δ_n	$a_{n1}^{(1)}$	$a_{n2}^{(1)}$	...	$a_{nn}^{(1)}$	$b_n^{(1)}$

Назвемо рядок таблиці, що відповідає змінній, яка вводиться до числа незалежних, *основним рядком*. У таблиці 1 основним є перший рядок (відповідає змінній δ_1).

Стовпчик, що відповідає змінній, яка виводиться з незалежних, назвемо *основним стовпчиком*. У розглянутому випадку основним є перший стовпчик (відповідає змінній x_1).

Елемент, що належить основному рядкові і основному стовпчикові, назвемо *основним елементом*. У таблиці 1 таким елементом є a_{11}.

Як видно з рівностей (9.2), (9.3), крок жорданового виключення (перехід від таблиці 1 до таблиці 2) треба виконувати за ***правилами:***

1) *Основний елемент* a_{ij} *замінюємо на* $\frac{1}{a_{ij}}$;

2) *Усі інші елементи основного рядка ділимо на основний елемент, змінивши знаки на протилежні*;

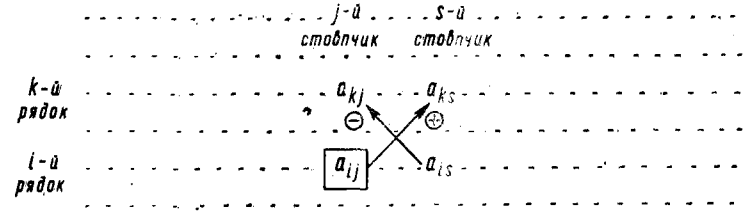
3) *Усі інші елементи основного стовпчика ділимо на основний елемент;*

4) *Усі інші елементи нової таблиці обчислюємо за формулою*

$$a'_{ks} = \frac{a_{ks}a_{ij} - a_{is}a_{kj}}{a_{ij}},$$

де a_{ij} *— основний елемент,* $k \neq i$, $s \neq j$, a_{ij}, a_{ks}, a_{is}, a_{kj} *— елементи вихідної таблиці,* a'_{ks} *— елемент нової таблиці.*

Схематично правило 4 подано на малюнку 4. Добуток, у який входить основний елемент, завжди беруть із знаком «+», а добуток двох інших елементів — із знаком «—».



Мал. 4

Примітка. При практичних обчисленнях головним критерієм вибору основного елемента є точність і зручність обчислень. Зауважимо, що не обов'язково визначати x_1 з першої рівності, x_2 — з другої рівності і т. д. Зрозуміло також, що, вилучивши в розглядуваній задачі деяку змінну x_j з незалежних, повертати цю змінну знову до незалежних немає потреби. Отже, при розв'язуванні розглядуваної задачі кожний рядок і кожний стовпчик беруться за основні лише один раз. Стовпчик вільних членів основним не береться ніколи. Очевидно, елементи a'_{ks} обчислюються найпростіше, коли $a_{ij} = 1$. Але такий вибір основного елемента не завжди можливий.

Нехай після r кроків жорданових виключень дістанемо таблицю.

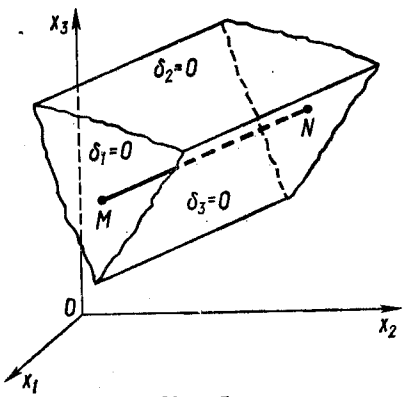
Таблиця 3

	δ_1	...	δ_r	x_{r+1}	...	x_n	1
x_1	$a_{11}^{(r)}$	...	$a_{1r}^{(r)}$	$a_{1r+1}^{(r)}$	...	$a_{1n}^{(r)}$	$b_1^{(r)}$
...	...	...	...	...	...	...	...
x_r	$a_{r1}^{(r)}$	...	$a_{rr}^{(r)}$	$a_{rr+1}^{(r)}$	...	$a_{rn}^{(r)}$	$b_r^{(r)}$
δ_{r+1}	$a_{r+11}^{(r)}$	...	$a_{r+1r}^{(r)}$	0	...	0	$b_{r+1}^{(r)}$
...	...	...	...	...	...	...	...
δ_n	$a_{n1}^{(r)}$	...	$a_{nr}^{(r)}$	0	...	0	$b_n^{(r)}$

У цьому випадку вилучити змінні x_{r+1}, ..., x_n з незалежних і ввести замість цих незалежних змінні δ_{r+1}, ..., δ_n не вдається.

Це означає, що визначити положення точки в просторі відхиленнями лише від площин $\delta_1 = 0$, $\delta_2 = 0$, ..., $\delta_n = 0$ неможливо (на мал. 5 для всіх точок прямої MN відхилення δ_1^0, δ_2^0, δ_3^0 від площин відповідно $\delta_1 = 0$, $\delta_2 = 0$, $\delta_3 = 0$ одні й ті самі, а тому трійка чисел $(\delta_1^0, \delta_2^0, \delta_3^0)$ не визначає якусь певну точку).

Надамо змінним $\delta_1, \delta_2, \ldots, \delta_r$ наперед заданих значень δ_1^0, $\delta_2^0, \ldots, \delta_r^0$, обчислимо значення, яких набувають при цьому змінні $\delta_{r+1}, \ldots, \delta_n$. Якщо вони збігаються з наперед заданими значеннями $\delta_{r+1}^0, \ldots, \delta_n^0$, то, очевидно, можна визначити такі значення змінних $x_1, x_2, \ldots, x_n$, при яких змінні δ_1^0, $\delta_2^0, \ldots, \delta_n^0$ набувають наперед заданих значень $\delta_1^0, \delta_2^0, \ldots, \delta_n^0$ (значення змінних $x_{r+1}, \ldots, x_n$ можна вибирати довільно). Задача виявляється, таким чином, невизначеною і має безліч розв'язків.



Мал. 5

Якщо після підстановки замість змінних $\delta_1, \delta_2, \ldots, \delta_r$ їх наперед заданих значень $\delta_1^0, \delta_2^0, \ldots, \delta_r^0$ змінні $\delta_{r+1}, \ldots, \delta_n$ не набувають своїх наперед заданих значень $\delta_{r+1}^0, \ldots, \delta_n^0$, то розглядувана задача виявляється суперечливою (мал. 5).

Нарешті, якщо $r = n$, тобто коли вдається вилучити всі змінні x_j, $(j = 1, 2, \ldots, n)$ з незалежних, замінивши їх змінними δ_i, $(i = 1, 2, \ldots, n)$, то задача має єдиний розв'язок.

У цьому випадку таблиця матиме вигляд:

Таблиця 4

	δ_1	δ_2	...	δ_n	1
x_1	$a_{11}^{(n)}$	$a_{12}^{(n)}$	...	$a_{1n}^{(n)}$	$b_1^{(n)}$
x_2	$a_{21}^{(n)}$	$a_{22}^{(n)}$	...	$a_{2n}^{(n)}$	$b_2^{(n)}$
...	...	...	...	...	...
x_n	$a_{n1}^{(n)}$	$a_{n2}^{(n)}$	...	$a_{nn}^{(n)}$	$b_n^{(n)}$

тобто

$$x_1 = a_{11}^{(n)}\delta_1 + a_{12}^{(n)}\delta_2 + \cdots + a_{1n}^{(n)}\delta_n + b_1^{(n)},$$
$$x_2 = a_{21}^{(n)}\delta_1 + a_{22}^{(n)}\delta_2 + \cdots + a_{2n}^{(n)}\delta_n + b_2^{(n)} \text{ і т. д.}$$

Підставивши замість змінних $\delta_1, \delta_2, \ldots, \delta_n$ їх задані значення, дістанемо шукані значення змінних $x_1, x_2, \ldots, x_n$.

Приклад.

За допомогою жорданових виключень знайти значення змінних x_1, x_2, x_3, при яких змінні

$$\delta_1 = 3x_1 + x_2 - x_3 + 1,$$
$$\delta_2 = 2x_1 - x_2 + 2x_3 - 2,$$
$$\delta_3 = -x_1 - x_2 + 4x_3 - 5$$

набувають відповідно значень 2, —1, 0.

Розмістивши вхідні дані у вигляді таблиці та виконавши кроки жорданових виключень (основні елементи обведено), матимемо:

1)

	x_1	x_2	x_3	1
δ_1	3	[1]	—1	1
δ_2	2	—1	2	—2
δ_3	—1	—1	4	—5

2)

	x_1	δ_1	x_3	1
x_2	—3	1	1	—1
δ_2	5	—1	[1]	—1
δ_3	2	—1	3	—4

3)

	x_1	δ_1	δ_2	1
x_2	—8	2	1	0
x_3	—5	1	1	1
δ_3	[—13]	2	3	—1

4)

	δ_3	δ_1	δ_2	1
x_2	$^8/_{13}$	$^{10}/_{13}$	$-^{11}/_{13}$	$^8/_{13}$
x_3	$^5/_{13}$	$^3/_{13}$	$-^2/_{13}$	$-^{18}/_{13}$
x_1	$-^1/_{13}$	$^2/_{13}$	$^3/_{13}$	$-^1/_{13}$

Підставивши замість $\delta_1, \delta_2, \delta_3$ відповідно 2, —1, 0, матимемо:

$$x_1 = -\frac{1}{13}\delta_3 + \frac{2}{13}\delta_1 + \frac{3}{13}\delta_2 - \frac{1}{13} =$$
$$= -\frac{1}{13}\cdot 0 + \frac{2}{13}\cdot 2 + \frac{3}{13}\cdot(-1) - \frac{1}{13} = 0,$$
$$x_2 = \frac{8}{13}\delta_3 + \frac{10}{13}\delta_1 - \frac{11}{13}\delta_2 + \frac{8}{13} =$$
$$= \frac{8}{13}\cdot 0 + \frac{10}{13}\cdot 2 - \frac{11}{13}\cdot(-1) + \frac{8}{13} = 3,$$
$$x_3 = \frac{5}{13}\delta_3 + \frac{3}{13}\delta_1 - \frac{2}{13}\delta_2 + \frac{18}{13} =$$
$$= \frac{5}{13}\cdot 0 + \frac{3}{13}\cdot 2 - \frac{2}{13}(-1) + \frac{18}{13} = 2.$$

Вправи.

1. Знайти значення x_1, x_2, при яких змінні

$$\delta_1 = x_1 + x_2 + 1, \quad \delta_2 = -x_1 - x_2 + 5$$

набувають відповідно значень: а) 3 і 2; б) 1 і 4. (Виконати малюнки.) Чи можна прямі $\delta_1 = 0$ і $\delta_2 = 0$ взяти за координатні?

2. Знайти значення x_1, x_2, при яких змінні

$$\delta_1 = x_1 + x_2 + 2, \quad \delta_2 = x_1 - 2x_2 - 1$$

набувають значень: а) 5 і 4; б) —3 і 1; в) 0 і 2; г) 0 і 0. Виконати малюнки. Чи можна прямі $\delta_1 = 0$, $\delta_2 = 0$ взяти за координатні? Якщо можна, то подати координати x_1, x_2 через координати δ_1 і δ_2 і навпаки. Чи будуть взаємно ортогональними нові координатні площини $\delta_1 = 0$, $\delta_2 = 0$?

3. Знайти значення x_1, x_2, x_3, при яких змінні

$$\begin{aligned}\delta_1 &= 1 \cdot x_1 + 1 \cdot x_2 + 1 \cdot x_3 + 1,\\ \delta_2 &= 0 \cdot x_1 + 1 \cdot x_2 + 1 \cdot x_3 + 3,\\ \delta_3 &= 2 \cdot x_1 + 1 \cdot x_2 + 0 \cdot x_3 - 1\end{aligned}$$

набувають відповідно значень 1, 1, 1. Чи можна площини $\delta_1 = 0$, $\delta_2 = 0$, $\delta_3 = 0$ взяти за координатні? Якщо можна, то визначити, як виражаються координати x_1, x_2, x_3 будь-якої точки M $(x_1; x_2; x_3)$ через відхилення δ_1, δ_2, δ_3 від площин $\delta_1 = 0$, $\delta_2 = 0$, $\delta_3 = 0$. Чи будуть взаємно ортогональними нові координатні площини?

§ 10. Деякі застосування жорданових виключень в лінійній алгебрі

Позначимо через δ, x, b, a_i відповідно вектори з координатами $(\delta_1; \delta_2; \ldots; \delta_n)$, $(x_1; x_2; \ldots; x_n)$, $(b_1; b_2; \ldots; b_n)$, $(a_{i1}; a_{i2}; \ldots; a_{in})$. Матрицю коефіцієнтів системи (9.1) позначимо через A.

Тоді рівності (9.1) можна подати у вигляді:

$$\delta_i = (a_i, x) + b_i \quad (i = 1, 2, \ldots, n), \tag{10.1}$$

де через (a_i, x) позначено скалярний добуток векторів a_i та x. У векторно-матричній формі систему (9.1) можна подати у вигляді:

$$\delta = Ax + b. \tag{10.2}$$

Надалі користуватимемось тією формою запису (10.1) та (10.2) системи (9.1), яка буде зручнішою.

1. Вважаємо, що в системі (10.1) $b_i = 0$ $(i = 1, 2, \ldots, n)$ (усі координати вектора b — нулі), тобто розглядаємо систему рівнянь виду:

$$\delta = Ax. \tag{10.3}$$

Нехай через n кроків жорданових виключень дістали таблицю 4. Очевидно, в усіх таблицях після перетворень вільні члени дорівнюватимуть нулю. Тоді можна записати:

$$x = A^{(n)}\delta.$$

Але згідно (10.3)

$$x = A^{-1}\delta,$$

де A^{-1} — матриця, обернена матриці A. Таким чином, $A^{(n)}$ є не що інше, як A^{-1} (відомо, коли обернена матриця існує, то вона єдина). Отже, за допомогою жорданових виключень за n кроків можна знайти матрицю, обернену матриці A, якщо обернена матриця існує (матриця A неособлива).

2. Вважатимемо, що $b = 0$ (і що виключити всі змінні x_1, x_2, ..., x_n з незалежних не вдається). Нехай через r кроків жорданових виключень дістали таблицю 3. Тоді, врахувавши, що всі $b_i = 0$, запишемо:

$$\begin{aligned}\delta_{r+1} &= a^{(r)}_{r+11}\delta_1 + a^{(r)}_{r+12}\delta_2 + \cdots + a^{(r)}_{r+1r}\delta_r \qquad (10.4)\\ &\cdots\cdots\cdots\cdots\cdots\cdots\\ \delta_n &= a^{(r)}_{n1}\delta_1 + a^{(r)}_{n2}\delta_2 + \cdots + a^{(r)}_{nr}\delta_r.\end{aligned}$$

Підставимо тепер замість δ_1, δ_2, ..., δ_n в рівностях (10.4) їх вирази (10.1), врахувавши, що $b_i = 0$.

Матимемо для будь-яких x:

$$\begin{aligned}(a_{r+1}, x) &= a^{(r)}_{r+11}(a_1, x) + a^{(r)}_{r+12}(a_2, x) + \cdots + a^{(r)}_{r+1r}(a_r, x),\\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ (a_n, x) &= a^{(r)}_{n1}(a_1, x) + a^{(r)}_{n2}(a_2, x) + \cdots + a^{(r)}_{nr}(a_r, x),\end{aligned}$$

тобто

$$\begin{aligned}a_{r+1} &= a^{(r)}_{r+11}a_1 + a^{(r)}_{r+12}a_2 + \cdots + a^{(r)}_{r+1r}a_r,\\ &\cdots\cdots\cdots\cdots\cdots\cdots\\ a_n &= a^{(r)}_{n1}a_1 + a^{(r)}_{n2}a_2 + \cdots + a^{(r)}_{nr}a_r.\end{aligned}$$

Останні рівності означають, що рядки $a_{r+1}, \ldots, a_n$ матриці A є лінійними комбінаціями перших r рядків $a_1, \ldots, a_r$ матриці A з коефіцієнтами лінійної залежності $a^{(r)}_{r+11}, \ldots, a^{(r)}_{r+1r}, \ldots$ $\ldots, a^{(r)}_{n1}, \ldots, a^{(r)}_{nr}$.

Отже, за допомогою жорданових виключень можна визначити ранг матриці A чи системи векторів a_i $(i = 1, 2, \ldots, n)$, а також систему r лінійно незалежних рядків матриці A чи векторів a_i, виразити всі інші лінійно залежні рядки матриці A (чи вектори a_i) через лінійно незалежні і визначити коефіцієнти лінійної залежності. Зауважимо, що при цьому кількість рядків і стовпчиків матриці A не обов'язково мають бути рівними між собою (кількість векторів a_i не обов'язково дорівнює розмірності цих векторів).

3. Можна довести, що добуток всіх основних елементів при виконанні кроків жорданових виключень дорівнює визначникові неособливої квадратної матриці A.

Отже, за допомогою жорданових виключень можна обчислити визначник матриці A. Якщо при цьому виявиться, що ранг r матриці A менший ніж n, то визначник такої матриці, як відомо, дорівнює нулю.

Приклади.

1. Знайти обернену матрицю і визначник матриці

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Обчислення розміщуємо в таблиці:

	x_1	x_2	x_3
δ_1	[1]	0	1
δ_2	1	1	1
δ_3	0	1	1

	δ_1	x_2	x_3
x_1	1	0	−1
δ_2	1	[1]	0
δ_3	0	1	1

	δ_1	δ_2	x_3
x_1	1	0	−1
x_2	−1	1	0
δ_3	−1	1	[1]

	δ_1	δ_2	δ_3
x_1	0	1	−1
x_2	−1	1	0
x_3	1	−1	1

Добуток основних елементів (вони обведені) дорівнює $1 \cdot 1 \times \times 1 = 1$. Отже, визначник цієї матриці дорівнює 1. В останній таблиці дістали матрицю A^{-1}, обернену A. Справді, добуток A на A^{-1} дорівнює

$$A \cdot A^{-1} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 & -1 \\ -1 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 1\cdot 0 + 0\cdot(-1) + 1\cdot 1 & 1\cdot 1 + 0\cdot 1 + 1\cdot(-1) \\ 1\cdot 0 + 1\cdot(-1) + 1\cdot 1 & 1\cdot 1 + 1\cdot 1 + 1\cdot(-1) \\ 0\cdot 0 + 1\cdot(-1) + 1\cdot 1 & 0\cdot 1 + 1\cdot 1 + 1\cdot(-1) \end{pmatrix}$$

$$\begin{pmatrix} 1\cdot(-1) + 0\cdot 0 + 1\cdot 1 \\ 1\cdot(-1) + 1\cdot 0 + 1\cdot 1 \\ 0\cdot(-1) + 1\cdot 0 + 1\cdot 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Отже, $A \cdot A^{-1} = E$ (E — одинична матриця).

2. Чи залежні вектори $a_1 = (1; 0; -1)$; $a_2 = (2; 1; 0)$; $a_3 = (0; -1; -2)$?

Якщо залежні, то визначити ранг цієї системи векторів і коефіцієнти лінійної залежності.

Маємо:

	x_1	x_2	x_3
δ_1	[1]	0	−1
δ_2	2	1	0
δ_3	0	−1	−2

	δ_1	x_2	x_3
x_1	1	0	1
δ_2	2	[1]	2
δ_3	0	−1	−2

	δ_1	δ_2	x_3
x_1	1	0	1
x_2	−2	1	−2
δ_3	2	−1	0

Як бачимо, x_3 вилучити з незалежних змінних не вдається. Отже, ранг матриці дорівнює 2. $\delta_3 = 2\delta_1 - \delta_2$, тобто вектор a_3 виражається через a_1 і a_2 з коефіцієнтами відповідно 2 і −1. Таким чином, $a_3 = 2a_1 - a_2$.

3. Визначити ранг системи векторів та коефіцієнти лінійної залежності:

$$a_1 = (-1; -1), \quad a_2 = (-1; 1), \quad a_3 = (4; 5).$$

Вектори a_1, a_2, a_3 задані своїми розкладами за базисними векторами $e_1 = (1; 0)$, $e_2 = (0; 1)$. Записавши ці розклади в таблицю і виконавши кроки жорданових виключень (основні елементи обведено), матимемо:

	e_1	e_2
a_1	[−1]	−1
a_2	−1	1
a_3	4	5

	a_1	e_2
e_1	−1	−1
a_2	1	[2]
a_3	−4	1

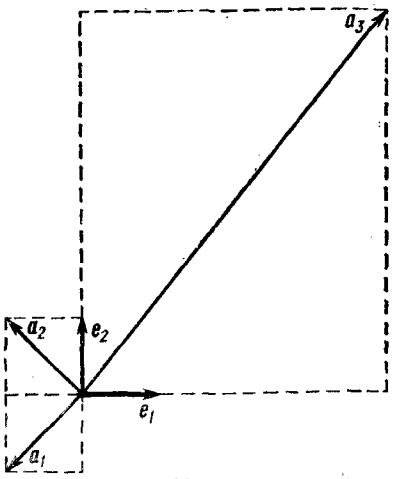
	a_1	a_2
e_1	$-1/2$	$-1/2$
e_2	$-1/2$	$1/2$
a_3	$-9/2$	$1/2$

Отже, дана система векторів має ранг 2, вектори a_1 та a_2 між собою лінійно незалежні, їх можна взяти за базисні, а

розклад вектора a_3 за векторами a_1 та a_2 матиме вигляд:

$$a_3 = \frac{-9}{2} a_1 + \frac{1}{2} a_2.$$

Зауважимо, що $-\frac{1}{2} a_1 - \frac{1}{2} a_2$ дає вектор (1, 0), а $\frac{-1}{2} a_1 + \frac{1}{2} a_2$ дає вектор (0; 1). Ці вектори і були спочатку взяті як базисні, а вектори a_1, a_2, a_3 були задані відповідно розкладами за базисними векторами e_1, e_2 (мал. 6).



Мал. 6

Таким чином, жорданові виключення дають змогу переходити від одного базису з n лінійно незалежних векторів до іншого, якщо задано розклад вектора, що вводитиметься до нового базису за базисними векторами вихідного базису. У щойно розглянутому прикладі вихідний базис у першій таблиці утворений векторами e_1, e_2, у другій векторами a_1, e_2, у третій — a_1, a_2. У кожному рядку таблиць записано коефіцієнти розкладів небазисних векторів (лівий стовпчик таблиці) через базисні (верхній рядок таблиці). Заміна кожного базисного вектора виконується за допомогою одного кроку жорданового виключення.

Вправи.

1. Знайти обернену матрицю і визначник матриці A:

а) $A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$; б) $A = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 2 & 2 \end{pmatrix}$; в) $A = \begin{pmatrix} 1 & 3 & -5 \\ 2 & 1 & 3 \\ 5 & 5 & 1 \end{pmatrix}$;

г) $A = \begin{pmatrix} 1 & 5 & 3 & 1 \\ 0 & 1 & 2 & -1 \\ 0 & -1 & 0 & 2 \\ 2 & 4 & 1 & 0 \end{pmatrix}$; д) $A = \begin{pmatrix} 2 & 4 & 3 & 2 \\ 0 & 0 & 0 & 1 \\ -2 & 2 & 1 & -1 \\ 1 & 1 & 0 & 2 \end{pmatrix}$.

2. Визначити ранг системи векторів та коефіцієнти лінійних залежностей. У двовимірному просторі виконати відповідні малюнки:

а) $a_1 = (1; 0)$, $a_2 = (1; 1)$;

б) $a_1 = (1; 0; 5)$; $a_2 = (1; 1; -2)$; $a_3 = (0; -1; -1)$;
в) $a_1 = (1; -1)$; $a_2 = (-1; 1)$; $a_3 = (2; 3)$;
г) $a_1 = (2; 5; 1; 3)$; $a_2 = (-1; 1; 4; 0)$; $a_3 = (0; 5; -1; -1)$; $a_4 = (0; 0; 3; 5)$;
д) $a_1 = (1; 0; 1)$; $a_2 = (0; 1; 1)$; $a_3 = (2; 1; 0)$; $a_4 = (2; 4; -5)$.

§ 11. Розв'язування систем лінійних алгебраїчних рівнянь методом Жордана—Гаусса

Нехай у (9.1) задані значення δ_i^0 ($i = 1, 2, \ldots, n$) дорівнюють нулю, тобто потрібно знайти розв'язок системи рівнянь

$$Ax + b = 0. \qquad (11.1)$$

Цю систему рівнянь можна розв'язати за допомогою жорданових виключень. Нехай після r кроків жорданових виключень дістали таблицю 3. Якщо в рядках, які відповідають змінним $\delta_{r+1}, \ldots, \delta_n$, усі вільні члени $b_{r+1}^{(r)}, \ldots, b_n^{(r)}$ дорівнюють нулеві, то, очевидно, поклавши $\delta_1 = 0$, $\delta_2 = 0$, …, $\delta_r = 0$ і надавши змінним $x_{r+1}, \ldots, x_n$ будь-яких значень, матимемо: $\delta_{r+1} = 0, \ldots, \delta_n = 0$, тобто $(r+1)$-е, $(r+2)$-е, …, n-е рівняння випливають з перших r рівнянь і якщо задовольняються перші r рівнянь, то задовольняється й решта $(n - r)$ рівнянь (див. рівності (10.4)). Отже, система (11.1) у цьому випадку має безліч розв'язків. Розмірність підпростору розв'язків, очевидно, дорівнює $(n - r)$, оскільки можна вказати $(n - r)$ лінійно незалежних наборів для довільних змінних $x_{r+1}, x_{r+2}, \ldots, x_n$ (див. таблицю 5).

Таблиця 5

x_{r+1}	x_{r+2}	⋯	x_{n-1}	x_n
1	0	⋯	0	0
0	1	⋯	0	0
⋯	⋯	⋯	⋯	⋯
0	0	⋯	1	0
0	0	⋯	0	1

Будь-який інший набір змінних $x_{r+1}, x_{r+2}, \ldots, x_n$ можна подати як лінійну комбінацію наборів з таблиці 5.

Зауважимо, що в розглянутому випадку ранги основної і розширеної матриць збігаються.

Якщо серед вільних членів $b_{r+1}^{(r)}, \ldots, b_n^{(r)}$ є такі, що відрізняються від нуля, то система рівнянь (11.1) виявляється несумісною — не можна задати такий набір значень $x_1, x_2, \ldots, x_n$,

щоб усі змінні $\delta_1, \delta_2, \ldots, \delta_n$ набули нульових значень. У цьому випадку ранг основної матриці менший від рангу розширеної матриці.

Примітка. Нагадуємо, що *основною* називається матриця A. *Розширеною* називається матриця, яку дістаємо приєднанням до матриці A ще одного стовпчика b вільних членів. Як бачимо, жорданові виключення дають можливість визначити одночасно ранги основної і розширеної матриць.

Якщо матриця A має обернену (див. табл. 4), то, очевидно, система (11.1) має єдиний розв'язок. А саме: поклавши в таблиці 4

$$\delta_1 = 0, \quad \delta_2 = 0, \ \ldots, \ \delta_n = 0,$$

дістанемо:

$$x_1 = b_1^{(n)}, \ x_2 = b_2^{(n)}, \ \ldots, \ x_n = b_n^{(n)}.$$

При практичному розв'язуванні системи рівнянь виду (11.1) після введення до числа незалежних деякої змінної δ_i стовпчик коефіцієнтів, що стоять проти цієї змінної в таблиці, обчислювати і зберігати потреби немає, оскільки змінна δ_i має набути нульового значення, і добуток її з будь-яким коефіцієнтом дорівнює нулю. Тому після виключення кожної змінної x_j стовпчики під змінними δ_i до розрахункової таблиці не вносяться. Залишаються лише стовпчики під ще не виключеними змінними x_j.

Примітка. Задачу розв'язування системи рівнянь

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + x_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases} \qquad (11.2)$$

можно тлумачити як задачу відшукання коефіцієнтів $x_1, x_2, \ldots, x_n$ розкладу вектора-стовпчика, складеного з вільних членів $b_1, b_2, \ldots, b_n$ системи (11.2) за векторами-стовпчиками матриці коефіцієнтів системи (11.2), тобто за векторами $a_1 = (a_{11}; a_{21}; \ldots; a_{n1})$, $a_2 = (a_{12}; a_{22}; \ldots; a_{n2})$, $\ldots$, $a_n = (a_{1n}; a_{2n}; \ldots; a_{nn})$ (порівняйте з прикладом 3 попереднього параграфа). Як ми вже переконались, ця задача повністю розв'язується за допомогою жорданових виключень.

Контроль обчислень при розв'язуванні системи рівнянь здійснюється так. Надамо змінним $x_1, x_2, \ldots, x_n$ довільних, що відрізняються від нуля, значень $x_1^*, x_2^*, \ldots, x_n^*$ і обчислимо відхилення $\delta_1^*, \delta_2^*, \ldots, \delta_n^*$ за початковою таблицею 1. Як ми бачили (див. § 9 та мал. 3), положення точки в просторі не змінюється внаслідок переходу від одного базису до іншого. Тому відхилення $x_1^*, x_2^*, \ldots, x_n^*, \delta_1^*, \delta_2^*, \ldots, \delta_n^*$ від площин $x_1 = 0$, $x_2 = 0, \ldots, x_n = 0, \ \delta_1 = 0, \ \delta_2 = 0, \ \ldots, \ \delta_n = 0$ повинні залишатись незмінними, якими б не були базисні площини. Надавши, наприклад, в таблиці 3 змінним $\delta_1, \ldots, \delta_r, x_{r+1}, \ldots, x_n$, що стоять у верхньому рядку таблиці, значень $\delta_1^*, \ldots, \delta_r^*, x_{r+1}^*, \ldots, x_n^*$, ми повинні дістати $x_1 = x_1^*, \ \ldots, \ x_r = x_r^*, \ \delta_{r+1} = \delta_{r+1}^*, \ldots, \delta_n = \delta_n^*$. Якщо останні рівності не справджуються, то це означає, що в обчисленнях припустилися помилок.

Приклад.

Відшукати розв'язок або встановити несумісність системи:

$$\begin{cases} x_1 - x_2 + x_3 - x_4 + 2 = 0, \\ x_2 - x_3 - 3 = 0, \\ 2x_1 + x_2 + 2x_3 + x_4 - 5 = 0, \\ 4x_1 - 2x_2 + 2x_3 - 4x_4 + 12 = 0. \end{cases}$$

Вихідною буде таблиця 1). За основний елемент візьмемо $a_{11} = 1$. Після кроку жорданового виключення дістанемо таблицю 2). Продовжуючи виключення змінних x_i з незалежних (основні елементи взято в рамку), дістанемо таблиці 3), 4).

1)

	x_1	x_2	x_3	x_4	1
δ_1	[1]	−1	1	−1	2
δ_2	0	1	−1	0	−3
δ_3	2	1	2	1	−5
δ_4	4	−2	2	−4	12

2)

	δ_1	x_2	x_3	x_4	1
x_1	1	1	−1	1	−2
δ_2	0	[1]	−1	0	−3
δ_3	2	3	0	3	−9
δ_4	4	2	−2	0	4

3)

	δ_1	δ_2	x_3	x_4	1
x_1	1	1	0	1	1
x_2	0	1	1	0	3
δ_3	2	3	[3]	3	0
δ_4	4	2	0	0	10

4)

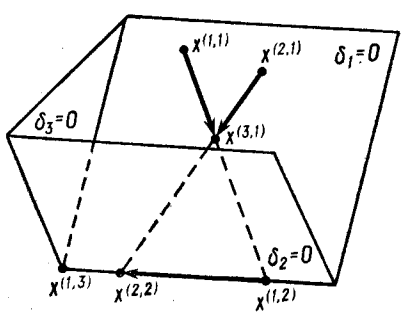
	δ_1	δ_2	δ_3	x_4	1
x_1	1	1	0	1	1
x_2	$-\frac{2}{3}$	0	$\frac{1}{3}$	-1	3
x_3	$-\frac{2}{3}$	-1	$\frac{1}{3}$	-1	0
δ_4	4	2	0	0	10

Ця система рівнянь несумісна, оскільки останній рядок вихідної матриці є лінійною комбінацією перших двох рядків $a_4 = 4a_1 + 2a_2$. Проте між вільними членами такої самої лінійної залежності не існує, бо, як видно з останньої таблиці, $b_4 = 4b_1 + 2b_2 + 10$.

§ 12. Векторний метод

Ідею цього методу розглянемо на прикладі системи трьох рівнянь з трьома змінними:

$$\begin{cases} \delta_1(x) = (a_1,\ x) + b_1 = 0, \\ \delta_2(x) = (a_2,\ x) + b_2 = 0, \\ \delta_3(x) = (a_3,\ x) + b_3 = 0, \end{cases} \tag{12.1}$$



Мал. 7

де a_1, a_2, a_3, x — вектори тривимірного простору $\mathbf{R}^3$, (a_i, x) — скалярний добуток векторів a_i та x, тобто в координатній формі

$$(a_i,\ x) = a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 \quad (i = 1,\ 2,\ 3).$$

Зауважимо, що кожній точці x відповідає радіус-вектор цієї точки і навпаки, а отже, множини всіх точок тривимірного простору і радіусів-векторів цих точок еквівалентні між собою. Тому один і той самий елемент x ми в одних випадках називатимемо *точкою*, а в інших — *вектором*.

Кожне з рівнянь (12.1) у просторі $\mathbf{R}^3$, як відомо, визначає деяку площину. Нехай (мал. 7) на першому кроці на площині $\delta_1 = 0$ взято три точки $x^{(1,1)}$, $x^{(2,1)}$, $x^{(3,1)}$, що не лежать на одній прямій (перший індекс у дужках означає номер точки, другий — номер кроку).

Візьмемо будь-які дві точки, наприклад $x^{(1,1)}$ і $x^{(3,1)}$ і рухатимемось з точки $x^{(1,1)}$ у напрямі вектора $(x^{(3,1)} - x^{(1,1)})$ доти, поки не потрапимо в точку $x^{(1,2)}$, що лежить на площині $\delta_2(x) = 0$. Для визначення точки $x^{(1,2)}$ запишемо параметричне рівняння прямої, що проходить через точки $x^{(1,1)}$ і $x^{(3,1)}$.

Маємо:

$$x = x^{(1,1)} + t(x^{(3,1)} - x^{(1,1)}), \tag{12.2}$$

де x — довільна точка прямої, що відповідає параметру t. Щоб визначити значення параметра t, при якому x належить площині $\delta_2(x) = 0$, підставимо вираз для x у рівняння площини $\delta_2(x) = 0$. Маємо:

$$(a_2,\ x^{(1,1)}) + t(x^{(3,1)} - x^{(1,1)}) + b_2 = 0.$$

Звідси

$$t = \frac{(a_2,\ x^{(1,1)}) + b_2}{-(a_2,\ x^{(1,1)}) + (a_2,\ x^{(3,1)})} =$$

$$= \frac{(a_2,\ x^{(1,1)}) + b_2}{-(a_2,\ x^{(1,1)}) - b_2 + (a_2,\ x^{(3,1)}) + b_2} =$$

$$= -\frac{\delta_2(x^{(1,1)})}{\delta_2(x^{(3,1)}) - \delta_2(x^{(1,1)})}. \tag{12.3}$$

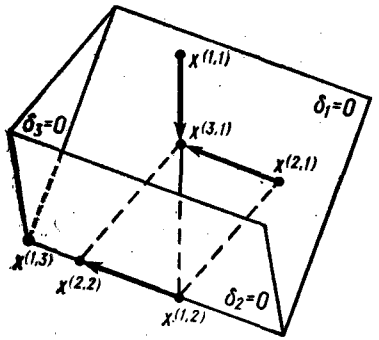
Як бачимо з (12.3), точки $x^{(1,1)}$ і $x^{(3,1)}$ не повинні лежати на прямій, паралельній площині $\delta_2(x) = 0$, тобто не повинні мати однакові відхилення $\delta_2(x^{(3,1)})$ і $\delta_2(x^{(1,1)})$ від площини $\delta_2(x) = 0$, бо інакше пряма (12.2) не перетинатиме площини $\delta_2(x) = 0$. Оскільки всі три точки $x^{(1,1)}$, $x^{(2,1)}$, $x^{(3,1)}$ не лежать на одній прямій, то принаймні дві з них мають різні відхилення від площини $\delta_2(x) = 0$. Нехай $\delta_2(x^{(3,1)}) \neq \delta_2(x^{(1,1)})$. Тоді, визначивши t з (12.3) і підставивши його значення в (12.2), дістанемо точку $x^{(1,2)}$, яка, очевидно, належить площинам $\delta_1(x) = 0$ і $\delta_2(x) = 0$.

Аналогічно за точками $x^{(2,1)}$ і $x^{(3,1)}$ визначимо точку $x^{(2,2)}$, що також належатиме як площині $\delta_1(x) = 0$, так і площині $\delta_2(x) = 0$.

Потім запишемо рівняння прямої, що проходить через точки $x^{(1,2)}$ і $x^{(2,2)}$. Очевидно, всі точки такої прямої належать площинам $\delta_1(x) = 0$ та $\delta_2(x) = 0$. Рухаючись по такій прямій від точки $x^{(1,2)}$ в напрямі вектора $(x^{(2,2)} - x^{(1,2)})$, знайдемо точку, що належить площині $\delta_3(x) = 0$, тобто знайдемо точку, що належить одночасно трьом площинам: $\delta_1(x) = 0$, $\delta_2(x) = 0$, $\delta_3(x) = 0$, а отже, і розв'язок системи (12.1).

Нехай тепер, наприклад, $\delta_2(x^{(3,1)}) \neq \delta_2(x^{(1,1)})$, але $\delta_2(x^{(2,1)}) = \delta_2(x^{(3,1)}) \neq 0$ (мал. 8). Тоді знайдемо точку $x^{(1,2)}$, за точка-

ми $x^{(1,1)}$ і $x^{(3,1)}$. Щоб знайти ще одну точку, що належить площинам $\delta_1(x) = 0$ та $\delta_2(x) = 0$, досить до вектора (точки) $x^{(1,2)}$ додати вектор $(x^{(3,1)} - x^{(2,1)})$. Отже, у випадку $\delta_2(x^{(2,1)}) = \delta_2(x^{(3,1)}) \neq 0$ задача лише спрощується (хоч вираз (12.3) і втрачає зміст).



Мал. 8

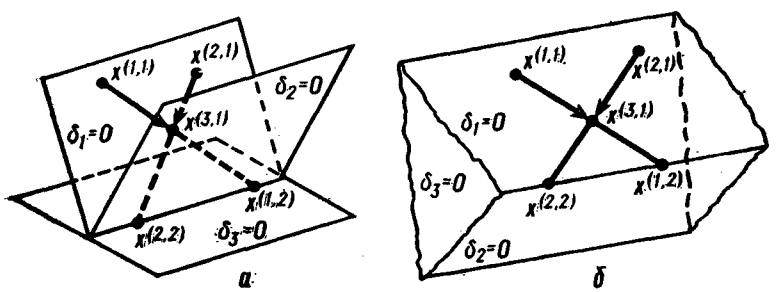
Якби відразу виявилось, наприклад, $\delta_2(x^{(2,1)}) = \delta_2(x^{(3,1)}) = 0$, то для другого кроку набір з двох точок, що належать площинам $\delta_1(x) = 0$ і $\delta_2(x) = 0$, вже готовий і додаткові обчислення не потрібні. Тому щоразу включаємо в новий набір насамперед ті точки, які вже належать розглядуваній площині.

Очевидно, якби виявилось, що $\delta_3(x^{(1,2)}) = \delta_3(x^{(2,2)}) = 0$, то система (12.1) мала б нескінченну одновимірну множину розв'язків (мал. 9, *а*). У випадку $\delta_3(x^{(1,2)}) = \delta_3(x^{(2,2)}) \neq 0$ система була б несумісною (мал. 9, *б*).

Перейдемо тепер до загального n-вимірного випадку.

Нехай задано неоднорідну систему лінійних алгебраїчних рівнянь

$$\delta_i(x) \equiv (a_i,\ x) + b_i = a_{i1}x_1 + a_{i2}x_2 + \ \cdots\ + a_{in}x_n + b_i = 0$$
$$(i = 1,\ 2,\ \ldots,\ n). \qquad (12.4)$$



Мал. 9

Припустимо спочатку, що система сумісна і має єдиний розв'язок. Візьмемо набір з n лінійно незалежних векторів так, щоб ці вектори задовольняли одне з рівнянь системи (12.4). Нехай, наприклад, $b_1 \neq 0$ і всі коефіцієнти першого рівняння $a_{1j} \neq 0$ $(j = 1,\ 2,\ \ldots,\ n)$.

Тоді перший набір векторів $x^{(k,1)}$ може бути таким:

$$x^{(1,1)} = \left(-\frac{b_1}{a_{11}};\ 0;\ \ldots;\ 0\right),$$
$$x^{(2,1)} = \left(0;\ -\frac{b_1}{a_{12}};\ 0;\ \ldots;\ 0\right),\ \ldots,\ x^{(n,1)} =$$
$$= \left(0;\ 0;\ \ldots;\ 0;\ -\frac{b_1}{a_{1n}}\right).$$

Зауважимо, що коли якийсь з коефіцієнтів $a_{1j} = 0$, то відповідні координати в усіх векторів $x^{(k,1)}$ можна вибрати довільно, зокрема так, щоб вектори $x^{(k,1)}$ задовольняли принаймні ще одне рівняння, крім першого, і таким чином спростити знаходження розв'язку.

Знайдемо тепер новий набір $(n-1)$ лінійно незалежних векторів так, щоб вони задовольняли перші два рівняння системи. Для цього обчислимо спочатку всі відхилення $\delta_2(x^{(k,1)})$ $(k = 1,\ 2,\ \ldots,\ n)$.

Оскільки за припущенням система має єдиний розв'язок, то серед чисел $\delta_2(x^{(k,1)})$ знайдуться принаймні два різних числа, бо інакше площини $\delta_1(x) = 0$ і $\delta_2(x) = 0$ були б паралельними, а система (12.4) або несумісною (якщо $\delta_2(x^{(k,1)}) \neq 0$), або мала б безліч розв'язків.

До набору $x^{(k,2)}$ включаємо насамперед ті вектори набору $x^{(k,1)}$, для яких $\delta_2(x^{(k,1)}) = 0$. Нехай $\delta_2(x^{(k,1)}) = 0$ для $k = 1,\ 2,\ \ldots\ \ldots,\ r_0$ $(r_0 \leqslant n-1)$. Тоді покладемо $x^{(k,2)} = x^{(k,1)}$ для $k = 1,\ 2,\ \ldots\ \ldots,\ r_0$. Якщо $r_0 = n-1$, то новий набір $x^{(k,2)}$ вже побудовано.

Нехай $r_0 < n-1$. Тоді, крім уже знайдених векторів $x^{(k,2)}$ $(k = 1,\ 2,\ \ldots,\ r_0)$, решту $(n-1-r_0)$ векторів знаходимо так. Якщо числа $\delta_2(x^{(k,1)})$ $(k = r_0+1,\ \ldots,\ n)$ не всі рівні між собою, наприклад $\delta_2(x) = (x^{(r_0+1,1)}) \neq \delta_2(x^{(r_0+2,1)})$, $\ldots\ \ldots,\ \delta_2(x^{(r_0+1,1)}) \neq \delta_2(x^{(n,1)})$, то вектори

$$x^{(r_0+j,2)} = x^{(r_0+1,2)} - \frac{\delta_2(x^{(r_0+1,1)})}{\delta_2(x^{(r_0+1+j,1)}) - \delta_2(x^{(r_0+1,1)})} \times$$
$$\times\ (x^{(r_0+1+j,1)} - x^{(r_0+1,1)}) \quad (j = 1,\ 2,\ \ldots,\ n-1-r_0) \qquad (12.5)$$

задовольняють перші два рівняння системи (12.4).

Якщо $r_0 = 0$, то таких векторів є $n-1$ (порівняйте (12.5) з (12.3)).

Нехай тепер числа $\delta_2(x^{(k,1)})$ $(k = r_0+1,\ \ldots,\ n)$ всі рівні між собою (це можливо, якщо $r_0 \neq 0$). Тоді вектори $x^{(r_0+1,2)},\ \ldots,\ x^{(n-1,2)}$ можна дістати так:

$$x^{(r_0+j,2)} = x^{(r_0,2)} + (x^{(r_0+1+j,1)} - x^{(r_0+1,1)})\ (j = 1,\ 2,\ldots,\ n-1-r_0). \qquad (12.6)$$

Вони очевидно також задовольняють обидва перші рівняння системи (12.1). Легко переконатись, що побудовані

таким чином вектори $x^{(k,2)}$ ($k = 1, 2, \ldots, n-1$) між собою лінійно незалежні.

Примітка. Якщо не всі числа $\delta_2(x^{(k,1)})$ ($k = r_0+1, \ldots, n$) рівні між собою, то можна використати як формули (12.5), так і формули (12.6). Однак при цьому має виконуватися вимога: вектори $x^{(k,2)}$ повинні бути між собою лінійно незалежними. Ця вимога завжди виконується, якщо додержувати такого правила: коли за векторами $x^{(k_1,1)}$ і $x^{(k_2,1)}$ дістали вектор $x^{(r_1,2)}$, а за векторами $x^{(k_1,1)}$ і $x^{(k_3,1)}$ вектор $x^{(r_2,2)}$, то для визначення $x^{(r_3,2)}$ не слід вибирати пару векторів $x^{(k_2,1)}$ і $x^{(k_3,1)}$, бо інакше вектори $x^{(r_1,2)}$, $x^{(r_2,2)}$, $x^{(r_3,2)}$ будуть лінійно залежними між собою.

Діставши другий набір із $(n-1)$ векторів $x^{(k,2)}$, які задовольняють два перших рівняння, за кожними двома векторами з набору $x^{(k,2)}$ ($k = 1, 2, \ldots, n-1$) визначаємо вектори $x^{(k,3)}$, які задовольнятимуть перші три рівняння. Очевидно, вибираючи будь-який напрям $(x^{(k_1,2)} - x^{(k_2,2)})$, ми завжди з точок $x^{(k_2,2)}$ ($k = 1, 2, \ldots, n-1$) рухатимемось у площинах $\delta_1(x) = 0$, $\delta_2(x) = 0$.

Аналогічно будуємо третій набір $x^{(k,3)}$ ($k = 1, 2, \ldots, n-2$) з $(n-2)$ лінійно незалежних векторів. Продовжуючи цей процес, за скінченну кількість кроків дістанемо вектор, що задовольняє всі рівняння системи.

Коли система несумісна або має безліч розв'язків, то за скінченну кількість кроків описаного процесу ці її властивості будуть з'ясовані.

Обчислення за описаною схемою зручно розмістити у вигляді таблиці.

$\delta_1(x) = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + b_1$					
. . . . .					
$\delta_n(x) = a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n + b_n$					
	$x^{(1,1)}$	$x^{(2,1)}$	...	$x^{(n-1,1)}$	$x^{(n,1)}$
x_1	$x_1^{(1,1)}$	$x_1^{(2,1)}$	...	$x_1^{(n-1,n)}$	$x_1^{(n,1)}$
. . .	. . . . .				
x_n	$x_n^{(1,1)}$	$x_n^{(2,1)}$	...	$x_n^{(n-1,1)}$	$x_n^{(n,1)}$
$\delta_2(x^{(k,1)})$	δ_2^1	δ_2^2	...	δ_2^{n-1}	δ_2^n
x_1	$x_1^{(1,2)}$		...	$x_1^{(n-1,2)}$	
...	. . . . .				
x_n	$x_n^{(1,2)}$		...	$x_n^{(n-1,2)}$	
$\delta_3(x^{(k,2)})$	δ_3^1		...	δ_3^{n-1}	
. . . . .					
$\delta_n(x^{(k,n-1)})$	δ_n^1			δ_n^2	
x_1	$x_1^{(1,n)}$				
. . .	. . . . .				
x_n	$x_n^{(1,n)}$				

Приклади.

1. Розв'яжемо систему рівнянь:

$$\begin{cases} \delta_1(x) = x_1 - x_2 + x_3 - 2 = 0, \\ \delta_2(x) = x_1 + x_2 - x_3 - 2 = 0, \\ \delta_3(x) = x_1 - x_2 - x_3 - 0 = 0 \end{cases}$$

за наведеною схемою. Перший набір трьох векторів має вигляд (таблиця 1)):

1)

	$x^{(1,1)}$	$x^{(2,1)}$	$x^{(3,1)}$
x_1	2	0	0
x_2	0	−2	0
x_3	0	0	2
$\delta_2(x)$	0	−4	−4

2)

	$x^{(1,2)}$	$x^{(2,2)}$
x_1	2	2
x_2	0	2
x_3	0	2
$\delta_3(x)$	2	−2

3)

	$x^{(1,3)}$
x_1	2
x_2	1
x_3	1

Обчисливши $\delta_2(x^{(1,1)})$, $\delta_2(x^{(2,1)})$, $\delta_2(x^{(3,1)})$, дістанемо відповідно $\delta_2(x^{(1,1)}) = 0$, $\delta_2(x^{(2,1)}) = -4$, $\delta_2(x^{(3,1)}) = -4$. Оскільки $\delta_2(x^{(1,1)}) = 0$, то $x^{(1,1)}$ відразу включаємо в новий набір, тобто покладаємо $x^{(1,2)} = x^{(1,1)} = (2;\ 0;\ 0)$. Оскільки $\delta_2(x^{(2,1)}) = \delta_2(x^{(3,1)})$, то другий вектор $x^{(2,2)}$ за формулами (12.6) дорівнюватиме (див. табл. 2):

$$x^{(2,2)} = x^{(1,2)} + (x^{(3,1)} - x^{(2,1)}) = (2;\ 0;\ 0) + \\ + [(0;\ 0;\ 2) - (0;\ -2;\ 0)] = (2;\ 2;\ 2).$$

Обчисливши відхилення точок $x^{(1,2)}$, $x^{(2,2)}$ від площини $\delta_3(x) = 0$, матимемо $\delta_3(x^{(1,2)}) = 2$, $\delta_3(x^{(2,2)}) = -2$. Оскільки $\delta_3(x^{(1,2)}) \neq \delta_3(x^{(2,2)})$, то за формулами (12.5) знайдемо (див.

табл. 3):

$$x^{(1,3)} = x^{(1,2)} - \frac{\delta_3(x^{(1,2)})}{\delta_3(x^{(2,2)}) - \delta_3(x^{(1,2)})}(x^{(2,2)} - x^{(1,2)}) =$$

$$= \begin{pmatrix}2\\0\\0\end{pmatrix} - \frac{2}{-2-2}\begin{pmatrix}2-2\\2-0\\2-0\end{pmatrix} = \begin{pmatrix}2\\0\\0\end{pmatrix} + \frac{1}{2}\begin{pmatrix}0\\2\\2\end{pmatrix} = \begin{pmatrix}2\\1\\1\end{pmatrix}.$$

Отже, $x_1 = 2$, $x_2 = 1$, $x_3 = 1$.

2. Нехай дано систему рівнянь:

$$\begin{cases} 1x_1 + 1x_2 + 0x_3 + 0x_4 - 2 = 0, \\ -1x_1 + 1x_2 + 1x_3 + 0x_4 - 1 = 0, \\ 0x_1 + 1x_2 - 1x_3 + 1x_4 - 1 = 0, \\ 0x_1 + 0x_2 + 1x_3 + 1x_4 - 2 = 0. \end{cases}$$

Оскільки коефіцієнти при x_3 і x_4 в першому рівнянні дорівнюють нулю, то координати x_3 та x_4 можна вибрати довільно. Крім того, коефіцієнт при x_4 в другому рівнянні також дорівнює нулю. Отже, можна задовольнити друге рівняння за рахунок відповідного добору координати x_3 при вибраних координатах x_1, x_2 і залишити x_4 довільним. Оскільки x_4 залишається довільним, можна при вибраних координатах x_1, x_2, x_3 за рахунок відповідного добору координати x_4 задовольнити третє рівняння. Отже, одразу можна вказати два вектори, які задовольняють три перші рівняння, а саме:

$$x^{(1,1)} = (2;\ 0;\ 3;\ 4),\ x^{(2,1)} = (0;\ 2;\ -1;\ -2).$$

Маємо: $\delta_4(x^{(1,1)}) = 5$, $\delta_4(x^{(2,1)}) = -5$. За формулами (12.5) дістанемо розв'язок розглядуваної системи:

$$x = x^{(1,1)} - \frac{\delta_4(x^{(1,1)})}{\delta_4(x^{(2,1)}) - \delta_4(x^{(1,1)})}(x^{(2,1)} - x^{(1,1)}) =$$

$$= \begin{pmatrix}2\\0\\3\\4\end{pmatrix} - \frac{5}{-5-5}\begin{pmatrix}0-2\\2-0\\-1-3\\-2-4\end{pmatrix} = \begin{pmatrix}2\\0\\3\\4\end{pmatrix} +$$

$$+ \frac{1}{2}\begin{pmatrix}-2\\2\\-4\\-6\end{pmatrix} = \begin{pmatrix}1\\1\\1\\1\end{pmatrix}.$$

Геометрія векторного методу така сама, як і геометрія методу Жордана — Гаусса. А саме спочатку потрапляємо на одну

з площин, наприклад $\delta_1 = 0$. Потім, рухаючись у цій площині, потрапляємо ще на одну площину, наприклад $\delta_2 = 0$. Після цього, рухаючись так, щоб відхилення δ_1 і δ_2 залишались рівними нулю, досягаємо третьої площини (якщо це можливо), наприклад $\delta_3 = 0$, і т. д. Однак у векторному методі коефіцієнти вихідної системи залишаються без змін, вони не підлягають ніяким перетворенням, що іноді буває зручно при обчисленнях.

Вправи.

Методом Жордана — Гаусса і векторним методом розв'язати системи рівнянь або встановити їх суперечливість.

1. $\begin{cases} 3x_1 - 2x_2 + 5x_3 - 2x_4 + 1x_5 = 23, \\ 0x_1 + 4x_2 - 6x_3 + 5x_4 - 6x_5 = -36, \\ 3x_1 + 6x_2 - 7x_3 - 9x_4 + 4x_5 = 24, \\ -2x_1 - 3x_2 + 1x_3 - 4x_4 + 6x_5 = 16, \\ 1x_1 - 4x_2 + 6x_3 + 0x_4 - 3x_5 = 4. \end{cases}$

2. $\begin{cases} 1{,}5x_1 + 2{,}5x_2 - 4x_3 + 2x_4 - 6x_5 = 10, \\ -1x_1 - 1{,}5x_2 + 2{,}5x_3 - x_4 + 2x_5 = 8, \\ 0x_1 - 2x_2 + 4x_3 - 6x_4 + 4{,}5x_5 = 6, \\ 3x_1 + 0x_2 - 3x_3 - 2x_4 + 1x_5 = 1, \\ 3x_1 + 4x_2 - 7x_3 - 4x_4 - 3x_5 = -8. \end{cases}$

3. $\begin{cases} 0{,}5x_1 + 3x_2 - 4x_3 + 5x_4 + 0x_5 = 2{,}5, \\ 1{,}5x_1 + 2x_2 + 0x_3 - 1x_4 + 6x_5 = 4, \\ 2x_1 + 1x_2 - 1x_3 + 0x_4 - 0{,}5x_5 = 6, \\ 1x_1 - 2{,}5x_2 + 2x_3 - 3x_4 + x_5 = 8{,}5, \\ 0x_1 - 1x_2 - 1x_3 + 1x_4 + 2{,}5x_5 = 4{,}5. \end{cases}$

4. $\begin{cases} 0{,}5x_1 - 2x_2 - 1x_3 - 3x_4 + 2{,}5x_5 = 4{,}5, \\ 2x_1 - 2{,}5x_2 + 0{,}5x_3 + 3{,}5x_4 + 0{,}5x_5 = 0{,}5, \\ 0x_1 + 0x_2 + 0x_3 + 3x_4 + 6x_5 = 8, \\ x_1 - x_2 + 2{,}5x_3 - 1x_4 + 4{,}5x_5 = 12, \\ 2{,}5x_1 + 1{,}5x_2 - 1{,}5x_3 - 3x_4 + 5x_5 = 36. \end{cases}$

§ 13. Розв'язування систем лінійних алгебраїчних рівнянь методом простої ітерації і методом Зейделя

У попередніх параграфах ми розглянули точні методи знаходження розв'язків систем лінійних рівнянь. Другу групу методів розв'язування систем лінійних рівнянь становлять наближені методи. Суть їх полягає в тому, що, починаючи з деякого початкового наближення, на кожному кроці обчислювального процесу це наближення до розв'язку поліпшують, дістаючи при цьому нове наближення. Цей процес послідовних наближень продовжують доти, поки розв'язок буде знай-

дено з потрібною точністю. До найпростіших наближених методів належить *метод простої ітерації*.

Нехай дано систему n лінійних рівнянь з n змінними

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \quad (i = 1, 2, \ldots, n), \tag{13.1}$$

або у векторно-матричній формі

$$Ax = b. \tag{13.2}$$

Подамо цю систему у вигляді еквівалентної їй системи:

$$x_i = \sum_{j=1}^{n} \alpha_{ij}x_j + \beta_i \quad (i = 1, 2, \ldots, n). \tag{13.3}$$

Позначивши

$$\alpha = \begin{pmatrix} \alpha_{11}\alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21}\alpha_{22} & \ldots & \alpha_{2n} \\ \alpha_{n1}\alpha_{n2} & \ldots & \alpha_{n-n} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix},$$

запишемо останню систему у векторно-матричній формі

$$x = \alpha x + \beta. \tag{13.4}$$

Систему (13.4) розв'язуватимемо методом послідовних наближень. За початкове наближення розв'язку системи (13.4) візьмемо вектор $x^{(0)} = \begin{pmatrix} x_1^{(0)} \\ \vdots \\ x_n^{(0)} \end{pmatrix}$, де $x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}$ — довільні числа.

Тоді послідовно знаходимо перше наближення

$$x^{(1)} = \alpha x^{(0)} + \beta,$$

друге наближення

$$x^{(2)} = \alpha x^{(1)} + \beta \text{ і т. д.}$$

Якщо вже побудовано k-те наближення, то наступне $(k + 1)$-е наближення будуємо за формулою

$$x^{(k+1)} = \alpha x^{(k)} + \beta, \tag{13.5}$$

або в розгорнутому вигляді

$$x_i^{(k+1)} = \sum_{j=1}^{n} \alpha_{ij}x_j^{(k)} + \beta_i \quad (i = 1, 2, \ldots, n). \tag{13.6}$$

Якщо кожна з послідовностей $\{x_j^{(k)}\}$ $(j = 1, 2, \ldots, n)$ має границю x_j^*, тобто

$$x_j^* = \lim_{k\to\infty} x_j^{(k)} \quad (j = 1, 2, \ldots, n), \tag{13.7}$$

то сукупність $x_1^*, x_2^*, \ldots, x_n^*$ є розв'язком системи (13.3), а отже, і системи (13.1).

Справді, перейшовши до границі в рівностях (13.6), матимемо:

$$\lim_{k\to\infty} x_i^{(k+1)} = \lim_{k\to\infty}\left(\sum_{j=1}^{n} \alpha_{ij}x_j^{(k)} + \beta_i\right) = \lim_{k\to\infty} \sum_{j=1}^{n} \alpha_{ij}x_j^{(k)} +$$

$$+ \lim_{k\to\infty} \beta_i = \sum_{j=1}^{n} \alpha_{ij} \lim_{k\to\infty} x_j^{(k)} + \beta_i,$$

або

$$x_i^* = \sum_{j=1}^{n} \alpha_{ij}x_j^* + \beta_i \quad (i = 1, 2, \ldots, n).$$

Твердження доведено.

Метод послідовних наближень, який визначається формулами (13.5) або (13.6), називається *методом простої ітерації*. Цей метод зручний для обчислень на ЕОМ.

Звичайно, методом простої ітерації можна визначити розв'язок системи лінійних рівнянь лише у випадку, коли цей метод є збіжним. Умови збіжності методу простої ітерації можна дістати з принципу стискуючих відображень.

Розглянемо відображення $y = Tx$ n-вимірного простору в себе, задаючи координати y_i вектора y формулами

$$y_i = \sum_{j=1}^{n} \alpha_{ij}x_j + \beta_i \quad (i = 1, 2, \ldots, n).$$

Розв'язками системи (13.4) є нерухомі точки відображення T.

Встановимо умови, при яких відображення T буде стискуючим. Ці умови залежать від вибору метрики в просторі.

Розглянемо три випадки (див. § 7):

1) $\rho(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$ (простір $\mathbf{R}^n$).

Тоді

$$\rho(y', y'') = \rho(Tx', Tx'') = \sqrt{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} \alpha_{ij}(x_j' - x_j'')\right)^2} \leqslant$$

$$\leqslant \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_{ij}^2 \sum_{j=1}^{n} (x_j' - x_j'')^2} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_{ij}^2}\, \rho(x', x'').$$

Ми тут скористалися нерівністю Коші.

Отже, відображення T буде стискуючим відображенням при умові, що

$$q^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij}^2 < 1.$$

2) $\rho_1(x, y) = \max\limits_{1 \leqslant i \leqslant n} |y_i - x_i|$ (простір $\mathbf{R}_0^n$).

$$\rho_1(y', y'') = \rho_1(Tx', Tx'') = \max_i \left| \sum_{j=1}^{n} \alpha_{ij} (x_j' - x_j'') \right| \leqslant$$

$$\leqslant \max_i \sum_{j=1}^{n} |\alpha_{ij}| |x_j' - x_j''| \leqslant \max_i \sum_{j=1}^{n} |\alpha_{ij}| \max_j |x_j' - x_j''| =$$

$$= \left(\max_i \sum_{j=1}^{n} |\alpha_{ij}| \right) \rho_1(x', x'').$$

Звідси дістанемо умову, при якій відображення T є стискуючим:

$$q_0 = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |\alpha_{ij}| < 1.$$

3) $\rho_2(x, y) = \sum\limits_{i=1}^{n} |y_i - x_i|$ (простір $\mathbf{R}_1^n$).

$$\rho_2(y', y'') = \sum_{i=1}^{n} |y_i' - y_i''| = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} \alpha_{ij} (x_j' - x_j'') \right| \leqslant$$

$$\leqslant \sum_{i=1}^{n} \sum_{j=1}^{n} |\alpha_{ij}| |x_j' - x_j''| \leqslant \max_j \sum_{i=1}^{n} |\alpha_{ij}| \sum_{j=1}^{n} |x_j' - x_j| =$$

$$= \left(\max_j \sum_{i=1}^{n} |\alpha_{ij}| \right) \rho_2(x', x'').$$

Таким чином, відображення T стискуюче, якщо

$$q_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} |\alpha_{ij}| < 1.$$

Теорема. ***Якщо матриця α така, що задовольняє хоч одну з умов***

$$\textit{а)}\ q_0 = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |\alpha_{ij}| < 1;$$

$$\textit{б)}\ q_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} |\alpha_{ij}| < 1; \qquad (13.8)$$

$$\textit{в)}\ q^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij}^2 < 1,$$

то система рівнянь

$$x_i = \sum_{j=1}^{n} \alpha_{ij} x_j + \beta_i \quad (i = 1, 2, \ldots, n)$$

має єдиний розв'язок: $(x^* = x_1^*; x_2^*; \ldots; x_n^*)$. Цей розв'язок є границею послідовності, знайденої методом простої ітерації, виходячи з будь-якого початкового наближення) $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$.

На практиці часто за початкове наближення беруть стовпчик вільних членів $x^{(0)} = \beta$.

Із загальних оцінок для похибки n-го наближення (метод послідовних наближень, § 7), дістанемо аналогічні оцінки для наближень, отриманих методом простої ітерації. Для цього достатньо лише підставити вирази для відстані в конкретних просторах. Наприклад, з формули (7.9) для простору $\mathbf{R}_0^n$ дістаємо оцінку

$$\max_{1 \leqslant i \leqslant n} |x_i^{(k)} - x_i^*| \leqslant \frac{q_0^k}{1 - q_0} \max_{1 \leqslant i \leqslant n} |x_i^{(0)} - x_i^{(1)}|,$$

де q_0 виражається формулою (13.8, *а*).

З (7.11) дістаємо, що як тільки

$$|x_i^{(k)} - x_i^{(k-1)}| \leqslant \frac{1 - q_0}{q_0} \varepsilon = \varepsilon_1 \quad (i = 1, 2, \ldots, n),$$

то

$$|x_i^* - x_i^{(k)}| \leqslant \varepsilon \quad (i = 1, 2, \ldots, n).$$

З формули (7.10) для похибки наближення $x^{(k)}$ випливають і такі оцінки

$$\sum_{i=1}^{n} |x_i^* - x_i^{(k)}| \leqslant \frac{q_1}{1 - q_1} \sum_{i=1}^{n} |x_i^{(k)} - x_i^{(k-1)}|,$$

$$\sum_{i=1}^{n} (x_i^* - x_i^{(k)})^2 \leqslant \frac{q^2}{(1 - q)^2} \sum_{i=1}^{n} (x_i^{(k)} - x_i^{(k-1)})^2,$$

де константи q_1 і q обчислюють відповідно за формулами (13.8, *б*) і (13.8, *в*). Процес ітерацій закінчують, коли вказані оцінки свідчать про досягнення заданої точності.

Приклад.

Розв'язати систему рівнянь з точністю до $\varepsilon = 0{,}0001$.

$$\begin{cases} 2{,}1734x_1 - 0{,}3567x_2 + 1{,}0767x_3 = 1{,}2452, \\ 0{,}4528x_1 - 3{,}9862x_2 + 0{,}9875x_3 = 3{,}5166, \\ 0{,}4267x_1 + 0{,}7698x_2 + 4{,}4567x_3 = 1{,}2727. \end{cases}$$

Розв'яжемо перше рівняння відносно x_1, друге — відносно x_2, третє — відносно x_3. Дістанемо систему:

$$\begin{cases} x_1 = & 0{,}16412x_2 - 0{,}49540x_3 + 0{,}57297, \\ x_2 = 0{,}11359x_1 & + 0{,}24773x_3 - 0{,}88219, \\ x_3 = -0{,}09574x_1 - 0{,}17273x_2 & + 0{,}28557. \end{cases}$$

Перевіримо одну з достатніх умов (13.8), а саме умову (13.8, *а*).

$$\sum_{j=1}^{n} |\alpha_{1j}| = 0{,}65952; \quad \sum_{j=1}^{n} |\alpha_{2j}| = 0{,}36137; \quad \sum_{j=1}^{n} |\alpha_{3j}| = 0{,}26847.$$

Отже,

$$q_0 = \max_{i=1,2,3} \sum_{j=1}^{n} |\alpha_{ij}| = 0{,}65952 < 1.$$

Метод ітерації збігатиметься. За початкове наближення розв'язку візьмемо $x_1^{(0)} = 0{,}573$; $x_2^{(0)} = -0{,}882$; $x_3^{(0)} = 0{,}286$;

$$\varepsilon_1 = \frac{1 - q_0}{q_0}\varepsilon = \frac{1 - 0{,}65952}{0{,}65952}\,0{,}0001 \approx 0{,}00005.$$

Процес ітерації закінчуватимемо, якщо

$$\max_{i=1,2,3} |x_i^{(k)} - x_i^{(k-1)}| \leqslant 0{,}00005.$$

Результати обчислень подано в таблиці.

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\max\limits_{i=1,2,3} \lvert x_i^{(k)} - x_i^{(k-1)} \rvert$
0	0,573	—0,882	0,286	
1	0,286	—0,746	0,383	0,287
2	0,261	—0,775	0,387	0,029
3	0,257	—0,757	0,391	0,018
4	0,255	—0,756	0,392	0,002
5	0,25470	—0,75611	0,39174	0,00030
6	0,25481	—0,75621	0,39179	0,00010
7	0,25477	—0,75619	0,39179	0,00004

Розв'язок системи: $x_1 \approx 0{,}2548$; $x_2 \approx -0{,}7562$; $x_3 \approx 0{,}3918$.

Контроль обчислень у методі ітерації не має такого принципового значення, як в точних методах. Справді, метод ітерації збіжний для будь-якого початкового значення $x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}$.

Помилка, допущена під час обчислень за формулами (13.6), не впливає на кінцевий результат, бо знайдене значення можна розглядати як нове початкове наближення. Тут ми маємо справу з важливою властивістю *самокорегування* ітераційного процесу.

Зауважимо, що систему рівнянь (13.1) завжди можна подати у вигляді (13.3), причому так, щоб виконувались умови (13.8). Наприклад, якщо модуль діагонального елемента матриці системи більший від суми модулів інших елементів рядка, тобто

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \quad (i = 1, 2, \ldots, n), \tag{13.11}$$

то, розв'язавши перше рівняння відносно x_1, друге — відносно x_2 і т. д., дістанемо систему:

$$x_i = \sum_{j=1}^{n} \alpha_{ij}x_j + \beta_i \quad (i = 1, 2, \ldots, n),$$

де

$$\alpha_{ii} = 0, \ \beta_i = \frac{b_i}{a_{ii}} \quad (i = 1, 2, \ldots, n).$$

Умова (13.8, а) виконується.

Якщо система (13.1) така, що умова (13.11) не виконується, то за допомогою спеціальних перетворень можна домогтися її виконання.

Наприклад, нехай маємо систему:

$$\begin{cases} 5x_1 + x_2 - 2x_3 + x_4 = 0, \\ 2x_1 + x_2 - x_3 + 6x_4 = -6, \\ x_1 + x_2 - 3x_3 + x_4 = -6, \\ x_1 + 2x_2 + x_3 + x_4 = 2. \end{cases} \tag{13.12}$$

Побудуємо нову систему, в якій перше і четверте рівняння є відповідно першим і другим рівняннями системи (13.12), коефіцієнти яких задовольняють умови (13.11). Додавши до другого рівняння системи (13.12), третє, помножене на (—1), і четверте, помножене на (—3), дістанемо: $-2x_1 - 6x_2 - x_3 + 2x_4 = -6$, де коефіцієнти задовольняють умови (13.11). Щоб визначити третє рівняння нової системи, від третього рівняння (13.12) віднімаємо четверте. Матимемо систему:

$$\begin{cases} 5x_1 + x_2 - 2x_3 + x_4 = 0, \\ -2x_1 - 6x_2 - x_3 + 2x_4 = -6, \\ -x_2 - 4x_3 = -8, \\ 2x_1 + x_2 - x_3 + 6x_4 = -6, \end{cases}$$

коефіцієнти якої задовольняють умови (13.11).

Метод Зейделя відрізняється від методу простої ітерації тільки тим, що при обчисленні x_i на k-му кроці враховуються значення $x_1, x_2, \ldots, x_{i-1}$, обчислені на цьому самому кроці. Формули для знаходження послідовних наближень мають вигляд:

$$x_1^{(k+1)} = \sum_{j=1}^{n} \alpha_{1j} x_j^{(k)} + \beta_1,$$

. . . . . .

$$x_i^{(k+1)} = \sum_{j=1}^{i-1} \alpha_{ij} x_j^{(k+1)} + \sum_{j=i}^{n} \alpha_{ij} x_j^{(k)} + \beta_i,$$

. . . . . .

$$x_n^{(k+1)} = \sum_{j=1}^{n-1} \alpha_{nj} x_j^{(k+1)} + \alpha_{nn} x_n^{(k)} + \beta_n.$$

Якщо виконуються умови (13.8), то процес Зейделя збігається.

Приклад.

Розв'язати систему рівнянь, якщо всі коефіцієнти при невідомих і вільні члени системи — точні числа:

$$\begin{cases} 7{,}0561x_1 + 0{,}8237x_2 - 1{,}6238x_3 = -2{,}2387, \\ 0{,}9384x_1 - 6{,}9408x_2 + 2{,}0465x_3 = 12{,}8984, \\ 1{,}8472x_1 + 0{,}2709x_2 - 5{,}6707x_3 = -4{,}7832. \end{cases}$$

Розв'яжемо перше рівняння відносно x_1, друге — відносно x_2, третє — відносно x_3. Дістанемо систему:

$$\begin{cases} x_1 = \quad\quad - 0{,}11674x_2 + 0{,}23013x_3 - 0{,}31727, \\ x_2 = 0{,}13520x_1 \quad\quad + 0{,}29485x_3 - 1{,}85834, \\ x_3 = 0{,}32576x_1 + 0{,}04777x_2 \quad\quad + 0{,}84354. \end{cases}$$

Легко бачити, що

$$q_0 = \max_{1 \leqslant i \leqslant 3} \sum_{j=1}^{3} |\alpha_{ij}| = 0{,}43005 < 1.$$

За початкове наближення візьмемо $x_1^{(0)} = 0$; $x_2^{(0)} = -1$; $x_3^{(0)} = 1$ і покладемо $\varepsilon = 0{,}00001$.

Результати обчислень подано в таблиці.

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0	—1	1
1	0,030	—1,559	0,780
2	0,044	—1,622	0,780
3	0,052	—1,621	0,783
4	0,052	—1,620	0,783
5	0,0504	—1,62044	0,78308
6	0,05211	—1,62040	0,78311
7	0,05211	—1,62039	0,78311

Таким чином, $x_1 \approx 0{,}0521$; $x_2 \approx -1{,}6204$; $x_3 \approx 0{,}7831$.

Вправи.

Розв'язати системи рівнянь методами простої ітерації і Зейделя, якщо всі коефіцієнти при невідомих і вільні члени — точні числа ($\varepsilon = 0{,}0001$):

а) $\begin{cases} 4{,}072x_1 - 0{,}781x_2 + 0{,}389x_3 = -3{,}855, \\ 0{,}482x_1 + 5{,}083x_2 + 0{,}326x_3 = 0{,}087, \\ 0{,}830x_1 - 0{,}321x_2 - 9{,}458x_3 = -7{,}501; \end{cases}$

б) $\begin{cases} 7{,}939x_1 + 0{,}042x_2 - 0{,}010x_3 = 1{,}626, \\ 0{,}237x_1 - 4{,}246x_2 + 0{,}703x_3 = -2{,}687, \\ 0{,}156x_1 + 0{,}346x_2 - 5{,}611x_3 = 5{,}824. \end{cases}$

Розділ IV

СИСТЕМИ ЛІНІЙНИХ АЛГЕБРАЇЧНИХ НЕРІВНОСТЕЙ. ЗАДАЧА ЛІНІЙНОГО ПРОГРАМУВАННЯ. СИМПЛЕКС-МЕТОД

§ 14. Загальна задача математичного програмування

Загальна задача математичного програмування формулюється так. У деякому просторі задано тим чи іншим способом множину G і функцію f елементів множини G, яка кожному елементу G ставить у відповідність деяке дійсне число. Потрібно знайти найменше (чи найбільше) значення функції f, заданої на множині G, а також елементи з G, на яких таке значення досягається.

Розглянемо дійсний n-вимірний простір $\mathbf{R}^n$, нехай $x = (x_1; x_2; \ldots; x_n)$ є елементом (точкою) з $\mathbf{R}^n$. Множину G можна визначити, наприклад, як множину елементів з $\mathbf{R}^n$, таких, що задовольняють умови:

$$\varphi_i(x) \leqslant 0 \quad (i = 1, 2, \ldots, m),$$

де $\varphi_i(x)$ — деякі функції елементів з $\mathbf{R}^n$, що мають значення в $\mathbf{R}^1$.

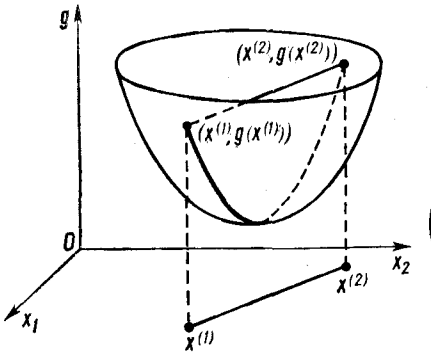
Якщо при цьому функції $\varphi_i(x)$ і $f(x)$ опуклі, то задача називається *задачею опуклого програмування*. Нагадаємо, що функція $g(x)$ називається *опуклою*, якщо для будь-яких двох точок x^1 та x^2 справджується нерівність:

$$g(\alpha x^{(1)} + (1-\alpha)x^{(2)}) \leqslant \alpha g(x^{(1)}) + (1-\alpha)g(x^{(2)}), \quad (0 \leqslant \alpha \leqslant 1).$$

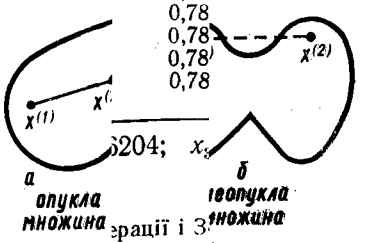
З геометричного погляду це означає, що графік функції $g(x)$ над прямою, що з'єднує точки $x^{(1)}$ і $x^{(2)}$, лежить нижче від хорди, що сполучає точки $(x^{(1)}; g(x^{(1)}))$ і $(x^{(2)}; g(x^{(2)}))$ (мал. 10). Для опуклих функцій $\varphi_i(x)$ кожна нерівність $\varphi_i(x) \leqslant 0$ визначає деяку опуклу множину G_i. Нагадаємо, що *множина називається опуклою, якщо разом з будь-якими двома точками $x^{(1)}$ і $x^{(2)}$ їй належать також усі точки відрізка $x^{(1)} + \alpha \times (x^{(2)} - x^{(1)})$, $0 \leqslant \alpha \leqslant 1$, який сполучає точки $x^{(1)}$ і $x^{(2)}$* (мал. 11).

Перерізом опуклих множин G_i є опукла множина G, $G = \bigcap_{i=1}^{m} G_i$. Отже, *задача опуклого програмування полягає в тому, щоб знайти найменше значення опуклої функції $f(x)$ на опуклій множині G* (мал. 12). Як бачимо на мал. 13, точки, в яких $\varphi_1(x) = 0$, визначають межу множини G_1 (лінія нульового рів-

ня функції $\varphi_1(x)$), а у внутрішніх точках G_1 ється нерівність $\varphi_1(x) < 0$. Аналогічно в точках мно- ється нерівність $\varphi_2(x) \leqslant 0$, причому на ме вого рі $\varphi_2(x)$ = точках $x \in$ $\in G_1 \cap$ обидві н і $\varphi_2(x)$



Мал. 10

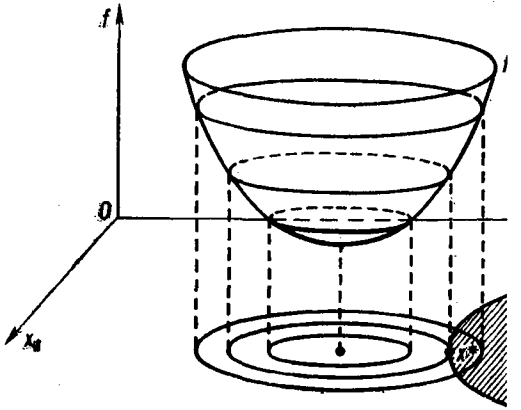


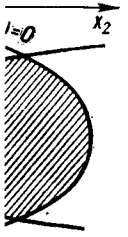
конуються ті $\varphi_1(x) \leqslant 0$ На мал. 12

0,78
0,78
0,78
0,78
0,78
0,78
0,78

204;

ерації і З ні числа 1





Мал. 12

показано точку x^*, у якій досягається ня функції $f(x)$ на множині G.

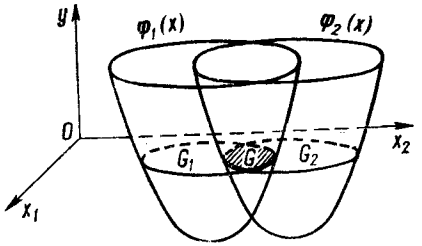
ле значення

Задача опуклого програмування має та вість: будь-який локальний мінімум оп на опуклій множині є водночас і *глобальн*

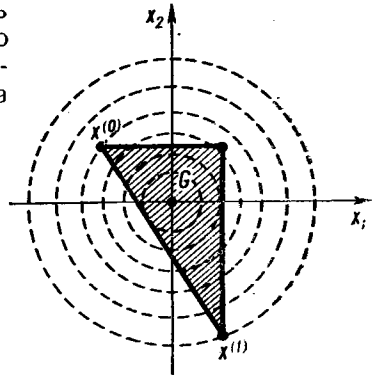
иву особли- ункції $f(x)$ *мумом*.

Зауважимо, що задача на знаходженн лького зна- ення опуклої функції $f(x)$ на опуклій мно не є задачею опуклого програмування. Вона не має за ї вище вла-

стивості. Наприклад, функція $f(x_1, x_2) = (x_1^2 + x_2^2)^{\frac{1}{2}}$ (мал. 14) у точці $x^{(0)}$ множини G досягає локального максимуму, однак це значення $f(x)$ не є найбільшим у множині G. Очевидно, $f(x)^{(1)} > f(x^{(0)})$. Пунктиром на мал. 14 показано лінії рівня функції $f(x)$, тобто такі множини, у кожній точці яких функція $f(x)$ набуває одного й того самого значення (мал. 12, 14). $f(x_1, x_2) = (x_1^2 + x_2^2)^{1/2}$, очевидно відстань точки $x = (x_1; x_2)$ від початку координат, і тому чим більший радіус кола на мал. 14, тим більшого значення набуває функція $f(x_1, x_2)$ у точках такого кола. (Відсутність зазначеної властивості значно ускладнює задачу відшукання найбільшого значення функції $f(x)$ на множині G).



Мал. 13



Мал. 14

Якщо функції $f(x)$, $\varphi_i(x)$ квадратичні, то задачу називають *задачею квадратичного програмування.*

Множина G в загальному випадку може бути дискретною. Тоді задачу називають *задачею дискретного програмування.*

Будь-яку точку $x \in G$ називають *допустимою точкою.* Точку $x^* \in G$, в якій досягається найменше (найбільше) значення функції $f(x)$ на множині G, називають *оптимальною точкою,* а $f(x^*)$ — *оптимальним значенням функції* $f(x)$ на множині G. Саму функцію $f(x)$ часто називають *цільовою функцією.*

Якщо функції $f(x)$, $\varphi_i(x)$ лінійні, тобто вирази для $f(x)$, $\varphi_i(x)$ мають вигляд $f(x) = (a, x) + b$, $\varphi_i(x) = (a_i, x) + b_i$, де a, a_i $(i = 1, 2, \ldots, n)$ — сталі вектори з $\mathbf{R}^n$, b, b_i — сталі числа, то задачу називають *задачею лінійного програмування.* Задача лінійного програмування є окремим випадком задачі опуклого програмування.

Одним із загальних методів розв'язування задачі лінійного програмування є так званий *симплекс-метод.* Обчислювальна схема цього методу грунтується на розглянутих раніше жорданових виключеннях.

Розглянемо практичний приклад, що підводить до задачі лінійного програмування.

Для виготовлення продукції двох видів потрібно чотири види сировини. Кількість наявної сировини обмежена і становить відповідно b_1, b_2, b_3, b_4 умовних одиниць. Нехай для виготовлення однієї одиниці продукції першого виду потрібно a_{11} одиниць сировини першого виду, a_{21} одиниць сировини другого виду, a_{31} одиниць сировини третього виду і a_{41} одиниць сировини четвертого виду, а для виготовлення однієї одиниці продукції другого виду потрібно відповідно a_{12} одиниць сировини першого виду, a_{22} одиниць сировини другого виду, a_{32} одиниць сировини третього виду і a_{42} одиниць сировини четвертого виду (при цьому не обов'язково, щоб усі числа a_{ij} відрізнялись від нуля, тобто для виготовлення деякої продукції можуть використовуватись не всі види сировини).

Нехай прибуток від одиниці продукції першого виду становить p_1 одиниць, а від одиниці продукції другого виду — p_2 одиниць. Треба виготовити по стільки одиниць продукції першого і другого видів, щоб за даних умов прибуток від їх реалізації був якнайбільшим. Іншими словами, програма випуску продукції повинна бути оптимальною.

Припустимо, що заплановано випустити x_1 одиниць продукції першого виду і x_2 одиниць продукції другого виду. Для цього потрібно витратити $a_{11}x_1 + a_{12}x_2$ одиниць сировини першого виду і відповідно $a_{21}x_1 + a_{22}x_2$, $a_{31}x_1 + a_{32}x_2$, $a_{41}x_1 + a_{42}x_2$ одиниць сировини другого, третього і четвертого видів. При цьому, очевидно, мають виконуватися обмеження:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &\leqslant b_1, \\ a_{21}x_1 + a_{22}x_2 &\leqslant b_2, \\ a_{31}x_1 + a_{32}x_2 &\leqslant b_3, \\ a_{41}x_1 + a_{42}x_2 &\leqslant b_4, \end{aligned} \tag{14.1}$$

тобто кількість сировини будь-якого виду, використаної для виготовлення тієї чи іншої продукції, не повинна перевищувати наявної кількості цієї сировини. При такому плані випуску продукції буде одержано прибуток $Z(x_1, x_2) = p_1x_1 + p_2x_2$. Очевидно, величини x_1 і x_2 мають задовольняти умови $x_1 \geqslant 0$, $x_2 \geqslant 0$, що випливають із змісту величин x_1 і x_2. Останні обмеження слід приєднати до обмежень (14.1).

Отже, остаточно задачу можна сформулювати так: *серед усіх величин* x_1, x_2, *що задовольняють зазначені обмеження, треба знайти такі, щоб функція* $Z(x_1, x_2) = p_1x_1 + p_2x_2$ *досягла максимального значення.* (Із змісту задачі зрозуміло, що

такі значення існують, але для загальної задачі лінійного програмування це не обов'язково).

Звичайно, якщо збільшити кількість видів продукції чи сировини, то задача ускладниться.

Загальна задача лінійного програмування формулюється так.

Якщо система лінійних нерівностей

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ \cdots\ + a_{1n}x_n + b_1 \geqslant 0, \\ a_{21}x_1 + a_{22}x_2 + \ \cdots\ + a_{2n}x_n + b_2 \geqslant 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \ \cdots\ + a_{mn}x_n + b_m \geqslant 0 \end{cases}$$

має розв'язки, то серед розв'язків цієї системи треба знайти такий, при якому лінійна функція

$$Z(x_1, x_2, \ \ldots\ , x_n) = p_1x_1 + p_2x_2 + \ \ldots\ + p_nx_n$$

набуває найбільшого (найменшого) значення.

Примітка. Оскільки кожне рівняння виду

$$a_{i1}x_1 + a_{i2}x_2 + \ \ldots\ + a_{in}x_n + b_i = 0$$

можна замінити еквівалентними йому двома нерівностями

$$a_{i1}x_1 + a_{i2}x_2 + \ \ldots\ + a_{in}x_n + b_i \geqslant 0,$$
$$-a_{i1}x_1 - a_{i2}x_2 - \ \ldots\ - a_{in}x_n - b_i \geqslant 0,$$

ми не виділятимемо окремо випадку, коли серед обмежень, крім нерівностей, є також рівняння.

Вправи.

1. Чи сумісні нерівності $(x-a)^2 + y^2 - 4 \leqslant 0$, $x^2 + (y-b)^2 - 4 \leqslant 0$, якщо $a = 0$, $b = 5$; $a = 1$, $b = 1$; $a = -\frac{1}{2}$, $b = \frac{1}{2}$; $a = \frac{1}{2}$, $b = 0$?

Зробити відповідні малюнки. Знайти (за малюнком) найменше значення функції $Z = x^2 + y^2$ на кожній з непорожніх розглянутих множин.

2. Знайти (графічно) найбільше значення функції $Z = x^2 + y^2$ на множині G, що визначається нерівностями: $x \leqslant 1$; $-x \leqslant 1$; $y \leqslant 1$; $-y \leqslant 1$.

3. Знайти найменше і найбільше значення функції xy на множині пар $(x; y)$, якщо $x \in \left\{-1; \frac{1}{2}; 0; \frac{1}{2}; 1\right\}$, $y \in \left\{-1; 0; 1; 2\right\}$.

4. Запаси сировини, що надходить від двох постачальників, становлять 40 і 80 одиниць, два її споживачі потребують відповідно 60 і 60 одиниць. Доставка однієї одиниці сировини від першого постачальника до першого споживача коштує 5, а до другого споживача — 10 одиниць вартості; відповідно від другого постачальника до першого споживача — 3 одиниці, а до другого 8 одиниць вартості. Скільки одиниць сировини повинен доставити кожний постачальник кожному споживачеві, щоб загальна вартість доставки сировини була найменшою?

5. Під посіви двох сільськогосподарських культур відведено чотири земельні ділянки, площа яких відповідно 2, 2, 4 та 1 га, причому середня врожайність першої культури на кожній з ділянок дорівнює відповідно 15, 18, 12, 20 ц з 1 га, а другої — 40, 50, 45, 60 ц з 1 га. Вартість одного центнера першої культури 40 крб., а другої 25 крб. Яку площу на кожній ділянці потрібно відвести під кожну культуру, щоб вартість зібраного врожаю була найбільшою, якщо потрібно першої культури зібрати не менш як 50, а другої — не менш як 150 ц?

§ 15. Системи лінійних нерівностей

Задача на знаходження розв'язку системи лінійних нерівностей відіграє важливу роль у розв'язуванні багатьох задач математичного програмування, а не лише лінійного. Ця задача має і самостійний інтерес.

Розглянемо спочатку такий приклад. Нехай дано систему нерівностей:

$$\begin{cases} \delta_1(x_1, x_2) \equiv -1x_1 - 1x_2 + 11 \geqslant 0, \\ \delta_2(x_1, x_2) \equiv -1x_1 - 2x_2 + 16 \geqslant 0, \\ \delta_3(x_1, x_2) \equiv -3x_1 + 0x_2 + 24 \geqslant 0, \\ \delta_4(x_1, x_2) \equiv \ \ 0x_1 - 2x_2 + 14 \geqslant 0, \\ \delta_5(x_1, x_2) \equiv \ \ 1x_1 + 0x_2 \ \ \geqslant 0, \\ \delta_6(x_1, x_2) \equiv \ \ 0x_1 + 1x_2 \ \ \geqslant 0. \end{cases} \qquad (15.1)$$

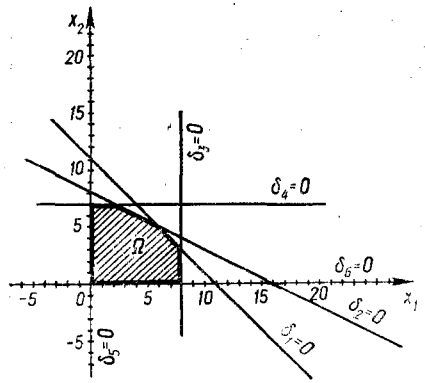
Геометрично кожну нерівність системи (15.1) можна тлумачити як деяку півплощину, обмежену прямою, рівняння якої можна дістати з відповідної нерівності, замінивши знак нерівності знаком рівності. Через $\delta_i(x_1, x_2)$ $(i = 1, 2, \ldots, 6)$ позначено відхилення довільної точки з координатами x_1, x_2 від відповідної прямої. Так, відхилення точки з координатами $x_1 = 3$, $x_2 = 8$ від прямої $\delta_1(x_1, x_2) = -1x_1 - 1x_2 + 11 = 0$ становить $\delta_1(3, 8) = -1 \cdot 3 - 1 \cdot 8 + 11 = 0$, тобто точка (3; 8) лежить на цій прямій; відхилення точки (3; 8) від прямої $\delta_2(x_1, x_2) \equiv -1x_1 - 2x_2 + 16 = 0$ становить $\delta_2(3; 8) = -1 \cdot 3 - 2 \cdot 8 + 16 = -3$; від прямої $\delta_3(x_1, x_2) \equiv -3x_1 + 0x_2 + 24 = 0$ відхилення становить $\delta_3(3, 8) = -3 \cdot 3 + 0 \cdot 8 + 24 = 15$ і т. д.

У точці з координатами $(x_1; x_2)$ нерівність, наприклад $\delta_1(x_1, x_2) \equiv -1x_1 - 1x_2 + 11 \geqslant 0$, виконується, якщо відхилення такої точки від прямої $\delta_1(x_1, x_2) = -1x_1 - 1x_2 + 11 = 0$ невід'ємне.

На мал. 15 зображено множину точок Ω (її заштриховано), у кожній точці якої виконуються всі нерівності системи (15.1).

Отже, множина Ω є множиною розв'язків системи нерівностей (15.1). Якщо точка x належить множині Ω розв'язків системи нерівностей (15.1), то відхилення цієї точки від усіх прямих, які обмежують множину Ω, будуть невід'ємними. А якщо точка не належить Ω, то принаймні одна з нерівностей не справджуватиметься і координати цієї точки не задовольнятимуть систему нерівностей, які визначають множину Ω. Очевидно, Ω є опуклим многокутником.

Отже, якщо всі відхилення деякої точки від прямих, що обмежують множину Ω, невід'ємні, то набір координат такої точки є одним з розв'язків даної системи нерівностей. Такий розв'язок називатимемо *допустимим*. Точку, що відповідає допустимому розв'язку, називатимемо *допустимою*.



Мал. 15

Як бачимо, у двовимірному просторі задачу можна розв'язати графічним способом, побудувавши множину розв'язків для кожної нерівності і взявши переріз таких множин. Якщо цим перерізом є порожня множина, то система нерівностей несумісна.

Зауважимо, що в n-вимірному (навіть в тривимірному) просторі графічно розв'язати довільну систему нерівностей вже неможливо. Проте за допомогою жорданових виключень досить легко розв'язати таку задачу.

Нехай задано систему лінійних нерівностей

$$\delta_i(x) \equiv (a_i, x) + b_i \geqslant 0 \quad (i = 1, 2, \ldots, m), \qquad (15.2)$$

де $x = (x_1; x_2; \ldots; x_n) \in \mathbf{R}^n$, $a_i = (a_{i1}; a_{i2}; \ldots; a_{in}) \in \mathbf{R}^n$, b_i — дійсні числа.

Розв'язком системи нерівностей (15.2) називається будь-яка точка $x \in \mathbf{R}^n$, яка задовольняє всі нерівності (15.2). У геометричному тлумаченні розв'язком системи (15.2) є будь-яка точка $x \in \mathbf{R}^n$, що має невід'ємні відхилення від усіх площин $\delta_1(x) = 0$, $\delta_2(x) = 0$, ..., $\delta_m(x) = 0$.

Припустимо спочатку, що ранг матриці коефіцієнтів системи (15.2) дорівнює n.

Запишемо систему (15.2) у вигляді таблиці

Таблиця 6

	x_1	x_2	...	x_n	1
δ_1	a_{11}	a_{12}	...	a_{1n}	b_1
δ_2	a_{21}	a_{22}	...	a_{2n}	b_2
...	...	...	...	...	...
δ_m	a_{m1}	a_{m2}	...	a_{mn}	b_m

Очевидно, що коли всі $b_i \geqslant 0$, то $x = (0; 0; \ldots; 0)$ є розв'язком системи нерівностей. Припустимо, що серед чисел b_i є від'ємні. Нехай, наприклад, $b_1 < 0$. За допомогою кроку жорданового виключення замість якоїсь із змінних x_j введемо до числа незалежних змінну δ_1. Нехай після кроку жорданового виключення дістанемо таблицю

Таблиця 7

	δ_1	x_2	...	x_n	1
x_1	$a_{11}^{(1)}$	$a_{12}^{(1)}$	...	$a_{1n}^{(1)}$	$b_1^{(1)}$
δ_2	$a_{21}^{(1)}$	$a_{22}^{(1)}$	...	$a_{2n}^{(1)}$	$b_2^{(1)}$
...	...	...	...	...	...
δ_m	$a_{m1}^{(1)}$	$a_{m2}^{(1)}$	...	$a_{mn}^{(1)}$	$b_m^{(1)}$

Якщо серед чисел $b_2^{(1)}$, $b_3^{(1)}$, ..., $b_m^{(1)}$ немає від'ємних, то, поклавши $\delta_1 = 0$, $x_2 = 0$, ..., $x_n = 0$, дістанемо: $x_1 = b_1^{(1)}$, $\delta_2 = b_2^{(1)}$, ..., $\delta_m = b_m^{(1)}$. Точка $x = (b_1^{(1)}; 0; \ldots; 0)$, очевидно, буде розв'язком системи (15.2), бо в цій точці $\delta_1 = 0$, $\delta_2 \geqslant 0$, ..., $\delta_m \geqslant 0$. Нехай серед чисел $b_2^{(1)}$, ..., $b_m^{(1)}$ є від'ємні, наприклад $b_2^{(1)} < 0$. Тоді замість якоїсь з решти незалежних змінних x_j вводимо змінну δ_2. Якщо таким чином вдається виключити всі змінні x_j, то через n кроків жорданових виключень дістанемо таблицю.

Таблиця 8

	δ_1	δ_2	...	δ_n	1
x_1	$a_{11}^{(n)}$	$a_{12}^{(n)}$	...	$a_{1n}^{(n)}$	$b_1^{(n)}$
...	...	...	...	...	...
x_n	$a_{n1}^{(n)}$	$a_{n2}^{(n)}$	...	$a_{nn}^{(n)}$	$b_n^{(n)}$
δ_{n+1}	$a_{n+11}^{(n)}$	$a_{n+12}^{(n)}$	...	$a_{n+1n}^{(n)}$	$b_{n+1}^{(n)}$
...	...	...	...	...	...
δ_m	$a_{m1}^{(n)}$	$a_{m2}^{(n)}$	...	$a_{mn}^{(n)}$	$b_m^{(n)}$

$\ldots, \delta_n = 0$, дістанемо: $x_1 = b_1^{(n)}$; $x_2 = b_2^{(n)}$, ...; $x_n = b_n^{(n)}$; $\delta_{n+1} = b_{n+1}^{(n)}$; ...; $\delta_m = b_m^{(n)}$*. Якщо серед чисел $b_{n+1}^{(n)}, \ldots, b_m^{(n)}$ немає від'ємних, то точка $x = (b_1^{(n)}; b_2^{(n)}; \ldots; b_n^{(n)})$ є розв'язком системи нерівностей (15.2). Ця точка належить площинам $\delta_1(x) = 0$, $\delta_2(x) = 0$, ..., $\delta_n(x) = 0$, оскільки має нульові відхилення від цих площин. Від площини $\delta_{n+1}(x) = 0, \ldots, \delta_m(x) = 0$ точка $x = (b_1^{(n)}; b_2^{(n)}; \ldots; b_n^{(n)})$ має відповідно невід'ємні відхилення $b_{n+1}^{(n)} \geqslant 0, \ldots, b_m^{(n)} \geqslant 0$. Нехай серед вільних членів $b_{n+1}^{(n)}, \ldots, b_m^{(n)}$ є від'ємні, наприклад $b_{n+1}^{(n)} < 0$*. Серед коефіцієнтів δ-рядка з від'ємним вільним членом знайдемо додатний. Якщо додатних коефіцієнтів в δ-рядку з від'ємним вільним членом немає, то система нерівностей (15.2) несумісна. Справді, нехай, наприклад,

$$a_{n+11}^{(n)} \leqslant 0, \ \ldots, \ a_{n+1n}^{(n)} \leqslant 0, \ b_{n+1}^{(n)} < 0,$$

тоді

$$\delta_{n+1}(x) = a_{n+11}^{(n)}\delta_1(x) + \ \ldots \ + a_{n+1n}^{(n)}\delta_n(x) + b_{n+1}^{(n)} \leqslant b_{n+1}^{(n)}$$

при будь-якому x, для якого $\delta_1(x) \geqslant 0$, $\delta_2(x) \geqslant 0$, ..., $\delta_n(x) \geqslant 0$. Отже, $b_{n+1}^{(n)}$ — найбільше значення функції $\delta_{n+1}(x)$ на множині розв'язків системи нерівностей $\delta_1(x) \geqslant 0$, $\delta_2(x) \geqslant 0$, ..., $\delta_n(x) \geqslant 0$ і в множині розв'язків цієї системи збільшити і зробити невід'ємним $\delta_{n+1}(x)$ неможливо. Таким чином, нерівності $\delta_1(x) \geqslant 0$, $\delta_2(x) > 0$, ..., $\delta_n(x) \geqslant 0$ і $\delta_{n+1}(x) \geqslant 0$ несумісні, а тому і вся система (15.2) не має розв'язку.

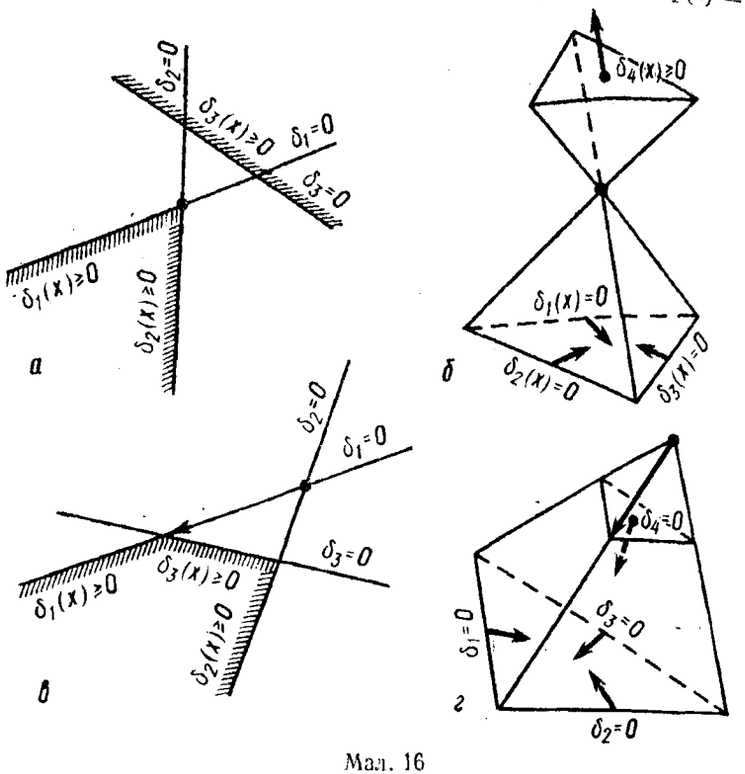
На мал. 16 показано сумісні системи нерівностей для двовимірного (мал. 16, *в*) та тривимірного (мал. 16, *г*) випадків і несумісні системи для двовимірного (мал. 16, *а*) та тривимірного випадку (мал. 16, *б*).

Нехай серед коефіцієнтів δ-рядка з від'ємним вільним членом є додатні, наприклад $b_{n+1}^{(n)} < 0$, $a_{n+11}^{(n)} > 0$. Тоді, збільшуючи відхилення $\delta_1(x)$, тобто віддаляючись від площини $\delta_1(x) = 0$ в напрямі збільшення $\delta_1(x)$, збільшуватимемо і від'ємне відхилення $\delta_{n+1}(x)$ за умови, що при цьому відбуватиметься рух по ребру $\delta_2(x) = 0, \ldots, \delta_n(x) = 0$ у напрямі збільшення $\delta_1(x)$ і $\delta_{n+1}(x)$. По ребру $\delta_2(x) = 0, \ldots, \delta_n(x) = 0$ рухатимемось у вказаному напрямі до найпершої із площин $\delta_{n+1}(x) = 0$, $\delta_{n+2}(x) = 0, \ldots, \delta_m(x) = 0$. Якщо $\delta_2(x) = 0, \ldots, \delta_n(x) = 0$, точка x досягне площини $\delta_{n+1}(x) = 0$ за умови, що

---

* Вирази для змінних x_i надалі з розрахункових таблиць можна вилучити. Визначивши потрібні значення змінних $\delta_1, \delta_2, \ldots, \delta_n$, можна легко знайти відповідні значення змінних $x_1, x_2, \ldots, x_n$ за таблицею 8.

$\delta_{n+1}(x) = a_{n+11}^{(n)} \cdot \delta_1(x) + b_{n+1}^{(n)} = 0$, тобто коли $\delta_1(x)$ набуде значення $-\dfrac{b_{n+1}^{(n)}}{a_{n+11}^{(n)}}$, очевидно, додатного. Площини $\delta_m(x) = 0$ точка x, рухаючись по ребру $\delta_2(x) = 0, \ldots, \delta_m(x) = 0$, досягне при умові $\delta_m(x) = a_{m1}^{(n)}\delta_1(x) + b_m^{(n)} = 0$, тобто при $\delta_1(x) = -\dfrac{b_m^{(n)}}{a_{m1}^{(n)}}$. Зауважимо, що коли відношення $\dfrac{b_i^{(n)}}{a_{i1}^{(n)}} > 0$ для деякого $i \in \{n+1, \ldots, m\}$, то площини $\delta_i(x) = 0$ із збільшенням $\delta_1(x)$ досягти неможливо. Справді, якщо $b_i^{(n)} > 0$ і $a_{i1}^{(n)} > 0$, то із збільшенням $\delta_1(x)$ значення $\delta_i(x) = a_{i1}^{(n)}\delta_1(x) + b_i^{(n)} \geqslant b_i^{(n)}$ збільшується, тобто із збільшенням $\delta_1(x)$ відповідна точка x не наближатиметься до площини $\delta_i(x) = 0$, а, навпаки, віддалятиметься. А коли $b_i^{(n)} < 0$ і $a_{i1}^{(n)} < 0$, то із збільшенням $\delta_1(x)$ значення $\delta_i(x) = a_{i1}^{(n)}\delta_1(x) + b_i^{(n)} \leqslant b_i^{(n)}$ лише зменшується (від'ємне й збільшується за модулем), відповідна точка x знову віддалятиметься від площини



Мал. 16

$\delta_i(x) = 0$. Отже, рухаючись по ребру $\delta_2(x) = 0$, ..., $\delta_n(x) = 0$ в напрямі збільшення $\delta_1(x)$, точка x може досягти лише тих площин $\delta_i(x) = 0$ $(i = n+1, \ldots, m)$, для яких $\frac{b_i^{(n)}}{a_{i1}^{(n)}} < 0$. При цьому $\delta_1(x)$ набуватиме значень відповідно $\delta_1(x) = -\frac{b_i^{(n)}}{a_{i1}^{(n)}} > 0$. Першою серед площин $\delta_{n+1}(x) = 0, \ldots, \delta_m(x) = 0$, очевидно, буде досягнуто ту площину, для якої відношення $\frac{b_i^{(n)}}{a_{i1}^{(n)}}$ за модулем найменше. Нехай таке найменше відношення досягнеться при $i = n+1$. Тоді дістанемо $\delta_{n+1}(x) = 0$, $\delta_1(x) = -\frac{b_{n+1}^{(n)}}{a_{n+11}^{(n)}}$. Виключимо змінну δ_1 із незалежних, ввівши замість неї змінну δ_{n+1}. Після кроку жорданового виключення матимемо таблицю:

	δ_{n+1}	δ_2	...	δ_n	1
x_1	$a_{11}^{(n+1)}$	$a_{12}^{(n+1)}$	...	$a_{1n}^{(n+1)}$	$b_1^{(n+1)}$
...	...	...	...	...	...
x_n	$a_{n1}^{(n+1)}$	$a_{n2}^{(n+1)}$	...	$a_{nn}^{(n+1)}$	$b_n^{(n+1)}$
δ_1	$a_{n+11}^{(n+1)}$	$a_{n+12}^{(n+1)}$	...	$a_{n+1n}^{(n+1)}$	$b_{n+1}^{(n+1)}$
δ_{n+2}	$a_{n+21}^{(n+1)}$	$a_{n+22}^{(n+1)}$	...	$a_{n+2n}^{(n+1)}$	$b_{n+2}^{(n+1)}$
...	...	...	...	...	...
δ_m	$a_{m1}^{(n+1)}$	$a_{m2}^{(n+1)}$	...	$a_{mn}^{(n+1)}$	$b_m^{(n+1)}$

Точка $x = (b_1^{(n+1)}; b_2^{(n+1)}; \ldots; b_n^{(n+1)})$ тепер уже належить площинам

$$\delta_{n+1}(x) = 0,\ \delta_2(x) = 0,\ \delta_3(x) = 0,\ \ldots,\ \delta_n(x) = 0.$$

Точку $x \in \mathbf{R}^n$, яка належить деяким n площинам

$$\delta_{i_1}(x) = 0,\ \ldots,\ \delta_{i_n}(x) = 0,\ i_j \in \{1, 2, \ldots, m\},$$

називатимемо *вершиною*. Площини $\delta_i(x) = 0$, яким належить вершина, називатимемо *базовими*. Вершину, яка є розв'язком системи (15.2), називатимемо *допустимою*.

Переходячи за допомогою жорданових виключень від однієї вершини до іншої, за скінченну кількість кроків знайдемо допустиму вершину або з'ясуємо, що система (15.2) несумісна.

Отже, можна сформулювати ***правила знаходження допустимої вершини системи нерівностей*** (15.2):

1. *Виключаємо всі змінні x_j $(j = 1, 2, \ldots, n)$ з незалежних.*

2. *Якщо всі вільні члени δ-рядків невід'ємні, то точка x, що належить базовим площинам $\delta_i(x) = 0$, допустима.*

3. *Якщо серед вільних членів δ-рядків є від'ємні і є δ-рядок з від'ємним вільним членом, де всі коефіцієнти недодатні, то система нерівностей* (15.2) *несумісна.*

4. *Якщо в кожному δ-рядку з від'ємним вільним членом є принаймні один додатний коефіцієнт, то, вибравши будь-який з таких коефіцієнтів, візьмемо стовпчик, в якому міститься цей коефіцієнт, за основний.*

5. *Знайдемо від'ємні відношення вільних членів δ-рядків до відповідних коефіцієнтів основного стовпчика. Рядок, на який припадає найменше за модулем відношення, візьмемо за основний.*

6. *З визначеним таким чином основним елементом виконаємо крок жорданового виключення. Дістанемо нову таблицю, знову повернемось до правила 2.*

Слід зазначити, що коли є стовпчик, всі елементи δ-рядків у якому додатні, відповідну незалежну змінну δ можна необмежено збільшувати, в результаті чого зрештою всі залежні змінні δ стануть додатними. Множина розв'язків у цьому випадку виявляється необмеженою.

Примітка. Коли найменше відношення, визначене в п. 5, дорівнює нулю, то ми не виходимо з попередньої точки x, оскільки крок вздовж відповідного ребра буде нульовим. При цьому змінюється лише набір базових площин. У такому разі можливе так зване *зациклювання*, коли через певну кількість кроків повторюватимуться ті самі таблиці. Є спеціальні засоби, щоб подолати зациклювання, на яких ми тут не зупиняємося.

У наведених щойно міркуваннях ми припускали, що ранг матриці a_{ij} $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n)$ дорівнює n. Якщо ранг цієї матриці менший за n, то виключити всі x_j $(j = 1, 2, \ldots, n)$ з незалежних не вдається.

Нехай ранг матриці (a_{ij}) $(i = \overline{1, m};\ j = \overline{1, n})$ дорівнює s $(s < n)$, тоді через s кроків жорданових виключень дістанемо таблицю.

	δ_1	...	δ_s	x_{s+1}	...	x_n	1
x_1	$a_{11}^{(s)}$	...	$a_{1s}^{(s)}$	$a_{1s+1}^{(s)}$	...	$a_{1n}^{(s)}$	$b_1^{(s)}$
...	...	...	...	...	...	...	...
x_s	a_{s1}^{s}	...	$a_{ss}^{(s)}$	$a_{ss+1}^{(s)}$	...	$a_{sn}^{(s)}$	$b_s^{(s)}$

Продовження

	δ_1	...	δ_s	x_{s+1}	...	x_n	1
δ_{s+1}	$a^{(s)}_{s+11}$	...	$a^{(s)}_{s+1s}$	0	...	0	$b^{(s)}_{s+1}$
...	...	...	...	...	...	...	...
δ_m	$a^{(s)}_{m1}$	...	$a^{(s)}_{ms}$	0	...	0	$b^{(s)}_m$

Змінним x_j $(j = s + 1, \ldots, n)$, що залишаться серед незалежних, можна надавати довільних значень. Якщо при цьому незалежні змінні δ_i $(i = \overline{1, s})$ покласти рівними нулю, то всі залежні змінні δ_i $(i = s + 1, \ldots, m)$ не змінюватимуться при зміні незалежних змінних x_j $(j = s + 1, \ldots, n)$. Змінюватимуться лише залежні змінні $x_1, \ldots, x_s$. При цьому відбуватиметься рух по $(n - s)$-вимірному «ребру», тобто по множині точок, що належать одночасно площинам $\delta_1(x) = 0, \ldots, \delta_s(x) = 0$. Площини $\delta_{s+1}(x) = 0, \ldots, \delta_m(x) = 0$ виявляються паралельними такому «ребру».

Правила розв'язування системи (15.2) при цьому залишаються незмінними. Геометричне тлумачення таке саме з тією лише відмінністю, що тепер зручно перейти до s-вимірного підпростору $\mathbf{R}^n$, оскільки в усіх перерізах, ортогональних до згаданого s-вимірного підпростору, картина залишається однаковою. Отже, фактично задача лише спрощується.

Вправи.

Розв'язати системи лінійних нерівностей. У двовимірному просторі подати відповідні малюнки.

1. $\begin{cases} -x_1 - x_2 + 1 \geqslant 0, \\ -x_1 + x_2 + 4 \geqslant 0, \\ 2x_1 - x_2 + 4 \geqslant 0. \end{cases}$

2. $\begin{cases} 1x_1 + 0x_2 - 1x_3 + 4 \geqslant 0, \\ 0x_1 - 1x_2 + 2x_3 - 5 \geqslant 0, \\ 1x_1 - 2x_2 - 5x_3 + 8 \geqslant 0, \\ 2x_1 + 1x_2 - 4x_3 - 6 \geqslant 0. \end{cases}$

3. $\begin{cases} 2x_1 - x_2 - 4 \geqslant 0, \\ x_1 + x_2 - 5 \geqslant 0, \\ x_1 + 2x_2 - 8 \geqslant 0, \\ -2x_1 + 2x_2 - 6 \geqslant 0. \end{cases}$

4. $\begin{cases} 2x_1 - x_2 + x_3 - 5 \geqslant 0, \\ x_1 - 2x_2 - 8 \geqslant 0, \\ x_2 - x_3 + 4 \geqslant 0, \\ x_1 + x_3 - 6 \geqslant 0, \\ x_1 + x_2 - 6x_3 + 6 \geqslant 0. \end{cases}$

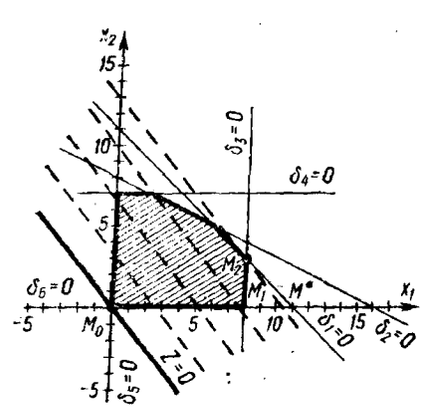
5. $\begin{cases} x_1 + x_2 + x_3 + x_4 - 8 \geqslant 0, \\ -x_1 + x_2 - 2x_3 - 4x_4 + 10 \geqslant 0. \end{cases}$

§ 16. Розв'язування задачі лінійного програмування

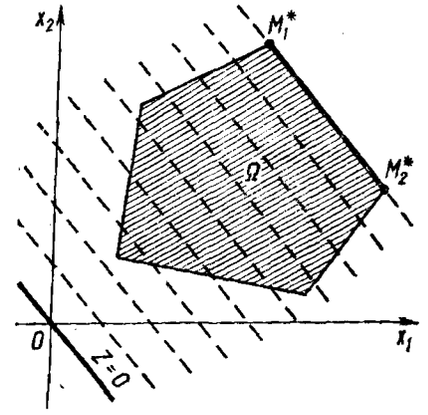
Нехай потрібно знайти максимум функції

$$Z(x_1, x_2) = 3x_1 + 2{,}5x_2$$

на множині Ω, що описується системою нерівностей (15.1), яка розглядається в § 15 (див. також § 14). Значення функції $Z(x_1, x_2) = 3x_1 + 2{,}5x_2$ в деякій точці з координатами x_1, x_2 також можна сприймати як відхилення цієї точки від прямої



Мал. 17



Мал. 18

$Z(x_1, x_2) = 3x_1 + 2{,}5x_2 = 0$. Тому чим більше точка відхиляється від прямої $Z(x_1, x_2) = 0$, тим більше значення має в цій точці функція $Z(x_1, x_2)$. Крім того, точки будь-якої прямої, паралельної прямій $Z(x_1, x_2) = 0$, однаково відхилені від прямої $Z(x_1, x_2) = 0$. Отже, точка, в якій функція $Z(x_1, x_2)$ досягає найбільшого значення, повинна бути точкою області Ω, яка лежатиме на прямій, що має з Ω спільні точки і відхилена від прямої $Z(x_1, x_2) = 0$ найбільше. Таким чином, шукана точка може лежати лише на межі області Ω, тобто належати деяким прямим, що обмежують область Ω (мал. 17, 18). Відхилення такої точки від кожної з межових прямих, яким вона належить, очевидно, дорівнює нулю.

Допустимий розв'язок, в якому функція Z досягає найбільшого значення, називатимемо *оптимальним розв'язком*. Точку, що відповідає оптимальному розв'язку, називатимемо *оптимальною точкою*. Якщо оптимальна точка існує і єдина (на мал. 17 — точка M_2), то це — деяка вершина многокутника Ω. А якщо оптимальних точок безліч (мал. 18), то до їх множини належать і деякі вершини многокутника Ω (на мал. 18 — точки

M_1^* і M_2^*). Тому для знаходження оптимальної точки достатньо розглядати лише вершини многокутника Ω.

Отже, з геометричного погляду нашу задачу можна сформулювати так: серед вершин многокутника Ω знайти таку, в якій би функція $Z(x_1, x_2)$ досягла найбільшого значення.

Симплекс-метод якраз і дає можливість знайти спочатку якусь вершину многокутника Ω, а потім від неї перейти до тієї, в якій значення функції Z не менше, ніж у попередній.

Перейдемо тепер до загального випадку. Нехай потрібно знайти максимум функції

$$Z(x) = (p,\ x) = p_1x_1 + p_2x_2 + \cdots + p_nx_n \qquad (16.1)$$

(або, що те саме, мінімум функції $Z_1(x) = -Z(x)$) на множині Ω, що визначається системою лінійних нерівностей

$$\delta_i(x) = (a_i,\ x) + b_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n + b_i \geqslant 0$$
$$(i = 1,\ 2,\ \ldots,\ m). \qquad (16.2)$$

Запишемо дані нашої задачі у вигляді таблиці.

Таблиця 9

	x_1	x_2	$\cdots$	x_n	1
δ_1	a_{11}	a_{12}	$\cdots$	a_{1n}	b_1
δ_2	a_{21}	a_{22}	$\cdots$	a_{2n}	b_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
δ_m	a_{m1}	a_{m2}	$\cdots$	a_{mn}	b_m
Z	p_1	p_2	$\cdots$	p_n	0

Не порушуючи загальності міркувань, припустимо, що ранг матриці (a_{ij}) $(i = \overline{1, m};\ j = \overline{1, n})$ дорівнює n. Знайдемо спочатку якусь допустиму вершину множини розв'язків Ω. Нехай через n кроків жорданових виключень матимемо таблицю:

Таблиця 10

	δ_1	δ_2	$\cdots$	δ_n	1
x_1	$a_{11}^{(n)}$	$a_{12}^{(n)}$	$\cdots$	$a_{1n}^{(n)}$	$b_1^{(n)}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_n	$a_{n1}^{(n)}$	$a_{n2}^{(n)}$	$\cdots$	$a_{nn}^{(n)}$	$b_n^{(n)}$

Продовження

	δ_1	δ_2	$\cdots$	δ_n	1
δ_{n+1}	$a_{n+11}^{(n)}$	$a_{n+12}^{(n)}$	$\cdots$	$a_{n+1n}^{(n)}$	$b_{n+1}^{(n)}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
δ_m	$a_{m1}^{(n)}$	$a_{m2}^{(n)}$	$\cdots$	$a_{mn}^{(n)}$	$b_m^{(n)}$
Z	$p_1^{(n)}$	$p_2^{(n)}$	$\cdots$	$p_n^{(n)}$	$Z^{(n)}$

Нехай вершина $\delta_1 = 0$, $\delta_2 = 0$, ..., $\delta_n = 0$ допустима, тобто $b_{n+1}^{(n)} \geqslant 0$, ..., $b_m^{(n)} \geqslant 0$. Якщо при цьому всі коефіцієнти в Z-рядку недодатні, тобто $p_1^{(n)} \leqslant 0$, $p_2^{(n)} \leqslant 0$, ..., $p_n^{(n)} \leqslant 0$, то в вершині $\delta_1 = 0$, $\delta_2 = 0$, ..., $\delta_n = 0$ досягається максимум функції $Z(x)$.

Справді, в такому разі $Z(x) = p_1^{(n)}\delta_1(x) + p_2^{(n)}\delta_2(x) + \cdots$ $\cdots + p_n^{(n)}\delta_n(x) + Z^{(n)} \leqslant Z^{(n)}$ для будь-якого x, для якого $\delta_1(x) \geqslant 0$, $\delta_2(x) \geqslant 0$, ..., $\delta_n(x) \geqslant 0$.

Нехай тепер серед чисел $p_1^{(n)}$, $p_2^{(n)}$, ..., $p_n^{(n)}$ є додатні, наприклад $p_1^{(n)} > 0$. Тоді, збільшуючи відхилення $\delta_1(x)$ і залишаючи $\delta_2(x)$, ..., $\delta_n(x)$ рівними нулю, тобто рухаючись по «ребру» $\delta_2(x) = 0$, ..., $\delta_n(x) = 0$ в напрямі збільшення $\delta_1(x)$, збільшуватимемо і функцію $Z(x)$. При цьому, вийшовши з вершини $\delta_1 = 0$, $\delta_2 = 0$, ..., $\delta_n = 0$, рухатимемось по згаданому ребру доти, поки вперше зіткнемось з якоюсь з «площин», що обмежують множину Ω. У результаті цього потрапимо в нову вершину множини Ω. Міркуючи так само, як і при знаходженні допустимої вершини, ми приходимо до таких ***правил:***

1. *За основний вибираємо який-небудь із стовпчиків, що містять додатні елементи Z-рядка. (Якщо таких елементів в Z-рядку немає, точка, що належить базовим площинам, є оптимальною).*

2. *У δ-рядках знаходимо від'ємні елементи в основному стовпчику. (Якщо таких елементів немає, то множина Ω необмежена, функція Z (x) на цій множині необмежена і оптимальної точки н існує).*

3. *Знаходимо найменше за абсолютною величиною від'ємне відношення вільних членів δ-рядків до відповідних елементів основного стовпчика. Рядок, який відповідає такому відношенню, беремо за основний.*

4. *З вибраним таким чином основним елементом виконуємо крок жорданового виключення. Дістави нову таблицю (нову допустиму вершину), знову повертаємось до правила 1.*

Через скінченну кількість кроків знайдемо оптимальну точку або встановимо, що такої не існує. Можливе зациклювання при цьому долаємо так само, як і при розв'язуванні системи нерівностей.

Щойно описаний метод розв'язування задачі лінійного програмування (і системи лінійних нерівностей) називається *симплекс-методом*.

Повернемось тепер до прикладу, розглянутого на початку цього параграфа. Дані задачі запишемо у вигляді таблиці:

Таблиця 11

	x_1 (δ_5)	x_2 (δ_6)	1
δ_1	−1	−1	11
δ_2	−1	−2	16
δ_3	**−3**	0	24
δ_4	0	−2	14
($x_1 \equiv$) δ_5	1	0	0
($x_2 \equiv$) δ_6	0	1	0
Z	3	2,5	0

Оскільки $\delta_5 \equiv x_1$, $\delta_6 \equiv x_2$, щоб виключити x_1 та x_2, достатньо замість x_1 зверху таблиці написати δ_5, замість x_2 — δ_6, а зліва в таблиці замість δ_5 та δ_6 написати відповідно x_1 і x_2. Якщо, виконуючи кроки жорданових виключень, мінятимемо місцями x_1 з δ_5 і x_2 з δ_6, то все одно дістанемо таку саму таблицю. Після такої зміни в таблиці всі вільні члени напроти змінних δ_1, δ_2, δ_3, δ_4 будуть додатними. Отже, точка, що лежить на прямих $\delta_5 = 0$, $\delta_6 = 0$, є вершиною многогранника (на мал. 17 це точка M_0, координати якої, за таблицею, $x_1 = 0$, $x_2 = 0$). Проте у Z-рядку обидва елементи додатні (вільний член не враховується). Тому функція $Z(x_1, x_2)$ у знайденій точці має не найбільше значення.

Виберемо стовпчик проти δ_5 основним. Найменше (за абсолютною величиною) серед відношень вільних членів до відповідних коефіцієнтів основного стовпчика $\frac{11}{-1}$, $\frac{16}{-1}$, $\frac{24}{-3}$ стоїть у δ_3-рядку. Отже, δ_3-рядок треба взяти основним, інакше якийсь вільний член стане від'ємним і ми вийдемо за межі області Ω. Наприклад, потрапимо в точку M^* (мал. 17), яка належить прямим $\delta_6 = 0$ і $\delta_1 = 0$, якщо основним візьмемо не δ_3-рядок, а δ_1-рядок, а основним стовпчиком буде той самий стовпчик напроти δ_5.

Після кроку жорданового виключення (основний елемент обведено) дістанемо таблицю:

Таблиця 12

	δ_3	δ_6	1
δ_1	$\frac{1}{3}$	**−1**	3
δ_2	$\frac{1}{3}$	−2	8
δ_5	$\frac{-1}{3}$	0	8
δ_4	0	−2	14
x_1	$\frac{1}{3}$	0	8
x_2	0	1	0
Z	−1	2,5	24

Маємо вершину, яка належить прямим $\delta_3 = 0$, $\delta_6 = 0$ і має координати $x_1 = 8$, $x_2 = 0$ (на мал. 17 — точка M_1). Оскільки в Z-рядку напроти δ_6 коефіцієнт знову додатний, то беремо тепер основним стовпчик напроти δ_6. Основним рядком треба взяти δ_1-рядок, оскільки найменше за модулем серед від'ємних відношень вільних членів до відповідних коефіцієнтів основного стовпчика припадає саме на цей рядок. Після кроку жорданового виключення матимемо таблицю:

Таблиця 13

	δ_3	δ_1	1
δ_6	$\frac{1}{3}$	−1	3
δ_2	$\frac{1}{3}$	2	2
δ_5	$-\frac{1}{3}$	0	8
δ_4	$-\frac{2}{3}$	2	8

Продовження

	δ_3	δ_1	1
x_1	$-\frac{1}{3}$	0	8
x_2	$\frac{1}{3}$	-1	3
Z	$-\frac{1}{6}$	$-\frac{5}{2}$	$31\frac{1}{2}$

Ми дістали вершину, яка належить прямим $\delta_3 = 0$, $\delta_1 = 0$ і координати якої у вихідній системі координат $x_1 = 8$, $x_2 = 3$ (на мал. 17 — точка M_2). Оскільки тепер в Z-рядку обидва коефіцієнти від'ємні, то точка, яка належить прямим $\delta_3 = 0$, $\delta_1 = 0$, є оптимальною. У цій точці функція $Z(x_1, x_2) = 3x_1 + 2{,}5x_2$ досягає найбільшого значення $Z(8, 3) = 31\frac{1}{2}$ (табл. 13)

Отже, щоб мати найбільший прибуток в умовах задачі § 14 (с. 67), потрібно виготовити 8 одиниць продукції першого виду і 3 одиниці продукції другого виду. При цьому матимемо максимально можливий за даних умов прибуток: 31,5 одиниці.

Зауважимо, що практичні задачі, як правило, мають значно більший обсяг, ніж розглянута задача. У загальному випадку про використання малюнків і геометричних побудов не може бути й мови. Проте симплекс-методом такі задачі розв'язувати порівняно легко.

Вправи.

1. Знайти найбільше значення функції

$$Z = (2 - t)x_1 + x_2$$

на множині Ω, що визначається нерівностями:

$$-2x_1 - x_2 + 8 \geqslant 0;$$
$$x_1 + x_2 - 1 \geqslant 0;$$
$$3x_1 - 2x_2 + 3 \geqslant 0;$$
$$-x_1 + x_2 + 4 \geqslant 0; \qquad x_1 \geqslant 0; \quad x_2 \geqslant 0,$$

для значень $t = 0; 1; 3; 5; 7; 9$.

2. Знайти найбільше значення функції:

$$Z = 5x_1 - x_2 - (4 - t)x_3$$

на множині, що визначається нерівностями:

$$0 \cdot x_1 + 1 \cdot x_2 + 2 \cdot x_3 - 8 \geqslant 0;$$
$$-1 \cdot x_1 + 1 \cdot x_2 + 0 \cdot x_3 - 2 \geqslant 0;$$
$$1 \cdot x_1 + 1 \cdot x_2 - 2 \cdot x_3 - 7 \geqslant 0;$$
$$-1 \cdot x_1 + 0 \cdot x_2 + 1 \cdot x_3 + 5 \geqslant 0$$

для значень $t = 0; 1; 2; 3; 4; 5$.

3. Знайти $\max\limits_{x \in \Omega} Z = -1x_1 + 2x_2$, де Ω визначається вимогами:

$$x_1 + 2x_2 + 2 \geqslant 0,$$
$$2x_1 + x_2 - 4 \geqslant 0,$$
$$x_1 - x_2 + 2 \geqslant 0,$$
$$x_1 - 4x_2 + 13 \geqslant 0,$$
$$4x_1 + x_2 + 20 \geqslant 0.$$

§ 17. Двоїстий симплекс-метод

Повернемось знову до задач (16.1), (16.2). Як ми бачили, в оптимальній точці вектор p — градієнт цільової функції $Z(x) = (p, x)$ виражається через градієнти (нормальні вектори) a_i базових площин $\delta_i(x) = 0$ з недодатними коефіцієнтами. Якщо, наприклад, у таблиці 10 точка x^*, що належить площинам $\delta_1(x) = 0$, $\delta_2(x) = 0$, ..., $\delta_n(x) = 0$, оптимальна, тобто числа $p_1^{(n)}$, $p_2^{(n)}$, ..., $p_n^{(n)}$ недодатні, то

$$p = p_1^{(n)}a_1 + p_2^{(n)}a_2 + \cdots + p_n^{(n)}a_n. \qquad (17.1)$$

Така рівність є достатньою умовою того, що точка x^*, яка належить площинам $\delta_1(x) = 0$, $\delta_2(x) = 0$, ..., $\delta_n(x) = 0$, оптимальна на множині точок, які задовольняють нерівності $\delta_1(x) \geqslant 0$, $\delta_2(x) \geqslant 0$, ..., $\delta_n(x) \geqslant 0$.

Справді, для будь-якого x, що задовольняє нерівності $\delta_1(x) \geqslant 0$, $\delta_2(x) \geqslant 0$, ..., $\delta_n(x) \geqslant 0$, якщо $p_i^{(n)} \leqslant 0$, маємо:

$$Z(x) = (p, x) \leqslant (p, x) - p_1^{(n)}\delta_1(x) - p_2^{(n)}\delta_2(x) - \cdots$$
$$\cdots - p_n^{(n)}\delta_n(x) = (p, x) - p_1^{(n)}[(a_1, x) + b_1] -$$
$$- p_2^{(n)}[(a_2, x) + b_2] - \cdots - p_n^{(n)}[(a_n, x) + b_n] =$$
$$= (p - p_1^{(n)}a_1 - p_2^{(n)}a_2 - \cdots - p_n^{(n)}a_n, x) -$$
$$- (p_1^{(n)}b_1 + p_2^{(n)}b_2 + \cdots + p_n^{(n)}b_n) =$$

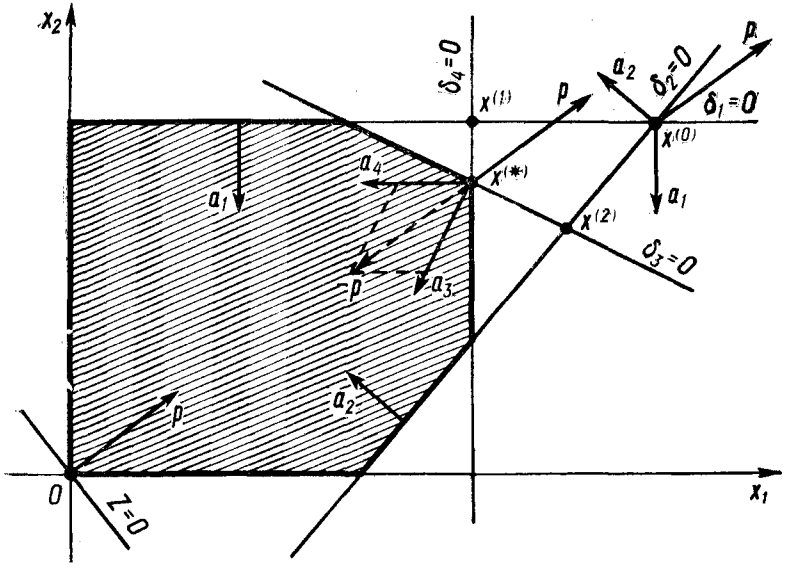
$$= (p - p_1^{(n)}a_1 - p_2^{(n)}a_2 - \cdots - p_n^{(n)}a_n,\ x^*) -$$

$$- (p_1^{(n)}b_1 + p_2^{(n)}b_2 + \cdots + p_n^{(n)}b_n) = (p,\ x^*) -$$

$$- \{p_1^{(n)}[(a_1,\ x^*) + b_1] + p_2^{(n)}[(a_2,\ x^*) + b_2] + \cdots$$

$$\cdots + p_n^{(n)}[(a_n,\ x^*) + b_n]\} = (p,\ x^*) - p_1^{(n)}\delta_1(x^*) -$$

$$- p_2^{(n)}\delta_2(x^*) - \cdots - p_n^{(n)}\delta_n(x^*) = (p,\ x^*) = Z(x^*). \quad (17.2)$$



Мал. 19

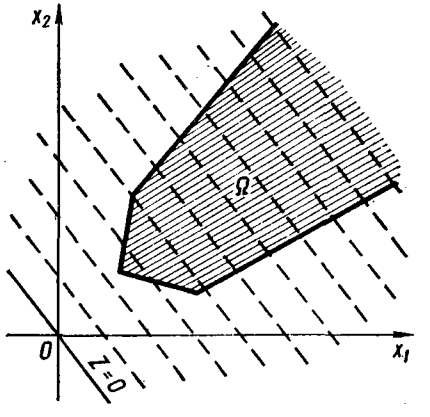
Тут ми використали умови $p - (p_1^{(n)}a_1 + p_2^{(n)}a_2 + \cdots + p_n^{(n)}a_n) = 0$, а також умови $\delta_i(x^*) = (a_i,\ x^*) + b_i = 0$ $(i = 1, 2, \ldots, n)$, оскільки точка x^* належить площинам $\delta_1(x) = 0$, $\delta_2(x) = 0, \ldots, \delta_n(x) = 0$.

Точку $\bar{x}$, яка належить деяким s $(s \leqslant n)$ площинам $\delta_{i_1}(x) = 0, \ldots, \delta_{i_s}(x) = 0$, називатимемо *двоїсто-допустимою*, якщо вектор-градієнт p цільової функції $Z(x) = (p,\ x)$ виражається через вектори-градієнти a_{i_k} $(k = 1, 2, \ldots, s)$ площин $\delta_{i_k}(x) = (a_{i_k},\ x) + b_{i_k} = 0$ з недодатними коефіцієнтами. Як видно з (17.2), якщо двоїсто-допустима точка $\bar{x}$ є водночас розв'язком системи нерівностей, тобто допустимою точкою, то вона оптимальна (мал. 19).

Можна показати, що умова (17.1) є також необхідною, щоб допустима точка $x^{(*)} \in \Omega$ була оптимальною. Наприклад, на мал. 19 точки $x^{(0)}$, $x^{(1)}$, $x^{(2)}$, $x^{(*)}$ двоїсто-допустимі, але допустимою серед них є лише точка $x^{(*)}$. Якщо зняти, наприклад, обмеження $\delta_4(x) \geqslant 0$, $\delta_3(x) \geqslant 0$, то точка $x^{(0)}$ стане допустимою, а отже, і оптимальною. На відміну від прямого симплекс-методу в двоїстому симплекс-методі насамперед знаходять двоїсто-допустиму точку.

Обчислення в двоїстому симплекс-методі розміщують так само в таблиці, як і в прямому симплекс-методі. Вихідною є таблиця 9.

Перехід від однієї таблиці до іншої виконується методом жорданових виключень за тими самими правилами, що і в прямому симплекс-методі.



Мал. 20

Перед тим як відшукувати двоїсто-допустиму точку, виключимо з незалежних змінні $x_1, x_2, \ldots, x_n$ (ми припускаємо, що ранг матриці $(a_{i,j})$ $(i = \overline{1, m};\ j = \overline{1, n})$ дорівнює n). Нехай через n кроків жорданових виключень дістанемо таблицю 10. Якщо при цьому всі $p_1^{(n)}$ недодатні, то точка $x_1 = b_1^{(n)}, \ldots, x_n = b_n^{(n)}$, яка належить площинам $\delta_1(x) = 0, \ldots, \delta_n(x) = 0$ (її нові координати $\delta_1 = 0$, $\delta_2 = 0, \ldots, \delta_n = 0$), є двоїсто-допустимою точкою. Нехай, не всі $p_i^{(n)} \leqslant 0$, наприклад $p_1^{(n)} > 0$. Тоді в δ-рядках у стовпчику, що містить $p_1^{(n)} > 0$, відшукаємо від'ємний елемент. Якщо такого немає, то двоїсто-допустимої точки не існує. При цьому відхилення δ_1 можна зробити як завгодно великим, і при досить великому δ_1 всі відхилення $\delta_{n+1}, \ldots, \delta_m$ стануть невід'ємними (якщо всі $a_{n+11}^{(n)} > 0$, $a_{n+21}^{(n)} > 0, \ldots, a_{m1}^{n} > 0$). Це означає, що область Ω необмежена і функція $Z(x)$ в цій області не досягає найбільшого значення (воно нескінченно велике). Оптимальної точки в такому випадку не існує (мал. 20).

Нехай у стовпчику, який містить $p_1^{(n)} > 0$, є від'ємний елемент, наприклад $a_{n+11}^{(n)} < 0$. Тоді рядок, що містить цей від'єм-

ний елемент, візьмемо основним і знайдемо всі від'ємні відношення елементів Z-рядка до відповідних елементів основного рядка (вільні члени до уваги не беруться). Найменше за модулем від'ємне відношення визначає основний стовпчик. Із знайденим таким чином основним елементом, що належить основному рядку і основному стовпчику, виконуємо один крок жорданового виключення. Якщо всі зазначені найменші від'ємні відношення відрізняються від нуля, то через скінченну кількість кроків або буде знайдено двоїсто-допустиму точку, або з'ясується, що її не існує.

Нехай тепер у таблиці 10 усі елементи Z-рядка недодатні, тобто ми маємо двоїсто-допустиму точку. Якщо при цьому всі вільні члени в δ-рядках невід'ємні, тобто $b_{n+1}^{(n)} \geqslant 0$, $b_{n+2}^{(n)} \geqslant 0, \ldots$ $\ldots, b_m^{(n)} \geqslant 0$, то точка також допустима, тобто оптимальна.

Припустимо, що серед вільних членів $b_{n+1}^{(n)}$, $b_{n+2}^{(n)}$, $\ldots$, $b_m^{(n)}$ є від'ємні. Тоді ми повинні з даної двоїсто-допустимої точки перейти до іншої двоїсто-допустимої так, щоб, по можливості, позбутися від'ємних вільних членів у δ-рядках. Отже, якщо в прямому симплекс-методі переходять від однієї допустимої вершини до іншої так, щоб функція $Z(x)$ при цьому збільшувалась, і рухаються по допустимих вершинах доти, поки досягнуть допустимої вершини, яка водночас буде й двоїсто-допустимою, то в двоїстому симплекс-методі рухаються по двоїсто-допустимих вершинах, наближаючись до області доти, поки не дістануть двоїсто-допустимої вершини, яка буде водночас і допустимою. При переході від однієї двоїсто-допустимої вершини до іншої значення функції $Z(x)$ зменшується (у всякому разі не збільшується). Отже, якщо $\Omega \neq \varnothing$ (множина Ω не порожня), то $\max\limits_{x \in \Omega} Z(x) = \min\limits_{x \in \Omega^*} Z(x)$, де Ω^* $(\neq \varnothing)$ множина всіх двоїсто-допустимих точок. Це є так званий принцип двоїстості.

Щоб перейти від однієї двоїсто-допустимої вершини до іншої, у δ-рядку з від'ємним вільним членом відшукуємо додатні елементи. Якщо таких немає, то система нерівностей несумісна, множина Ω порожня і оптимальної точки не існує (бо не існує допустимих точок). Нехай такі елементи є. Тоді цей δ-рядок беремо основним і знову відшукуємо найменше за модулем від'ємне відношення елементів Z-рядка до відповідних елементів основного рядка. Таке відношення визначає основний елемент. З визначеним таким чином основним елементом виконуємо крок жорданового виключення. Якщо всі зазначені найменші відношення не дорівнюють нулеві, то через скінченну кількість кроків або дістанемо допустиму, а отже, й оптимальну, вершину або з'ясуємо, що система (16.2) несумісна.

Приклад.

Двоїстим симплекс-методом розв'язати задачу лінійного програмування: знайти максимум лінійної функції

$$Z(x) = 1 \cdot x_1 + 1 \cdot x_2$$

при обмеженнях

$$\delta_1(x) = -1 \cdot x_1 + 0 \cdot x_2 + 4 \geqslant 0,$$
$$\delta_2(x) = 0 \cdot x_1 - 1 \cdot x_2 + 4 \geqslant 0,$$
$$\delta_3(x) = -1 \cdot x_1 - 1 \cdot x_2 + 6 \geqslant 0,$$
$$\delta_4(x) = 0 \cdot x_1 + 1 \cdot x_2 + 0 \geqslant 0,$$
$$\delta_5(x) = 1 \cdot x_1 + 1 x_2 - 1 \geqslant 0,$$
$$\delta_6(x) = 1 \cdot x_1 + 0 \cdot x_2 + 0 \geqslant 0.$$

Вихідною є таблиця 1).

1)

	x_1	x_2	1
δ_1	−1	0	4
δ_2	0	−1	4
δ_3	−1	−1	6
δ_4	0	1	0
δ_5	1	1	−1
δ_6	1	0	0
Z	1	1	0

2)

	δ_6	δ_4	1
δ_1	**−1**	0	4
δ_2	0	−1	4
δ_3	−1	−1	6
x_2	0	1	0
δ_5	1	1	−1
x_1	1	0	0
Z	1	1	0

3)

	δ_1	δ_4	
δ_6	−1	0	4
δ_2	0	**−1**	4
δ_3	1	−1	2
x_2	0	1	0
δ_5	−1	1	3
x_1	−1	0	4
Z	−1	1	4

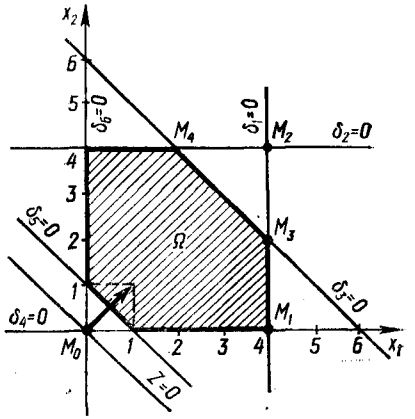
4)

	δ_1	δ_2	1
δ_6	−1	0	4
δ_4	0	−1	4
δ_3	**1**	1	−2
x_2	0	−1	4
δ_5	−1	−1	7
x_1	−1	0	4
Z	−1	−1	8

5)

	δ_3	δ_2	1
δ_6	−1	1	2
δ_4	0	−1	4
δ_1	1	−1	2
x_2	0	−1	4
δ_5	−1	0	5
x_1	−1	1	2
Z	−1	0	6

Оскільки $\delta_4 \equiv x_2$, $\delta_6 \equiv x_1$, то, виключивши x_1 та x_2 з незалежних змінних (достатньо поміняти місцями x_1 з δ_6 та δ_4 з x_2), дістанемо таблицю 2). На мал. 21 цій таблиці відповідає точка M_0. Як бачимо, у таблиці 2) точка $\delta_6 = 0$, $\delta_4 = 0$ не є двоїсто-допустимою, бо в Z-рядку є додатні елементи. Знаходимо напроти одного з них, наприклад напроти першого, від'ємний коефіцієнт. У першому стовпчику таких коефіцієнтів два: у δ_1-рядку і в δ_3-рядку. Беремо довільний з них, наприклад δ_1-рядок, основним. Найменше за модулем від'ємне відношення елементів Z-рядка до відповідних елементів основного рядка припадає на перший стовпчик. Отже, вибираємо його за основний. З визначеним таким чином основним елементом (у таблиці 2) його обведено) виконуємо крок жорданового виключення. Дістаємо таблицю 3). На мал. 21 цій таблиці відповідає точка M_1. Знайдена точка не є двоїсто-допустимою, оскільки в Z-рядку є ще додатні коефіцієнти. Беремо тепер основний елемент у δ_2-рядку і в δ_4-стовпчику. Дістаємо таблицю 4). На мал. 21 таблиці 4) відповідає точка M_2. Ця точка є двоїсто-допустимою (у Z-рядку всі коефіцієнти недодатні), однак точка не є допустимою, бо не задовольняє нерівність $\delta_3 \geqslant 0$ (у таблиці 4 в δ_3-рядку є від'ємний вільний член). Оскільки в δ_3-рядку є додатні коефіцієнти, то беремо δ_3-рядок основним. Основним можна взяти δ_1-стовпчик чи δ_2-стовпчик, бо в обох випадках відношення коефіцієнтів Z-рядка до коефіцієнтів δ_3-рядка однакове $\left(\frac{-1}{1}, \frac{-1}{1}\right)$. Візьмемо за основний δ_1-стовпчик. Після кроку жорданового виключення дістанемо таблицю 5).



Мал. 21

На мал. 21 таблиці 5) відповідає точка M_4. Оскільки тепер у δ-рядках усі вільні члени невід'ємні, то наявна двоїсто-допустима точка є водночас і допустимою, а отже, оптимальною. Як бачимо (див. табл., мал. 21), задача має безліч розв'язків — будь-яка точка відрізка $[M_3; M_4]$ оптимальна.

Вправи.

Двоїстим симплекс-методом розв'язати задачу на знаходження максимуму лінійної функції $Z(x)$ при заданих обмеженнях.

1. $Z(x_1, x_2) = p_1x_1 + p_2x_2$

p_1	1	1	1	1	2	−1	0	−1	0	1
p_2	1	2	−1	−2	1	1	1	0	−1	0

Обмеження:

$\delta_1(x_1, x_2) = x_1 \geqslant 0$; $\quad \delta_2(x_1, x_2) = x_2 \geqslant 0$;

$\delta_3(x_1, x_2) = -x_1 + 10 \geqslant 0$; $\quad \delta_4(x_1, x_2) = -x_2 + 10 \geqslant 0$;

$\delta_5(x_1, x_2) = -4x_1 + x_2 + 6 \geqslant 0$; $\quad \delta_6(x_1, x_2) = -x_1 - 4x_2 + 40 \geqslant 0$;

$\delta_7(x_1, x_2) = -5x_1 + 2x_2 + 40 \geqslant 0$; $\quad \delta_8(x_1, x_2) = x_1 - x_2 + 6 \geqslant 0$;

$\delta_9(x_1, x_2) = x_1 + x_2 - 4 \geqslant 0$.

2. $Z(x_1, x_2, x_3) = p_1x_1 + p_2x_2 + p_3x_3$

p_1	1	1	1	0	0	0	1	−1
p_2	1	0	1	1	0	1	0	−1
p_3	1	1	0	1	1	0	0	−1

Обмеження:

$-2 \leqslant x_1 \leqslant 4$; $\quad -4 \leqslant x_2 \leqslant 2$; $\quad -2 \leqslant x_3 \leqslant 2$;

$-1 \leqslant x_1 + x_2 + x_3 \leqslant 1$; $\quad -1 \leqslant x_1 - x_2 + x_3 \leqslant 1$;

$-1 \leqslant -x_1 + x_2 + x_3 \leqslant 1$; $\quad -1 \leqslant x_1 + x_2 - x_3 \leqslant 1$.

Розділ V
НАБЛИЖЕНЕ РОЗВ'ЯЗУВАННЯ РІВНЯНЬ

§ 18. Дослідження рівняння. Відокремлення коренів

Нехай задано неперервну функцію $f(x)$ і потрібно знайти деякі або всі корені рівняння

$$f(x) = 0. \qquad (18.1)$$

Здебільшого, як відомо, корені рівняння (18.1) можна знайти лише наближено. Для знаходження наближених значень коренів даного рівняння з потрібною точністю доводиться розв'язувати такі задачі: 1) досліджувати кількість коренів та їх розміщення; 2) знаходити наближені значення коренів; 3) обчислювати потрібні корені з необхідною точністю.

Перші дві задачі розв'язують аналітичними і графічними методами. Наприклад, складають таблицю функції $f(x)$, і якщо в двох сусідніх вузлах таблиці функція має різні знаки, то між ними лежить принаймні один корінь. Для відокремлення коренів інколи вдається використати властивості функції $f(x)$, її графік тощо.

Розглянемо графічний метод. Для наближеного розв'язання рівняння $f(x) = 0$ будують графік функції $y = f(x)$. Абсциси точок перетину цього графіка з віссю абсцис і є коренями рівняння $f(x) = 0$.

Оскільки практично графік функції $y = f(x)$ точно побудувати неможливо, то цим методом можна визначити лише наближені значення коренів рівняння $f(x) = 0$.

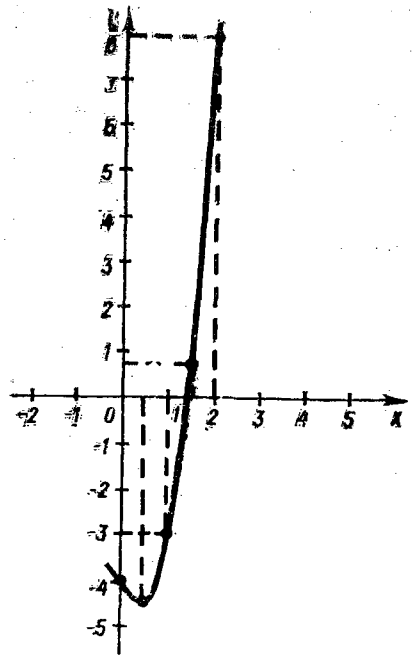
Приклад.

Знайти графічним методом додатний корінь рівняння

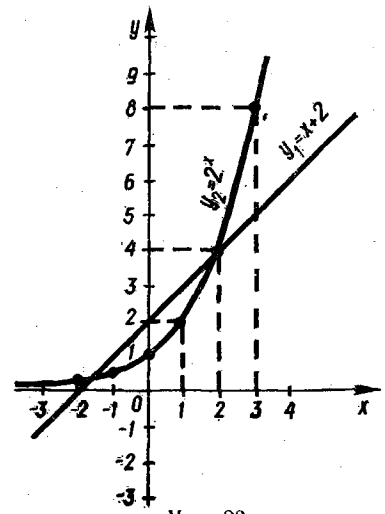
$$x^3 + 2x^2 - 2x - 4 = 0.$$

Складаємо таблицю значень функції $y = f(x) = x^3 + 2x^2 - 2x - 4$:

x	0	0,5	1	1,5	2
y	—4	—4,375	—3	0,875	8



Мал. 22



Мал. 23

Графік функції зображено на мал. 22. Як бачимо, наближене значення кореня 1,4 (точне значення $x = \sqrt{2} = 1{,}4142\ldots$).

У деяких випадках рівняння $f(x) = 0$ зручно подати у вигляді $f_1(x) = f_2(x)$, а потім побудувати графіки двох функцій: $y = f_1(x)$ і $y = f_2(x)$.

Коренями рівняння $f(x) = 0$ є абсциси точок перетину графіків функцій $y = f_1(x)$ і $y = f_2(x)$.

Приклад.

Графічним методом розв'язати рівняння $x + 2 - 2^x = 0$.

Подамо це рівняння у вигляді $x + 2 = 2^x$. Поклавши $f_1(x) = x + 2$, $f_2(x) = 2^x$, побудуємо таблиці значень обох функцій:

x	—3	—2	—1	0	1	2	3
$f_1(x)$	—1	0	1	2	3	4	5
$f_2(x)$	0,125	0,25	0,5	1	2	4	8

Графіки функцій $y = x + 2$ і $y = 2^x$ зображено на мал. 23. За графіком $x_1 = -1{,}7$; $x_2 = 2$ (при цьому $x_2 = 2$ — точне значення кореня).

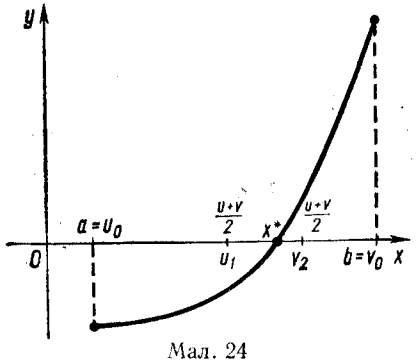
Графічний метод не дає можливості знайти корінь рівняння з наперед заданою точністю, і тому найчастіше його використовують для знаходження початкового, наближеного розв'язку та для визначення меж, в яких міститься розв'язок.

Наближені значення кореня уточнюють різними ітераційними методами. Розглянемо деякі з них.

§ 19. Метод поділу проміжку пополам

Нехай на проміжку $[a; b]$ функція $f(x)$ неперервна і набуває на кінцях проміжку значень різних знаків, тобто $f(a) \cdot f(b) < 0$. Це означає, що на $[a; b]$ рівняння (18.1) має принаймні один корінь. Цей корінь можна визначити з наперед заданою точністю *методом поділу проміжку пополам*.

Суть методу полягає в тому, що проміжок, на якому міститься корінь, поступово звужують, зменшуючи його щоразу вдвоє, поки не досягнуть потрібної точності визначення кореня.



Мал. 24

Позначимо лівий кінець проміжку, на якому міститься корінь, буквою u, правий — буквою v і знайдемо середину цього проміжку: $x = \frac{u+v}{2}$. Оскільки (за умовою) $f(u) \times f(v) < 0$, то $f(x) f(a) > 0$, або $f(x) f(a) < 0$ (мал. 24), або $f(x) = 0$. Якщо $f(x) = 0$, то корінь $x^* = x$.

Зрозуміло, що у випадку $f(x) f(a) > 0$ корінь міститься на проміжку $[x; v]$.

У випадку $f(x) f(a) < 0$ корінь міститься на проміжку $[u; x]$. Якщо довжина проміжку, на якому міститься корінь, не перевищує заданої величини ε, то це означає, що x^* знайдено з точністю до ε, бо $|x - x^*| \leqslant x - u$. Якщо заданої точності ще не досягнуто, то, позначивши x через u у випадку $f(x) f(a) > 0$ або через v у випадку $f(x) f(a) < 0$, знову знаходимо середину проміжку $[u; v]$ і повторюємо обчислення.

Переконатися в тому, що потрібна точність при обчисленні кореня x^* уже досягнута, можна й іншим способом. Якщо на

деякому проміжку $[u; v]$ функція $f(x)$ диференційовна і $0 < m \leqslant |f'(x)|$ (у цьому випадку на $[u; v]$ міститься єдиний корінь рівняння $f(x) = 0$) і якщо

$$\frac{|f(x)|}{\min\limits_{u \leqslant x \leqslant v} |f'(x)|} \leqslant \varepsilon, \qquad (19.1)$$

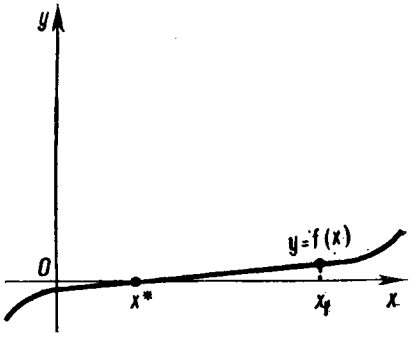
то можна вважати, що x — наближене значення кореня x^* з точністю до ε. Справді, за теоремою про середнє маємо:

$$|f(x) - f(x^*)| = |(x - x^*) f'(\xi)| \quad (x < \xi < x^*, \text{ або } x^* < \xi < x).$$

Враховуючи, що $f(x^*) = 0$, дістанемо:

$$|x - x^*| = \frac{|f(x)|}{|f'(\xi)|} \leqslant \frac{|f(x)|}{m},$$

де $m \leqslant \min\limits_{u \leqslant x \leqslant v} |f'(x)|$.



Мал. 25

Умову (19.1) можна використати для перевірки близькості x до x^*, якщо x знайдено будь-яким способом, а не тільки методом ділення проміжку пополам.

Зауважимо, що близькість до нуля $f(x)$ не означає близькість x до x^* (див. мал. 25, порівняйте з мал. 1).

Послідовність наближень, знайдених методом поділу проміжку пополам, збігається до кореня x^* рівняння $f(x) = 0$, причому щоразу маємо для x^* оцінки знизу і зверху: $u \leqslant x^* \leqslant v$.

При обчисленні значень $f(x)$ достатньо мати одну-дві правильні значущі цифри, оскільки нас цікавить лише знак $f(x)$ при даному x.

Приклад.

Методом поділу проміжку пополам знайти з точністю до 0,005 корінь рівняння $x^3 + 1{,}76439\,x^2 + 2{,}21584x - 3{,}31344 = 0$, що лежить на проміжку $[0; 1]$.

Легко з'ясувати, що в даному випадку

$$m = \min_{0 \leqslant x \leqslant 1} |f'(x)| = \min_{0 \leqslant x \leqslant 1} |3x^2 + 2 \cdot 1{,}76439x + 2{,}21584| =$$

$$= 2{,}21584.$$

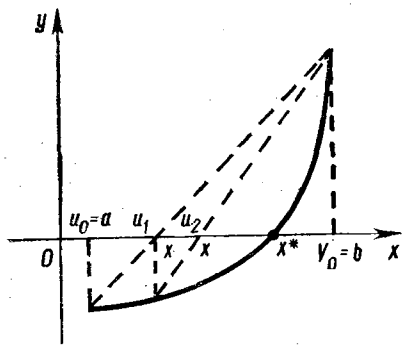
Результати обчислень подано в таблиці.

u	v	x	$x-u$	$f(x)$	$\frac{\lvert f(x)\rvert}{m}$	Знак $f(x)\cdot f(a)$
0,000	1,000	0,500	0,500	−1,639	0,74	+
0,500	1,000	0,750	0,250	−0,237	0,11	+
0,750	1,000	0,855	0,125	+0,646	0,29	−
0,750	0,875	0,812	0,062	+0,184	0,08	−
0,750	0,812	0,781	0,031	−0,030	0,014	+
0,7810	0,8120	0,7965	0,0155	+0,076	0,034	−
0,7810	0,7965	0,7888	0,0078	+0,023	0,0104	−
0,7810	0,7888	0,7849	0,0039	−0,003	0,0014	+

Як видно з таблиці, $x^* = 0{,}785$ (з точністю до 0,005).

§ 20. Метод хорд

Методом хорд (його ще називають методом лінійного інтерполювання) нове значення x знаходять як абсцису точки перетину хорди, що проходить через точки $(u;\ f(u))$, $(v;\ f(v))$, з віссю Ox (мал. 26). При цьому $f(u)\,f(v) < 0$.



Мал. 26

Рівняння згаданої хорди матиме вигляд:

$$y - f(u) = \frac{f(v)-f(u)}{v-u}(x-u).$$

Звідси знаходимо абсцису точки перетину хорди з віссю Ox (рівняння якої $y = 0$):

$$x = u - \frac{v-u}{f(v)-f(u)}\,f(u). \qquad (20.1)$$

Якщо $f(x)\,f(u) > 0$, то корінь міститься на проміжку $[x;\ v]$, тому щойно знайдений x беремо за нове значення лівого кінця проміжку, на якому міститься корінь. Якщо $f(x)\,f(u) < 0$, то корінь міститься на проміжку $[u;\ x]$, тому x беремо за нове значення правого кінця проміжку, на якому міститься корінь. Якщо $f(x) = 0$ або задовольняється умова $\frac{|f(x)|}{m} \leqslant \varepsilon$ (див. (19.1)), то можна вважати, що x — наближення кореня x^* з точністю до ε, інакше для визначення нових u і v повторюємо обчислення.

Оскільки

$$x - u = (v-u)\left(\frac{-f(u)}{f(v)-f(u)}\right) = (v-u)\,\Theta,$$

$$v - x = \left(v - u + \frac{(v-u)\,f(u)}{f(v)-f(u)}\right) = (v-u)\left(1 - \frac{-f(u)}{f(v)-f(u)}\right) = (v-u)(1-\Theta), \qquad (20.2)$$

де $\Theta = \frac{-f(u)}{f(v)-f(u)}$, $0 \leqslant \Theta \leqslant 1$, то процес послідовних наближень, знайдених методом хорд, збігається до розв'язку x^* рівняння $f(x) = 0$. Справді, нехай деякі члени послідовності чисел Θ, знайдених на кожному кроці обчислювального процесу, прямують до нуля. Неперервна на замкненому проміжку $[u;\ v]$ функція, як відомо, обмежена. Оскільки $\Theta = \frac{-f(u)}{f(v)-f(u)}$ (якщо $\Theta = 0$, $f(u) = 0$), то ці члени можуть прямувати до нуля тільки тоді, коли $f(u) \to 0$.

Отже, відповідні члени послідовності лівих кінців u збігаються до розв'язку x^* рівняння $f(x) = 0$. А якщо деякі члени послідовності чисел Θ прямують до 1, то для відповідних v $f(v) \to 0$. У цьому разі відповідні члени послідовності правих кінців збігаються до x^*. Якщо ж для всіх членів послідовності чисел Θ виконуються нерівності $0 < \alpha \leqslant \Theta \leqslant \beta < 1$, то $f(u) \to 0$ і $f(v) \to 0$ одночасно, бо, як видно з формул (20.2), на кожному кроці обчислювального процесу звужуються і проміжок $[u;\ x]$, і проміжок $[x;\ v]$. Оскільки $0 < \alpha \leqslant \Theta \leqslant \beta < 1$, то послідовність довжин таких проміжків прямує до нуля, а самі проміжки стягуються в точку. Таким чином, в усіх випадках за скінченну кількість кроків буде знайдено шукане x^* з точністю до $\varepsilon > 0$.

Якщо припустити, що на деякому проміжку $[u;\ v]$ виконується умова $0 < m \leqslant |f'(x)| \leqslant M < +\infty$, причому друга похідна не змінює знака, то можна показати, що справджується нерівність

$$|x^* - x| \leqslant \frac{M-m}{m}\,|x - x'|,$$

де x і x' — два послідовні наближення розв'язку x^*, знайдені методом хорд.

Таким чином, ми дістали ще одну оцінку близькості x до розв'язку x^*:

$|x^* - x| < \varepsilon$, якщо $|x - x'| < \frac{m}{M-m}\,\varepsilon = \varepsilon_1$, де x' і x — два послідовні значення змінної x, знайдені методом хорд. Зрозуміло, що коли $v - u < \varepsilon$, то x відрізнятиметься від x^* менше ніж на ε.

Приклад.

Методом хорд знайти корінь рівняння

$$x^3 + 1{,}76439x^2 + 2{,}21584x - 3{,}31344 = 0$$

з точністю до 0,00005 на проміжку [0; 1].

Результати обчислень розміщуємо в таблиці.

u	v	$f(u)$	$f(v)$	$\frac{f(v)-f(u)}{}$	x	$f(x)$	$\left\|\frac{1}{m}\right\| f'(x)$	Знак $f(u)f(x)$
0,00	1,00	—3,31	1,67	4,98	0,66	—0,80	0,36	+
0,66	1,00	—0,80	1,67	2,47	0,75	—0,24	0,11	+
0,75	1,00	—0,24	1,67	1,91	0,770	—0,11	0,05	+
0,770	1,000	—0,11	1,67	1,78	0,784	—0,01	0,0045	+
0,784	1,000	—0,01	1,67	1,68	0,7853	—0,001	0,00045	+
0,7853	1,0000	—0,001	1,667	1,668	0,78543	—0,00007	0,00003	+

З таблиці видно, що шукане значення $x^* = 0{,}78543$.

§ 21. Метод дотичних

Нехай функція $f(x)$ двічі неперервно диференційовна на проміжку $[a; b]$, причому похідні $f'(x)$ і $f''(x)$ не дорівнюють нулеві і зберігають знак на цьому проміжку, а $f(a)\,f(b) < 0$ (мал. 27).

За методом дотичних (методом Ньютона) наближене значення кореня x знаходять як абсцису точки перетину дотичної до кривої $y = f(x)$ в одній з точок $(u; f(u))$ чи $(v; f(v))$ з віссю Ox (мал. 27). При цьому змінна x наближається до кореня лише з одного боку. Оцінку близькості x до кореня x^*, як і раніше, можна знайти за формулою (19.1).

Першу дотичну треба проводити в тій точці $(u; f(u))$ або $(v; f(v))$, для якої виконується умова:

$$f(u)\,f''(u) > 0 \text{ чи } f(v)\,f''(v) > 0.$$

Малюнок 27 ілюструє такі випадки:

а) $f''(x) > 0,\ f'(x) > 0,\ f(u_0) < 0 \quad (u_0 \leqslant x \leqslant v_0)$;

б) $f''(x) > 0,\ f'(x) < 0,\ f(u_0) > 0 \quad (u_0 \leqslant x \leqslant v_0)$;

в) $f''(x) < 0,\ f'(x) < 0,\ f(u_0) > 0 \quad (u_0 \leqslant x \leqslant v_0)$;

г) $f''(x) < 0,\ f'(x) > 0,\ f(u_0) < 0 \quad (u_0 \leqslant x \leqslant v_0)$.

Штриховою лінією зображено дотичну, яка відповідає тому випадку, коли початкову точку вибрано невдало. Точка перетину такої дотичної з віссю Ox не належить проміжку $[u_0; v_0]$. Нехай дотична проводиться в точці $(u; f(u))$. Тоді рівняння дотичної в цій точці матиме вигляд:

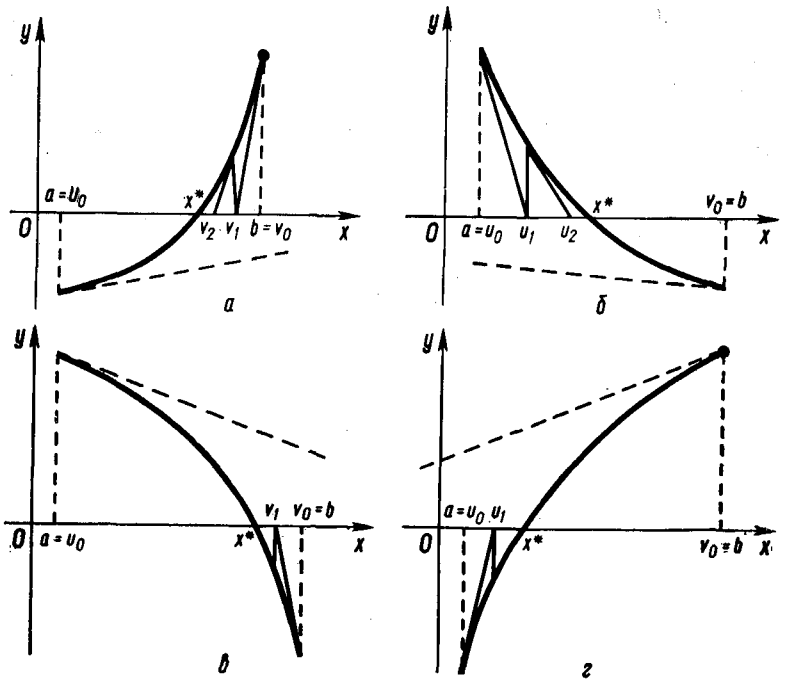
$$y - f(u) = f'(u)\,(x - u).$$

Звідси знаходимо абсцису точки перетину дотичної з віссю Ox:

$$x = u - \frac{f(u)}{f'(u)}.$$

Якщо значення u вважати за початкове (нульове, позначимо його через x_0) наближення до кореня x^*, то значення x, знайдене за останньою формулою, вважатимемо новим наближенням (першим, позначимо його x_1). Значення x_1 можна уточнити, а саме: наступне наближення до кореня x^* знову визначити як точку перетину дотичної до кривої $y = f(x)$, проведеної в точці $(x_1; f(x_1))$, з віссю абсцис. Отже, за методом Ньютона кожне наступне наближення до кореня x^* знаходять за формулою

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \quad (k = 0,\ 1,\ 2,\ \ldots).$$



Мал. 27

Покажемо, що при виконанні згаданих вище умов послідовність значень x_k, знайдених методом дотичних, збігається до розв'язку x^* рівняння $f(x) = 0$.

Розглянемо випадок $f''(x) > 0$, $f(a) > 0$ (мал. 27, *б*). Тоді $u_1 = x_1 = u_0 - \frac{f(u_0)}{f'(u_0)} > u_0$, оскільки $f(u_0) > 0$, $f'(u_0) < 0$.

Крім того, $x_1 < x^*$. Справді, щоб потрапити в точку $(x^*; 0)$, треба з точки $(u_0; f(u_0))$ рухатись в напрямі $(x^*; 0)$ вздовж хорди, яка проходить через ці дві точки. Рівняння такої хорди має вигляд

$$y - f(u_0) = \frac{f(u_0) - f(x^*)}{u_0 - x^*}(x - u_0).$$

За теоремою про середнє знайдеться така точка ξ, $u_0 < \xi < x^*$, що дотична до кривої в цій точці паралельна хорді, яка сполучає точки $(u_0; f(u_0))$ і $(x^*; 0)$. Отже, рівняння цієї хорди можна подати у вигляді

$$y - f(u_0) = f'(\xi)(x - u_0).$$

Оскільки $f''(x) > 0$ на $[u_0; v_0]$, то $f'(x)$ — монотонно зростаюча функція на цьому проміжку, і тому на $[u_0; x]$ існує єдина точка ξ, для якої справджується остання рівність.

Поклавши в рівнянні хорди $y = 0$, дістанемо:

$$x^* = u_0 - \frac{f(u_0)}{f'(\xi)}.$$

Нагадаємо, що за методом дотичних:

$$x_1 = u_0 - \frac{f(u_0)}{f'(u_0)}.$$

Крім того, врахувавши, що $f'(x)$ — зростаюча функція на $[u_0; v_0]$, маємо: $f'(\xi) > f'(u_0)$ (якщо $\xi > u_0$). Звідси

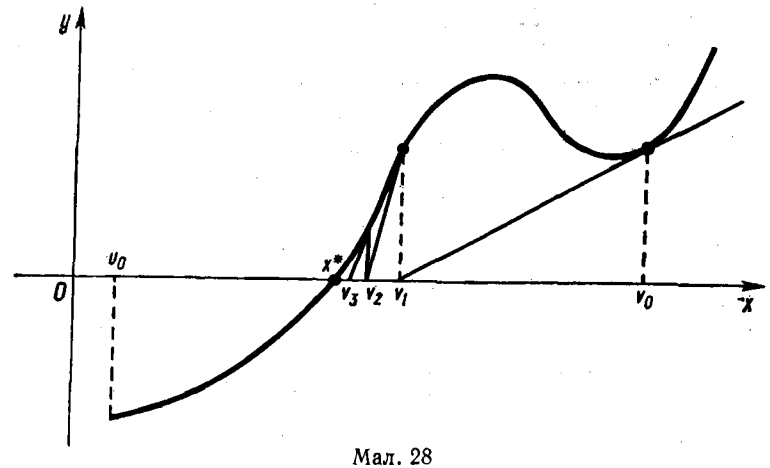
$$x_1 = u_0 - \frac{f(u_0)}{f'(u_0)} < u_0 - \frac{f(u_0)}{f'(\xi)} = x^*.$$

Отже, кожний член послідовності наближених значень кореня більший за попередній, і всі вони обмежені зверху значенням x^* (бо для них виконуються ті самі умови, що й для u_0). Але, як відомо, монотонна обмежена послідовність має границю. Позначимо її через δ, тобто $\delta = \lim\limits_{n\to\infty} u_k$. Тоді з рівності

$$u_{k+1} = u_k - \frac{f(u_k)}{f'(u_k)}$$

дістанемо:

$$\delta = \delta - \frac{f(\delta)}{f'(\delta)}, \text{ якщо } k \to \infty.$$



Мал. 28

Звідси $f(\delta) = 0$ (бо $f'(\delta) < 0$ і $f'(x)$ на $[a; b]$ обмежена як неперервна функція). Оскільки на проміжку $[a; b]$ рівняння $f(x) = 0$ має єдиний корінь, то $\delta = x^*$, тобто $\lim\limits_{k\to\infty} x_k = \lim\limits_{k\to\infty} u_k = x^*$.

Аналогічно досліджуються інші випадки.

Можна довести, що для методу дотичних справджується оцінка

$$|x^* - x_{k+1}| \leqslant \frac{1}{2}\frac{L}{m}|x^* - x_k|^2, \left(L = \max_{u_0 \leqslant x \leqslant v_0} |f''(x)|\right),$$

яка показує, що послідовні наближення за методом дотичних досить швидко збігаються до розв'язку рівняння, бо похибка кожного наступного наближення пропорційна квадрату похибки попереднього наближення (у такому випадку говорять, що метод збігається з квадратичною швидкістю). Цю оцінку ми тут не обгрунтовуємо.

Зауважимо, що умови, за яких досліджувалася збіжність методу дотичних, достатні, але не необхідні. Наприклад, на відрізку $[u_0; v_0]$ (мал. 28) жодна з умов а) — г) не виконується, та провівши лише одну дотичну, дістанемо проміжок $[u_0; v_1]$, на якому достатні умови збіжності виконуються. Отже, метод збігатиметься і в цьому випадку.

Приклади.

1. Написати рекурентні співвідношення для послідовних наближень до кореня рівняння $f(x) = x^2 - a = 0$ методом дотичних.

Маємо: $f'(x) = 2x$,

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^2 - a}{2x_k} = \frac{2x_k^2 - x_k^2 + a}{2x_k} =$$

$$= \frac{1}{2}\left(x_k + \frac{a}{x_k}\right).$$

Ми записали процес послідовних наближень для розв'язання рівняння $x^2 - a = 0$, тобто процес знаходження $x^* = \sqrt{a}$. Оскільки в нашому прикладі $f''(x) = 2 > 0$ при будь-якому x, $f'(x) = 2x > 0$, якщо $x > 0$, то процес послідовних наближень збігається до розв'язку рівняння $x^2 - a = 0$, починаючи з будь-якого початкового наближення $x_0 > 0$. (Якщо взяти $x_0 < 0$, то процес збігатиметься до $x_1^* = -\sqrt{a}$.)

2. Методом дотичних знайти корінь рівняння

$$x^3 + 1{,}76439x^2 + 2{,}21584x - 3{,}31344 = 0$$

з точністю до 0,00005 на проміжку [0; 1].

З виразів для $f'(x)$ і $f''(x)$

$$f'(x) = 3x^2 + 3{,}52878x + 2{,}21584;$$

$$f''(x) = 6x + 3{,}52878$$

знаходимо, що $f'(x) > 0$ і $f''(x) > 0$ при довільних $x \in [0; 1]$, $m = \min\limits_{0 \leqslant x \leqslant 1} |f'(x)| = 2{,}21584$.

Обчислення подано в таблиці

k	x_k	$f(x_k)$	$f'(x_k)$	$\frac{1}{m}\lvert f(x_k)\rvert$
0	1,00	1,67	8,74	0,75
1	0,81	0,17	7,04	0,08
2	0,790	0,03	6,88	0,01
3	0,78560	0,0011	6,8396	0,0005
4	0,78544	0,00000		0,00000

З таблиці знаходимо, що $x^* = 0{,}78544$ з точністю до 0,00005.

Примітка. У даному випадку всі значення функції $f(x)$ для всіх x_k, знайдених за методом дотичних, повинні мати однакові знаки. Якщо при деякому x_k внаслідок похибки обчислень дістанемо протилежний знак, потрібно збільшити точність обчислень.

Вправи.

Написати рекурентні співвідношення для послідовних наближень до кореня за методом дотичних:

а) $f(x) = a - \frac{1}{x} = 0$;

б) $f(x) = a - \frac{1}{x^2} = 0$;

в) $f(x) = x^3 - a = 0$.

§ 22. Метод ітерацій

Нехай задано рівняння $f(x) = 0$, де $f(x)$ — неперервна функція. Потрібно обчислити дійсні корені цього рівняння. Замінимо рівняння $f(x) = 0$ еквівалентним йому рівнянням

$$x = \varphi(x). \tag{22.1}$$

Виберемо деяке початкове наближення кореня x_0 і застосуємо до рівняння (1) метод послідовних наближень (метод ітерацій):

$$x_{k+1} = \varphi(x_k) \quad (k = 0, 1, 2, \ldots). \tag{22.2}$$

Якщо послідовність $\{x_k\}$ має границю x^*, тобто $\lim\limits_{k\to\infty} x_k = x^*$, то ця границя є коренем рівняння (22.1).

Справді, перейшовши до границі в рівності (22.2) і припустивши, що функція $\varphi(x)$ неперервна, знайдемо:

$$\lim_{k\to\infty} x_{k+1} = \varphi\left(\lim_{k\to\infty} x_k\right),$$

або

$$x^* = \varphi(x^*).$$

Достатні умови збіжності методу ітерацій дістанемо на основі принципу стискуючих відображень.

Нехай функція $\varphi(x)$ визначена на відрізку $[a; b]$ і відображає відрізок $[a; b]$ в себе. Множина точок відрізка $[a; b]$ з відстанню $\rho(x, y) = |y - x|$ є повним метричним простором.

Нехай функція $\varphi(x)$ на $[a; b]$ задовольняє умову Ліпшиця

$$|\varphi(x_2) - \varphi(x_1)| \leqslant K|x_2 - x_1|$$

з константою $K < 1$. При цих умовах маємо:

$$\rho(\varphi(x_2), \varphi(x_1)) = |\varphi(x_2) - \varphi(x_1)| \leqslant K|x_2 - x_1| = K\rho(x_2, x_1),$$

і тому відображення φ є стискуючим. З принципу стискуючих відображень випливає, що на відрізку $[a; b]$ існує лише одна точка x^*, така, що

$$\varphi(x^*) = x^*,$$

і послідовність $x_k = \varphi(x_{k-1})$ $(k = 1, 2, \ldots)$, де x_0 — довільна точка відрізка $[a; b]$, збігається до єдиного кореня x^* рівняння $x = \varphi(x)$.

Умова Ліпшиця для функції $\varphi(x)$, як відомо, задовольнятиметься, якщо на $[a; b]$ існує $\varphi'(x)$, причому

$$|\varphi'(x)| < K.$$

Отже, справджується таке твердження:

Нехай функція $\varphi(x)$ визначена і диференційовна на відрізку $[a; b]$, причому всі значення $\varphi(x) \in [a; b]$. Якщо існує правильний дріб q, такий, що $|\varphi'(x)| \leqslant q < 1$ для всіх $x \in [a; b]$, то

1) процес ітерацій

$$x_k = \varphi(x_{k-1}) \quad (k = 1, 2, \ldots)$$

збігається для будь-якого початкового наближення $x_0 \in [a; b]$;

2) граничне значення $x^* = \lim\limits_{k\to\infty} x_k$ є єдиним коренем рівняння $x = \varphi(x)$ на відрізку $[a; b]$.

Очевидно, твердження залишиться справедливим, якщо відрізок $[a; b]$ замінити на $]-\infty; +\infty[$.

Оцінки для наближень, отримані методом ітерацій, можна дістати з оцінок для методу послідовних наближень (див. § 7). Зокрема,

$$|x_k - x^*| \leqslant \frac{q^k}{1-q} |\varphi(x_0) - x_0|; \tag{22.3}$$

$$|x_k - x^*| \leqslant \frac{q}{1-q} |x_k - x_{k-1}|. \tag{22.4}$$

Отже, якщо

$$|x_k - x_{k-1}| \leqslant \varepsilon \frac{1-q}{q}, \tag{22.5}$$

то

$$|x_k - x^*| \leqslant \varepsilon.$$

Таким чином, процес ітерацій слід продовжувати доти, поки для двох послідовних наближень x_k і x_{k-1} не справджуватиметься нерівність (22.5), де ε — гранична абсолютна похибка наближення для кореня x^*.

На мал. 29 *а*, *б*, *в*, *г* показано випадки:

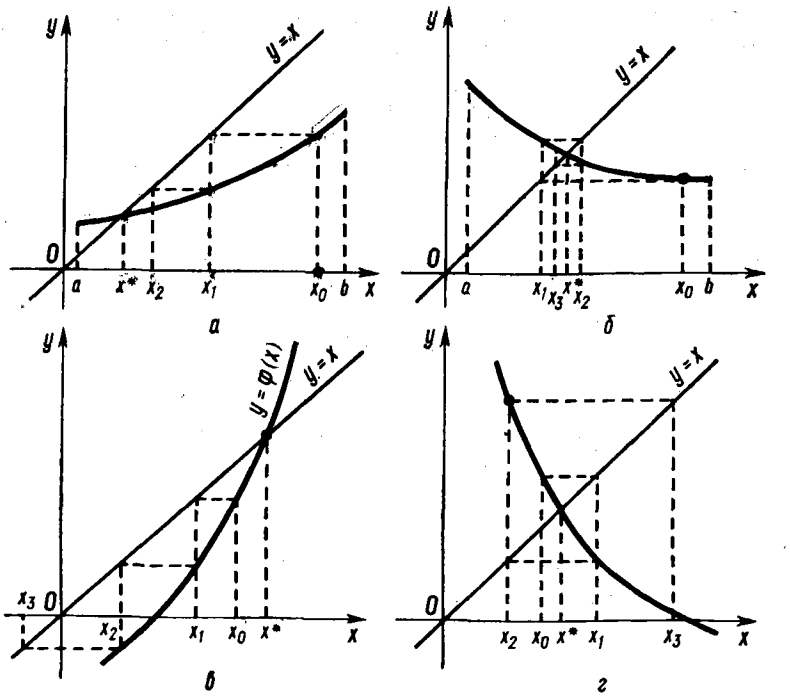
а) $0 < \varphi'(x) < 1$; *в)* $1 < \varphi'(x)$;

б) $-1 < \varphi_1(x) < 0$; *г)* $\varphi'(x) < -1$.

З мал. 29 *в*, *г* видно, що коли не виконуються умови $|\varphi'(x)| < 1$, послідовні наближення віддаляються від розв'язку x^*.

Рівняння $f(x) = 0$ завжди можна подати у вигляді (22.1), наприклад так:

$$x = x - Mf(x) = \varphi(x), \tag{22.6}$$



Мал. 29

де M — стала. Якщо $M \neq 0$, рівняння $f(x) = 0$ і (22.6) еквівалентні.

Сталу M добирають так, щоб в околі $[a; b]$ кореня x^*

$$|\varphi'(x)| = |(x - Mf(x))'| = |1 - Mf'(x)| \leqslant q < 1,$$

тобто щоб виконувались умови:

$$-1 < 1 - Mf'(x) < 1$$

або

$$0 < Mf'(x) < 2.$$

Отже, стала M повинна мати той самий знак, що й $f'(x)$, і задовольняти умову $M < \frac{2}{\max\limits_{a\leqslant x\leqslant b} |f'(x)|}$. (Звідси випливає ще одна вимога: $f'(x)$ не повинна змінювати знак).

Як видно з попередніх міркувань, процес ітерації збігається тим швидше, чим ближче q до нуля. Отже, M слід добирати так, щоб добуток $Mf'(x)$ був по можливості ближчий до 1 для всіх $x \in [a; b]$.

Іноді замість сталої M у формулі (22.6) беруть деяку функцію $\psi(x)$, тоді рівняння (22.6) матиме вигляд:

$$x = x - \psi(x) f(x) = \varphi(x)$$

($\psi(x) \neq 0$ і неперервна, якщо $x \in [a; b]$). При цьому $\psi(x)$ добирають так, щоб на $[a; b]$ виконувалась умова $|\varphi'(x)| < 1$, або $0 < \psi'(x) f(x) + f'(x) \psi(x) < 2$. Наприклад, за методом Ньютона взято $\psi(x) = \frac{1}{f'(x)}$, а за методом хорд (з нерухомим кінцем ω) $\psi(x) = \frac{\omega - x}{f(\omega) - f(x)}$.

Зауважимо, що ітераційний процес автоматично виправляє незначні помилки, допущені в процесі обчислень (якщо вони не виводять за область, в якій виконуються умови, достатні для збіжності процесу). Тому немає потреби весь час виконувати обчислення з великою кількістю значущих цифр. Точність треба поступово збільшувати з наближенням до x^*.

Приклад.

Методом ітерацій знайти на відрізку $[0; 1]$ корінь рівняння з точністю до $\varepsilon = 0{,}00005$.

$$x^3 + 1{,}76439x^2 + 2{,}21584x - 3{,}31344 = 0.$$

Подамо це рівняння в такому вигляді:

$$x = x - \frac{x^3 + 1{,}76439x^2 + 2{,}21584x - 3{,}31344}{10}.$$

Тоді $\varphi'(x) = 1 - \frac{1}{10}(3x^2 + 3{,}52878x + 2{,}21584)$,

$$0 < \varphi'(x) < 0{,}8, \text{ якщо } x \in [0; 1].$$

Отже, $q = 0{,}8$.

Тоді $\frac{1-q}{q}\varepsilon = \frac{1 - 0{,}8}{0{,}8} 0{,}00005 = 0{,}0000125$.

Отже, ітерації треба продовжувати доти, поки виконається $|x_k - x_{k-1}| < 0{,}0000125$.

За початкове наближення x_0 візьмемо 0.

Обчислення записуємо в таблицю

k	x_k	$\psi(x_k) = x_{k+1}$	$\lvert x_k - x_{k+1} \rvert$
0	0	0,331	0,331
1	0,331	0,566	0,235
2	0,566	0,696	0,130
3	0,696	0,754	0,058
4	0,754	0,775	0,021

Продовження

k	x_k	$\varphi(x_k) = x_{k+1}$	$\lvert x_k - x_{k+1} \rvert$
5	0,775	0,782	0,007
6	0,782	0,7841	0,0024
7	0,7844	0,78511	0,00071
8	0,78511	0,78534	0,00023
9	0,78534	0,78541	0,00007
10	0,78541	0,78543	0,00002
11	0,78543	0,78544	0,00001

З цієї таблиці випливає, що $x^* = 0{,}78544$ з точністю до 0,00005.

Вправи.

1. Розв'язати графічним методом рівняння:

а) $x^3 - 1{,}3x^2 + 2x - 2{,}6 = 0$;
б) $2\sin x + 1{,}5x - 3 = 0$;
в) $x^2 - \cos x = 0$;
г) $x + 3^x - 1 = 0$.

2. Знайти корінь рівняння (з точністю до 0,0005) на проміжку $[a; b]$ методом поділу проміжку пополам, методами хорд, дотичних та методом ітерацій:

а) $x^3 + 0{,}673x^2 - 0{,}921x - 5{,}732 = 0$; $a = 0{,}5$; $b = 2{,}5$;
б) $x^3 + 2{,}212x^2 + 3{,}400x - 0{,}901 = 0$; $a = 0{,}0$; $b = 2{,}0$;
в) $x^3 + x - 3{,}12 = 0$; $a = 0{,}5$; $b = 1{,}5$;
г) $2^x - 8x + 2 = 0$; $a = 0$; $b = 1$.

Розділ VI
АПРОКСИМАЦІЯ ФУНКЦІЙ

§ 23. Постановка задачі наближення функцій

Вважають, що на множині дійсних чисел X визначено деяку дійсну функцію $y = f(x)$, якщо кожному числу x з цієї множини поставлено у відповідність одне дійсне число y з множини Y. На практиці часто трапляються випадки, коли знайти значення y для відповідних x досить важко. Крім того, часто аналітичний вираз функції $y = f(x)$ взагалі невідомий, а відомі лише її значення в скінченній кількості точок. Ці значення можуть бути знайдені в результаті спостережень чи вимірювань в якому-небудь експерименті, або в результаті обчислень. Тому виникає потреба вихідну функцію $y = f(x)$ наближено замінити (апроксимувати) деякою іншою функцією $\varphi(x)$, в певному розумінні близько до $f(x)$ і такою, що простіше обчислюється чи досліджується. Тоді при всіх значеннях аргументу з множини X покладають $f(x) \approx \varphi(x)$. Функцію $\varphi(x)$ називають *апроксимуючою*. Близькість функцій $f(x)$ і $\varphi(x)$ можна, зокрема, оцінювати в метричних просторах за допомогою відстані $\rho(f(x), \varphi(x))$. По-різному вводячи відстань, дістають різні конкретні випадки задачі апроксимації.

Часто апроксимуючу функцію $\varphi(x)$ беруть у вигляді лінійної комбінації функцій деякого класу, які утворюють скінченну чи зчисленну множину $\{\varphi_i(x)\}$, причому будь-яка скінченна система елементів $\varphi_i(x)$ лінійно незалежна. Тобто $\varphi(x)$ беруть у вигляді

$$\varphi(x) = a_0\varphi_0(x) + a_1\varphi_1(x) + \cdots + a_n\varphi_n(x), \qquad (23.1)$$

де a_i $(i = 0, 1, \ldots, n)$ — сталі коефіцієнти.

Функцію $\varphi(x)$ в цьому випадку називають *узагальненим многочленом*. Надалі розглядатимемо наближення функцій узагальненими многочленами. У цьому випадку задачу апроксимації можна сформулювати так.

Задано функцію $f(x)$. Потрібно знайти такий узагальнений многочлен $\varphi(x)$ (підібрати його коефіцієнти a_i $(i = 0, 1, \ldots, n)$), щоб відхилення (в деякому розумінні) функції $f(x)$ від $\varphi(x)$ на заданій множині X було найменшим.

Нехай у точках $x_0, x_1, \ldots, x_n$ $(x_i \neq x_j$, якщо $i \neq j)$ з відрізка $[a; b]$ відомі значення функції $y = f(x)$:

$$y_0 = f(x_0), \quad y_1 = f(x_1), \quad \ldots, \quad y_n = f(x_n).$$

Розглянемо один з випадків апроксимації, що називається *інтерполюванням*, або *інтерполяцією*. Суть його полягає в тому, що коефіцієнти a_i $(i = 0, 1, \ldots, n)$ многочлена (23.1) добирають так, щоб у точках x_i $(i = 0, 1, \ldots, n)$ значення функцій $\varphi(x)$ і $f(x)$ збігалися, тобто

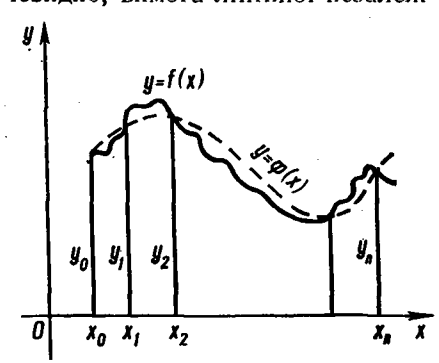
$$f(x_i) = \sum_{j=0}^{n} a_j\varphi_j(x_i) \quad (i = 0, 1, \ldots, n). \qquad (23.2)$$

Точки x_i $(i = 0, 1, \ldots, n)$ називаються *вузлами інтерполювання*, а многочлен $\varphi(x)$ — *інтерполяційним многочленом*. Формулу $y = \varphi(x)$, знайдену для обчислення значень функції $y = f(x)$, називають *інтерполяційною*.

Задача інтерполювання матиме єдиний розв'язок, якщо при будь-якому розміщенні вузлів (серед яких немає таких, що збігаються) визначник системи (23.2) не дорівнюватиме нулю, тобто

$$\begin{vmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \ldots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \ldots & \varphi_n(x_1) \\ \ldots & \ldots & \ldots & \ldots \\ \varphi_0(x_n) & \varphi_1(x_n) & \ldots & \varphi_n(x_n) \end{vmatrix} \neq 0.$$

Системи функцій, які задовольняють таку умову, називають *системами Чебишова*. Очевидно, вимога лінійної незалежності системи $\{\varphi_i(x)\}$ є необхідною для того, щоб ця система функцій була системою Чебишова. При інтерполюванні узагальнений многочлен будують за деякою чебишовською системою функцій.



Мал. 30

З геометричного погляду задача інтерполювання полягає в знаходженні лінії $y = \varphi(x)$, що проходить через точки площини з координатами $(x_0; y_0)$, $(x_1; y_1)$, …, $(x_n; y_n)$ (мал. 30).

На практиці систему $\{\varphi_i(x)\}$ часто беруть у вигляді послідовності цілих невід'ємних степенів змінної x, тобто

$$\varphi_i(x) = x^i \quad (i = 0, 1, 2, \ldots).$$

Тут узагальнені многочлени є звичайними алгебраїчними многочленами:

$$\varphi(x) = P_n(x) = \sum_{j=0}^{n} a_j x^j.$$

Інтерполювання в цьому випадку називається *поліноміальним,* або *параболічним.*

Для інтерполювання періодичних функцій природно використати не алгебраїчні, а тригонометричні многочлени. Зокрема, якщо період функції $f(x)$ дорівнює 2π, то інтерполяційний многочлен шукають у вигляді

$$T_n(x) = a_0 + \sum_{i=1}^{n} (a_i \cos ix + b_i \sin ix).$$

Коефіцієнти $T_n(x)$ визначають з умов:

$$T_n(x_k) = f(x_k) \quad (k = 0, 1, \ldots, 2n).$$

§ 24. Інтерполяційний многочлен Лагранжа

Нехай у точках x_i $(i = 0, 1, \ldots, n)$ $(x_i \neq x_j$, якщо $i \neq j)$ з відрізка $[a; b]$ задано значення функції $y = f(x)$: $y_i = f(x_i)$. Треба побудувати такий поліном $L_n(x) = a_0 + a_1 x + \ldots + a_n x^n$ (степеня, не вищого за n), який у вузлах $x_0, x_1, \ldots, x_n$ набуває тих самих значень, що й функція $y = f(x)$, тобто

$$L_n(x_i) = f(x_i) \quad (i = 0, 1, \ldots, n). \tag{24.1}$$

З умови (24.1) для знаходження коефіцієнтів a_i $(i = 0, 1, \ldots, n)$ дістанемо систему рівнянь:

$$\begin{cases} a_0 + a_1 x_0 + a_2 x_0^2 + \cdots + a_n x_0^n = f(x_0), \\ a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n = f(x_1), \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \cdots + a_n x_n^n = f(x_n). \end{cases} \tag{24.2}$$

Визначником цієї системи є так званий визначник Ван дер Монда:

$$\Delta = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{vmatrix},$$

який, як відомо, дорівнює

$$\Delta = \prod_{0 \leqslant j < i \leqslant n} (x_i - x_j) \neq 0.$$

Це означає, що функції $1, x, x^2, \ldots, x^n$ утворюють систему Чебишова. Отже, система (24.2) має єдиний розв'язок при будь-яких значеннях

$$f(x_i) \quad (i = 0, 1, \ldots, n).$$

Таким чином, коли серед вузлів немає таких, що збігаються, то алгебраїчний інтерполяційний многочлен існує і єдиний (з точністю до форми запису). Деякі форми інтерполяційних многочленів розглянемо в цьому розділі.

Щоб дістати інтерполяційний многочлен у явному вигляді, не розв'язуватимемо систему (24.2), а побудуємо безпосередньо многочлен $L_n(x)$, який задовольняє умову (24.1).

Многочлен $L_n(x)$ шукатимемо у вигляді лінійної комбінації деяких многочленів степеня n, причому коефіцієнтами цієї лінійної комбінації будуть задані значення функції $y = f(x)$ у вузлах, тобто

$$L_n(x) = \sum_{i=0}^{n} y_i L_i^{(n)}(x), \tag{24.3}$$

де $L_i^{(n)}(x)$ $(i = 0, 1, \ldots, n)$ — поки що невідомі многочлени степеня n. З умови (24.1) випливає, що многочлени $L_i^{(n)}(x)$ мають задовольняти умову

$$L_i^{(n)}(x_j) = \begin{cases} 1, & \text{якщо } j = i, \\ 0, & \text{якщо } j \neq i, \end{cases}$$

тобто

$$L_i^{(n)}(x) = \frac{(x - x_0)(x - x_1) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n)}{(x_i - x_0)(x_i - x_1) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)} =$$

$$= \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}.$$

Підставивши вирази $L_i^{(n)}(x)$ у формулу (24.3), дістанемо вираз інтерполяційного многочлена $L_n(x)$:

$$L_n(x) =$$

$$= \sum_{i=0}^{n} \frac{(x - x_0)(x - x_1) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n)}{(x_i - x_0)(x_i - x_1) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_1 - x_n)} y_i. \tag{24.4}$$

Інтерполяційний многочлен, записаний у вигляді (24.4), називається *інтерполяційним многочленом Лагранжа.* Многочлени $L_i^{(n)}(x)$ називаються *коефіцієнтами Лагранжа.*

Досить поширеними випадками інтерполювання є *лінійне* і *квадратичне інтерполювання.* Формули лінійного і квадратичного інтерполювання дістанемо із загальної формули, якщо $n = 1$ і $n = 2$.

Так, якщо $n = 1$, маємо многочлен першого степеня, який є рівнянням прямої, що проходить через дві точки $(x_0; y_0)$

і $(x_1;\ y_1)$:

$$y = L_1(x) = \frac{x - x_1}{x_0 - x_1} y_0 + \frac{x - x_0}{x_1 - x_0} y_1. \qquad (24.5)$$

Ця формула є формулою лінійної інтерполяції.

Поклавши у формулі Лагранжа $n = 2$, дістанемо:

$$y = L_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} y_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} y_1 + \\ + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} y_2, \qquad (24.6)$$

де $y = L_2(x)$ — рівняння параболи, що проходить через три задані точки: $(x_0;\ y_0)$, $(x_1;\ y_1)$, $(x_2;\ y_2)$. Отже, суть квадратичного інтерполювання полягає в тому, що функцію $y = f(x)$ на відрізку $[a;\ b]$ замінюють многочленом другого степеня (параболою), значення якого в трьох точках (вузлах) дорівнюють значенням функції. Формула (24.6) є формулою квадратичного інтерполювання.

Приклади.

1. Для функції, значення якої в чотирьох точках задано таблицею

x	0	1	2	4
$f(x)$	1	3	2	4

побудувати інтерполяційний многочлен Лагранжа.

В даному випадку $n = 3$. Тоді за формулою (24.4) знаходимо:

$$L_3(x) = 1\,\frac{(x-1)(x-2)(x-4)}{(0-1)(0-2)(0-4)} + 3\,\frac{(x-0)(x-2)(x-4)}{(1-0)(1-2)(1-4)} + \\ + 2\,\frac{(x-0)(x-1)(x-4)}{(2-0)(2-1)(2-4)} + 4\,\frac{(x-0)(x-1)(x-2)}{(4-0)(4-1)(4-2)}.$$

Розклавши цей многочлен за степенями x, матимемо:

$$L_3(x) = \frac{13}{24} x^3 - \frac{75}{24} x^2 + \frac{55}{12} x + 1.$$

Неважко побачити, що формулі Лагранжа можна надати і такого вигляду:

$$L_n(x) = \Pi_{n+1}(x) \sum_{i=0}^{n} \frac{y_i}{(x - x_i)\,\Pi'_{n+1}(x_i)}, \qquad (24.7)$$

де $\Pi_{n+1}(x) = (x - x_0)(x - x_1) \ldots (x - x_n)$, а $\Pi'_{n+1}(x_i)$ — значення похідної від функції $\Pi_{n+1}(x)$ у точці $x = x_i$.

Зауважимо, що форма коефіцієнтів Лагранжа інваріантна відносно лінійної підстановки $x = at + b$, де a і b — сталі, причому $a \neq 0$. Інакше кажучи, форма коефіцієнтів $L_i^{(n)}(x)$ $(i = 0, 1, \ldots, n)$ після підстановки $x = at + b$ залишається такою самою, тільки x замінюється на t, а x_i на t_i. Справді, записавши формулу для коефіцієнтів Лагранжа

$$L_i^{(n)}(x) = \frac{(x - x_0)(x - x_1) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n)}{(x_i - x_0)(x_i - x_1) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)}$$

і поклавши в ній $x = at + b$ і $x_k = at_k + b$ $(k = 0, 1, \ldots, n)$, знайдемо:

$$L_i^{(n)}(t) = \\ = \frac{a(t - t_0)\,a(t - t_1) \ldots a(t - t_{i-1})\,a(t - t_{i+1}) \ldots a(t - t_n)}{a(t_i - t_0)\,a(t_i - t_1) \ldots a(t_i - t_{i-1})\,a(t_i - t_{i+1}) \ldots a(t_i - t_n)}.$$

Скоротивши чисельник і знаменник на a^n, матимемо:

$$L_i^{(n)}(t) = \frac{(t - t_0)(t - t_1) \ldots (t - t_{i-1})(t - t_{i+1}) \ldots (t - t_n)}{(t_i - t_0)(t_i - t_1) \ldots (t_i - t_{i-1})(t_i - t_{i+1}) \ldots (t_i - t_n)},$$

що й треба було довести.

Цю властивість коефіцієнтів Лагранжа часто використовують для спрощення обчислень.

2. Використовуючи значення $\cos x$ для $x = 0°$, $30°$, $45°$, $60°$, $90°$, $120°$, $135°$, обчислити $\cos 63°$.

x	0	30°	45°	60°	90°	120°	135°
$y = \cos x$	1,00000	0,86603	0,70711	0,50000	0,00000	−0,50000	−0,70711

Для даного прикладу $x_0 = 0°$, $x_1 = 30°$, $x_2 = 45°$, $x_3 = 60°$, $x_4 = 90°$, $x_5 = 120°$, $x_6 = 135°$, а $x = 63°$.

Щоб спростити обчислення, покладемо $x = 15t$. Тоді значення нової змінної t, які відповідають вузлам інтерполювання, дорівнюватимуть: $t_0 = 0$, $t_1 = 2$, $t_2 = 3$, $t_3 = 4$, $t_4 = 6$, $t_5 = 8$, $t_6 = 9$. Значенню $x = 63°$ відповідає значення $t = 4{,}2$. Для обчислення шуканого значення $\cos 63°$ підставимо дані в формулу (24.7), в якій треба лише замінити x і x_i відповідно на t і t_i. Тоді $\Pi_7(4{,}2) = (4{,}2 - 0)(4{,}2 - 1)(4{,}2 - 3)(4{,}2 - 4)(4{,}2 - 6)(4{,}2 - 8)(4{,}2 - 9) = 72{,}80824$.

Обчислимо $S = \sum_{i=0}^{6} \frac{y_i}{(t - t_i)\,\Pi'_7(t_i)}$, якщо $t = 4{,}2$, і знайдемо

(у проміжних результатах беремо дві запасні цифри порівняно з табличними значеннями y_i, в остаточному результаті запасні

цифри округляємо):

$$S = 6\,235\,437 \cdot 10^{-9} \text{ і } L_7(4,2) = \Pi_7(4,2) \cdot S = 0{,}45399.$$

Звідси $\cos 63° = 0{,}45399$.

У результаті, знайденому за допомогою інтерполювання, всі цифри правильні.

Якщо вузли рівновіддалені ($x_1 - x_0 = x_2 - x_1 = \ldots = x_n - x_{n-1} = h$), доцільно зробити лінійну підстановку виду $x = x_0 + ht$.

§ 25. Оцінка похибки інтерполяційного многочлена Лагранжа

Побудований для функції $y = f(x)$ інтерполяційний многочлен Лагранжа n-го степеня $L_n(x)$ у вузлах інтерполювання $x_0, x_1, \ldots, x_n$ збігається з функцією $y = f(x)$. В інших точках відрізка $[a; b]$ він відрізняється від функції $y = f(x)$.

Величина

$$R_n(x) = f(x) - L_n(x),$$

яка характеризує близькість полінома $L_n(x)$ до функції $f(x)$ у деякій точці x відрізка $[a; b]$, називається *залишковим членом інтерполяційної формули* $L_n(x)$. Знайдемо вираз для залишкового члена $R_n(x)$ для функцій $f(x)$, які на відрізку $[a; b]$, що містить вузли інтерполювання, мають неперервні похідні до $(n+1)$-го порядку включно.

Оскільки відомо, що залишковий член у вузлах інтерполювання дорівнює нулю, то можна записати:

$$f(x) - L_n(x) = \Pi_{n+1}(x)\, k(x), \tag{25.1}$$

$\Pi_{n+1}(x) = (x - x_0)(x - x_1)(x - x_2) \ldots (x - x_n)$, а $k(x)$ — невідома функція. Для знаходження $k(x)$ розглянемо допоміжну функцію

$$u(t) = f(t) - L_n(t) - \Pi_{n+1}(t)\, k(x),$$

де x відіграє роль параметра. Тут за x візьмемо точку, в якій потрібно обчислити залишковий член інтерполяційного многочлена.

За рівністю (25.1) функція $u(t)$ дорівнює нулю, якщо $t = x_0, x_1, \ldots, x_n$ і якщо $t = x$, тобто має на відрізку $[a; b]$ принаймні $n+2$ нулі. Отже, на кінцях кожного з відрізків

$$[x_0; x_1],\ [x_1; x_2],\ \ldots,\ [x_i; x],\ [x; x_{i+1}],\ \ldots,\ [x_{n-1}; x_n]$$

функція $u(t)$ набуває значень, що дорівнюють нулю. За теоремою Ролля всередині кожного з цих відрізків знайдеться

принаймні одна точка, яка буде нулем похідної $u'(t)$. Отже, похідна $u'(t)$ має на відрізку $[a; b]$ принаймні $n+1$ нуль. Міркуючи аналогічно, встановлюємо, що $u''(t)$ на $[a; b]$ має принаймні n нулів. Нарешті, $u^{(n+1)}(t)$ має на $[a; b]$ принаймні один нуль ξ: $u^{(n+1)}(\xi) = 0$.

Оскільки $L_n^{(n+1)}(t) = 0$, $\Pi_{n+1}^{(n+1)}(t) = (n+1)!$, то для $u^{(n+1)}(t)$ маємо:

$$u^{(n+1)}(t) = f^{(n+1)}(t) - (n+1)!\, k(x).$$

Отже, при $t = \xi$ дістаємо:

$$0 = f^{(n+1)}(\xi) - (n+1)!\, k(x).$$

Звідси

$$k(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Підставивши в рівність (25.1) значення $k(x)$, дістанемо вираз для залишкового члена інтерполяційного многочлена $L_n(x)$:

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\Pi_{n+1}(x), \tag{25.2}$$

де $\xi \in [a; b]$, $a = \min(x_0, x_1, \ldots, x_n, x)$; $b = \max(x_0, x_1, \ldots, x_n, x)$. З формули (25.2) можна записати таку оцінку для $R_n(x)$:

$$|R_n(x)| = |f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!}\,|\Pi_{n+1}(x)|, \tag{25.3}$$

де $M_{n+1} = \max\limits_{x \in [a,b]} |f^{(n+1)}(x)|$.

Якщо вузли рівновіддалені, тобто $x_i - x_{i-1} = h$, $i = 1, 2, \ldots, n$, то оцінка для залишкового члена набирає вигляду:

$$|R_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!}\,h^{n+1}\,|t(t-1)(t-2)\ldots(t-n)|, \tag{25.4}$$

$$t = \frac{x - x_0}{h}.$$

Таким чином, для випадку рівновіддалених вузлів $|R_n(x)| = O(h^{n+1})$*.

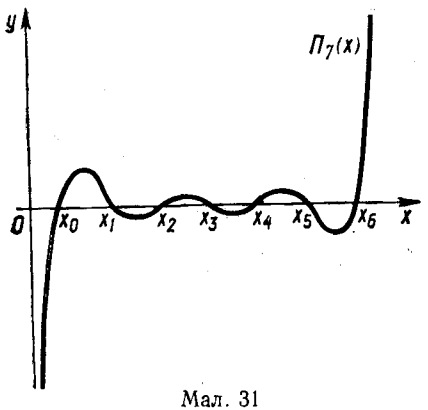
З оцінки (25.3) видно, що величина похибки многочлена Лагранжа залежить від того, як вибрано вузли інтерполювання (вони визначають функцію $\Pi_{n+1}(x)$). Оцінити $\Pi_{n+1}(x)$ при довільному розміщенні вузлів інтерполювання досить

---

* Нагадаємо, що дві нескінченно малі $\alpha(x)$ і $\beta(x)$ називаються *величинами одного порядку* при $x \to x_0$ (записують: $\alpha(x) = O(\beta(x))$, якщо скінченна границя їх відношення $\lim\limits_{x \to x_0} \frac{\alpha(x)}{\beta(x)}$ існує і не дорівнює нулю.

складно. Коли вузли рівновіддалені, то поблизу центрального вузла інтерполяції екстремуми функції $\Pi_{n+1}(x)$ невеликі, поблизу крайніх вузлів дещо більші, а якщо x виходить за крайні вузли, то $\Pi_{n+1}(x)$ швидко зростає. Наприклад, графік $\Pi_7(x)$ подано на мал. 31.

Знаходження значень $y = f(x)$, якщо значення аргументу x лежать між крайніми вузлами, називають *інтерполюванням*, а якщо x лежить поза крайніми вузлами,— то *екстраполюванням*. З попереднього випливає, що при екстраполяції далеко за крайні вузли похибка може бути досить великою. Екстраполяцію доцільно застосовувати тоді, коли функція $f(x)$ біля кінців таблиці змінюється плавно і відстань від кінців таблиці, на якій екстраполюють, невелика (менша від кроку h).



Мал. 31

Примітка. Звичайно, якщо аналітичний вираз функції $f(x)$ невідомий, то оцінками (25.3) і (25.4) користуватися не можна. Та й тоді, коли аналітичний вираз для $f(x)$ відомий, цими оцінками користуватися важко через складність знаходження M_{n+1}. На практиці похідні замінюють скінченними різницями. Розглянута вище похибка є похибкою методу.

Оскільки значення y_i часто даються наближено, то навіть і тоді, коли всі проміжні обчислення виконуються точно, результат дістаємо з похибкою. Це неусувна похибка. Якщо формулу Лагранжа взяти у вигляді (24.7), то, скориставшись результатами I розділу, встановимо, що неусувна похибка формули Лагранжа не перевищуватиме величини

$$\Delta_L = \Delta_f \,|\, \Pi_{n+1}(x) \,|\, \sum_{i=0}^{n} \frac{1}{|\,(x - x_i)\,\Pi'_{n+1}(x_i)\,|}, \qquad (25.5)$$

де Δ_f — максимальне значення абсолютних похибок величин

$$y_i = f(x_i) \quad (i = 0, 1, 2, \ldots, n).$$

Зауважимо, що остання похибка при $x \in [x_0; x_n]$ (відповідно $t = \frac{x - x_0}{h} \in [0; n]$) порівняно мала і дуже повільно зростає із збільшенням n.

У процесі обчислень виникає також похибка при округленнях проміжних результатів. Практично цю похибку можна зробити значно меншою порівняно з неусувною, якщо проміжні обчислення виконувати з більшою кількістю цифр, ніж це вимагається для табличних значень $f(x)$, тобто коли обчислення виконувати із запасними цифрами. В остаточному результаті запасні цифри округлюють. При оцінці похибки результату слід враховувати і похибку остаточного округлення. Повна похибка інтерполювання дорівнює сумі всіх перелічених вище похибок.

Приклад.

Користуючись інтерполяційним многочленом Лагранжа, обчислити sin 0,86, якщо відомі значення sin x при $x = 0{,}5$; 0,7; 0,9; 1,1; 1,3 з п'ятьма правильними значущими цифрами. Оцінити похибку результату.

Для спрощення обчислень робимо заміну $x = ht + x_0$, тобто $x = 0{,}2t + 0{,}5$. Тоді значенню $x = 0{,}86$ відповідає $t = \frac{0{,}86 - 0{,}5}{0{,}2} = 1{,}8$.

У проміжних результатах зберігаємо дві запасних цифри порівняно з табличними даними y_i, в остаточному результаті округлюємо їх.

Виконавши обчислення за формулою (24.7), дістанемо:

$$L_4(1{,}8) = 1{,}8 \cdot 0{,}8\,(-0{,}2)\,(-1{,}2)\,(-2{,}2)\left(\frac{0{,}47943}{43{,}2} + \frac{0{,}64422}{-4{,}8} + \right.$$

$$\left. + \frac{0{,}78333}{-0{,}8} + \frac{0{,}89121}{7{,}2} + \frac{0{,}96356}{-52{,}8}\right) = -0{,}76032\,(-0{,}9967471) =$$

$$= 0{,}7578468.$$

Отже,

$$\sin 0{,}86 \approx 0{,}75785.$$

Визначимо точність інтерполяції для sin 0,86.

Залишковий член оцінимо за формулою (25.4). Оскільки $f^{(5)}(x) = \cos x$, то

$$M_5 = \max_{0{,}5 \leqslant x \leqslant 1{,}3} |\cos x| = 0{,}87758 < 1,$$

$$|R_4(0{,}86)| < \frac{1\,(0{,}2)^5}{5!}\,|\,1{,}8\,(1{,}8 - 1)\,(1{,}8 - 2)\,(1{,}8 - 3)\,(1{,}8 - 4)\,| =$$

$$= \frac{(0{,}2)^5}{120}\,0{,}76032 = 0{,}2 \cdot 10^{-5}.$$

Неусувну похибку формули Лагранжа обчислимо за формулою (25.5). Оскільки табличні значення функції $y = \sin x$ задані з п'ятьма правильними значущими цифрами, то $\Delta_f =$

$= 0{,}5 \cdot 10^{-5}$. Підставивши значення необхідних величин у формулу (25.5), дістанемо:

$$\Delta_L = 0{,}5 \cdot 10^{-5} \cdot 0{,}76032 \left(\frac{1}{43{,}2} + \frac{1}{4{,}8} + \frac{1}{0{,}8} + \frac{1}{7{,}2} + \frac{1}{52{,}8}\right) = 0{,}62 \cdot 10^{-5}.$$

Отже, значення $y = 0{,}7578468$ (результат до округлення) має абсолютну похибку, не більшу як $0{,}2 \cdot 10^{-5} + 0{,}62 \times \times 10^{-5} = 0{,}82 \cdot 10^{-5}$.

Оскільки проміжні обчислення виконувались із запасними знаками, то похибкою округлення проміжних результатів ми нехтуємо. Похибка округлення остаточного результату $0{,}32 \cdot 10^{-5}$.

Отже, похибка значення 0,75785, знайденого інтерполюванням, не перевищує $0{,}82 \cdot 10^{-5} + 0{,}32 \cdot 10^{-5} = 1{,}14 \cdot 10^{-5}$, тобто в результаті буде принаймні чотири правильні значущі цифри.

Для порівняння наведемо значення sin 0,86 з п'ятьма правильними значущими цифрами: sin 0,86 = 0,75784.

Вправи.

1. Для таблично заданої функції побудувати інтерполяційний многочлен Лагранжа:

x	5	6	8	10
y	1,66	1,43	1,11	0,91

2. Використовуючи інтерполяційний поліном Лагранжа, обчислити значення функції в даних точках і оцінити похибку результату. Функції задано таблицями значень:

1)

x	0,30	0,40	0,50	0,60	0,70
$y = x + \sin x$	0,59552	0,78942	0,97943	1,16464	1,34422

2)

x	15°	30°	45°	55°	60°
$y = \cos x$	0,96593	0,86603	0,70711	0,57358	0,50000

3. Дано таблицю функції $y = e^x$ на проміжку $[0; 1]$ з кроком $h = 0{,}01$. Яка найбільша похибка лінійної інтерполяції? Яка найбільша похибка квадратичної інтерполяції?

4. Оцінити, з якою точністю можна обчислити за формулою Лагранжа lg 2,5, коли відомі значення lg 1; lg 2; lg 3; lg 4; lg 5.

§ 26. Скінченні різниці. Зв'язок скінченних різниць з похідними

Нехай y_0, y_1, y_2, ...— значення деякої функції $y = f(x)$, що відповідають рівновіддаленим значенням аргументу x_0, x_1, x_2... ($x_{i+1} - x_i = h > 0$, $i = 0, 1, 2, \ldots$). Величина $y_{i+1} - y_i$ називається *скінченною різницею першого порядку* функції $f(x)$ в точці x_i (з кроком h) і позначається Δy_i або $\Delta f(x_i)$, тобто $\Delta y_i = y_{i+1} - y_i$. Скінченна різниця другого порядку в точці x_i визначається рівностями

$$\Delta^2 y_i = \Delta(\Delta y_i) = \Delta y_{i+1} - \Delta y_i = (y_{i+2} - y_{i+1}) - (y_{i+1} - y_i) = y_{i+2} - 2y_{i+1} + y_i.$$

Взагалі, скінченна різниця n-го порядку функції $y = f(x)$ в точці x_i визначається за такою рекурентною формулою:

$$\Delta^n y_i = \Delta(\Delta^{n-1} y_i) = \Delta^{n-1} y_{i+1} - \Delta^{n-1} y_i,$$

де

$$n \geqslant 1, \quad \Delta^0 y_i = y_i, \quad \Delta^1 y_i = \Delta y_i.$$

Скінченні різниці зручно розміщувати у вигляді таблиці.

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	...
x_0	y_0	Δy_0	$\Delta^2 y_0$	$\Delta^3 y_0$	
$x_1 = x_0 + h$	y_1	Δy_1	$\Delta^2 y_1$	$\Delta^3 y_1$	
$x_2 = x_0 + 2h$	y_2	Δy_2	$\Delta^2 y_2$		
$x_3 = x_0 + 3h$	y_3	Δy_3			
$x_4 = x_0 + 4h$	y_4				
...	...	...	...	...	...

Така таблиця називається *діагональною таблицею різниць*. Всі різниці записуватимемо цілими числами в одиницях молодшого розряду значень функції у вузлах інтерполювання.

Скориставшись методом математичної індукції, легко довести формулу

$$\Delta^n y_i = \sum_{k=0}^{n} (-1)^k C_n^k y_{n+i-k}, \tag{26.1}$$

де C_n^k — біномні коефіцієнти. Остання формула виражає скінченні різниці через значення функції у вузлових точках.

Вкажемо тепер на зв'язок між різницями і похідними. *Якщо функція $f(x)$ має на відрізку $[x_i; x_{i+k}]$ неперервну похідну $f^{(k)}(x)$, то існує така точка $\xi \in]x_i; x_{i+k}[$, що*

$$\Delta^k y_i = f^{(k)}(\xi) h^k. \tag{26.2}$$

Формула (26.2) при $n = 1$ є формулою Лагранжа для скінченних приростів

$$\Delta y_i = f(x_{i+1}) - f(x_i) = f'(\xi)(x_{i+1} - x_i) = f'(\xi)\, h,$$

де $x_i < \xi < x_{i+1}$. Доведемо її при $n = 2$. Нехай $\varphi(x) = f(x + h) - f(x)$. Тоді $\Delta^2 y_i = \varphi(x_i + h) - \varphi(x_i)$ і за формулою Лагранжа маємо:

$$\Delta^2 y_i = h\varphi'(\xi_1),$$

де $\xi_1 \in]x_i;\ x_i + h[$ — деяка точка, $\varphi'(\xi_1) = f'(\xi_1 + h) - f'(\xi_1)$. Застосовуючи ще раз формулу Лагранжа до функції f', дістанемо:

$$\Delta^2 y_i = h(f'(\xi_1 + h) - f'(\xi_1)) = h^2 f''(\xi),$$

де $\xi \in]\xi_1;\ \xi_1 + h[\subset]x_i;\ x_{i+2}[$. Для $n > 2$ твердження доводиться аналогічно.

Звідси випливає такий наслідок.

Скінченна різниця n-го порядку алгебраїчного многочлена n-го степеня постійна, тобто не залежить від i, а скінченні різниці вищих порядків дорівнюють нулеві.

Зауважимо, що звідси не випливає: якщо при деякому h скінченні різниці n-го порядку, функції $f(x)$ сталі, то $f(x)$ буде многочленом n-го степеня. У цьому разі справедливе таке твердження. *Якщо різниці n-го порядку функції $f(x)$ сталі для будь-якого кроку h, то ця функція є многочленом n-го степеня.*

Зауважимо, що сформульовані властивості скінченних різниць стосуються точних різниць функції. Проте значення функції, як правило, в таблиці подаються наближено з деякою точністю і при обчисленні різниць похибка зростає. У таблицях значення функції в усіх точках обчислюють з однаковою абсолютною похибкою і подають лише правильні цифри значень функції, тобто похибка кожного значення не перевищує 0,5 одиниці останнього збереженого розряду табличних значень функції. Тоді похибка перших різниць дорівнює одиниці молодшого розряду (0,5 + 0,5 = 1), похибка других різниць — двом одиницям молодшого розряду (1 + 1 = = 2), третіх — чотирьом одиницям (2 + 2 = 4) і т. д. Отже, похибка k-тих різниць дорівнює 2^{k-1} одиниць останнього розряду.

Таким чином, можна сформулювати означення:

Різниці k-го порядку на деякій ділянці таблиці називаються практично сталими, якщо всі вони відрізняються між собою не більш як на 2^k одиниць останнього розряду табличних значень функції.

Якщо різниці k-го порядку практично сталі, то різниці $k + 1$)-го порядку обчислювати немає потреби, бо вони складатимуться лише з неправильних цифр, і в межах даної точності можна вважати, що вони дорівнюють нулю.

Якщо на деякій ділянці таблиці із сталим кроком різниці порядку k практично сталі, то можна вважати, що функція в межах даної точності поводиться як многочлен k-го степеня. Тому, коли знаходять проміжні значення функції на цій ділянці, обмежуються обчисленням інтерполяційних многочленів k-го степеня.

З формули (26.2) випливає, що коли похідна k-го порядку функції змінюється мало, то зменшення кроку в r раз призведе до зменшення різниці k-го порядку в r^k раз, а отже при заданій точності таблиці крок можна вибрати таким, щоб скінченні різниці порядку k були практично сталі.

§ 27. Інтерполяційні многочлени Ньютона

Нехай для функції $y = f(x)$ задано її значення $y_i = f(x_i)$ для значень аргументу, які утворюють арифметичну прогресію $x_i = x_0 + ih$ $(i = 0, 1, \ldots, n)$, де h — крок таблиці. Треба побудувати многочлен $P_n(x)$, степінь якого був би не більшим за n, а значення його у вузлах інтерполювання збігалися б із значенням функції $y = f(x)$, тобто

$$P_n(x_i) = y_i \quad (i = 0, 1, 2, \ldots, n). \tag{27.1}$$

Многочлен $P_n(x)$ визначатимемо у вигляді

$$P_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + a_3(x - x_0)(x - x_1)(x - x_2) + \cdots + a_n(x - x_0)(x - x_1) \cdots (x - x_{n-1}). \tag{27.2}$$

Коефіцієнти $a_0, a_1, \ldots, a_n$ визначимо так, щоб виконувалися умови (27.1). Для цього підставимо в (27.2) замість x значення $x_0, x_1, x_2, \ldots, x_n$.

Покладемо $x = x_0$. Тоді $P_n(x_0) = y_0 = a_0$, або $a_0 = y_0$.

$$P_n(x_1) = y_1 = a_0 + a_1(x_1 - x_0) = y_0 + a_1 h,$$

звідси

$$a_1 = \frac{y_1 - y_0}{h} = \frac{\Delta y_0}{h}.$$

Аналогічно знайдемо:

$$y_2 = P_n(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) = y_0 + \frac{y_1 - y_0}{h} 2h + a_2 \cdot 2h \cdot h.$$

Тоді

$$a_2 = \frac{y_2 - 2y_1 + y_0}{2h^2}.$$

Скориставшись формулою (26.1), яка виражає різниці через значення функції, дістанемо:

$$a_2 = \frac{\Delta^2 y_0}{2h^2}.$$

Продовжуючи процес, знайдемо:

$$a_3 = \frac{\Delta^3 y_0}{3!\,h^3}, \quad a_4 = \frac{\Delta^4 y_0}{4!\,h^4}, \quad \ldots, \quad a_n = \frac{\Delta^n y_0}{n!\,h^n}.$$

Підставивши значення a_0, a_1, ..., a_n у формулу (27.2), матимемо:

$$P_n(x) = y_0 + \frac{\Delta y_0}{1!\,h}(x - x_0) + \frac{\Delta^2 y_0}{2!\,h^2}(x - x_0)(x - x_1) + \cdots$$

$$\cdots + \frac{\Delta^n y_0}{n!\,h^n}(x - x_0)(x - x_1)(x - x_2) \ \ldots \ (x - x_{n-1}). \quad (27.3)$$

Формула (27.3) називається *першою інтерполяційною формулою Ньютона*. У § 23 ми довели, що існує лише один інтерполяційний многочлен n-го степеня, значення якого у вузлах інтерполяції x_0, x_1, ..., x_n дорівнюють значенням функції y_0, y_1, ..., y_n.

Тому інтерполяційний поліном Ньютона (27.3) збігатиметься з інтерполяційним многочленом Лагранжа, побудованим для цього випадку. Ці многочлени тільки записані в різних формах.

Як бачимо, у першій інтерполяційній формулі Ньютона використовують скінченні різниці, розміщені в діагональній таблиці різниць по діагоналі зверху вниз.

Приклади.

1. Побудувати 1-й інтерполяційний многочлен Ньютона за такими даними:

x	0	1	2	3
$y = f(x)$	1	3	2	−1

Складемо таблицю різниць функції

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
0	1			
		2		
1	3		−3	
		−1		1
2	2		−2	
		−3		
3	−1			

У цьому прикладі $h = 1$ і $n = 3$. За формулою (27.3) маємо:

$$P_3(x) = 1 + \frac{2}{1!\,1}(x - 0) - \frac{3}{2!\,1^2}(x - 0)(x - 1) +$$

$$+ \frac{1}{3!\,1^3}(x - 0)(x - 1)(x - 2).$$

Записавши многочлен за степенями x, дістанемо:

$$P_3(x) = \frac{1}{6}x^3 - 2x^2 + \frac{23}{6}x + 1.$$

Якщо за цими даними побудувати інтерполяційний многочлен Лагранжа, то, записавши його за степенями x, переконаємось, що він тотожно збігається з многочленом Ньютона.

У випадку рівновіддалених вузлів многочлен Ньютона зручніший для обчислень, ніж многочлен Лагранжа. На відміну від інтерполяційного многочлена Лагранжа, де новий вузол потребує повного переобчислення многочлена, в інтерполяційному многочлені Ньютона із збільшенням кількості вузлів зростає лише кількість нових доданків, а всі обчислення, виконані раніше, залишаються без змін. Крім того, формула Ньютона дає змогу досить просто оцінювати точність результату, добутого інтерполюванням.

Подамо формулу Ньютона (27.3) в зручнішому для користування вигляді. Для цього введемо нову безрозмірну змінну t за формулою $t = \frac{x - x_0}{h}$, або $x = ht + x_0$. Тоді формула (27.3) набирає вигляду

$$\bar{P}_n(t) = P_n(x_0 + ht) = y_0 + t\Delta y_0 + \frac{t(t-1)}{2!}\Delta^2 y_0 +$$

$$+ \frac{t(t-1)(t-2)}{3!}\Delta^3 y_0 + \cdots$$

$$\cdots + \frac{t(t-1)(t-2)\ldots(t-n+1)}{n!}\Delta^n y_0. \quad (27.4)$$

Це остаточний вигляд першої інтерполяційної формули Ньютона. Формулу (27.4) називають ще *формулою Ньютона для інтерполювання вперед*. Таку назву вона дістала тому, що в ній використовуються значення функції, дані у вузлах, які містяться вправо від x_0 (вперед, вниз по стовпчику).

Формулою (27.4) можна користуватися для інтерполювання в таблиці $y = f(x)$, якщо $x \in]x_0; x_n[$; $0 < t < n$. Проте доцільно користуватися цією формулою тоді, коли $0 < t < 1$, бо коли $f(x)$ не є многочленом степеня, не вищого за n, то наближена рівність $P_n(x) \approx f(x)$ буде тим точнішою, чим менші h і t. Оскільки за x_0 можна взяти будь-яке значення аргументу x_i, то при $x \in]x_k; x_{k+1}[$ за x_0 беруть x_k, а за $x_1 - x_{k+1}$. Тоді $0 < t < 1$.

Якщо знаходити вигляд інтерполяційного многочлена потреби немає, а потрібно тільки обчислити його значення в деякій точці $t = \frac{x - x_0}{h}$, то обчислення зручно виконувати за схемою, подібною до схеми Горнера. Наприклад, для многочлена Ньютона 5-го степеня матимемо:

$$\bar{P}_5(t) = y_0 + t\left(\Delta y_0 + \frac{t-1}{2}\left(\Delta^2 y_0 + \frac{t-2}{3}\left(\Delta^3 y_0 + \right.\right.\right.$$

$$\left.\left.\left. + \frac{t-3}{4}\left(\Delta^4 y_0 + \frac{t-4}{5}\Delta^5 y_0\right)\right)\right)\right).$$

Звідси, послідовно обчислюючи

$$S_0 = \Delta^5 y_0, \quad S_1 = S_0\frac{t-4}{5} + \Delta^4 y_0,$$

$$S_2 = S_1\frac{t-3}{4} + \Delta^3 y_0, \quad S_3 = S_2\frac{t-2}{3} + \Delta^2 y_0,$$

$$S_4 = S_3\frac{t-1}{2} + \Delta y_0, \quad S_5 = S_4 t + y_0,$$

знайдемо: $S_5 = \bar{P}_5(t)$.

2. За даними значеннями $\sin x$ у точках $x = 10°$, $12°$, $14°$, $16°$, $18°$ знайти значення $\sin 10° 20'$ і $\sin 12° 20'$.

x	$10°$	$12°$	$14°$	$16°$	$18°$
$\sin x$	0,17365	0,20791	0,24192	0,27564	0,30902

В обчисленнях використовуємо перший інтерполяційний многочлен Ньютона. Спочатку складемо таблицю різниць функції.

Таблиця 14

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
10°	0,17365	3426		
12°	0,20791	3401	—25	—4
14°	0,24192	3372	—29	—5
16°	0,27564	3338	—34	
18°	0,30902			

Треті різниці практично сталі, тому інтерполюватимемо многочленом третього степеня. Для обчислення $\sin 10° 20'$ за x_0 беремо $10°$. Крок $h = 2°$, $t = \frac{10°20' - 10°}{2°} = \frac{1}{6}$. Проміжні значення знайдемо з однією запасною цифрою, яку в остаточному результаті округлимо:

$$P_3(10°20') = 0{,}17365 + \frac{1}{6}\cdot 0{,}03426 + \frac{\frac{1}{6}\left(\frac{1}{6}-1\right)}{2}(-0{,}00025) +$$

$$+ \frac{\frac{1}{6}\left(\frac{1}{6}-1\right)\left(\frac{1}{6}-2\right)}{6}(-0{,}00004) = 0{,}179374.$$

Отже, $\sin 10° 20' \approx 0{,}17937$.

Значення $\sin 10° 20'$ з п'ятьма правильними значущими цифрами дорівнює 0,17937. Отже, в результаті всі цифри правильні.

Тепер обчислимо $\sin 12° 20'$.

Оскільки x лежить у проміжку $]12°; 14°[$, то покладемо $x_0 = 12°$. Тоді $y_0 = 0{,}20791$, $\Delta y_0 = 0{,}03401$, $\Delta^2 y_0 = -0{,}00029$, $\Delta^3 y_0 = -0{,}00005$, $t = \frac{12°\,20' - 12°}{2°} = \frac{1}{6} = 0{,}166667$ (різниці в таблиці підкреслено хвилястою лінією).

$$P_3(12°\,20') = 0{,}20791 + \frac{1}{6}\,0{,}03401 +$$

$$+ \frac{1}{2}\cdot\frac{1}{6}\left(\frac{1}{6}-1\right)(-0{,}00029) +$$

$$+ \frac{1}{6}\cdot\frac{1}{6}\left(\frac{1}{6}-1\right)\left(\frac{1}{6}-2\right)(-0{,}00005) = 0{,}213598.$$

Отже, $\sin 12° 20' \approx 0{,}21360$.

І цього разу всі знайдені цифри правильні.

Оскільки у першій інтерполяційній формулі Ньютона використовуються скінченні різниці, які йдуть по діагоналі вниз,

то цією формулою користуються переважно для значень x, близьких до початкової точки таблиці. Для значень x, близьких до кінцевої точки таблиці, формула непридатна, бо там не вистачає потрібної кількості різниць. Наприклад, у табл. 14 для $x_0 = 14°$ є тільки перша і друга різниці, а для $x_0 = 16°$ — тільки перша.

Як уже зазначалося, важливими випадками інтерполювання є лінійне і квадратичне інтерполювання. Формули лінійного і квадратичного інтерполювання можна вивести як окремі випадки першої інтерполяційної формули Ньютона. Поклавши у формулі (27.4) $n = 1$, дістанемо формулу лінійного інтерполювання:

$$\bar{P}_1(t) = P_1(x_0 + ht) = y_0 + t\Delta y_0,$$

де $t = \frac{x - x_0}{h}$.

Якщо $n = 2$, дістанемо формулу параболічної (квадратичної) інтерполяції:

$$\bar{P}_2(t) = P_2(x_0 + ht) = y_0 + t\Delta y_0 + \frac{t(t-1)}{2}\Delta^2 y_0. \qquad (27.5)$$

Із сказаного випливає, що квадратичне інтерполювання слід застосовувати тоді, коли другі різниці функції практично сталі.

Для інтерполювання в кінці таблиці користуються іншою формулою. Виведемо цю формулу.

Запишемо шуканий інтерполяційний многочлен у вигляді

$$P_n(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1}) + \cdots$$
$$\cdots + a_n(x - x_n)(x - x_{n-1}) \ldots (x - x_1). \qquad (27.6)$$

Як і раніше, коефіцієнти $a_0, a_1, a_2, \ldots, a_n$ визначимо так, щоб

$$P_n(x_i) = y_i \quad (i = 0, 1, 2, \ldots, n).$$

Підставивши послідовно у формулу (27.6) замість x значення $x_n, x_{n-1}, \ldots, x_0$, знайдемо:

$$a_0 = y_n, \quad a_1 = \frac{\Delta y_{n-1}}{h}, \quad a_2 = \frac{\Delta^2 y_{n-2}}{2!\,h^2}, \ldots, a_n = \frac{\Delta^n y_0}{n!\,h^n}.$$

Отже, формула (27.6) набирає вигляду:

$$P_n(x) = y_n + \frac{\Delta y_{n-1}}{1!\,h}(x - x_n) + \frac{\Delta^2 y_{n-2}}{2!\,h^2}(x - x_n)(x - x_{n-1}) + \cdots$$
$$\cdots + \frac{\Delta^n y_0}{n!\,h^n}(x - x_n)(x - x_{n-1}) \ldots (x - x_1). \qquad (27.7)$$

Ця формула називається *другою інтерполяційною формулою Ньютона,* або *формулою Ньютона для інтерполювання назад.*

Запишемо останню формулу у зручному для користування вигляді, аналогічно першій інтерполяційній формулі Ньютона. Поклавши $t = \frac{x - x_n}{h}$, дістанемо:

$$\bar{P}_n(t) = P_n(x_n + ht) = y_n + t\Delta y_{n-1} + \frac{t(t+1)}{2!}\Delta^2 y_{n-2} +$$
$$+ \frac{t(t+1)(t+2)}{3!}\Delta^3 y_{n-3} + \cdots$$
$$\cdots + \frac{t(t+1)(t+2)\ldots(t+n-1)}{n!}\Delta^n y_0. \qquad (27.8)$$

Це і є звичайний вигляд другої інтерполяційної формули Ньютона для інтерполювання назад. У ній використовуються скінченні різниці, розміщені в діагональній таблиці різниць по діагоналі знизу вгору.

3. Побудувати другу інтерполяційну формулу Ньютона за даними прикладу 1.

У цьому випадку $y_3 = -1$, $\Delta y_2 = -3$, $\Delta^2 y_1 = -2$, $\Delta^3 y_0 = 1$. Отже,

$$\bar{P}_3(t) = -1 + t(-3) + \frac{t(t+1)}{2}(-2) + \frac{t(t+1)(t+2)}{3!}\,1 =$$
$$= -1 - 3t - t(t+1) + \frac{1}{6}t(t+1)(t+2), \text{ де } t = \frac{x-3}{1} =$$
$$= x - 3.$$

Оскільки інтерполяційні формули Лагранжа і Ньютона — різні форми запису інтерполяційного многочлена, то оцінки залишкових членів формул Ньютона будуть такими, як і для формули Лагранжа, побудованої за такими самими даними.

На практиці інтерполяційні многочлени Ньютона обриваються на членах, які містять практично сталі різниці, а точка x_0 чи x_n добирається так, щоб $|t| < 1$. При цьому результат інтерполювання, як правило, має стільки правильних цифр, скільки їх в табличних значеннях функції. Тому в таких випадках немає потреби обчислювати залишкові члени інтерполяційних формул. Оцінку залишкового члена знаходять тільки тоді, коли різниці не стають практично сталими або коли незручно користуватися різницями вище деякого порядку.

Як і у випадку інтерполяційної формули Лагранжа, у повну похибку результату, знайденого за формулами Ньютона, крім похибки методу, входитиме неусувна похибка, а також похибка округлення.

Повну похибку можна знайти так, як вона визначалась у § 25.

Формулу Ньютона, як і формулу Лагранжа, можна використати і для екстраполювання, тобто для обчислення значення функції в точках, які лежать за межами таблиці.

Якщо $x < x_0$ і x близьке до x_0, то доцільно застосовувати першу інтерполяційну формулу Ньютона, якщо x близьке до x_n і $x > x_n$, то зручно скористатися другою інтерполяційною формулою Ньютона.

Отже, перша інтерполяційна формула Ньютона застосовується для інтерполювання вперед ($t > 0$) і екстраполювання назад ($t < 0$), а друга — для інтерполювання назад ($t < 0$) та екстраполювання вперед ($t > 0$).

4. За даними значеннями $\sin x$ у точках $x = 10°$, $12°$, $14°$, $16°$, $18°$ знайти значення $\sin 9° 30'$ і $\sin 18° 30'$.

Складемо таблицю різниць (табл. 14). Треті різниці практично сталі, тому екстраполюватимемо за допомогою многочлена третього степеня.

Обчислимо $\sin 9° 30'$ за першою інтерполяційною формулою Ньютона, поклавши $x_0 = 10°$, $x = 9° 30'$, $t = \frac{9° 30' - 10°}{2°} = -0{,}25$.

$$P_3(9° 30') = 0{,}17365 + (-0{,}25) \cdot 0{,}03426 + \frac{(-0{,}25)(-0{,}25-1)}{2}(0{,}00025) + \frac{(-0{,}25)(-0{,}25-1)(-0{,}25-2)}{6}(-0{,}00004) = 0{,}16505.$$

Отже, $\sin 9° 30' \approx 0{,}16505$. Всі цифри результату правильні.

Щоб знайти $\sin 18° 30'$, покладемо $x_n = 18°$ і $x = 18° 30'$. Звідси

$$t = \frac{18° 30' - 18°}{2°} = 0{,}25.$$

Обчислення виконаємо за другою інтерполяційною формулою Ньютона:

$$P_3(18° 30') = 0{,}30902 + 0{,}25 \cdot 0{,}03338 + \frac{0{,}25 \cdot (0{,}25+1)}{2}(-0{,}00034) + \frac{0{,}25\,(0{,}25+1)\,(0{,}25+2)}{6}(-0{,}00005) = 0{,}31731.$$

За шестизначними таблицями $\sin 18° 30' = 0{,}317305$.

§ 28. Застосування інтерполювання. Інтерполювання під час роботи з таблицями. Обернене інтерполювання

Таблиці функцій — один з найпоширеніших і найважливіших допоміжних засобів обчислень, які значно полегшують роботу обчислювача і широко використовуються на практиці. Таблиці можна складати для функції одного, двох і більше аргументів. Тому вони бувають з одним, двома і більше входами. Ми розглядатимемо таблиці функцій, що залежать від одного аргументу, тобто таблиці з одним входом.

У таблиці деякої функції $y = f(x)$ подають значення цієї функції для значення аргументу x_i, які, як правило, пов'язані між собою співвідношенням $x_{i+1} = x_i + h$.

При складанні таблиці додержують таких правил: 1) значення функції $y = f(x)$ у табличних точках обчислюють з однаковою абсолютною похибкою; 2) у таблицях даються лише правильні цифри значень функції, тобто похибка кожного значення не перевищує 0,5 одиниці молодшого розряду табличних значень функції.

Одним із застосувань загальної теорії інтерполювання функцій є використання інтерполяційних формул у процесі роботи з таблицями. Значення функції для значень аргументу, не наведених у таблиці, визначаються інтерполюванням. Звичайно, немає потреби будувати один інтерполяційний многочлен для всієї таблиці. Для кожної ділянки в разі потреби будують свій інтерполяційний многочлен якомога нижчого степеня. Степінь інтерполяційного многочлена добирають так, щоб похибка інтерполяції на даній ділянці була меншою, ніж похибка таблиці, тобто меншою від половини одиниці молодшого розряду табличних значень функції. Як уже зазначалося, степінь інтерполяційного многочлена має бути не вищим від порядку практично сталих різниць.

Вважають, що *таблиця (на деякій ділянці) допускає інтерполювання порядку k (лінійне, квадратичне і т. д.), якщо похибка цього інтерполювання не перевищує похибки таблиці.*

Отже, якщо таблиця допускає інтерполювання деякого порядку, то за його допомогою можна обчислити в нетабличних точках значення функції з тією самою точністю, що й табличні значення. Порядок інтерполювання, яке допускає таблиця, залежить від її призначення. Якщо таблиця використовується при ручних розрахунках, то вона повинна допускати найпростіші способи інтерполювання (лінійне або квадратичне). При розрахунках на ЕОМ зручними є таблиці з великим кроком,

хоча вони й потребують складних способів інтерполювання (інтерполяційних многочленів вищих порядків). Таблиці з дрібним кроком займатимуть багато місця в пам'яті ЕОМ, на пошук потрібного значення в таблиці витрачатиметься більше часу, ніж на обчислення цього значення.

Враховуючи, що при роботі з таблицями найчастіше застосовують лінійну і квадратичну інтерполяцію, розглянемо це питання докладніше.

Виведемо формулу для похибки лінійної інтерполяції, вважаючи, що вузли в таблиці рівновіддалені, а саме:

$$x_{i+1} - x_i = h \quad (i = 0, 1, 2, \ldots).$$

З оцінки (25.4), якщо $n = 1$, дістанемо:

$$|R_1(x)| \leqslant \frac{h^2 M_2}{2} |t(t-1)|, \tag{28.1}$$

де $$M_2 = \max_{x \in [x_0; x_1]} |f''(x)|, \quad t = \frac{x - x_0}{h}.$$

З (28.1) випливає, що абсолютна похибка Δ методу лінійного інтерполювання для будь-якого $x \in [x_0; x_1]$ задовольняє нерівність

$$\Delta \leqslant \frac{h^2 M_2}{2} \max_{0 \leqslant t \leqslant 1} |t(t-1)| = \frac{h^2 M_2}{8}.$$

Очевидно, для будь-якого з відрізків $[x_i; x_{i+1}]$ оцінка похибки аналогічна.

Звідси робимо висновок:

Якщо $\frac{h^2 M_2}{8}$ (*де* $M_2 = \max_{x \in [a;b]} |f''(x)|$, $a = \min(x_0, x_1, \ldots, x_n)$, $b = \max(x_0, \ldots, x_n)$) *не перевищує половини одиниці молодшого розряду табличних значень функції, то лінійна інтерполяція допустима.*

Якщо друга похідна $f''(x)$ змінюється мало на $[x_i; x_{i+2}]$, то за формулою (26.2) наближено можна вважати

$$|\Delta^2 y_i| = M_2 h^2.$$

Тоді для максимальної похибки лінійної інтерполяції дістанемо таку наближену оцінку:

$$\Delta \leqslant \frac{|\Delta^2 y_i|}{8}.$$

Тому для допустимості лінійної інтерполяції потрібно, щоб $\frac{|\Delta^2 y_i|}{8}$ не перевищувала $\frac{1}{2}$ одиниці молодшого розряду табличних значень функцій.

Отже, *якщо модулі других різниць функції на заданому відрізку не перевищують чотирьох одиниць молодшого розряду*

табличних значень функції, то на цьому відрізку таблиці лінійна інтерполяція допустима.

Виведемо тепер просту формулу для максимальної похибки, яку дістанемо при квадратичному інтерполюванні за таблицями із сталим кроком. Нагадаємо, що квадратичне інтерполювання полягає у наближеній заміні функції $y = f(x)$ на деякому відрізку $[x_0; x_2]$ многочленом другого степеня, який у вузлах x_0, x_1, x_2 набуває значень, що дорівнюють значенням функції $y = f(x)$.

З оцінки (25.4), якщо $n = 2$, дістанемо:

$$|R_2(x)| \leqslant \frac{h^3 M_3}{3!} |t(t-1)(t-2)|, \tag{28.2}$$

де

$$M_3 = \max_{x \in [x_0; x_2]} |f'''(x)|, \quad t = \frac{x - x_0}{h}.$$

Абсолютна похибка Δ методу квадратичного інтерполювання для будь-якого x з відрізка $[x_0; x_2]$ за (28.2) задовольняє нерівність

$$\Delta \leqslant \frac{h^3 M_3}{3!} \max_{0 \leqslant t \leqslant 2} |t(t-1)(t-2)|.$$

Виконавши обчислення, знайдемо:

$$\max_{0 \leqslant t \leqslant 2} |t(t-1)(t-2)| = \frac{2\sqrt{3}}{9}.$$

Отже, формула для максимальної похибки квадратичної інтерполяції матиме вигляд:

$$\Delta \leqslant \frac{h^3 M_3}{3!} \cdot \frac{2\sqrt{3}}{9} = h^3 M_3 \frac{\sqrt{3}}{27}.$$

Оскільки $\frac{\sqrt{3}}{27} < \frac{1}{15}$, то користуватимемося оцінкою

$$\Delta < \frac{1}{15} M_3 h^3.$$

Звідси робимо висновок:

Якщо $\frac{1}{15} M_3 h^3$ *на заданій ділянці таблиці не перевищує половини одиниці молодшого розряду табличного значення функції, то на цій ділянці квадратичне інтерполювання допустиме.*

Якщо скористатися наближеною рівністю $M_3 h^3 = |\Delta^3 y_0|$ (коли $f'''(x)$ на $[x_0; x_3]$ змінюється мало і h мале), то для допустимості квадратичного інтерполювання треба, щоб величина $\frac{1}{15} |\Delta^3 y_0|$ не перевищувала половини одиниці молод-

шого розряду табличних значень функції, тобто щоб $|\Delta^3 y_0|$ не перевищувала семи одиниць молодшого розряду табличних значень функції.

Отже, *якщо модулі третіх табличних різниць значень функції не перевищують семи одиниць молодшого розряду табличних значень функції, то квадратична інтерполяція допустима.*

Приклад.

Дано таблицю функції:

x	3,60	3,65	3,70	3,75	3,80	3,85	3,90	3,95	4,00
$y = e^x$	36,598	38,475	40,447	42,521	44,701	46,993	49,402	51,935	54,598

Чи допускає ця таблиця квадратичне інтерполювання? Щоб відповісти на запитання, обчислимо величину $\frac{1}{15} M_3 h^3$. Тут $h = 0{,}05$, для кожної ділянки цієї таблиці $[x_i;\ x_{i+2}]$ величина $M_3 = \max\limits_{x_i \leqslant x \leqslant x_{i+2}} |e^x|$ не перевищуватиме 54,598. Тому покладемо:

$$M_3 = 54{,}598. \quad \text{Тоді } \frac{1}{15} M_3 h^3 = \frac{54{,}598 \cdot (0{,}05)^3}{15} < 0{,}00046.$$

Оскільки похибка квадратичного інтерполювання $\Delta < 0{,}00046$ менша від похибки табличних значень $(0{,}5 \cdot 10^{-3})$, то наведена таблиця допускає квадратичне інтерполювання. Лінійного інтерполювання таблиця не допускає, оскільки

$$\frac{M_2 h^2}{8} = \frac{54{,}598\,(0{,}05)^2}{8} = 0{,}0171 > 0{,}5 \cdot 10^{-3}.$$

Складемо таблицю різниць цієї функції.

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
3,60	36,598			
		1877		
3,65	38,475		95	
		1972		7
3,70	40,447		102	
		2074		4
3,75	42,521		106	
		2180		6
3,80	44,701		112	
		2292		5
3,85	46,993		117	
		2409		7
3,90	49,402		124	
		2533		6
3,95	51,935		130	
		2663		
4,00	54,598			

Другі сусідні табличні різниці відрізняються між собою не більш як на 7 одиниць молодшого розряду, тому і за другим правилом таблиця допускає квадратичне інтерполювання.

Оскільки треті різниці практично сталі, то, звичайно, для інтерполювання можна було б користуватися многочленом третього степеня. Проте многочлен другого степеня також забезпечує необхідну точність.

За формулою квадратичного інтерполювання визначимо, наприклад, $e^{3{,}73}$. Покладемо $x_0 = 3{,}70$; $x_1 = 3{,}75$; $x_2 = 3{,}80$. Тоді $t = 0{,}6$. Підставивши дані в формулу квадратичної інтерполяції (27.5), дістанемо:

$$P_2(3{,}73) = 40{,}447 + 0{,}6 \cdot 2{,}074 + \frac{0{,}6 \cdot (0{,}6 - 1)}{2} 0{,}106 =$$

$$= 41{,}6787 \approx 41{,}679.$$

Отже, $e^{3{,}73} \approx 41{,}679$. У знайденому результаті всі цифри правильні.

Розглянемо приклад на складання таблиці функції.

Скласти таблицю значень функції

$$y = \frac{e^{-x^2} + x + 1}{\sqrt{\pi} \cdot x^2 + e^{-x^2}}$$

на відрізку $[0{,}2;\ 1{,}0]$ з точністю до $0{,}5 \cdot 10^{-3}$, причому так, щоб таблиця допускала лінійну інтерполяцію.

Таблиця 15

x	x^2	$\sqrt{\pi x^2}$	e^{-x^2}	$\sqrt{\pi x^2} + e^{-x^2}$	$e^{-x^2} + x + 1$	y	Δy	$\Delta^2 y$	$\Delta^3 y$
0,2	0,04	0,07090	0,96079	1,03169	2,16079	2,094			
							—032		
0,3	0,09	0,15952	0,91393	1,07345	2,21393	2,062		—047	
							—079		+008
0,4	0,16	0,28359	0,85214	1,13574	2,25214	1,983		—039	
							—118		+012
0,5	0,25	0,44311	0,77880	1,22191	2,27880	1,865		—027	
							—145		+013
0,6	0,36	0,63808	0,69768	1,33576	2,29768	1,720		—014	
							—159		+013
0,7	0,49	0,86850	0,61263	1,48113	2,31263	1,561		—001	
							—160		+007
0,8	0,64	0,13437	0,52729	1,66166	2,32729	1,401		+006	
							—154		+007
0,9	0,81	1,43569	0,44486	1,88055	2,34486	1,247		+013	
							—141		
1,0	1,00	1,77245	0,36788	2,14033	2,36788	1,106			

Крок таблиці нам невідомий, визначити його з умови $\frac{h^2 M_2}{8} \leqslant 0{,}5 \cdot 10^{-3}$ важко, оскільки важко знайти $\max |f''(x)|$. Тому ми спочатку складемо таблицю значень даної функції з точністю до $0{,}5 \cdot 10^{-3}$ з кроком $h = 0{,}1$ (табл. 15).

Якщо побудована таблиця не допускатиме лінійної інтерполяції (модулі других різниць більші за $4 \cdot 10^{-3}$), то відповідним чином зменшимо крок так, щоб таблиця допускала лінійну інтерполяцію.

Перед тим як обчислювати, визначимо, з якою точністю треба брати x, x^2, $\sqrt{\pi}$, e^{-x^2}, щоб мати y із заданою точністю. Таким чином, треба розв'язати обернену задачу теорії похибок. Позначивши $r = e^{-x^2} + x + 1$, $s = \sqrt{\pi}x^2 + e^{-x^2}$, дістанемо: $y = \frac{r}{s}$.

Тоді, як відомо,

$$\Delta_y = \frac{\Delta_r}{|s|} + \Delta_s \frac{|r|}{s^2}.$$

Неважко встановити, що функції r і s на [0,2; 1,0] монотонно зростають (їх похідні додатні). Тому на заданому відрізку справедливі оцінки $1 < r < 3$, $1 < s$.

Таким чином, справджується нерівність

$$\Delta_y = \frac{\Delta_r}{|s|} + \Delta_s \frac{|r|}{s^2} < \Delta_r + 3 \cdot \Delta_s.$$

Щоб гарантувати в табличних значеннях функції після округлення точність $0{,}5 \cdot 10^{-3}$, похибки Δ_r і Δ_s потрібно вибрати так, щоб $\Delta_y \leqslant 0{,}5 \cdot 10^{-4}$ (див. примітку § 2).

Для цього достатньо, щоб $\Delta_r \leqslant 0{,}00002$; $\Delta_s \leqslant 0{,}00001$.

Отже, якщо похибки значень e^{-x^2} і $\sqrt{\pi}$ не перевищуватимуть 0,000005, то при точних значеннях x і x^2 $\Delta_s \leqslant 0{,}00001$; $\Delta_r \leqslant 0{,}000005$. У таблиці з кроком $h = 0{,}1$ ми маємо змогу брати точні значення x і x^2.

Якщо значення x і x^2 наближені, то

$$\Delta_r = \Delta_{e^{-x^2}} + \Delta_x.$$

Аналогічно $\Delta_s = \sqrt{\pi} \cdot \Delta_{x^2} + x^2 \Delta_{\sqrt{\pi}} + \Delta_{e^{-x^2}}$.

Врахувавши значення, яких можуть набувати величини x, e^{-x^2}, $\sqrt{\pi}$, знайдемо, що в цьому випадку треба брати: $\Delta_{x^2} \leqslant 0{,}0000005$, $\Delta_{e^{-x^2}} \leqslant 0{,}0000005$, $\Delta_{\sqrt{\pi}} \leqslant 0{,}0000005$. Тоді $\Delta_r \leqslant 0{,}000001$, $\Delta_s \leqslant 0{,}000002$ і похибки значень функції не перевищуватимуть $0{,}5 \cdot 10^{-4}$, оскільки $\Delta_s \leqslant 2 \cdot 10^{-5}$, $\Delta_r \leqslant 1 \times 10^{-5}$.

Отже, в проміжних обчисленнях у даному випадку треба брати три запасні цифри, якщо значення x і x^2 наближені, і дві запасні цифри, якщо значення x і x^2 точні.

Для обчислення значень e^{-x^2} з потрібною точністю скористаємося десятизначними таблицями e^x і e^{-x}.

Врахувавши, що $\sqrt{\pi} \approx 1{,}772454$, складаємо таблицю значень функції з кроком $h = 0{,}1$. Результати обчислень розміщуємо за схемою таблиці 15.

Щоб ця таблиця допускала лінійну інтерполяцію, треба на проміжку [0,20; 0,60] зменшити крок у 4 рази, а на проміжку [0,60; 1,00] — вдвічі. При цьому на проміжку [0,20; 0,60] другі різниці зменшаться приблизно в 16 раз, а на проміжку [0,60; 1,00] — у 4 рази. Після цього другі різниці не перевищуватимуть $4 \cdot 10^{-3}$, і таблиця значень функції, взятих з точністю до $0{,}5 \cdot 10^{-3}$, допускатиме лінійну інтерполяцію.

Якщо значення x^2 взяти наближено, то проміжні обчислення треба виконувати з точністю до 0,0000005. Складаємо таблицю.

Таблиця 16

x	x^2	$\sqrt{\pi}\,x^2$	e^{-x^2}	$\sqrt{\pi}x^2 + e^{-x^2}$	$e^{-x^2} + x + 1$	y	Δy	$\Delta^2 y$
0,200	0,040000	0,070898	0,960789	1,03169	2,16079	2,094		
							—003	
0,225	0,050625	0,089730	0,950635	1,04037	2,17564	2,091		—003
							—006	—004
0,250	0,062500	0,110783	0,939413	1,05019	2,18941	2,085		
							—010	—003
0,275	0,075625	0,134042	0,927164	1,06121	2,20216	2,075		
							—013	—002
0,300	0,090000	0,159521	0,913931	1,07345	2,21393	2,062		
							—015	—004
0,325	0,105625	0,187215	0,899762	1,08698	2,22476	2,047		
							—019	—002
0,350	0,122500	0,217126	0,884706	1,10183	2,23471	2,028		
							—021	—003
0,375	0,140625	0,249251	0,868815	1,11807	2,24382	2,007		
							—024	—002
0,400	0,160000	0,283593	0,852144	1,13574	2,25214	1,983		
							—026	—003
0,425	0,180625	0,320149	0,834748	1,15490	2,25975	1,957		
							—029	—002
0,450	0,202500	0,358992	0,816686	1,17561	2,26669	1,928		
							—031	—001
0,475	0,225625	0,399910	0,798017	1,19793	2,27302	1,897		
							—032	—002
0,500	0,250000	0,443113	0,778801	1,22191	2,27880	1,865		
							—034	—002
0,525	0,275625	0,488533	0,759097	1,24763	2,28410	1,831		
							—035	—001
0,550	0,302500	0,536167	0,738968	1,27514	2,28897	1,795		
							—037	—001
0,575	0,330625	0,586018	0,718475	1,30449	2,29348	1,758		
							—038	
0,600	0,360000	0,638083	0,697676	1,33576	2,29768	1,720		
							—078	—003
0,650	0,422500	0,748862	0,655406	1,40427	2,30541	1,642		
							—081	+001
0,700	0,490000	0,868502	0,612626	1,48113	2,31263	1,561		
							—080	+000
0,750	0,562500	0,997005	0,569783	1,56679	2,31978	1,481		
							—080	+001
0,800	0,640000	1,134370	0,527292	1,66166	2,32729	1,401		
							—079	+004
0,850	0,722500	1,280598	0,485537	1,76614	2,33554	1,332		
							—075	+003
0,900	0,810000	1,435688	0,444858	1,88055	2,34486	1,247		
							—072	+003
0,950	0,902500	1,599640	0,405555	2,00520	2,35556	1,175		
							—069	
1,000	1,000000	1,772454	0,367879	2,14033	2,36788	1,106		

Оскільки тепер другі різниці за модулем не перевищують чотирьох одиниць останнього розряду, то таблиця значень функції, взятих з точністю до $0{,}5 \cdot 10^{-3}$, допускає лінійну інтерполяцію.

Зауважимо, що, взявши в таблиці 16 значення x^2, $\sqrt{\pi}x^2$, e^{-x^2} з точністю до 0,000005, дістанемо такі самі остаточні результати (хоч значення $\sqrt{\pi}x^2 + e^{-x^2}$ і $e^{-x^2} + x + 1$ будуть дещо іншими). Це свідчить про те, що надто багато запасних правильних знаків у проміжних обчисленнях брати не треба.

Адже оцінки похибок результатів, як правило, перебільшені, тому може трапитись, що похибка знайденого результату набагато менша, ніж та, якої можна було сподіватися, і що в такому результаті може бути ще кілька правильних цифр після цифри, правильність якої гарантовна.

Обернене інтерполювання. Часто на практиці виникає така задача: Знайти значення аргументу, яке відповідає заданому значенню функції, що міститься між двома табличними значеннями. Така задача називається *задачею оберненого інтерполювання*.

Якщо задана функція $y = f(x)$ монотонна, то при оберненому інтерполюванні достатньо значення $y_i = f(x_i)$ вважати значеннями незалежної змінної, а значення x_i — значеннями функції $x_i = f^{-1}(y_i)$ і тоді інтерполювати, як і раніше. Якщо функція $f(x)$ неперервна і монотонна, то за теоремою про існування оберненої функції існує обернена $x = f^{-1}(y)$, яка буде також неперервною і монотонною. Оскільки значення нового аргументу y не рівновіддалені, для оберненого інтерполювання використовується інтерполяційний многочлен Лагранжа. Зауважимо, що для досягнення заданої точності пряме та обернене інтерполювання потребують, взагалі кажучи, різної кількості вузлів.

Якщо $f(x)$ не монотонна, то для неї записують інтерполяційний многочлен $P_n(x)$, і розв'язують рівняння $P_n(x) = y$ відносно x при заданому y.

Приклади.

1. Функція $y = e^x$ задана таблицею:

x	1,0	1,1	1,2	1,3	1,4	1,5
y	2,7183	3,0042	3,3201	3,6693	4,0552	4,4817

Знайти таке x, при якому $e^x = 3{,}5$, тобто знайти ln 3,5.

Як видно з таблиці, шукане значення аргументу міститься між 1,2 і 1,3. Щоб знайти його, використаємо значення функції лише в точках 1,1; 1,2; 1,3, тобто складемо інтерполяційний многочлен Лагранжа другого степеня. ln 3,5 знайдемо за алгоритмом обчислення значення многочлена Лагранжа, де треба лише поміняти місцями x і y. У даному прикладі $x_0 = 1{,}1$; $x_1 = 1{,}2$; $x_2 = 1{,}3$; $y_0 = 3{,}0042$; $y_1 = 3{,}3201$; $y_2 = 3{,}6693$; $y = 3{,}5$.

$$x = L_2(3{,}5) = \frac{(3{,}5 - 3{,}3201)(3{,}5 - 3{,}6693)}{(3{,}0042 - 3{,}3201)(3{,}0042 - 3{,}6693)} 1{,}1 +$$

$$+ \frac{(3{,}5 - 3{,}0042)(3{,}5 - 3{,}6693)}{(3{,}3201 - 3{,}0042)(3{,}3201 - 3{,}6693)} 1{,}2 +$$

$$+ \frac{(3{,}5 - 3{,}0042)(3{,}5 - 3{,}3201)}{(3{,}6693 - 3{,}0042)(3{,}6693 - 3{,}3201)} 1{,}3 = 1{,}2527.$$

За п'ятизначними таблицями ln 3,5 = 1,2528.

Розглянемо випадок, коли таблиця функції $y = f(x)$ допускає лінійне інтерполювання. При оберненому інтерполюванні, тобто під час знаходження значень оберненої функції, можна також вважати, що в інтервалі між сусідніми табличними значеннями приріст x змінюється пропорційно до приросту y.

Якщо задане значення y лежить між табличними значеннями y_0 і y_1, тоді маємо:

$$x = x_0 + \frac{y - y_0}{\Delta y_0} h. \qquad (28.3)$$

Формулу (28.3) дістанемо з формули лінійної інтерполяції, помінявши місцями x і y.

Отже, щоб визначити x, треба до табличного значення x_0 внести поправку $\frac{y - y_0}{\Delta y_0} h$.

2. Знайти значення arcsin 0,3295, користуючись таблицею для функції sin x:

x	0,30	0,31	0,32	0,33	0,34	0,35
sin x	0,2955	0,3051	0,3146	0,3240	0,3335	0,3429

Значення $\sin x = 0{,}3295$ лежить між двома табличними 0,3240 і 0,3335. Тому покладемо $y_0 = 0{,}3240$; $y_1 = 0{,}3335$; $x_0 = 0{,}33$; $h = 0{,}01$. За формулою (28.3) знайдемо:

$$x = 0{,}33 + \frac{0{,}3295 - 0{,}3240}{0{,}0095} 0{,}01 = 0{,}33 + 0{,}0058 = 0{,}3358.$$

Отже, arc sin 0,3295 $\approx$ 0,3358.

У таблицях, як правило, подаються табличні різниці і є табличка пропорційних частин, тому такі обчислення виконувати досить легко.

Вправи.

1. Побудувати перший і другий інтерполяційні многочлени Ньютона для функції, що задана таблицею значень:

x	0	1	2	3
y	1	−1	0	2

2. Побудувати інтерполяційний поліном Ньютона для функції $f(x) = \lg x$ з вузлами $x = 2, 3, 4, 5$ і за його допомогою обчислити $\lg 2{,}5$. Оцінити похибку результату.

Значення функції $y = \lg x$ у вузлах інтерполяції такі:

x	2	3	4	5
$\lg x$	0,3010	0,4771	0,6021	0,6990

3. Чи допускає на відрізку $\left[0; \frac{\pi}{4}\right]$ чотиризначна таблиця значень функції $y = \sin x$, крок якої $h = 0{,}1$ рад., квадратичне і лінійне інтерполювання?

4. Для функції, заданої таблицею, знайти її значення в точках x'. Результати обчислити з максимально можливою точністю.

x	0.1	0.2	0,3	0.4	0.5	0.6	0,7	0.8	0.9	1,0
$f(x)$	0,10017	0,18099	0,26302	0,34692	0,43344	0,52360	0,61873	0,72082	0,83307	0,96141
x'	0,036	0,117	0,161	0,242	0,251	0,862	0,831	0,920	0,982	1,025

5. За таблицею вправи 2 знайти x, якщо $\lg x = 0{,}5$.

6. За таблицею прикладу 3 (с. 135) знайти $\arcsin 0{,}3001$; $\arcsin 0{,}3205$; $\arcsin 0{,}3392$.

7. Скласти таблицю значень функції на відрізку $[a; b]$ з точністю до $0{,}5 \cdot 10^{-3}$ так, щоб таблиця допускала лінійну інтерполяцію. Першу таблицю скласти з точністю до $0{,}5 \cdot 10^{-3}$ з кроком h:

1) $y = \dfrac{e^{-x^2} + 2x}{\sqrt{\pi x + e^{-x^2}}}$, $a = 0{,}2$; $b = 1{,}0$; $h = 0{,}1$;

2) $y = \dfrac{0{,}043x + \pi}{2{,}489x + 0{,}253x^2}$, $a = 1{,}2$; $b = 2{,}6$; $h = 0{,}1$.

§ 29. Рівномірні і середньоквадратичні наближення

1. Задача найкращого наближення. З матеріалу попередніх параграфів ми знаємо, що апроксимувати функції можна інтерполяційними многочленами. Такий спосіб апроксимації широко застосовується. Проте в ряді випадків інтерполювання не ефективне, бо внаслідок цього часто утворюються многочлени високого степеня, з якими працювати незручно. Правда, цьому можна запобігти. Для цього відрізок, на якому потрібно апроксимувати задану функцію $f(x)$, розбивають на частини невеликої довжини і для кожної такої частини будують свій інтерполяційний многочлен невисокого степеня. Проте цей метод

має також недоліки, оскільки в багатьох випадках бажано мати єдину наближену формулу, яку б можна було застосовувати для всього відрізка.

При інтерполюванні, як відомо, інтерполяційний многочлен будують при умові, що у вузлах його значення дорівнюють значенням функції $f(x)$. А якщо значення функції $f(x)$ у вузлах знайдені неточно (наприклад, визначені в процесі деякого експерименту), то, очевидно, точно пристосовувати значення многочлена і функції у вузлах недоцільно. Корисно в цьому випадку підібрати для апроксимації таку функцію, яка хоча і не дорівнює заданій функції у заданих вузлах, проте найкраще наближає цю функцію на всьому відрізку.

Перед тим як розглядати конкретні випадки теорії апроксимації, наведемо деякі загальні міркування.

Нехай M — деяка множина в просторі $\mathbf{R}$. Зокрема, це може бути відрізок $[a; b]$ або скінченна сукупність точок $x_0, x_1, \ldots, x_n$ з відрізка $[a; b]$. І нехай на M задано деякий простір (множину) функцій (наприклад, простір неперервних функцій, простір функцій, сумовних з квадратом тощо), який позначимо через $H(M)$. Нехай цей простір лінійний і нормований. Візьмемо в $H(M)$ $m + 1$ лінійно незалежних елементів $\varphi_0(x)$, $\varphi_1(x), \ldots, \varphi_m(x)$ і розглянемо всі можливі лінійні комбінації

$$\varphi(x) = \sum_{j=0}^{m} a_j \varphi_j(x). \qquad (29.1)$$

Множина таких функцій $\varphi(x)$ утворить, очевидно, $(m + 1)$-вимірний лінійний нормований простір, який позначимо через W_{m+1}.

Як відомо (див. § 8), нормований простір $H(M)$ можна зробити метричним, якщо ввести відстань за формулою $\rho(f, g) = \|f - g\|$, де $f, g \in H(M)$.

Нехай $f(x)$ фіксована функція з $H(M)$ і така, що $f(x) \notin W_{m+1}$. Числова множина $\rho(f, \varphi)$ обмежена знизу (норми — невід'ємні числа). Тому існує точна нижня межа значень $\rho(f, \varphi)$.

Означення.

Число

$$\rho_0 = \inf_{\varphi \in W_{m+1}} \rho(f, \varphi)$$

називається відстанню від f до класу W_{m+1}, а елемент $\hat{\varphi}$ такий, що

$$\rho(f, \hat{\varphi}) = \rho_0,$$

називається елементом найкращого наближення (найближчим елементом).

Розглядаючи різні простори і вводячи по-різному норму в них, діставатимемо різні конкретні випадки теорії апроксимації.

Природно виникає питання про існування і єдиність елемента найкращого наближення. Для довільних нормованих просторів ми цього питання не розглядатимемо. Зауважимо лише, що при апроксимації функціями виду (29.1) найкращий елемент завжди існує, але не завжди єдиний.

2. Найкраще рівномірне наближення функцій. Нехай маємо простір $C[a; b]$. Введемо норму $\|f\| = \max\limits_{x\in[a;b]} |f(x)|$. Ця норма в просторі $C[a; b]$ визначає метрику

$$\rho(f, g) = \|f - g\| = \max_{x\in[a;b]} |f(x) - g(x)|.$$

Отже, побудова елемента найкращого наближення в цьому просторі зводиться до відшукування таких коефіцієнтів a_0, a_1, ..., a_m лінійної комбінації $\sum\limits_{j=0}^{m} a_j\varphi_j(x)$, де $\{\varphi_j(x)\}$ — деякі лінійно незалежні функції із $C[a; b]$, що

$$\max_{x\in[a;b]} \left| f(x) - \sum_{j=0}^{m} a_j\varphi_j(x) \right|$$

набуває найменшого значення.

Задача апроксимації в цьому випадку називається *задачею найкращого рівномірного або чебишовського наближення на проміжку* $[a; b]$.

Часто за систему $\{\varphi_j(x)\}$ беруть степеневі функції $1, x, x^2, \ldots, x^m$, тобто виконують найкраще рівномірне наближення алгебраїчними многочленами.

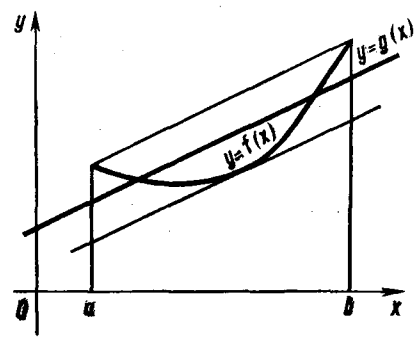
Зауважимо, що коли функція $f(x)$ задана лише в скінченній кількості точок $Q = \{x_1, x_2, \ldots, x_n\}$ (дискретний випадок), то норма $\|f\| = \max\limits_{x\in Q} |f(x)|$ і знаходження елемента найкращого наближення в цьому випадку зводиться до відшукування такої функції $g(x)$, для якої $\|f(x) - g(x)\| = \max\limits_{x\in Q} |f(x) - g(x)|$ набуває найменшого значення. Це задача дискретного чебишовського наближення.

Чебишовські наближення відіграють важливу роль у практичних питаннях, зокрема коли задаються функції в ЕОМ. Вводити таблицю значень функції в пам'ять ЕОМ не вигідно, оскільки для цього потрібно багато місця в пам'яті і багато часу для відшукування потрібного значення функції. Як правило, дану функцію $f(x)$ заміняють деякою іншою $g(x)$, яка обчислюється простіше; при цьому функцію $g(x)$ часто визначають так, щоб вона на заданій множині відхилялася від функції $f(x)$ не більш як на задану величину ε.

Приклад.

Побудувати найкраще рівномірне наближення виду $g(x) = a_0 + a_1x$ для двічі неперервно диференційовної функції $f(x)$ на $[a; b]$, для якої $f''(x) > 0$.

З геометричних міркувань легко встановити, що в даному випадку многочлен $g(x) = a_0 + a_1x$ найкращого рівномірного наближення є прямою, що рівновіддалена від хорди, яка проходить через точки $(a; f(a))$ і $(b; f(b))$, і дотичної до кривої $y = f(x)$, що паралельна хорді (мал. 32).



Мал. 32

Наближено задачу чебишовського наближення функції $f(x)$ узагальненим поліномом $\sum\limits_{j=0}^{m} a_j\varphi_j(x)$ на відрізку $[a; b]$ можна розв'язати так. Виберемо на $[a; b]$ досить густу сітку точок $G = \{x_1, x_2, \ldots, x_n\}$ і розглядатимемо задачу відшукування таких $a_0, a_1, \ldots, a_m$, при яких величина

$$\rho^* = \max_{1\leqslant i\leqslant n} \left| f(x_i) - \sum_{j=0}^{m} a_j\varphi_j(x_i) \right| \qquad (29.2)$$

досягала б найменшого значення. Якщо сітка G досить густа, то величина ρ^* близька до $\rho(f, \varphi)$ в $C[a; b]$, а поліном $\varphi^*(x)$, який найкраще наближає функцію $f(x)$ на множині G, досить близький до шуканого полінома $\varphi(x)$, який найкраще наближає $f(x)$ на всьому відрізку $[a; b]$.

Розглянемо тепер задачу відшукання найменшого значення величини η при обмеженнях

$$\left| f(x_i) - \sum_{j=0}^{m} a_j\varphi_j(x_i) \right| \leqslant \eta \quad (i = 1, 2, \ldots, n),$$

або, що те саме, задачу мінімізації η при обмеженнях

$$-f(x_i) + \sum_{j=0}^{m} a_j\varphi_j(x_i) + \eta \geqslant 0,$$

$$+f(x_i) - \sum_{j=0}^{m} a_j\varphi_j(x_i) + \eta \geqslant 0. \qquad (29.3)$$

Тут $f(x_i)$, $\varphi_j(x_i)$ $(i = 1, 2, \ldots, n)$ — відомі числа, причому значення $\varphi_j(x_i)$ є коефіцієнтами при невідомих a_j. Отже, обмеження (29.3) є системою $2n$-лінійних нерівностей щодо змінних $a_0, a_1, \ldots, a_m, \eta$. Функція $Z = \eta$, мінімум якої потрібно знайти, також лінійна відносно змінних $a_0, a_1, \ldots, a_m, \eta$. Таким чином, ми прийшли до задачі лінійного програмування, яку можна розв'язати симплекс-методом.

Очевидно, що коли найкраще чебишовське наближення функції $f(x)$ відшукається не на відрізку $[a; b]$, а на дискретній множині G (наприклад, якщо функція $f(x)$ задана таблично), то таку задачу розглянутим способом можна розв'язати точно (з точністю до похибок обчислень).

3. Середньоквадратичні наближення. Дещо простіше розв'язується задача найкращого наближення в евклідових просторах (див. § 8).

Нехай $f(x)$ — деяка функція з евклідового простору E. Будемо наближати її узагальненим многочленом

$$\varphi(x) = \sum_{j=0}^{m} a_j \varphi_j(x),$$

де $\varphi_0(x)$, $\varphi_1(x)$, ..., $\varphi_m(x)$ — лінійно незалежні функції з простору E. Задача найкращого наближення формулюється так: підібрати коефіцієнти $a_0, a_1, \ldots, a_m$ узагальненого многочлена $\varphi(x)$ так, щоб величина

$$\rho^2(f, \varphi) = \|f(x) - \varphi(x)\|^2 = \left\| \varphi(x) - \sum_{j=0}^{m} a_j \varphi_j(x) \right\|^2$$

була найменша (тут для зручності мінімізуємо не відстань, а її квадрат). Многочлен $\varphi(x)$, який мінімізує $\rho^2(f, \varphi)$, називається *многочленом найкращого середньоквадратичного наближення* функції $f(x)$, а розглянутий тут метод для відшукання такого многочлена — *методом найменших квадратів*. Величина, яка мінімізується, є функцією коефіцієнтів $a_0, a_1, \ldots, a_m$; позначимо її через $S(a_0, a_1, \ldots, a_m)$. Отже (нижче для зручності записів аргумент x опускаємо),

$$S(a_0, a_1, \ldots, a_m) = \left\| f - \sum_{j=0}^{m} a_j \varphi_j \right\|^2 =$$

$$= \left(f - \sum_{j=0}^{m} a_j \varphi_j, f - \sum_{j=0}^{m} a_j \varphi_j \right) =$$

$$= (f, f) - 2 \sum_{j=0}^{m} a_j (f, \varphi_j) + \sum_{j=0}^{m} \sum_{k=0}^{m} a_j a_k (\varphi_j, \varphi_k).$$

Для знаходження невідомих коефіцієнтів $a_0, a_1, \ldots, a_m$, при яких вираз $S(a_0, a_1, \ldots, a_m)$ досягає мінімуму, знайдемо похідні від S по коефіцієнтах a_i і прирівняємо їх до нуля. Таким чином дістанемо систему:

$$\frac{\partial S}{\partial a_i} = -2 \cdot (f, \varphi_i) + 2 \sum_{j=0}^{m} a_j (\varphi_i, \varphi_j) = 0 \quad (i = 0, 1, \ldots, m),$$

тобто для визначення коефіцієнтів $a_0, a_1, \ldots, a_m$ шуканого многочлена маємо систему лінійних алгебраїчних рівнянь (яка називається *нормальною системою*):

$$\sum_{j=0}^{m} a_j (\varphi_i, \varphi_j) = (f, \varphi_i) \quad (i = 0, 1, \ldots, m), \qquad (29.4)$$

де

$$(\varphi_i, \varphi_j) = \int_a^b \varphi_i(x) \varphi_j(x)\, dx, \quad (f, \varphi_i) = \int_a^b f(x) \varphi_i(x)\, dx. \qquad (29.5)$$

Визначник системи (29.4)

$$\begin{vmatrix} (\varphi_0, \varphi_0) & (\varphi_0, \varphi_1) & \ldots & (\varphi_0, \varphi_m) \\ (\varphi_1, \varphi_0) & (\varphi_1, \varphi_1) & \ldots & (\varphi_1, \varphi_m) \\ \ldots & \ldots & \ldots & \ldots \\ (\varphi_m, \varphi_0) & (\varphi_m, \varphi_1) & \ldots & (\varphi_m, \varphi_m) \end{vmatrix}$$

є так званий *визначник Грама* функцій $\varphi_k(x)$. Як відомо, цей визначник не дорівнює нулеві (він завжди додатний), якщо система $\varphi_0(x), \varphi_1(x), \ldots, \varphi_m(x)$ лінійно незалежна. Отже, система (29.4) має єдиний розв'язок. Оскільки функція $S(a_0, a_1, \ldots, a_m)$ мінімуму досягає, то єдина її стаціонарна точка може бути лише точкою мінімуму.

Таким чином, знаходження многочлена $\varphi(x)$, який дає найкраще середньоквадратичне наближення функції $f(x)$, зводиться до розв'язування лінійної алгебраїчної системи (29.4).

В кожному конкретному евклідовому просторі інший вигляд матимуть лише скалярні добутки.

Нехай, наприклад, маємо простір $C_2[a; b]$, що складається з неперервних на $[a; b]$ функцій (див. § 8) *. Скалярний

* Читач, ознайомлений з інтегралом Лебега, може замість простору $C_2[a; b]$ розглядати простір $L_2[a; b]$. Взагалі для середньоквадратичних наближень вимога неперервності зайва. Тут потрібно накладати лише більш слабку вимогу існування інтеграла $\int_a^b f^2(x)\, dx$.

добуток введемо за формулою

$$(f, g) = \int_a^b f(x)\, g(x)\, dx.$$

При цьому

$$\|f\| = \sqrt{(f, f)} = \sqrt{\int_a^b f^2(x)\, dx}, \quad \text{у}\rho(f, g) = \|f - g\| = \sqrt{\int_a^b (f(x) - g(x))^2\, dx}.$$

Тоді відповідні скалярні добутки в нормальній системі (29.4) матимуть вигляд:

$$(\varphi_i, \varphi_j) = \int_a^b \varphi_i(x)\, \varphi_j(x)\, dx; \quad (f, \varphi_i) = \int_a^b f(x)\, \varphi_i(x)\, dx.$$

Часто за систему $\varphi_0(x)$, $\varphi_1(x)$, ..., $\varphi_m(x)$ беруть функції $1, x, x^2, \ldots, x^m$, які належать простору $C_2[a; b]$ і лінійно незалежні на $[a; b]$ при будь-якому m. У цьому випадку $\varphi(x)$ буде звичайним алгебраїчним многочленом $P_m(x) = a_0 + a_1x + a_2x^2 + \ldots + a_mx^m$.

Елементом найкращого наближення для функції $f(x) \in C_2[a; b]$ буде такий многочлен $P_m(x)$, для якого $\|f(x) - P_m(x)\|^2$ набуває найменшого значення.

Тут $(\varphi_i, \varphi_j) = \int_a^b x^i x^j dx = \int_a^b x^{i+j} dx$, $(f, \varphi_i) = \int_a^b f(x)\, x^i dx$.

При обчисленні інтегралів $\int_a^b f(x)\, x^i dx$ можуть виникнути деякі труднощі, оскільки знайти первісну через елементарні функції часто не вдається. В такому разі інтеграли обчислюють наближено, використовуючи чисельні методи, про які йтиметься в розділі VII.

Приклад. Знайти найкращу середньоквадратичну апроксимацію функції $y = \sin x$ на відрізку $\left[0; \frac{\pi}{2}\right]$ алгебраїчним многочленом другого степеня $P_2(x) = a_0 + a_1x + a_2x^2$. Тут $\varphi_0(x) = 1$: $\varphi_1(x) = x$, $\varphi_2(x) = x^2$. Знайдемо потрібні скалярні добутки,

$$(\varphi_0, \varphi_0) = \int_0^{\frac{\pi}{2}} 1 \cdot dx = \frac{\pi}{2}, \quad (\varphi_0, \varphi_1) = \int_0^{\frac{\pi}{2}} 1 \cdot x dx = \frac{\pi^2}{8},$$

$$(\varphi_0, \varphi_2) = \int_0^{\frac{\pi}{2}} 1 \cdot x^2 dx = \frac{\pi^3}{24}, \quad (\varphi_1, \varphi_1) = \int_0^{\frac{\pi}{2}} x \cdot x dx = \frac{\pi^3}{24},$$

$$(\varphi_1, \varphi_2) = \int_0^{\frac{\pi}{2}} x \cdot x^2 dx = \frac{\pi^4}{64},$$

$$(\varphi_2, \varphi_2) = \int_0^{\frac{\pi}{2}} x^2 \cdot x^2 dx = \frac{\pi^5}{160}, \quad (f, \varphi_0) = \int_0^{\frac{\pi}{2}} \sin x \cdot 1 dx = 1;$$

$$(f, \varphi_1) = \int_0^{\frac{\pi}{2}} x \sin x dx = 1; \quad (f, \varphi_2) = \int_0^{\frac{\pi}{2}} x^2 \sin x dx = \pi - 2.$$

Система (29.4) має вигляд:

$$\begin{cases} 1{,}571a_0 + 2{,}234a_1 + 1{,}292a_2 = 1, \\ 2{,}234a_0 + 1{,}292a_1 + 1{,}522a_2 = 1, \\ 1{,}292a_0 + 1{,}522a_1 + 1{,}913a_2 = 1{,}142. \end{cases}$$

Звідси $a_0 = 0{,}0656$, $a_1 = 0{,}152$, $a_2 = 0{,}432$.

Отже, многочлен найкращого наближення записується так:

$$P_2(x) = 0{,}432x^2 + 0{,}152x + 0{,}0656.$$

Особливо просто розв'язується задача найкращого наближення, якщо система функцій $\varphi_0, \varphi_1, \ldots, \varphi_m$ є ортогональною (див. § 8), тобто коли

$$(\varphi_i, \varphi_j) = 0, \quad i \neq j \quad (\varphi_i, \varphi_i) > 0,$$

де $0 \leqslant i, j \leqslant m$.

Зауважимо, що ортогональна система функцій є лінійно незалежна. Справді, припустимо, що система $\varphi_0, \varphi_1, \ldots, \varphi_m$ ортогональна і для неї виконується рівність

$$\alpha_0\varphi_0 + \alpha_1\varphi_1 + \cdots + \alpha_m\varphi_m = \theta.$$

Помноживши останню рівність скалярно на φ_j, дістанемо $\alpha_j(\varphi_j, \varphi_j) = 0$. З останнього випливає, що $\alpha_j = 0$ при будь-якому j. Отже, система $\varphi_0, \varphi_1, \ldots, \varphi_m$ лінійно незалежна.

У випадку ортогональності системи $\varphi_0, \varphi_1, \ldots, \varphi_m$ матриця нормальної системи (29.4) стає діагональною і тому коефіцієнти узагальненого многочлена найкращого середньоквадра-

тичного наближення функції $f(x)$ обчислюються за формулами

$$a_j = \frac{(f, \varphi_j)}{(\varphi_j, \varphi_j)} \quad (j = 0, 1, \ldots, m).$$

Вони називаються *коефіцієнтами Фур'є* функції $f(x)$ за ортогональною системою $\varphi_0, \varphi_1, \ldots, \varphi_m$.

Приклад.

Знайти найкращу середньоквадратичну апроксимацію функції $y = \frac{e^x + e^{-x}}{2}$ на відрізку $[-\pi; \pi]$ тригонометричним многочленом

$$T_n(x) = a_0 + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx).$$

Як уже зазначалося (§ 8), система

$$1, \cos kx, \quad \sin kx \quad (k = 1, 2, 3, \ldots)$$

є ортогональною на $[-\pi; \pi]$, і тому коефіцієнти a_0, a_k, b_k $(k = 1, 2, \ldots, n)$ мають бути коефіцієнтами Фур'є функції $y = \frac{e^x + e^{-x}}{2}$ за цією системою.

Отже, для коефіцієнтів многочлена $T_n(x)$ маємо:

$$a_0 = \frac{(f, 1)}{(1, 1)} = \frac{\int_{-\pi}^{\pi} \frac{e^x + e^{-x}}{2} \cdot 1 dx}{\int_{-\pi}^{\pi} 1 \cdot 1 dx} = \frac{1}{2\pi} \cdot \frac{e^x - e^{-x}}{2} \Big|_{-\pi}^{\pi} =$$

$$= \frac{e^{\pi} - e^{-\pi}}{2\pi};$$

$$a_k = \frac{(f, \varphi_k)}{(\varphi_k, \varphi_k)} = \frac{\int_{-\pi}^{\pi} \frac{e^x + e^{-x}}{2} \cos kx dx}{\int_{-\pi}^{\pi} \cos^2 kx dx} = \frac{1}{\pi} \int_{-\pi}^{\pi} \frac{e^x + e^{-x}}{2} \cos kx dx =$$

$$= (-1)^k \frac{e^{\pi} - e^{-\pi}}{\pi (1 + k^2)},$$

$$b_k = \frac{(f, \varphi_k)}{(\varphi_k, \varphi_k)} = \frac{\int_{-\pi}^{\pi} \frac{e^x + e^{-x}}{2} \sin kx dx}{\int_{-\pi}^{\pi} \sin^2 kx dx} = \frac{1}{\pi} \int_{-\pi}^{\pi} \frac{e^x + e^{-x}}{2} \sin kx dx = 0.$$

Під останнім інтегралом стоїть непарна функція, і інтеграл дорівнює нулеві.

Таким чином, шуканий многочлен має такий вигляд:

$$T_n(x) = \frac{e^{\pi} - e^{-\pi}}{\pi} \left(\frac{1}{2} + \sum_{k=1}^{n} (-1)^k \frac{\cos kx}{1 + k^2} \right) \approx$$

$$\approx 7{,}35216 \left(\frac{1}{2} + \sum_{k=1}^{n} (-1)^k \frac{\cos kx}{1 + k^2} \right).$$

Примітка. Якщо система $\{\varphi_j(x)\}$ не ортогональна, то із зростанням m (m — порядок многочлена) визначник Грама, як правило, швидко прямує до нуля, а розв'язуються такі системи з великою неточністю (див. § 6). Тому брати m більшим 5 недоцільно. Ортогональних систем це не стосується.

4. Наближення функцій, заданих таблично методом найменших квадратів. Нехай функція $f(x)$ задана своїми значеннями в скінченному числі точок $x_0, x_1, \ldots, x_n$ з відрізка $[a; b]$. Як і раніше, будемо шукати найкраще наближення функції $f(x)$ узагальненими многочленами $\varphi(x) = \sum_{j=0}^{m} a_j \varphi_j(x)$, де $\varphi_j(x)$ $(i = 0, 1, \ldots, m)$ — лінійно незалежні функції, $m \leqslant n$.

У цьому випадку многочлен найкращого наближення шукаємо в просторі функцій, що задані на дискретній множині точок $Q = \{x_0, x_1, \ldots, x_n\}$, з скалярним добутком

$$(f, g) = \sum_{i=0}^{n} f(x_i) g(x_i). \qquad (29.6)$$

Позначимо цей евклідів простір через E_{n+1}. Норма елемента і відстань між двома елементами визначаються відповідно формулами (див. § 8, евклідові простори, приклад 3):

$$\|f\| = \sqrt{(f, f)} = \sqrt{\sum_{i=0}^{n} f^2(x_i)}; \quad \rho(f, g) = \sqrt{\sum_{i=0}^{n} (f(x_i) - g(x_i))^2}.$$

Отже, коефіцієнти a_j $(j = 0, 1, \ldots, m)$ многочлена найкращого наближення знайдемо з нормальної системи (29.4), де $(\varphi_i, \varphi_j) = \sum_{k=0}^{n} \varphi_i(x_k) \varphi_j(x_k)$; $(f, \varphi_i) = \sum_{k=0}^{n} f(x_k) \varphi_i(x_k)$.

При $m = n$ многочлен найкращого наближення, побудований методом найменших квадратів, збігається з інтерполяційним многочленом. Справді, для останнього маємо: $S(a_0, a_1, \ldots, a_n) = \rho^2(f, \varphi) = \sum_{i=0}^{n} (f(x_i) - \varphi(x_i))^2 = 0$, а для $\rho^2(f, \varphi)$ значення 0 є найменшим з усіх можливих.

Розглянемо наближення функцій алгебраїчними многочленами.

Нехай $\varphi_i(x) = x^i$ $(i = 0, 1, \ldots, m)$. Тоді $\varphi(x)$ є алгебраїчним многочленом степеня m: $\varphi(x) = P_m(x) = a_0 + a_1x + a_2x^2 + \cdots + a_mx^m$.

Функції $1, x, \ldots, x^m$ $(m \leqslant n)$ лінійно незалежні в E_{n+1}. Справді, лінійна комбінація $\alpha_0 + \alpha_1x + \ldots + \alpha_mx^m$, де не всі α_j дорівнюють нулеві, не може набувати нульових значень для всіх $x \in Q = \{x_0, x_1, \ldots, x_n\}$ (многочлен степеня m не може мати більше ніж m нулів). Отже, визначник системи (29.4), як визначник Грама лінійно незалежних функцій, не дорівнює нулеві.

Якщо $m \geqslant n + 1$, то система функцій $\{x^j\}$ $(j = 0, 1, \ldots, m)$ лінійно залежна в E_{n+1}. Це випливає з того, що многочлен $(x - x_0)(x - x_1) \ldots (x - x_n) = x^{n+1} + \alpha_nx^n + \cdots + \alpha_0$ на множині Q набуває нульових значень.

Приклад.

Побудувати за методом найменших квадратів многочлен другого степеня, якщо функція $y = f(x)$ задана таблицею:

x	0,37	0,74	1,02	2,34
y	— 1,60	— 0,59	0,05	0,95

Враховуючи вирази для скалярних добутків

$$(\varphi_i, \varphi_j) = \sum_{k=0}^{n} x_k^i x_k^j = \sum_{k=0}^{n} x_k^{i+j}, \quad (f, \varphi_i) = \sum_{k=0}^{n} f(x_k)\, x_k^i$$

та проводячи обчислення з трьома знаками після коми, для визначення коефіцієнтів a_0, a_1, a_2 дістанемо систему:

$$\begin{cases} 4a_0 + 4{,}47a_1 + 7{,}201a_2 = -1{,}19, \\ 4{,}47a_0 + 7{,}201a_1 + 14{,}330a_2 = 1{,}245, \\ 7{,}201a_0 + 14{,}330a_1 + 31{,}383a_2 = 4{,}712. \end{cases}$$

Розв'язавши її, матимемо: $a_0 = -2{,}877$; $a_1 = 3{,}793$; $a_2 = -0{,}9217$.

Таким чином, шуканий поліном

$$P_2(x) = -2{,}877 + 3{,}793x - 0{,}9217x^2.$$

При цьому квадратичне відхілення дорівнює $\sum_{i=0}^{3} (y_i - P_2(x_i))^2 = 3{,}1 \cdot 10^{-6}$.

Особливо просто відшукується многочлен найкращого наближення, якщо функції $\varphi_0(x)$, $\varphi_1(x)$, …, $\varphi_m(x)$ утворюють ортогональну або ортонормальну систему на множині $Q = \{x_0, x_1, \ldots, x_n\}$, тобто у випадку, коли функції $\{\varphi_k(x)\}$ задовольняють умову:

$$\sum_{i=0}^{n} \varphi_k(x_i)\, \varphi_j(x_i) = \begin{cases} 0, & k \neq j, \\ c > 0, & k = j. \end{cases}$$

Якщо $\varphi_j(x)$, $(j = 0, 1, \ldots, m)$ є многочленами, що задовольняють останню умову, вони називаються *ортогональними многочленами дискретного аргументу*, а коефіцієнти многочлена найкращого наближення є коефіцієнтами Фур'є функції $f(x)$ за цією системою.

В просторі E_{n+1} ортогональна система може містити не більше ніж $(n + 1)$-шу функцію. Якщо $m = n$, то ортогональна система утворює в E_{n+1} ортогональний базис. І, отже, будь-яку функцію $f(x) \in E_{n+1}$ можна подати у вигляді

$$f(x) = \varphi(x) = \sum_{j=0}^{n} a_j\varphi_j(x), \quad x \in Q = \{x_0, x_1, \ldots, x_n\}.$$

Тоді $\rho(f, \varphi) = 0$. Очевидно, многочлен $\varphi(x)$ збігатиметься з многочленом найкращого середньоквадратичного наближення функції $f(x) \in E_{n+1}$ при $m = n$.

5. Застосування методу найменших квадратів для побудови емпіричних формул. Під час проведення наукових експериментів, а також у виробничій практиці досить часто доводиться встановлювати функціональну залежність між деякими величинами. Нехай, наприклад, вивчаючи функціональну залежність між величинами y і x, ми провели ряд вимірювань, в результаті яких дістали таблицю значень величини y при $x = x_1, x_2, \ldots, x_n \in [a; b]$. На основі цих даних потрібно побудувати аналітичний вираз (формулу) $\bar{y}(x)$, який би наближав функцію $y(x)$ на всьому проміжку $[a; b]$. Формули, які дістанемо на основі аналізу дослідних даних, назвемо *емпіричними*.

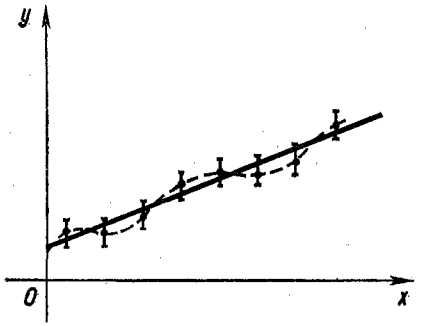
Емпіричну формулу подамо у вигляді

$$\bar{y}(x) = f(x, a_1, a_2, \ldots, a_m),$$

де $a_1, a_2, \ldots, a_m$ $(m < n)$ — деякі параметри.

Вигляд функції f і кількість параметрів в одних випадках відомі з деяких додаткових міркувань, в інших їх вдається визначити з графіка, який побудований за експериментальними значеннями y_i. Якщо графік побудовано, то вигляд функціональної залежності вдається приблизно передбачити, порівнявши даний графік з графіками відомих кривих (окремі непра-

вильності при цьому не враховуються). Кількість параметрів підбирають так, щоб емпірична формула найкраще відображала результати спостережень і була досить простою. Як уже зазначалось, для задач такого типу інтерполювання непридатне. З мал. 33 видно, що іноді значно краще взяти емпіричну формулу у вигляді лінійного виразу, ніж інтерполяційного многочлена, графік якого показано пунктирною лінією. На малюнку зображено експериментальні значення, знайдені наближено, та межі, в яких можуть знаходитися їх точні значення.



Мал. 33

Побудова емпіричної формули проводиться двома етапами:

1) з'ясування загального виду формули,

2) визначення її параметрів.

Параметри часто визначають методом найменших квадратів, тобто з умови мінімізації величини

$$S(a_1, a_2, \ldots, a_m) = \sum_{i=1}^{n} (y_i - f(x_i, a_1, a_2, \ldots, a_m))^2.$$

Використавши необхідні умови екстремуму для функцій багатьох змінних, дістанемо систему:

$$\frac{\partial S(a_1, a_2, \ldots, a_m)}{\partial a_i} = 0 \quad (i = 1, 2, \ldots, m). \qquad (29.7)$$

Знайшовши всі розв'язки системи, які належать області зміни параметрів $a_1, a_2, \ldots, a_m$, і вибравши той, для якого величина $S(a_1, a_2, \ldots, a_m)$ має абсолютний мінімум, матимемо шукані значення параметрів.

Система (29.7) значно спрощується, якщо функція $f(x, a_1, a_2, \ldots, a_m)$ лінійна відносно параметрів.

Зауважимо, що задачу наближення функцій, заданих таблично, за допомогою узагальнених чи алгебраїчних многочленів степеня m методом найменших квадратів можна інтерпретувати як задачу побудови емпіричних формул у вигляді многочлена степеня m.

Приклад.

Знайти об'єм тіла обертання, якщо для осьового перерізу тіла експериментально дістали такі дані (мал. 34):

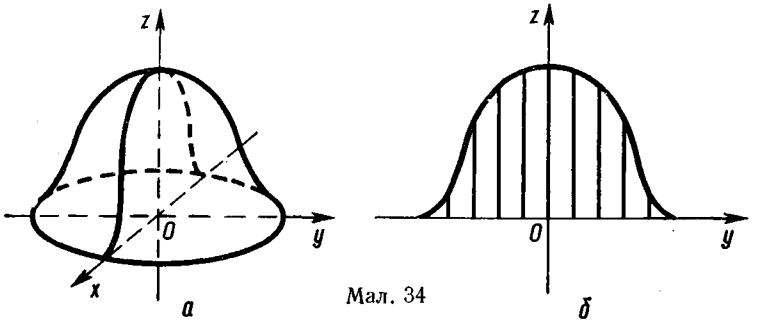
y	0	1	3	5	7	9
z	7,875	7,875	7,625	7,250	6,500	5,500

y	11	13	15	17	19	21
z	4,750	4,000	2,500	1,500	1,000	0,000

Об'єм тіла обертання легко визначити, якщо відомий аналітичний вираз кривої, від обертання якої утворюється дане тіло, тобто досить мати аналітичний вираз для функції, що задана таблицею значень.

Спочатку з'ясуємо загальний вигляд емпіричної формули. Побудувавши за табличними даними графік, переконаємося, що подібний графік має також функція

$$\overline{Z} = ay^4 + by^2 + c.$$



а Мал. 34 б

Тому шукану функцію візьмемо у такому ж вигляді. Коефіцієнти a, b і c визначимо методом найменших квадратів, тобто так, щоб величина

$$S = \sum_{i=1}^{12} (z_i - ay_i^4 - by_i^2 - c)^2$$

набувала найменшого значення. Тоді для визначення шуканих коефіцієнтів дістанемо систему:

$$\begin{cases} a\sum\limits_{i=1}^{12} y_i^8 + b\sum\limits_{i=1}^{12} y_i^6 + c\sum\limits_{i=1}^{12} y_i^4 = \sum\limits_{i=1}^{12} z_i y_i^4, \\ a\sum\limits_{i=1}^{12} y_i^6 + b\sum\limits_{i=1}^{12} y_i^4 + c\sum\limits_{i=1}^{12} y_i^2 = \sum\limits_{i=1}^{12} z_i y_i^2, \\ a\sum\limits_{i=1}^{12} y_i^4 + b\sum\limits_{i=1}^{12} y_i^2 + 12c = \sum\limits_{i=1}^{12} z_i. \end{cases}$$

Підставивши значення величин y_i і z_i, подані в таблиці, і розв'язавши систему, знайдемо: $a = -0{,}646 \cdot 10^{-4}$; $b = 0{,}428 \cdot 10^{-1}$; $c = 1{,}14$. Отже,

$$\bar{Z} = -0{,}646 \cdot 10^{-4} y^4 + 0{,}0428 y^2 + 1{,}14.$$

Знаючи тепер аналітичний вигляд кривої, легко можна визначити об'єм тіла (обчисліть самостійно).

Вправи.

1. Серед многочленів нульового степеня (констант) знайти многочлен найкращого рівномірного наближення для функції $f(x) = \frac{1}{1+x}$ на відрізку $[0; 1]$.

2. На площині задано три точки $P_i(x_i; y_i)$ $(i = 1, 2, 3)$. Провести пряму $g(x) = a_0 + a_1 x$ так, щоб величина $\max\limits_{1 \leqslant i \leqslant 3} |y_i - g(x_i)|$ була мінімальною.

3. Задано точки P_1 (—4; 1) і P_2 (5; 4). Знайти для них найкраще чебишовське наближення виду $g(x) = a_1 x$.

4. Знайти найкращу середньоквадратичну апроксимацію алгебраїчним многочленом другого степеня функції:

а) $f(x) = \cos x$ на відрізку $[1; 2]$;

б) $f(x) = e^x$ на відрізку $[0; 2]$.

5. Серед тригонометричних многочленів $T_2(x) = a_0 + a_1 \cos 2\pi x + b_1 \sin 2\pi x + a_2 \cos 4\pi x + b_2 \sin 4\pi x$ визначити многочлен, який на відрізку $[0; 1]$ дає найкраще середньоквадратичне наближення для функції $f(x) = x^3 - 2x^2 + x + 1$.

6. Знайти методом найменших квадратів наближене подання функції $f(x) = \frac{1}{1+x}$ за її значеннями в точках $x = 0, 1, 2, 3, 4$ многочленом першого і другого степеня.

7. Побудувати методом найменших квадратів многочлен другого степеня для функції, що задана таблицею:

x	0,4	0,8	1,1	1,5	1,9
$f(x)$	0,27	0,39	0,77	1,17	1,53

8. Встановити емпіричну формулу виду $y = a_1 e^{a_2 x}$, якщо результати експерименту подано таблицею:

x	0	1	2	3
y	2,01	1,21	0,74	0,45

Розділ VII
ЧИСЕЛЬНЕ ДИФЕРЕНЦІЮВАННЯ Й ІНТЕГРУВАННЯ

§ 30. Чисельне диференціювання

Задача чисельного диференціювання полягає в знаходженні значень похідних від функції за її даними значеннями. До чисельного диференціювання звертаються тоді, коли функція $y = f(x)$, для якої треба знайти похідну, задана таблично (наприклад, при знаходженні швидкості і прискорення, якщо закон руху заданий експериментально) або має складний аналітичний вираз. У першому випадку методи диференціального числення просто незастосовні, а в другому їх використання складне.

Похідна від функції $y = f(x)$ відносно x у точці x_0 визначається, як відомо, рівністю

$$\left(\frac{dy}{dx}\right)_{x=x_0} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

і характеризує нахил дотичної, що проведена до кривої $y = f(x)$ у точці x_0. Отже, якщо значення $f(x)$ при $x = x_0$ і $x = x_1$ відомі, то за наближене значення похідної можна взяти величину $\frac{f(x_1) - f(x_0)}{x_1 - x_0}$. Але цей спосіб не дуже зручний: не завжди можна взяти x_1 досить близьке до x_0 (наприклад, якщо функція задана таблично) і обчислити дріб з потрібною точністю. Тому функцію $f(x)$ на розглядуваному проміжку $[a; b]$ замінюють інтерполяційним многочленом $P_n(x)$ і вважають, що

$$f'(x) \approx P_n'(x), \quad x \in [a; b]. \tag{30.1}$$

Аналогічно знаходять похідні вищих порядків.

$$f^{(k)}(x) \approx P_n^{(k)}(x), \quad x \in [a; b].$$

Якщо для інтерполяційного многочлена відомий залишковий член

$$R_n(x) = f(x) - P_n(x),$$

то можна знайти залишковий член $r_n(x)$ похідної $P_n^{(k)}(x)$:

$$r_n(x) = f^{(k)}(x) - P_n^{(k)}(x) = R_n^{(k)}(x).$$

З того, що залишковий член $R_n(x)$ інтерполяційного полінома малий, зовсім не випливає, що залишковий член похід-

ної $r_n(x)$ також малий, бо похідні від малої функції можуть бути досить великими. Наприклад, нехай на відрізку $[a; b]$ задано дві неперервно диференційовані функції $f(x)$ і $g(x)$, такі, що

$$f(x) = g(x) + \frac{\sin n^2 x}{n}.$$

Для цих функцій

$$|f(x) - g(x)| \leqslant \max_{x\in[a;b]} \left|\frac{\sin n^2 x}{n}\right| \leqslant \frac{1}{n}.$$

Отже, при великому n ця різниця може бути як завгодно малою.

Для перших похідних від цих функцій маємо: $|f'(x) - g'(x)| = n|\cos n^2 x|$. Таким чином, різниця між похідними для деяких x може бути як завгодно великою при великих n. Отже, не існує неперервної залежності значень похідної від значень функції, тобто задача диференціювання некоректна (див. § 6). Саме цим задача диференціювання істотно відрізняється від задачі інтерполювання. Чисельне диференціювання, взагалі кажучи, є операцією менш точною, ніж інтерполювання. Особливо значні похибки при чисельному диференціюванні з'являються тоді, коли обчислюються похідні вищих порядків.

Залежно від форми інтерполяційного полінома $P_n(x)$ для функції $f(x)$, похідну від якої треба знайти, дістаємо різні формули чисельного диференціювання відповідно до наближеної рівності (30.1). Так, нехай функція $y = f(x)$ задана в точках x_i $(i = 0, 1, \ldots, n)$ з проміжку $[a; b]$ значеннями $y_i = f(x_i)$. Розглянемо випадок, коли вузли рівновіддалені, тобто

$$x_{i+1} - x_i = h \quad (i = 0, 1, \ldots, n-1).$$

Виведемо формули чисельного диференціювання, використавши інтерполяційні поліноми Ньютона. Замінимо, наприклад, функцію $y = f(x)$ першим інтерполяційним многочленом Ньютона, побудувавши його за вузлами $x_0, x_1, \ldots, x_n$. Цей многочлен, як відомо, має такий вигляд:

$$P_n(t) = P_n(x_0 + ht) = y_0 + t\cdot\Delta y_0 + \frac{t(t-1)}{2!}\Delta^2 y_0 + \frac{t(t-1)(t-2)}{3!}\Delta^3 y_0 + \cdots + \frac{t(t-1)(t-2)\ldots(t-n+1)}{n!}\Delta^n y_0, \qquad (30.2)$$

де $t = \frac{x - x_0}{h}$.

Вважатимемо, що похідна від функції $f(x)$ у будь-якій точці x наближено дорівнює похідній від цього многочлена. Щоб знайти похідну від многочлена $P_n(x)$, подамо його в іншому вигляді, а саме: запишемо за степенями t кожний доданок формули (30.2).

Відомо, що

$$t(t-1)(t-2)\ldots(t-k) = t^{k+1} - S_k^{(1)}t^k + S_k^{(2)}t^{k-1} - \cdots + (-1)^k S_k^{(k)}t, \qquad (30.3)$$

де $S_k^{(i)}$ — сума всіх можливих добутків натуральних чисел від 1 до k по i множників. Наприклад, для $n = 3$ матимемо:

$$t(t-1)(t-2)(t-3) = t^4 - (1+2+3)t^3 + (1\cdot 2 + 1\cdot 3 + 2\cdot 3)t^2 - 1\cdot 2\cdot 3t = t^4 - 6t^3 + 11t^2 - 6t.$$

На основі (30.3) многочлен $\overline{P}_n(t)$ запишемо так:

$$\overline{P}_n(t) = y_0 + t\Delta y_0 + \frac{t^2 - t}{2!}\Delta^2 y_0 + \frac{t^3 - 3t^2 + 2t}{3!}\Delta^3 y_0 + \cdots + \frac{t^n - S_{n-1}^{(1)}t^{n-1} + S_{n-1}^{(2)}t^{n-2} - \cdots + (-1)^{n-1}S_{n-1}^{(n-1)}}{n!}\Delta^n y_0.$$

Обчислимо тепер похідну по x від $P_n(t)$. Врахувавши, що $\frac{d}{dx}\times P_n(t) = \frac{d}{dt}\overline{P}_n(t)\frac{dt}{dx} = \frac{1}{h}\frac{d}{dt}\overline{P}_n(t)$, матимемо:

$$f'(x) \approx P_n'(x) = \frac{1}{h}\left(\Delta y_0 + \frac{2t-1}{2!}\Delta^2 y_0 + \frac{3t^2 - 6t + 2}{3!}\Delta^3 y_0 + \cdots + \frac{nt^{n-1} - (n-1)S_{n-1}^{(1)}t^{n-2} + \cdots + (-1)^{n-1}S_{n-1}^{(n-1)}}{n!}\Delta^n y_0\right). \qquad (30.4)$$

Рівність (30.4) — шукана формула чисельного диференціювання. Вона значно спрощується для знаходження похідної в табличних точках x_i. Наприклад, якщо $n = 2$, $t = 0$, $t = 1$, $t = 2$, дістанемо, такі формули відповідно:

$$y_0' = f'(x_0) \approx \frac{1}{h}\left(\Delta y_0 - \frac{1}{2}\Delta^2 y_0\right);$$

$$y_1' = f'(x_1) \approx \frac{1}{h}\left(\Delta y_0 + \frac{1}{2}\Delta^2 y_0\right);$$

$$y_2' = f'(x_2) \approx \frac{1}{h}\left(\Delta y_0 + \frac{3}{2}\Delta^2 y_0\right),$$

або, врахувавши вирази для скінченних різниць, подані через значення функції, матимемо:

$$y_0' = \frac{1}{2h}(-3y_0 + 4y_1 - y_2), \quad y_1' = \frac{1}{2h}(y_2 - y_0),$$

$$y_2' = \frac{1}{2h}(y_0 - 4y_1 + 3y_2). \tag{30.5}$$

Формула для похідної y_1' простіша, ніж для y_0' і y_2'; вона часто застосовується для обчислення похідних.

Якщо в таблиці функції задано її значення в достатній кількості вузлів, то, очевидно, кожне табличне значення можна вважати за початкове. Якщо $x = x_0$, формула (30.4) набере вигляду:

$$f'(x_0) \approx P_n'(x_0) = \frac{1}{h}\left(\Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \right.$$

$$\left. + \frac{1}{3}\Delta^3 y_0 - \cdots + (-1)^{n-1}\frac{1}{n}\Delta^n y_0\right). \tag{30.6}$$

А коли для виведення формули чисельного диференціювання використати другий інтерполяційний многочлен Ньютона, то дістанемо:

$$f'(x) \approx P_n'(x) = \frac{1}{h}\left(\Delta y_{n-1} + \frac{2t+1}{2!}\Delta^2 y_{n-2} + \right.$$

$$+ \frac{3t^2+6t+2}{3!}\Delta^3 y_{n-3} + \cdots$$

$$\left. \cdots + \frac{nt^{n-1} + (n-1)S_{n-1}^{(1)}t^{n-2} + \cdots + S_{n-1}^{(n-1)}}{h!}\Delta^n y_0\right), \tag{30.7}$$

де $t = \frac{x - x_n}{h}$.

Тоді для обчислення похідної в табличних точках (кожну табличну точку можна взяти за x_n) матимемо:

$$f'(x_n) \approx P_n'(x_n) = \frac{1}{h}\left(\Delta y_{n-1} + \frac{1}{2}\Delta^2 y_{n-2} + \right.$$

$$\left. + \frac{1}{3}\Delta^3 y_{n-3} + \ldots + \frac{1}{h}\Delta^n y_0\right). \tag{30.8}$$

Примітка. Зауважимо, що формули з різницями добирають так само, як і для інтерполювання. Наприклад, якщо треба знайти похідну в точці, близькій до початку ряду табличних значень, то користуються формулами, в основі яких лежить перший інтерполяційний многочлен Ньютона, тобто формулами (30.4) і (30.6). А коли потрібно знайти похідну в точці, близькій до кінця таблиці, то застосовують формули, в основі яких лежить другий інтерполяційний многочлен Ньютона, тобто формули (30.7), (30.8).

Залишковий член формул чисельного диференціювання, як уже зазначалося, можна обчислити, взявши похідну від залишкового члена відповідного інтерполяційного многочлена. Так, для інтерполяційного многочлена Ньютона залишковий член має вигляд:

$$R_n(x) = \bar{R}_n(t) = \frac{f^{n+1}(\xi)\,h^{n+1}}{(n+1)!}\Pi_{n+1}(t),$$

де

$$\Pi_{n+1}(t) = t(t-1)\ \ldots\ (t-n),$$

$$\xi \in [\min\{x, x_0, x_1, \ldots, x_n\};\ \max\{x, x_0, \ldots, x_n\}].$$

Отже, якщо функція $f(x)$ має на заданому відрізку неперервну $(n+2)$-гу похідну, то для залишкового члена формули чисельного диференціювання (30.4) дістанемо:

$$r_n(x) = R_n'(x) = \frac{d\bar{R}_n(t)}{dt}\frac{dt}{dx} =$$

$$= \frac{h^n}{(n+1)!}\left(f^{(n+1)}(\xi)\frac{d}{dt}\Pi_{n+1}(t) + \Pi_{n+1}(t)\frac{d}{dt}f^{(n+1)}(\xi)\right).$$

Звідси для похибки у вузлових точках x_i запишемо:

$$r_n(x_i) = (-1)^{n-i}h^n\frac{(n-i)!\,i!}{(n+1)!}f^{(n+1)}(\xi).$$

Зокрема, у точці $x = x_0$ маємо:

$$r_n(x_0) = (-1)^n h^n \frac{1}{n+1}f^{(n+1)}(\xi),$$

де ξ — проміжне значення між $x_0, x_1, \ldots, x_n$.

Отже, у вузлах x_i для залишкового члена $r_n(x)$ маємо таку оцінку:

$$|r_n(x_i)| \leqslant \frac{(n-i)!\,i!}{(n+1)!}M_{n+1}h^n = O(h^n), \tag{30.9}$$

де

$$M_{n+1} = \max_{x_0 \leqslant x \leqslant x_n}|f^{(n+1)}(x)|.$$

Звичайно, якщо аналітичний вираз функції $f(x)$ невідомий чи складний, то оцінкою (30.9) користуватися важко або й зовсім неможливо. А коли відомо, що функція $f(x)$ має на заданому відрізку неперервну похідну $f^{(n+1)}(x)$, то для M_{n+1} можна взяти наближену рівність $M_{n+1} \approx \frac{|\Delta^{n+1}f(x_0)|}{|h|^{n+1}}$ (див. § 26).

Розглянута похибка є похибкою методу. З (30.9) видно, що похибка методу прямує до нуля, якщо $h \to 0$. Крім похибки

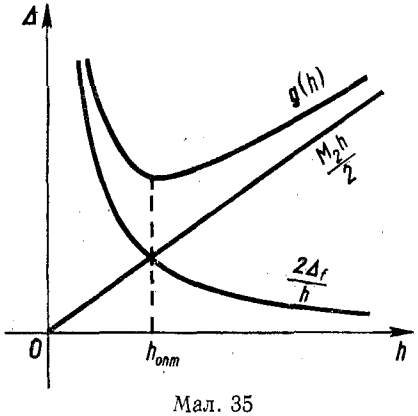
методу, у похибку результату входить неусувна похибка (зумовлена неточністю табличних значень функції), а також похибки округлень проміжних результатів.

Якщо припустити, що абсолютні похибки табличних значень функції не перевищують величини Δ_f і, врахувавши, що абсолютні похибки скінченних різниць $\Delta^k y_i$ не перевищуватимуть $2^k\Delta_f$ (див. § 26), то для неусувної похибки дістанемо оцінку:

$$\Delta(f'(x_0)) \leqslant \frac{1}{h}\left(2\Delta_f + \frac{1}{2}\cdot 4\Delta_f + \frac{1}{3}\cdot 8\Delta_f + \cdots + \frac{1}{n}\cdot 2^n\Delta_f\right) = \frac{\Delta_f}{h}\sum_{k=1}^{n}\frac{2^k}{k}. \qquad (30.10)$$

Отже, неусувна похибка має порядок $O\left(\frac{1}{h}\right)$, тобто обернено пропорційна кроку h. Наприклад, для формули $y_1' = \frac{1}{2h}\times$ $\times (y_2 - y_0)$ дістанемо неусувну похибку $\Delta y_1' = \frac{\Delta_f}{h}$.

Таким чином, із зменшенням кроку h похибка методу чисельного диференціювання зменшується, а неусувна похибка збільшується.



Мал. 35

Похибкою округлень проміжних результатів нехтуватимемо, вважаючи, що проміжні обчислення проводяться із запасними знаками. Повна похибка результату дорівнює сумі похибок методу і неусувної.

Якщо крок зменшується, але залишається ще великим, то повна похибка результату зменшується, бо неусувна похибка мала порівняно з похибкою методу. Якщо крок зменшувати й далі, неусувна похибка стане помітною. Нарешті, при досить малому кроці неусувна похибка буде значно більшою за похибку методу.

Із зменшенням кроку результат обчислень стає дедалі менш достовірним (позначається вплив некоректності задачі).

Природно виникає завдання: вибрати такий крок диференціювання, при якому похибка буде мінімальною.

Нехай формула чисельного диференціювання має вигляд: $f'(x_0) = \frac{\Delta y_0}{h} = \frac{f(x_1) - f(x_0)}{h}$ (формула (30.6), якщо $n = 1$).

Для похибки методу цієї формули дістанемо оцінку (формула (30.9)): $|r_1(x_0)| \leqslant \frac{M_2}{2}h$; неусувна похибка не перевищуватиме величини $\frac{2\Delta_f}{h}$. Таким чином, повна похибка не більша за $g(h) = \frac{M_2}{2}h + \frac{2\Delta_f}{h}$. Графік функції $g(h)$ зображений на мал. 35. Мінімізуючи $g(h)$, знайдемо оптимальний крок диференціювання. З умови $g'(h) = 0$ дістанемо:

$$h_{\text{опт}} = 2\sqrt{\frac{\Delta_f}{M_2}}.$$

При цьому повна похибка не перевищує величини $g(h_{\text{опт}}) = 2\sqrt{M_2\Delta_f}$.

Приклад.

У точках $x = 0{,}3$; $x = 0{,}45$; $x = 1{,}05$ обчислити першу похідну від функції, заданої таблицею:

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0,30	0,29552				
		13 945			
0,45	0,43497		— 978		
		12 967		— 289	
0,60	0,56464		— 1267		25
		11 700		— 264	
0,75	0,68164		— 1531		35
		11 169		— 229	
0,90	0,78333		— 1760		
		8409			
1,05	0,86742				

Оскільки четверті різниці практично сталі, то покладемо $n = 4$. Точка $x = 0{,}3$ розміщена на початку таблиці, тому похідну обчислимо за формулою (30.6). Матимемо: $x_0 = 0{,}30$; $\Delta y_0 = 0{,}13945$; $\Delta^2 y_0 = -0{,}00978$; $\Delta^3 y_0 = -0{,}00289$; $\Delta^4 y_0 = 0{,}00025$.

Виконавши обчислення, знайдемо:

$$f'(0{,}30) \approx \frac{1}{0{,}15}\left(0{,}13945 - \frac{1}{2}(-0{,}00978) + \frac{1}{3}(-0{,}00289) - \frac{1}{4}\cdot 0{,}00025\right) = 0{,}9554.$$

Значення похідної в точці $x = 0{,}45$ обчислимо за тією самою формулою, але покладемо $x_0 = 0{,}45$ і $\Delta y_0 = 0{,}12967$; $\Delta^2 y_0 = -0{,}01267$, $\Delta^3 y_0 = -0{,}00264$; $\Delta^4 y_0 = 0{,}00035$.

Виконавши обчислення, матимемо:

$$f'(0{,}45) \approx \frac{1}{0{,}15}\left(0{,}12967 - \frac{1}{2}(-0{,}01267) + \right.$$
$$\left. + \frac{1}{3}(-0{,}00264) - \frac{1}{4}0{,}00035\right) = 0{,}9002.$$

Похідну в точці $x = 1{,}05$ обчислимо за формулою (30.8), поклавши $x_n = 1{,}05$ $(n = 4)$; $\Delta y_3 = 0{,}08409$; $\Delta^2 y_2 = -0{,}01760$; $\Delta^3 y_1 = -0{,}00229$; $\Delta^4 y_0 = 0{,}00035$. Тоді

$$f'(1{,}05) \approx \frac{1}{0{,}15}\left(0{,}08409 + \frac{1}{2} - (0{,}01760) + \right.$$
$$\left. + \frac{1}{3}(-0{,}00229) + \frac{1}{4}0{,}00035\right) = 0{,}4974.$$

Обчислимо, наприклад, похибку результату для похідної $f'(0{,}30)$. У розглянутому прикладі таблицею задана функція $y = \sin x$. З формули (30.9) $(n = 4, i = 0)$ для похибки методу маємо оцінку $|r_4(x_0)| \leqslant \frac{1}{5}h^4$. Тут $M_5 = \max|\sin^{(5)}(x)| \leqslant 1$.

З формули (30.10) дістанемо, що неусувна похибка не перевищує величини $\frac{\Delta f}{h}\sum_{k=1}^{4}\frac{2^k}{k}$. Тому похибка результату не більша за

$$\frac{1}{5}h^4 + \frac{\Delta f}{h}\sum_{k=1}^{4}\frac{2^k}{k} =$$
$$= \frac{1\cdot(0{,}15)^4}{5} + \frac{0{,}5\cdot 10^{-5}}{0{,}15}\cdot\frac{32}{3} = 0{,}00010 + 0{,}00036 = 0{,}00046.$$

Таким чином, у знайденому результаті $f'(0{,}30)$ можна гарантувати три правильні значущі цифри.

Для порівняння наведемо значення шуканої похідної (тут $f'(x) = \cos x$) з п'ятьма правильними значущими цифрами:

$f'(0{,}30) = 0{,}95534$; $f'(0{,}45) = 0{,}90045$; $f'(1{,}05) = 0{,}49757$.

Отже, всі значення, знайдені методом чисельного диференціювання, справді мають по три правильні значущі цифри. Для збільшення точності треба брати табличні значення функції з більшою кількістю правильних значущих цифр, а у формулах (30.6) і (30.8) — більше членів.

Для ілюстрації впливу числа взятих членів у формулах диференціювання на точність результату наведемо, наприклад, значення $f'(0{,}30)$, обчислені за формулою (30.6), якщо відповідно $n = 1, 2, 3$; $f'(0{,}30) \approx 0{,}9297$; $f'(0{,}30) \approx 0{,}9623$; $f'(0{,}30) \approx 0{,}9558$.

Значення $f'(0{,}45)$ обчислимо ще за формулою $y'_1 = \frac{1}{2h}\times$ $\times (y_2 - y_0)$:

$$f'(0{,}45) = \frac{1}{2\cdot 0{,}15}(0{,}56464 - 0{,}29552) = 0{,}8971.$$

Воно має лише дві правильні значущі цифри.

Вправи.

1. Обчислити $f'(c)$, $f'(d)$, $f'(s)$, якщо функція $f(x)$ задана таблицею значень (таблицю взяти з вправи 4, § 28).
а) $c = 0{,}1$; $d = 0{,}9$; $s = 0{,}15$; б) $c = 0{,}2$; $d = 1{,}0$; $s = 0{,}95$.

2. За формулами (30.5) знайти $f'(c)$, $f'(d)$ і $f'(s)$ та оцінити похибки результатів, якщо функція $f(x)$ задана таблицею:

x	1,00	1,05	1,10	1,15	$c = 1{,}00$;
$f(x) = \frac{1}{x}$	1,00000	0,95238	0,90909	0,86957	$d = 1{,}15$; $s = 1{,}10$.

3. Встановити оптимальні кроки диференціювання для формул (30.5).

4. Вивести формули чисельного диференціювання для обчислення похідних другого порядку.

5. Вивести формули для знаходження у вузлових точках похідних першого і другого порядків, в яких використовуються табличні значення функції, а не їх різниці. Розглянути чотиричленні (вузли x_0, x_1, x_2, x_3) і п'ятичленні (вузли x_0, x_1, x_2, x_3, x_4) формули. Оцінити порядок точності цих формул.

§ 31. Чисельне інтегрування

Якщо для визначеної і неперервної на проміжку $[a; b]$ функції $f(x)$ відома первісна $F(x)$, то визначений інтеграл $\int_a^b f(x)\,dx$ можна обчислити за формулою Ньютона — Лейбніца:

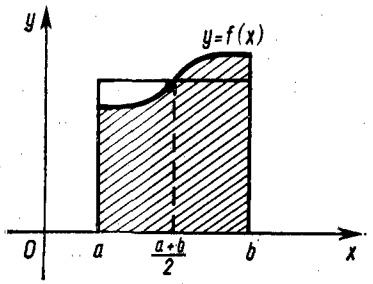
$$\int_a^b f(x)\,dx = F(b) - F(a), \qquad (31.1)$$

де $F'(x) = f(x)$.

Проте в багатьох випадках обчислити визначений інтеграл за цією формулою неможливо, оскільки знайти первісну $F(x)$ через елементарні функції, як правило, не вдається. Навіть тоді, коли її можна визначити, вона часто має досить складний і незручний для обчислень вигляд. Крім того, на практиці підінтегральна функція часто задається таблично і в такому разі аналітичні методи просто незастосовні. У цих випадках для

обчислення визначених інтегралів користуються чисельними методами.

Чисельне інтегрування — це обчислення значення визначеного інтеграла через ряд значень підінтегральної функції та її похідних. Оскільки знаходження числового значення визначеного інтеграла $\int_a^b f(x)\,dx$ (якщо $f(x) \geqslant 0$) з геометричного погляду можна тлумачити як обчислення площі криволінійної трапеції (її квадратури), обмеженої віссю Ox, прямими $x = a$, $x = b$ і лінією $y = f(x)$ (мал. 36), то формули для наближеного обчислення визначеного інтеграла називаються *квадратурними*.



Мал. 36

Найширше застосовуються квадратурні формули, які дають можливість наближено відшукувати значення інтеграла у вигляді лінійної комбінації кількох значень підінтегральної функції

$$\int_a^b f(x)\,dx \approx \sum_{k=0}^{n} A_k f(x_k), \qquad (31.2)$$

де A_k — *коефіцієнти формули* (дійсні числа), x_k — *вузли формули*.

Якщо задано деякий клас функцій і для нього будуємо квадратурну формулу типу (31.2), то коефіцієнти і вузли формули не повинні залежати від вибору функції $f(x)$ з даного класу функцій.

Величина

$$R(f) = \int_a^b f(x)\,dx - \sum_{k=0}^{n} A_k f(x_k) \qquad (31.3)$$

називається *залишковим членом квадратурної формули* (похибкою формули).

Можливі різні підходи до побудови квадратурних формул такого типу.

1. Квадратурні формули найвищого алгебраїчного степеня точності. Вузли x_k і коефіцієнти A_k формули (31.2) вибиратимемо так, щоб формула була точною (тобто щоб залишковий член дорівнював нулю) для алгебраїчних многочленів максимально високого степеня. Такі формули називаються *квадратурними формулами найвищого алгебраїчного степеня точності*.

Вважають, що *формула* (31.2) *має алгебраїчний степінь точності s, якщо вона точна для функцій* x^i $(i = 0, 1, \ldots, s)$ *і не точна для* x^{s+1}. Квадратурна формула, алгебраїчний степінь точності якої s, буде точною для всіх алгебраїчних многочленів степеня $\leqslant s$, а не тільки для многочленів виду x^i.

Зауважимо, що не завжди при побудові квадратурних формул вузли x_k і коефіцієнти A_k можна вибирати довільно. Часто вони повинні задовольняти деякі обмеження. Наприклад, при інтегруванні таблично заданих функцій за x_k можна брати лише табличні значення аргументу. У цьому випадку вузли наперед відомі і для побудови формули треба підібрати відповідним чином лише коефіцієнти A_k.

2. Квадратурні формули з найменшою оцінкою для залишкового члена. Нехай на відрізку $[a; b]$ задано деякий клас функцій H (наприклад, неперервних функцій, неперервно диференційовних функцій, функцій, що мають на відрізку $[a; b]$ неперервну другу похідну і т. ін.). Вузли x_k і коефіцієнти A_k формули (31.2) (з обмеженнями на них або без обмежень) треба вибрати так, щоб при заданому n величина

$$\sup_{f \in H} |R(f)| = \sup_{f \in H} \left| \int_a^b f(x)\,dx - \sum_{k=0}^{n} A_k f(x_k) \right|$$

була найменшою. Такі квадратурні формули називаються *формулами з найменшою або найкращою оцінкою залишкового члена*. Ідеться про такий вибір вузлів x_k і коефіцієнтів A_k квадратурної формули, при якому наближення, що дає квадратурна формула для всього класу функцій H, було б найкращим серед усіх можливих наближень. Звичайно, для кожної окремої функції $f \in H$ залишковий член квадратурної формули (31.2) має конкретне числове значення, яке виражається формулою (31.3).

§ 32. Залишковий член квадратурної формули. Оцінка похибки чисельного інтегрування

Однією з характеристик квадратурної формули є оцінка її залишкового члена. За нею можна визначити, яка з двох квадратурних формул точніша для даного класу функцій. Точнішою, як відомо, для функцій деякого класу H є та квадратурна формула, для якої величина $\sup_{f \in H} |R(f)|$ менша. Слід пам'ятати, що для одного класу функцій може виявитися точнішою одна формула, а для другого — інша. Для знаходження таких

оцінок можна використати характерне для даного класу подання функцій. Покажемо, як можна дістати вирази залишкового члена квадратурної формули для класів функцій різного порядку диференційовності.

Розглянемо функції $f(x)$, які визначені на відрізку $[a; b]$ і мають на ньому неперервні похідні до r-го порядку включно, тобто розглядатимемо функції класу $C^{(r)}[a; b]$.

Для функцій класу $C^{(r)}[a; b]$ справджується формула Тейлора із залишковим членом в інтегральній формі:

$$f(x) = \sum_{i=0}^{r-1} f^{(i)}(a)\,\frac{(x-a)^i}{i!} + \frac{1}{(r-1)!}\int_a^x f^{(r)}(t)\,(x-t)^{r-1}\,dt. \quad (32.1)$$

Формулу (32.1) легко дістати, послідовно застосувавши до інтеграла $\frac{1}{(r-1)!}\int_a^x f^{(r)}(t)\,(x-t)^{r-1}\,dt$ метод інтегрування частинами:

$$\frac{1}{(r-1)!}\int_a^x f^{(r)}(t)\,(x-t)^{r-1}\,dt = \frac{(x-t)^{r-1}}{(r-1)!}\,f^{(r-1)}(t)\Big|_a^x +$$

$$+\frac{1}{(r-2)!}\int_a^x f^{(r-1)}(t)(x-t)^{r-2}\,dt = -\frac{(x-a)^{r-1}}{(r-1)!}\,f^{(r-1)}(a) +$$

$$+\frac{(x-t)^{r-2}}{(r-2)!}\,f^{(r-1)}(t)\Big|_a^x + \frac{1}{(r-3)!}\int_a^x f^{(r-2)}(t)\,(x-t)^{r-3}\,dt = \cdots$$

$$\cdots = -\frac{(x-a)^{r-1}}{(r-1)!}\,f^{(r-1)}(a) - \frac{(x-a)^{r-2}}{(r-2)!}\,f^{(r-2)}(a) - \cdots$$

$$\cdots - \frac{x-a}{1!}\,f'(a) - f(a) + f(x).$$

Формулу (32.1) перепишемо так:

$$f(x) = \sum_{i=0}^{r-1} f^{(i)}(a)\,\frac{(x-a)^i}{i!} + \int_a^b f^{(r)}(t)\,E(x-t)\cdot\frac{(x-t)^{r-1}}{(r-1)!}\,dt, \quad (32.2)$$

де

$$E(x) = \begin{cases} 1, & \text{якщо } x \geqslant 0, \\ 0, & \text{якщо } x < 0. \end{cases}$$

Вираз (32.2) використаємо для того, щоб дістати характерне подання для залишкового члена квадратурної формули (31.2).

Позначивши через $R(\varphi)$ залишковий член квадратурної формули для функції φ і підставивши в формулу (31.3) вираз (32.2), дістанемо:

$$R(f) = \sum_{i=0}^{r-1}\frac{f^{(i)}(a)}{i!}\,R\left((x-a)^i\right) +$$

$$+ R\left(\int_a^b f^{(r)}(t)\,E(x-t)\,\frac{(x-t)^{r-1}}{(r-1)!}\,dt\right).$$

Знайдемо вираз для другого доданка в $R(f)$ (див. (31.3)):

$$R\left(\int_a^b f^{(r)}(t)\,E(x-t)\,\frac{(x-t)^{r-1}}{(r-1)!}\,dt\right) =$$

$$= \int_a^b\left(\int_a^b f^{(r)}(t)\,E(x-t)\,\frac{(x-t)^{r-1}}{(r-1)!}\,dt\right)dx -$$

$$-\sum_{k=0}^{n} A_k\int_a^b f^{(r)}(t)\,E(x_k-t)\,\frac{(x_k-t)^{(r-1)}}{(r-1)!}\,dt.$$

Змінивши порядок інтегрування в подвійному інтегралі, матимемо:

$$R\left(\int_a^b f^{(r)}(t)\,E(x-t)\,\frac{(x-t)^{r-1}}{(r-1)!}\,dt\right) =$$

$$= \int_a^b f^{(r)}(t)\int_a^b E(x-t)\,\frac{(x-t)^{r-1}}{(r-1)!}\,dx\,dt -$$

$$-\int_a^b f^{(r)}(t)\sum_{k=0}^{n} A_k E(x_k-t)\,\frac{(x_k-t)^{r-1}}{(r-1)!}\,dt =$$

$$= \frac{1}{(r-1)!}\int_a^b f^{(r)}(t)\left(\int_a^b E(x-t)\,(x-t)^{r-1}\,dx -\right.$$

$$\left. -\sum_{k=0}^{n} A_k E(x_k-t)\,(x_k-t)^{r-1}\right)dt.$$

Врахувавши, що

$$\int_a^b E(x-t)\,(x-t)^{r-1}\,dx = \int_t^b (x-t)^{r-1}\,dx = \frac{(b-t)^r}{r},$$

дістанемо:

$$R(f)=\sum_{i=0}^{r-1}\frac{f^{(i)}(a)}{i!}R\left((x-a)^i\right)+\frac{1}{(r-1)!}\int_a^b f^{(r)}(t)\,K_r(t)\,dt, \quad (32.3)$$

де

$$K_r(t)=\frac{(b-t)^r}{r}-\sum_{k=0}^{n}A_k E\left(x_k-t\right)(x_k-t)^{r-1}. \quad (32.4)$$

Формула (32.3) дає вираз для залишкового члена $R(f)$ будь-якої квадратурної формули в класі функцій, які мають на $[a;\ b]$ неперервні похідні до порядку r.

Якщо квадратурна формула (31.2) точна для многочленів степеня $\leqslant r-1$, то для них $R\left((x-a)^i\right)=0$ $(i=0,\ 1,\ \ldots,\ r-1)$ і залишковий член набирає вигляду:

$$R(f)=\frac{1}{(r-1)!}\int_a^b f^{(r)}(t)\,K_r(t)\,dt. \quad (32.5)$$

Звідси дістаємо таку оцінку для залишкового члена:

$$|R(f)|\leqslant\frac{M_r}{(r-1)!}\int_a^b|K_r(t)|\,dt, \quad (32.6)$$

де $M_r=\max\limits_{x\in[a;b]}|f^{(r)}(x)|$.

Причому ця оцінка є точною верхньою границею в класі $C^{(r)}[a;\ b]$:

$$\sup_{f\in C^{(r)}[a;b]}|R(f)|=\frac{M_r}{(r-1)!}\int_a^b|K_r(t)|\,dt.$$

Звичайно, якщо функція $K_r(t)$ на відрізку $[a;\ b]$ не змінює знака, то, скориставшись теоремою про середнє, для залишкового члена з (32.5) дістанемо такий вираз:

$$R(f)=\frac{1}{(r-1)!}f^{(r)}(\xi)\int_a^b K_r(t)\,dt,\quad \xi\in\,]a;\ b[.$$

Отже, *з двох квадратурних формул, точних для многочленів степеня, не вищого за* $r-1$, *у класі* $C^{(r)}[a;\ b]$ *точнішою буде та формула, для якої величина* $\int_a^b|K_r(t)|\,dt$ *є меншою.*

Таким чином, спосіб побудови квадратурних формул з найменшою оцінкою залишкового члена в класі $C^{(r)}[a;\ b]$

полягає в тому, що вузли x_k і коефіцієнти A_k $(k=0,\ 1,\ \ldots,n)$ вибирають так, щоб величина $\int_a^b|K_r(t)|\,dt$ була найменшою.

Розглянемо тепер питання про похибку чисельного інтегрування. Крім похибки, що виникає внаслідок відкидання залишкового члена (похибки методу), виникає похибка, зумовлена виконанням дій з наближеними числами (у процесі обчислень майже завжди доводиться мати справу з наближеними значеннями $f(x_k)$, в яких правильні тільки кілька значущих цифр). Нехай, наприклад, всі значення $f(x_k)$ обчислені наближено, причому абсолютні похибки їх не перевищують числа Δ_f. Обчисливши за допомогою наближених значень $f(x_k)$ квадратурну суму $\sum\limits_{k=0}^{n}A_k f(x_k)$, при точних значеннях A_k, дістанемо похибку

$$\Delta_\Sigma=\Delta_f\sum_{k=0}^{n}|A_k|. \quad (32.7)$$

Це неусувна похибка квадратурної формули.

Отже, якщо сума $\sum\limits_{k=0}^{n}|A_k|$ велика, то навіть незначні похибки в значеннях $f(x_k)$ можуть призвести до великої похибки в наближеному значенні інтеграла. Тому практичну цінність мають лише такі квадратурні формули, для яких сума $\sum\limits_{k=0}^{n}|A_k|$ невелика. Якщо квадратурна формула точна для $f(x)=1$, то неважко встановити умову, за якої сума $\sum\limits_{k=0}^{n}|A_k|$ набуває найменших значень. Справді, формула точна для $f(x)=1$ тоді й тільки тоді, коли $\int_a^b 1\cdot dx=b-a=\sum\limits_{k=0}^{n}A_k$. З останнього випливає, що $\sum\limits_{k=0}^{n}|A_k|$ матиме найменше значення, коли всі A_k будуть додатні. Тому квадратурні формули з додатними коефіцієнтами використовуються найчастіше.

При знаходженні квадратурної суми, як і раніше, виконуватимемо обчислення із запасними цифрами, тому похибками округлення проміжних результатів нехтуватимемо. Похибкою округлень проміжних результатів можна також нехтувати під час виконання обчислень на ЕОМ, оскільки в них числа записуються з достатньою кількістю знаків, і тому похибка окре-

мого округлення, як правило, значно менша від похибки методу і неусувної похибки. Отже, *повна похибка чисельного інтегрування дорівнює сумі трьох похибок: похибки методу, неусувної похибки* Δ_Σ *та заключної похибки округлення результату* Δ_0.

Надалі підінтегральну функцію у вузлах x_k обчислюватимемо (якщо це можливо) з такою точністю, щоб похибка квадратурної суми Δ_Σ була значно меншою від похибки методу, тобто, щоб обчислити значення інтеграла з наперед заданою точністю $\varepsilon > 0$, обчислюватимемо $f(x_k)$ так, щоб похибка (32.7) була значно меншою від ε (менша за ε принаймні в 10 раз).

§ 33. Інтерполяційні квадратурні формули

Для побудови квадратурних формул, так само як і для формул чисельного диференціювання, можна використати інтерполяційний многочлен, а саме: підінтегральну функцію $f(x)$ на відрізку інтегрування замінити інтерполяційним многочленом $P_n(x)$ і вважати, що інтеграл від інтерполяційного многочлена наближено дорівнює інтегралу від заданої функції

$$\int_a^b f(x)\,dx \approx \int_a^b P_n(x)\,dx. \tag{33.1}$$

Якщо для інтерполяційного многочлена відомий залишковий член

$$R_n(x) = f(x) - P_n(x),$$

то можна дістати вираз для залишкового члена квадратурної формули

$$R(f) = \int_a^b f(x)\,dx - \int_a^b P_n(x)\,dx = \int_a^b R_n(x)\,dx,$$

тобто залишковий член квадратурної формули дорівнює інтегралу від залишкового члена інтерполяційного многочлена.

Виберемо на відрізку інтегрування $[a; b]$ точки $x_0, x_1, \ldots, x_n$ ($x_i \neq x_j$, якщо $i \neq j$). Побудуємо інтерполяційний многочлен, значення якого в заданих точках дорівнюють значенням підінтегральної функції $f(x)$. Інтерполяційний многочлен запишемо у формі Лагранжа:

$$L_n(x) = \Pi_{n+1}(x) \sum_{k=0}^{n} \frac{f(x_k)}{(x - x_k)\,\Pi'_{n+1}(x_k)},$$

де

$$\Pi_{n+1}(x) = (x - x_0)(x - x_1) \ldots (x - x_n).$$

Підставивши в (33.1) вираз для многочлена Лагранжа, дістанемо таку квадратурну формулу:

$$\int_a^b f(x)\,dx \approx \sum_{k=0}^{n} A_k f(x_k), \tag{33.2}$$

де

$$A_k = \int_a^b \frac{\Pi_{n+1}(x)}{(x - x_k)\,\Pi'_{n+1}(x_k)}\,dx. \tag{33.3}$$

Квадратурні формули (33.2), коефіцієнти яких виражаються формулами (33.3), називаються *інтерполяційними*.

Якщо функція $f(x)$ на відрізку $[a; b]$ має неперервну похідну $(n+1)$-го порядку, тобто $f(x) \in C^{(n+1)}[a; b]$, то, скориставшись виразом (25.2) для залишкового члена інтерполяційного многочлена Лагранжа, запишемо вираз для залишкового члена інтерполяційної квадратурної формули (33.2), (33.3):

$$R(f) = \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\xi)\,\Pi_{n+1}(x)\,dx. \tag{33.4}$$

Звідси до похибки квадратурної формули дістанемо таку оцінку:

$$|R(f)| \leqslant \frac{M_{n+1}}{(n+1)!} \int_a^b |\Pi_{n+1}(x)|\,dx,$$

де $M_{n+1} = \max\limits_{x \in [a;b]} |f^{(n+1)}(x)|$.

Зауважимо, що ця оцінка не є точною для функцій класу $C^{(n+1)}[a; b]$. Рівність досягається лише за умови, що $\Pi_{n+1}(x)$ на $[a; b]$ не змінює знака. В інших випадках оцінка може бути далекою від оптимальної. Точну оцінку для будь-якої квадратурної формули, в тому числі й інтерполяційної, можна дістати за формулою (32.6).

З формули (33.4) випливає, що алгебраїчний степінь точності інтерполяційної квадратурної формули (33.2), (33.3) не нижчий за n. При вдалому виборі вузлів степінь точності може бути і більшим за n.

Так, якщо за вузли взяти корені ортогональних многочленів, то можна добитися, що формули (33.2), (33.3) матимуть алгебраїчний степінь точності $2n + 1$. Вони є формулами найвищого алгебраїчного степеня точності і називаються *квадратурними формулами Гаусса*.

§ 34. Квадратурні формули Ньютона—Котеса

Розглянемо інтерполяційні квадратурні формули, в яких вузли $x_k \in [a;\ b]$ рівновіддалені. Такі формули називаються *формулами Ньютона — Котеса*. Їх вперше розглянув Ньютон, а коефіцієнти для них при $n \leqslant 9$ знайшов Котес. Дослідження таких формул показали, що коли $n \geqslant 10$, то серед коефіцієнтів A_k є від'ємні і $\lim\limits_{n\to\infty} \sum\limits_{k=0}^{n} |A_k| = \infty$. Отже, при великих n похибка Δ_Σ квадратурної суми буде великою навіть при малих похибках в значеннях функції $f(x_k)$. Тому на практиці квадратурні формули Ньютона — Котеса для великих n не використовуються. Розглянемо частинні випадки формул Ньютона — Котеса.

Вважатимемо, що для формул Ньютона — Котеса, які містять не менш як два доданки ($n \geqslant 1$), вузли x_k розміщено так, що $x_0 = a$, $x_n = b$, $x_{k+1} = x_k + h$ $(k = 0,\ 1,\ \ldots,\ n-1)$. Крок h в даному випадку дорівнює $h = \frac{b-a}{n}$. У випадку $n = 0$ за єдиний вузол x_0 можна взяти будь-яку точку на відрізку $[a;\ b]$.

1. Квадратурні формули прямокутників.

Нехай $n = 0$, тоді формула (33.2), (33.3) набирає вигляду:

$$\int_a^b f(x)\,dx \approx (b-a) f(x_0). \tag{34.1}$$

Ця формула називається *формулою прямокутників*. Поклавши $x_0 = a$, або $x_0 = b$, або $x_0 = \frac{a+b}{2}$, дістанемо три формули (лівих, правих і середніх) прямокутників.

Формулу прямокутників можна легко вивести на основі геометричних міркувань.

Якщо для додатної неперервної на $[a;\ b]$ функції площу криволінійної трапеції, що дорівнює точному значенню інтеграла, наближено замінити площею прямокутника, висота якого $f(x_0)$, то дістанемо формулу (34.1). Ці формули мають зміст для будь-якої неперервної на $[a;\ b]$ функції, не обов'язково додатної. Мал. 36 є геометричною ілюстрацією формули середніх прямокутників. Залишкові члени формул прямокутників дістанемо з формули (33.4), якщо $n = 0$, тобто

$$R(f) = \int_a^b f'(\xi)(x - x_0)\,dx.$$

Для формул лівих ($x_0 = a$) і правих ($x_0 = b$) прямокутників функція $\Pi_1(x) = x - x_0$ на $[a;\ b]$ не змінює знака. Якщо $f'(x)$ неперервна, то, скориставшись теоремою про середнє, відповідно матимемо:

$$R(f) = \frac{(b-a)^2}{2} f'(\eta), \quad \eta \in]a;\ b[, \tag{34.2}$$

$$R(f) = -\frac{(b-a)^2}{2} f'(\eta_1), \quad \eta_1 \in]a;\ b[. \tag{34.3}$$

Для формули середніх прямокутників функція $\Pi_1(x) = x - \frac{a+b}{2}$ на $[a;\ b]$ змінює знак і тому скористатися теоремою про середнє не вдається. Формули прямокутників даватимуть точний результат, якщо функція $f(x)$ на $[a;\ b]$ константа.

Проте навіть з геометричних міркувань (мал. 36) видно, що формула середніх прямокутників даватиме точний результат не тільки тоді, коли $f(x)$ буде константою, а й тоді, коли функція $y = f(x)$ на відрізку $[a;\ b]$ буде довільною лінійною функцією $Ax + B$. Крім того, формула середніх прямокутників дає точний результат і для всіх непарних відносно середини відрізка $[a;\ b]$ функцій $f(x)$.

Для функцій $f(x)$, які мають на відрізку $[a;\ b]$ неперервну другу похідну (клас $C^{(2)}$ $[a;\ b]$) виведемо уточнений вираз для залишкового члена. Оскільки формула середніх прямокутників точна для многочленів першого степеня, то шуканий вираз для залишкового члена можна дістати за формулою (32.5), де слід покласти $r = 2$. За формулою (32.4) знайдемо функцію $K_2(t)$:

$$K_2(t) = \frac{(b-t)^2}{2} - (b-a) E\left(\frac{a+b}{2} - t\right)\left(\frac{a+b}{2} - t\right),$$

де

$$E(x) = \begin{cases} 1, & x \geqslant 0, \\ 0, & x < 0, \end{cases}$$

тобто

$$K_2(t) = \begin{cases} \dfrac{(b-t)^2}{2} - (b-a)\left(\dfrac{a+b}{2} - t\right), & \text{якщо } a \leqslant t \leqslant \dfrac{a+b}{2}, \\ \dfrac{(b-t)^2}{2}, & \text{якщо } \dfrac{a+b}{2} < t \leqslant b. \end{cases}$$

Легко встановити, що функція $K_2(t)$ симетрична відносно прямої $x = \frac{a+b}{2}$ і на відрізку $[a;\ b]$ не змінює знака. Графік її показаний на мал. 37.

Таким чином, маємо (формула (32.5)):

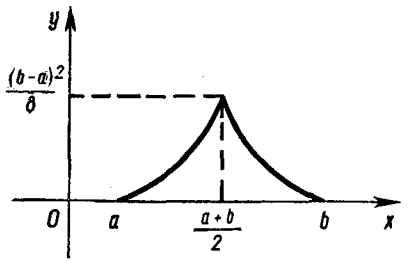
$$R(f) = \int_a^b f''(t)\, K_2(t)\, dt = f''(\xi) \int_a^b K_2(t)\, dt =$$

$$= 2f''(\xi) \int_{\frac{a+b}{2}}^{b} \frac{(b-t)^2}{2}\, dt, \quad \xi \in]a;\ b[.$$

Ми скористалися теоремою про середнє, оскільки $K_2(t)$ не змінює знака на $[a;\ b]$. Обчисливши останній інтеграл, остаточно знаходимо:

$$R(f) = \frac{(b-a)^3}{24} f''(\xi),$$

$$\xi \in]a;\ b[. \qquad (34.4)$$



Мал. 37

1.1. Загальні формули прямокутників. Якщо відрізок інтегрування $[a;\ b]$ великий, то похибка формули (34.1) може бути досить значною. Щоб цьому запобігти, розбивають відрізок на частини і застосовують формули до кожної з них. Розділимо, наприклад, відрізок інтегрування $[a;\ b]$ на m рівних частин: $[x_0;\ x_1]$, $[x_1;\ x_2]$,...,$[x_{m-1};\ x_m]$ завдовжки $h = \frac{b-a}{m}$, де $x_0 = a$, $x_m = b$. За властивістю інтеграла маємо:

$$\int_a^b f(x)\, dx = \sum_{i=1}^{m} \int_{x_{i-1}}^{x_i} f(x)\, dx.$$

Позначивши $y_{i-\frac{1}{2}} = f\left(x_i - \frac{h}{2}\right)$ $(i = 1,\ 2,\ ...,\ m)$ і обчисливши кожний з інтегралів $\int_{x_{i-1}}^{x_i} f(x)\, dx$ $(i = 1,\ 2,\ ...,\ m)$, наприклад за формулою середніх прямокутників, матимемо:

$$\int_{a=x_0}^{b=x_m} f(x)\, dx \approx h\left(y_{\frac{1}{2}} + y_{\frac{3}{2}} + \cdots + y_{m-\frac{1}{2}}\right) =$$

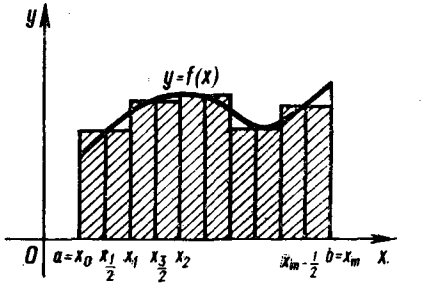
$$= h \sum_{i=1}^{n} f\left(a + \frac{2i-1}{2m}(b-a)\right). \qquad (34.5)$$

Формулу (34.5) називають *узагальненою формулою середніх прямокутників.*

Аналогічно дістанемо *узагальнені формули лівих і правих прямокутників.*

З геометричного погляду, якщо $f(x) \geqslant 0$, $x \in [a;\ b]$, права частина формули (34.5) виражає площу фігури, що складається із заштрихованих прямокутників (мал. 38).

Обчислимо тепер залишкові члени узагальнених формул прямокутників. Очевидно, залишковий член формули (34.5) дорівнюватиме сумі залишкових членів відповідних формул прямокутників, які ми застосували до кожного з інтегралів $\int_{x_{i-1}}^{x_i} f(x)\, dx\ (i = 1, 2, ..., m)$.



Мал. 38

Так, врахувавши формулу (34.4), для залишкового члена узагальненої формули середніх прямокутників матимемо:

$$R(f) = \sum_{i=1}^{m} \frac{h^3}{24} f''(\eta_i) = \frac{h^3}{24} \sum_{i=1}^{m} f''(\eta_i) =$$

$$= \frac{(b-a)\,h^2}{24} \cdot \frac{1}{m} \sum_{i=1}^{m} f''(\eta_i),$$

де $\eta_i \in]x_{i-1};\ x_i[$, $b - a = mh$.

Оскільки $f''(x)$ на відрізку $[a;\ b]$ неперервна, то знайдеться така точка $\xi \in]a;\ b[$, що $\frac{1}{m}(f''(\eta_1) + f''(\eta_2) + ... + f''(\eta_m)) = f''(\xi)$. Таким чином, для залишкового члена узагальненої квадратурної формули середніх прямокутників дістанемо такий вираз:

$$R(f) = \frac{(b-a)\,h^2}{24} f''(\xi), \quad \xi \in]a;\ b[, \quad h = \frac{b-a}{m}. \qquad (34.6)$$

Звідси дістанемо оцінку

$$|R(f)| \leqslant \frac{(b-a)\,h^2}{24} M_2, \qquad (34.7)$$

де $$M_2 = \max_{x \in [a;b]} |f''(x)|.$$

Ця формула дає оцінку залишкового члена квадратурної формули середніх прямокутників для функцій $f(x) \in C^{(2)}[a;\ b]$.

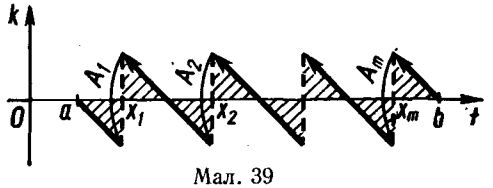
Для функцій $f(x) \in C^{(1)}[a; b]$ аналогічну оцінку можна дістати за формулою (32.6), поклавши $r = 1$, з (32.4) для формули (34.5), знайдемо вираз функції $K_1(t)$:

$$K_1(t) = b - t - \sum_{k=1}^{m} A_k E(x_k - t) =$$

$$= b - t - \frac{b-a}{m} \sum_{k=1}^{m} E\left(a + \frac{2k-1}{2m}(b-a) - t\right).$$

Графік цієї функції подано на мал. 39, де

$$x_k = a + \frac{2k-1}{2m}(b-a), \quad A_k = \frac{b-a}{m} \qquad (k = 1, 2, \ldots, m).$$



Мал. 39

Інтеграл $\int_a^b |K_1(t)|\,dt$, очевидно, дорівнює площі заштрихованих на мал. 39 трикутників. Усі $2m$ трикутники прямокутні й рівні між собою, а катети їх дорівнюють $\frac{b-a}{2m}$. Тому

$$\int_a^b |K_1(t)|\,dt = 2m \cdot \frac{1}{2}\left(\frac{b-a}{2m}\right)^2 = \frac{(b-a)^2}{4m}.$$

Отже, для залишкового члена формули середніх прямокутників в класі $C^{(1)}[a; b]$ маємо таку оцінку:

$$|R(f)| \leqslant M_1 \frac{(b-a)^2}{4m} = \frac{M_1 (b-a) h}{4}. \qquad (34.8)$$

Легко довести, що в класі $C^{(1)}[a; b]$ для будь-якої іншої квадратурної формули виду $\int_a^b f(x)\,dx \approx \sum_{k=1}^{m} A_k f(x_k)$, для якої $\sum_{k=1}^{m} A_k = b - a$ оцінка $|R(f)|$ буде більшою, тобто квадратурна формула середніх прямокутників в класі $C^{(1)}[a; b]$ серед формул такого типу є найкращою (з найменшою оцінкою залишкового члена).

Оскільки в квадратурних формулах прямокутників коефіцієнти додатні і сума їх для кожної з формул дорівнює $mh = b - a$, то похибка, зумовлена неточністю значень підінтегральної функції у вузлах (неусувна похибка), за (32.7) не перевищуватиме величини

$$\Delta_\Sigma = \Delta_f (b - a). \qquad (34.9)$$

Отже, гранична абсолютна похибка результату за квадратурною формулою середніх прямокутників дорівнює

$$\Delta_I = \frac{(b-a)\,h^2}{24} M_2 + \Delta_f (b - a) + \Delta_0, \qquad (34.10)$$

де M_2 — максимум другої похідної, Δ_0 — заключна похибка округлення результату.

Приклад.

Обчислити інтеграл $I = \int_0^1 \frac{dx}{1+x}$; $h = 0{,}1$. Оцінити похибку результатів.

У даному випадку $m = \frac{1-0}{0{,}1} = 10$.

Обчислимо інтеграл за формулою середніх прямокутників. Відповідні значення функції $\frac{1}{1+x}$ подано в таблиці.

$$I \approx 0{,}1\,(0{,}95238 + 0{,}86957 + 0{,}80000 + \cdots + 0{,}51282) =$$

$$= 0{,}692836 \approx 0{,}69284.$$

k	$x_{k+\frac{1}{2}}$	$x_{k+\frac{1}{2}} + 1$	$\frac{1}{x_{k+\frac{1}{2}} + 1}$
0	0,05	1,05	0,95238
1	0,15	1,15	0,86957
2	0,25	1,25	0,80000
3	0,35	1,35	0,74074
4	0,45	1,45	0,68966
5	0,55	1,55	0,64516
6	0,65	1,65	0,60606
7	0,75	1,75	0,57143
8	0,85	1,85	0,54054
9	0,95	1,95	0,51282

Оцінимо тепер похибку результату. Граничну абсолютну похибку результату визначимо за формулою (34.10).

У даному випадку

$$f''(x) = \frac{2}{(1+x)^3}, \quad M_2 = \max_{0 \leqslant x \leqslant 1} |f''(x)| = \max_{0 \leqslant x \leqslant 1} \left|\frac{2}{(1+x)^3}\right| = 2.$$

Значення підінтегральної функції у вузлах обчислювались з точністю до $0{,}5 \cdot 10^{-5}$. Тому гранична абсолютна похибка

значень $f(x_k)$ дорівнює $\Delta_f = 0{,}5 \cdot 10^{-5}$. Похибка округлення Δ_0 дорівнює $4 \cdot 10^{-6}$.

Підставивши потрібні дані у формулу (34.10), знаходимо:

$$\Delta_I = \frac{1 \cdot (0{,}1)^2}{24} \cdot 2 + 0{,}5 \cdot 10^{-5} \cdot 1 + 4 \cdot 10^{-6} = 0{,}00084.$$

Отже, у знайденому результаті можемо гарантувати правильність двох значущих цифр. Тому добутий результат потрібно записати так (залишаємо правильні цифри і одну сумнівну): 0,693.

Для порівняння наведемо значення шуканого інтеграла з вісьмома правильними значущими цифрами:

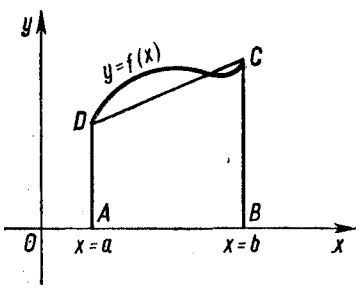
$$\int_0^1 \frac{dx}{1+x} = \ln 2 = 0{,}69314718.$$

2. Квадратурна формула трапецій. Розглянемо випадок квадратурної формули Ньютона — Котеса, яка має два вузли $x_0 = a$ і $x_1 = b$ $(n = 1)$.

Коефіцієнти A_k $(k = 1, 2)$ знаходимо з формули (33.3)

$$A_0 = \int_a^b \frac{(x-x_0)(x-x_1)}{(x-x_0)(x_0-x_1)}\,dx = \frac{b-a}{2};$$

$$A_1 = \int_a^b \frac{(x-x_0)(x-x_1)}{(x-x_1)(x_1-x_0)}\,dx = \frac{b-a}{2}.$$



Мал. 40

Тоді формула (33.2) набирає вигляду:

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2}(f(a)+f(b)).$$

(34.11)

Ми дістали квадратурну формулу, замінивши функцію $f(x)$ інтерполяційним многочленом першого степеня (лінійна інтерполяція), який в точках $x = a$ і $x = b$ набуває відповідно значень $f(a)$ і $f(b)$ (мал. 40).

Формула (34.11) називається *квадратурною формулою трапецій.*

Квадратурну формулу трапецій легко дістати також на основі геометричних міркувань. Так, у випадку додатної функції інтеграл $\int_a^b f(x)\,dx$ наближено замінюється числом, яке дорівнює площі трапеції $ABCD$ (мал. 40). Звідси і походить назва формули. Формула (34.11) застосовна до будь-якої неперервної функції (не обов'язково додатної).

Якщо $f(x)$ на відрізку $[a; b]$ має неперервну другу похідну, тобто $f(x) \in C^{(2)}[a; b]$, то з формули (33.4) дістанемо вираз для залишкового члена формули трапецій:

$$R(f) = \frac{1}{2}\int_a^b f''(\xi)(x-x_0)(x-x_1)\,dx.$$

Оскільки функція $(x-x_0)(x-x_1)$ не змінює знака на відрізку $[a; b]$, то до останнього інтеграла можна застосувати теорему про середнє:

$$R(f) = \frac{f''(\eta)}{2}\int_a^b (x-x_0)(x-x_1)\,dx = \frac{-(b-a)^3}{12} f''(\eta), \quad \eta \in]a; b[.$$

(34.12)

Формула трапецій точна для многочленів першого степеня, а для функції $f(x) = x^2$ вона вже не точна. Таким чином, алгебраїчний степінь точності формули трапецій, як і формули середніх прямокутників, дорівнює 1.

Якщо функція $f(x)$ має лише першу неперервну похідну $(f(x) \in C^{(1)}[a; b])$, то оцінку для залишкового члена можна дістати з формули (32.6). Функцію $K_1(t)$ знайдемо за формулою (32.4), якщо $r = 1$, $n = 1$, $A_0 = A_1 = \frac{b-a}{2}$, $x_0 = a$, $x_1 = b$.

$$K_1(t) = b - t - \frac{b-a}{2}E(a-t) - \frac{b-a}{2}E(b-t) =$$

$$= b - t - \frac{b-a}{2} = \frac{b+a}{2} - t.$$

Отже,

$$|R(f)| \leqslant M_1 \int_a^b |K_1(t)|\,dt = M_1 \int_a^b \left| -t + \frac{b+a}{2} \right| dt =$$

$$= M_1 \left(\int_a^{\frac{a+b}{2}} \left(\frac{b+a}{2} - t \right) dt + \int_{\frac{a+b}{2}}^b \left(t - \frac{b+a}{2} \right) dt \right) =$$

$$= M_1 \frac{(b-a)^2}{4}. \qquad (34.13)$$

Для формули трапецій у класі $C^{(1)}[a; b]$ ми дістали таку саму оцінку, як і для формули середніх прямокутників (див. формулу (34.8)).

2.1. Узагальнена квадратурна формула трапецій. Якщо відрізок $[a; b]$ великий, то й похибка формули (34.11) може бути великою. Щоб зменшити її, поділимо відрізок інтегрування $[a; b]$ на m рівних частин: $[x_0; x_1], [x_1; x_2], \ldots, [x_{m-1}; x_m]$ завдовжки $h = \frac{b-a}{m}$, причому $x_0 = a$, $x_m = b$. Тоді

$$\int_a^b f(x)\,dx = \sum_{k=1}^{m} \int_{x_{i-1}}^{x_i} f(x)\,dx.$$

До кожного з m інтегралів $\int_{x_{i-1}}^{x_i} f(x)\,dx$ $(i = 1, 2, \ldots, m)$ застосуємо формулу (34.11). Дістанемо:

$$\int_a^b f(x)\,dx \approx \frac{h}{2} \sum_{i=1}^{m} (y_{i-1} + y_i) =$$

$$= h\left(\frac{y_0}{2} + y_1 + y_2 + \cdots + y_{m-1} + \frac{y_m}{2}\right), \qquad (34\ 14)$$

де $y_i = f(x_i)$.

Формула (34.14) називається *узагальненою формулою трапецій.*

Залишковий член її, очевидно, дорівнює

$$R(f) = \sum_{i=1}^{m} R_i(f),$$

де $R_i(f)$ — залишковий член квадратурної формули на i-му відрізку. Врахувавши (34.12), маємо: $R_i(f) = -\frac{h^3}{12} f''(\eta_i)$, $\eta_i \in]x_{i-1}; x_i[$. Якщо $f(x) \in C^{(2)}[a; b]$, то для $R(f)$ дістанемо:

$$R(f) = -\frac{h^3}{12} \sum_{i=1}^{m} f''(\eta_i) = -\frac{h^3 \cdot m}{12} \cdot \frac{1}{m} \sum_{i=1}^{m} f''(\eta_i) =$$

$$= \frac{-h^3 m}{12} f''(\xi) = \frac{-(b-a)h^2}{12} f''(\xi), \quad \xi \in]a; b[. \qquad (34.15)$$

Позначивши $M_2 = \max\limits_{x \in [a;b]} |f''(x)|$, дістанемо:

$$|R(f)| \leqslant \frac{(b-a)h^2}{12} M_2. \qquad (34.16)$$

Як і в квадратурних формулах прямокутників, коефіцієнти формули трапецій додатні і сума їх дорівнює $mh = b - a$. Тому при обчисленнях за цією формулою неусувна похибка не перевищуватиме величини $\Delta_\Sigma = \Delta_f (b - a)$, де Δ_f — максимальна абсолютна похибка значень підінтегральної функції у вузлах.

Отже, для граничної абсолютної похибки результату, знайденого за узагальненою квадратурною формулою трапецій, матимемо:

$$\Delta_I = \frac{b-a}{12} h^2 M_2 + \Delta_f (b - a) + \Delta_0, \qquad (34.17)$$

де Δ_0 — похибка округлення остаточного результату.

Приклад.

Обчислити $\ln 2 = \int_0^1 \frac{dx}{1+x}$ за формулою трапецій з кроком $h = 0{,}1$ і оцінити похибку результату.

Для нашого випадку $m = 10$ (проміжок інтегрування поділено на 10 рівних частин). Значення підінтегральної функції у вузлових точках обчислюватимемо з п'ятьма правильними цифрами після коми:

$$\int_0^1 \frac{dx}{1+x} \approx h\left(\frac{y_0}{2} + y_1 + y_2 + y_3 + \cdots + y_9 + \frac{y_{10}}{2}\right) =$$

$$= 0{,}1\left(\frac{1}{2} + 0{,}90909 + 0{,}83333 + 0{,}76923 + 0{,}71429 +\right.$$

$$\left. + 0{,}66667 + 0{,}62500 + 0{,}58824 + 0{,}55556 + 0{,}52632 + \frac{0{,}5}{2}\right) =$$

$$= 0{,}693773 \approx 0{,}69377.$$

Обчислимо за формулою (34.17) похибку результату. Оскільки $f(x) = \frac{1}{1+x}$, то $f''(x) = \frac{2}{(1+x)^3}$.

Звідси $$M_2 = \max_{x \in [0;1]} \left|\frac{2}{(1+x)^3}\right| = 2.$$

Значення підінтегральної функції у вузлах обчислені з точністю до $0{,}5 \cdot 10^{-5}$, тому $\Delta_f = 0{,}5 \cdot 10^{-5}$. Похибка округлення $\Delta_0 = 0{,}3 \cdot 10^{-5}$. Підставивши значення потрібних величин у формулу (34.17), дістанемо:

$$\Delta_I = \frac{1}{12}(0{,}1)^2 \cdot 2 + 0{,}5 \cdot 10^{-5} \cdot 1 + 0{,}3 \cdot 10^{-5} =$$

$$= 0{,}001675 < 0{,}002.$$

Тому результат доцільно подати так: $I = 0{,}694$. Зрозуміло, що знайдена оцінка похибки завищена. Справді, абсолютна похибка результату $\ln 2 = |\,0{,}693147\ \ldots - 0{,}69377\,| < 0{,}000623$.

3. Квадратурна формула Сімпсона. Розглянемо ще один приклад квадратурної формули Ньютона — Котеса, яка широко використовується на практиці і називається *квадратурною формулою парабол*, або *формулою Сімпсона*. Цю формулу дістанемо з (33.2), (33.3), якщо $n = 2$. Вузлами тут є точки $x_0 = a$, $x_1 = \frac{a+b}{2}$, $x_2 = b$. Знайдемо коефіцієнти формули.

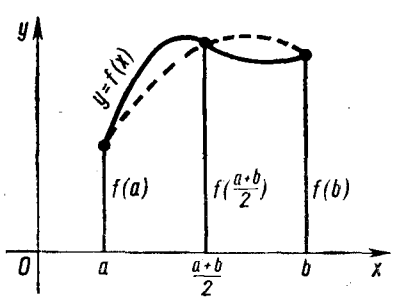
$$A_0 = \int_a^b \frac{(x-x_0)(x-x_1)(x-x_2)}{(x-x_0)\cdot\frac{1}{2}(a-b)(a-b)}\,dx = \frac{b-a}{6};$$

$$A_1 = \int_a^b \frac{(x-x_0)(x-x_1)(x-x_2)\,dx}{(x-x_1)\cdot 0{,}5\cdot(b-a)\cdot 0{,}5\,(a-b)} = \frac{2\,(b-a)}{3};$$

$$A_2 = \int_a^b = \frac{(x-x_0)(x-x_1)(x-x_2)}{(x-x_2)\,(b-a)\cdot 0{,}5\,(b-a)}\,dx = \frac{b-a}{6}.$$

Таким чином, квадратурна формула Сімпсона має вигляд:

$$\int_a^b f(x)\,dx \approx \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right). \quad (34.18)$$



Мал. 41

У випадку додатної функції $f(x)$ формула (34.18), як бачимо, зводиться до того, що інтеграл $\int_a^b f(x)\,dx$ наближено замінюється площею фігури, яка обмежена віссю Ox, прямими $x = a$ і $x = b$ і параболою другого степеня (з вертикальною віссю), що проходить через точки $(a;\ f(a))$, $\left(\frac{a+b}{2};\ f\left(\frac{a+b}{2}\right)\right)$, $(b;\ f(b))$. (На мал. 41 парабола зображена штриховою лінією).

Оскільки формула (34.18) інтерполяційна, то її алгебраїчний степінь точності не нижчий за 2 (див. § 33). Виявляється,

що формула Сімпсона точна також для всіх многочленів третього степеня. Щоб переконатися в цьому, достатньо з'ясувати, що формула точна для функції $f(x) = x^3$, інтеграл від якої дорівнює $\int_a^b x^3dx = \frac{b^4-a^4}{4}$. За формулою (34.18) дістанемо такий самий результат

$$\frac{b-a}{6}\left(a^3 + 4\left(\frac{a+b}{2}\right)^3 + b^3\right) =$$

$$= \frac{b-a}{4}(a^3 + a^2b + ab^2 + b^3) = \frac{b^4-a^4}{4}.$$

Знайдемо вираз для залишкового члена формули Сімпсона для функцій $f(x) \in C^{(4)}[a;\ b]$. Оскільки формула точна для всіх многочленів степеня $\leqslant 3$, то для залишкового члена справджується формула (32.5), де $r = 4$. За формулою (32.4) знайдемо функцію $K_4(x)$. Врахувавши, що $A_0 = \frac{b-a}{6}$; $A_1 = \frac{2}{3}(b-a)$; $A_2 = \frac{b-a}{6}$, $x_0 = a$, $x_1 = \frac{a+b}{2}$, $x_2 = b$, матимемо:

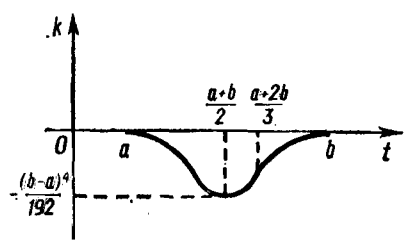
$$K_4(t) = \frac{(b-t)^4}{4} - \sum_{k=0}^{2} A_k E(x_k - t)(x_k - t)^3,$$

тобто

$$K_4(t) = \begin{cases} \dfrac{(b-t)^4}{4} - \dfrac{2}{3}(b-a)\left(\dfrac{a+b}{2} - t\right)^3 - \dfrac{b-a}{6}(b-t)^3, \\ \qquad\qquad \text{якщо } a \leqslant t \leqslant \dfrac{a+b}{2}, \\ \dfrac{(b-t)^4}{4} - \dfrac{b-a}{6}(b-t)^3, \quad \text{якщо } \dfrac{a+b}{2} < t \leqslant b. \end{cases}$$

Неважко перевірити, що $K_4\left(\frac{a+b}{2} + \alpha\right) = K_4\left(\frac{a+b}{2} - \alpha\right)$ для довільного $\alpha \in \left[0;\ \frac{a+b}{2}\right]$. Отже, графік функції $K_4(t)$ симетричний відносно прямої $x = \frac{a+b}{2}$. Тому достатньо дослідити криву лише на відрізку $\left[\frac{a+b}{2};\ b\right]$.

(Графік функції $K_4(t)$ зображений на мал. 42).



Мал. 42

Оскільки функція $K_4(t)$ на відрізку $[a;\ b]$ зберігає знак,

то до інтеграла в формулі (32.5) можна застосувати теорему про середнє. Таким чином,

$$R(f) = \frac{1}{3!} f^{(4)}(\xi) \int_a^b K_4(t)\,dt, \quad \xi \in]a; b[.$$

Обчисливши інтеграл $\int_a^b K_4(t)\,dt = 2 \int_{\frac{a+b}{2}}^b \left(\frac{(b-t)^4}{4} - \frac{b-a}{6}(b-t)^3 \right) dt = -\frac{(b-a)^5}{480}$, знаходимо остаточний вираз для залишкового члена формули Сімпсона:

$$R(f) = \frac{-(b-a)^5}{2880} f^{(4)}(\xi), \quad \xi \in]a; b[. \qquad (34.19)$$

3.1. Узагальнена квадратурна формула Сімпсона. Якщо відрізок $[a; b]$ великий, то його ділять на парну кількість $2m$ рівних частин: $[x_0; x_1]$, $[x_1; x_2]$, ..., $[x_{2m-1}; x_{2m}]$ завдовжки $h = \frac{b-a}{2m}$ (тут $x_0 = a$, $x_{2m} = b$) і до кожних двох сусідніх відрізків завдовжки $2h$ застосовують формулу Сімпсона (34.18). Тоді

$$\int_a^b f(x)\,dx = \sum_{i=1}^m \int_{x_{2i-2}}^{x_{2i}} f(x)\,dx \approx \sum_{i=1}^m \frac{h}{3} \times$$

$$\times \left(f(x_{2i-2}) + 4f(x_{2i-1}) + f(x_{2i})\right),$$

або

$$\int_{a=x_0}^{b=x_{2m}} f(x)\,dx \approx \frac{h}{3}\left(y_0 + 4y_1 + 2y_2 + 4y_3 + \cdots + 2y_{2m-2} + 4y_{2m-1} + y_{2m}\right), \qquad (34.20)$$

де $y_i = f(x_i)$, $h = \frac{b-a}{2m}$.

Формула (34.20) називається *узагальненою квадратурною формулою Сімпсона*, або *формулою парабол*.

Скориставшись формулою (34.19), для залишкового члена узагальненої формули Сімпсона матимемо:

$$R(f) = -\sum_{i=1}^m \frac{(2h)^5}{2880} f^{(4)}(\xi_i) =$$

$$= -\frac{h^5}{90} \sum_{i=1}^m f^{(4)}(\xi_i), \quad \xi_i \in]x_{2i-2}, x_{2i}[.$$

Оскільки $f^{(4)}(x)$ на відрізку $[a; b]$ неперервна, то існує така точка $\xi \in]a; b[$, що $f^{(4)}(\xi) = \frac{1}{m}\sum_{i=1}^m f^{(4)}(\xi_i)$. Тоді для залишкового члена формули Сімпсона дістанемо вираз:

$$R(f) = \frac{-h^5 m}{90} f^{(4)}(\xi) = -\frac{(2h \cdot m)\,h^4}{180} f^{(4)}(\xi) =$$

$$= -\frac{(b-a)\,h^4}{180} \cdot f^{(4)}(\xi), \quad \xi \in]a; b[. \qquad (34.21)$$

Звідси знаходимо оцінку для залишкового члена:

$$|R(f)| \leqslant \frac{(b-a)\,h^4}{180} M_4, \quad M_4 = \max_{x \in [a;b]} |f^{(4)}(x)|. \qquad (34.22)$$

Коефіцієнти квадратурної формули парабол додатні і сума їх дорівнює $2mh = b - a$. Отже, за формулою (32.7) знаходимо, що похибка дій, зумовлена неточністю значень y_i ($i = 0, 1, \ldots, 2m$), дорівнює $\Delta_\Sigma = \Delta_f (b - a)$ (як і для формул прямокутників та трапецій).

Таким чином, гранична абсолютна похибка значення інтеграла, обчисленого за квадратурною формулою Сімпсона, дорівнює

$$\Delta_I = \frac{(b-a)\,h^4}{180} M_4 + \Delta_f (b-a) + \Delta_0, \qquad (34.23)$$

де Δ_0 — похибка остаточного округлення результату.

Приклад.

Обчислити за формулою Сімпсона (з кроком $h = 0{,}1$) інтеграл $I = \int_0^1 \frac{dx}{1+x} = 0{,}693147\ldots$ і оцінити похибку результату.

Оскільки $h = 0{,}1$, то $2m = \frac{1}{0{,}1} = 10$. Потрібні значення підінтегральної функції у вузлових точках обчислимо з точністю до $0{,}5 \cdot 10^{-5}$. За формулою (34.20) дістанемо:

$$\int_0^1 \frac{dx}{1+x} \approx \frac{0{,}1}{3}(1 + 4 \cdot 0{,}90909 + 2 \cdot 0{,}83333 + 4 \cdot 0{,}76923 +$$

$$+ 2 \cdot 0{,}71429 + 4 \cdot 0{,}66667 + 2 \cdot 0{,}62500 + 4 \cdot 0{,}58824 +$$

$$+ 2 \cdot 0{,}55556 + 4 \cdot 0{,}52632 + 0{,}5) = 0{,}693152 \approx 0{,}69315.$$

Похибку результату обчислимо за формулою (34.23). Оскільки $f(x) = \frac{1}{1+x}$, то $f^{(4)}(x) = \frac{24}{(1+x)^5}$. Звідси

$$M_4 = \max_{0 \leqslant x \leqslant 1} \left| \frac{24}{(1+x)^5} \right| = 24.$$

Значення підінтегральної функції у вузлах обчислені з точністю до $0{,}5 \cdot 10^{-5}$, тобто $\Delta_f = 0{,}5 \cdot 10^{-5}$, а $\Delta_0 = 0{,}2 \times \times 10^{-5}$.

Отже, для граничної абсолютної похибки результату матимемо:

$$\Delta_I = \frac{1 \cdot (0{,}1)^4}{180} 24 + 0{,}5 \cdot 10^{-5} \cdot 1 + 0{,}2 \cdot 10^{-4} = 0{,}2 \cdot 10^{-4}.$$

Звідси випливає, що в знайденому результаті не менш як чотири значущі цифри правильні. Порівнявши цей результат з точним значенням інтеграла, переконаємось, що в ньому правильні всі п'ять значущих цифр. Зауважимо, що за формулою трапецій з тим самим кроком $h = 0{,}1$ дістанемо значення цього інтеграла, в якому правильні лише дві значущі цифри, а за формулою середніх прямокутників — три.

Формули середніх прямокутників і трапецій разом мають одну цінну властивість. А саме, якщо $f''(x)$ на $[a; b]$ не змінює знака, то вони дають для інтеграла $I = \int_a^b f(x)\,dx$ двосторонні наближення, оскільки їх залишкові члени мають протилежні знаки (див. (34.6) і (34.15)). Зокрема, для інтеграла $\int_0^1 \frac{dx}{1+x}$ $f''(x) = \frac{2}{(1+x)^3} > 0$, $x \in [0; 1]$. Тому $I_h^{\text{пр}} < I < I_h^{\text{тр}}$, де $I_h^{\text{пр}}$ і $I_h^{\text{тр}}$ — значення інтеграла, обчислені з кроком h відповідно за формулою середніх прямокутників і трапецій. В даному випадку доцільно покласти $I \approx \frac{I_h^{\text{пр}} + I_h^{\text{тр}}}{2} = I^*$. Звідси дістаємо оцінку для похибки значення I^*:

$$|I - I^*| \leqslant \frac{I_h^{\text{тр}} - I_h^{\text{пр}}}{2}.$$

Якщо $f(x) \in C^{(4)}[a; b]$ і $\int_a^b f''(x)\,dx \neq 0$, то можна показати, що при досить малому кроці h формули прямокутників і трапецій дають двосторонні наближення інтеграла, навіть якщо $f''(x)$ не зберігає знака на $[a; b]$.

§ 35. Практичні оцінки точності квадратурних формул. Вибір кроку інтегрування

У попередніх параграфах знайдено оцінки для залишкових членів квадратурних формул прямокутників, трапецій, парабол. Але практично оцінити залишковий член квадратурної формули за цими формулами іноді важко або й зовсім неможливо. Пояснюється це складністю знаходження максимумів модулів відповідних похідних, що входять до оцінок. У таких випадках доводиться користуватися іншими критеріями. Один з них, який полягає в одночасному застосуванні формули прямокутників і трапецій, розглянуто вище. Часто вдаються до подвійного перерахунку з кроками h і $2h$.

Розглянемо зручний спосіб визначення похибки квадратурних формул при подвійному перерахунку. Цей спосіб застосовують тоді, коли відомий порядок залишкового члена квадратурної формули відносно кроку h. Зауважимо, що точність квадратурної формули характеризується в основному порядком залишкового члена $R(f)$ відносно степенів h. Якщо залишковий член відносно h має порядок s, тобто $R(f) = O(h^s)$ (s — натуральне число), то квадратурна формула вважається тим точнішою, чим більше число s.

З оцінок для залишкових членів видно, що формули лівих і правих прямокутників відносно кроку h мають перший порядок точності ($R(f) = O(h)$), формула середніх прямокутників і трапецій — другий ($R(f) = O(h^2)$), а формула Сімпсона — четвертий ($R(f) = O(h^4)$).

Звичайно, якість формули виявляється при досить малому h.

Проте в деяких конкретних випадках менш точна формула (при тому самому кроці) може дати точніший результат, ніж точніша формула.

Нехай залишковий член квадратурної формули має порядок s відносно кроку h, тобто $R_h(f) = O(h^s)$ ($s \geqslant 1$), де $h = \frac{b-a}{m}$ (індекс h у виразі $R_h(f)$, ми ввели для того, щоб показати, з яким кроком обчислено інтеграл).

Якщо похідна від підінтегральної функції $f(x)$, що входить до виразу для залишкового члена, на відрізку інтегрування $[a; b]$ змінюється мало, то наближено можна покласти

$$R_h(f) = Mh^s, \tag{35.1}$$

де M — деяка, поки що невідома величина, яку для функції $f(x)$ вважатимемо на $[a; b]$ сталою.

Припустимо, що інтеграл обчислено за даною квадратурною формулою з кроком h і деяким іншим кроком $k = \frac{b-a}{n}$, причому $n \neq m$. Значення інтеграла I, обчислені за квадратурною формулою з кроками h і k, позначимо відповідно через I_h і I_k. Тоді $I = I_h + Mh^s$, $I = I_k + Mk^s$. Віднявши почленно від першої рівності другу, знайдемо невідому величину M:

$$M = \frac{I_h - I_k}{k^s - h^s}.$$

Підставивши значення M у формулу (35.1), для залишкового члена квадратурної формули, матимемо:

$$R_h(f) = \frac{I_h - I_k}{k^s - h^s} h^s = \frac{I_h - I_k}{\left(\frac{k}{h}\right)^s - 1}. \qquad (35.2)$$

Таким чином, ми наближено подали залишковий член через два значення інтеграла, обчислені за квадратурною формулою з кроками h і k. Така оцінка похибки чисельного інтегрування називається *правилом Рунге*. Здебільшого крок k вибирають так: $k = 2h$, де h — початковий крок. Це пояснюється тим, що при такому кроці для обчислення I_k за квадратурною формулою можна використати вже знайдені табличні значення функції, вибираючи значення через одне. У випадку, коли $k = 2h$, формула (35.2) набирає вигляду:

$$R_h(f) = \frac{I_h - I_{2h}}{2^s - 1}. \qquad (35.3)$$

Якщо вираз залишкового члена для інтеграла I відомий, можна відшукати уточнене значення $I_{h,2h}$:

$$I_{h,2h} = I_h + \frac{1}{2^s - 1}(I_h - I_{2h}). \qquad (35.4)$$

Обчислення уточненого значення $I_{h,2h}$ за формулою (35.4) називається *екстраполюванням за Річардсоном*. Якщо $I_h \neq I_{2h}$, то значення $I_{h,2h}$ завжди лежать поза відрізком $[I_h; I_{2h}]$. Тому обчислення $I_{h,2h}$ і називають *екстраполюванням*. З формул (35.3) і (35.4) ми дістанемо відповідні формули для квадратурних формул прямокутників, трапецій, формули Сімпсона. Так для формули середніх прямокутників і формули трапецій ($s = 2$) матимемо:

$$R_h(f) = \frac{I_h - I_{2h}}{3},$$

$$I_{h,2h} = I_h + \frac{I_h - I_{2h}}{3}, \qquad (35.5)$$

а для формули Сімпсона ($s = 4$):

$$R_h(f) = \frac{I_h - I_{2h}}{15}; \quad I_{h,2h} = I_h + \frac{1}{15}(I_h - I_{2h}). \qquad (35.6)$$

Приклад.

Обчислити значення інтеграла $\int_0^1 f(x)\,dx$ і оцінити похибку результату, якщо функція $f(x)$ задана таблицею:

x	0,0	0,1	0,2	0,3	0,4	0,5
$f(x)$	1,000000	0,990099	0,961538	0,917431	0,862069	0,800000

x	0,6	0,7	0,8	0,9	1,0
$f(x)$	0,735294	0,671101	0,609756	0,552486	0,500000

Інтеграл обчислимо за формулою трапецій з кроком $h = 0,1$.

$$\int_0^1 f(x) \approx 0,1\left(\frac{1}{2}\cdot 1,000000 + 0,990099 + 0,961538 + \right.$$

$$+ 0,917431 + 0,862069 + 0,800000 + 0,735294 + 0,671101 +$$

$$\left. + 0,609756 + 0,552486 + \frac{1}{2}\,0,500000\right) = 0,784981.$$

Щоб оцінити знайдений результат, виконаємо контрольний перерахунок з кроком, удвічі більшим, $h = 0,2$:

$$\int_0^1 f(x)\,dx \approx 0,2\left(\frac{1,000000}{2} + 0,961538 + 0,862069 + 0,735294 + \right.$$

$$\left. + 0,609756 + \frac{1}{2}\,0,500000\right) = 0,783731.$$

У наших позначеннях $I_h = 0,784981$, а $I_{2h} = 0,783731$. За формулою (35.5) для залишкового члена значення I_h дістанемо: $R_h(f) = \frac{1}{3}(0,784981 - 0,783731) = 0,000417$. Похибка дій $\Delta_\Sigma = \Delta_f(b-a) = 0,5 \cdot 10^{-6} \cdot 1 = 0,5 \cdot 10^{-6}$, а $\Delta_0 = 0,4 \cdot 10^{-6}$.

Отже, гранична абсолютна похибка результату $\Delta_{I_h} = 0,000417 + 0,5 \cdot 10^{-6} + 0,4 \cdot 10^{-6} = 0,0004179$. Оскільки

$\Delta_{I_h} < 0{,}0005$, то в результаті $I_h = 0{,}784981$ принаймні три значущі цифри правильні. Значенням шуканого інтеграла доцільно вважати уточнене значення, яке дістали екстраполюванням за Річардсоном (формула (35.5)):

$$I_{h,2h} = 0{,}784981 + 0{,}000417 = 0{,}785398.$$

У цьому прикладі таблично була задана функція $f(x) = \frac{1}{1+x^2}$. Для порівняння наведемо значення інтеграла, обчисленого за формулою Ньютона — Лейбніца:

$$I = \int_0^1 \frac{dx}{1+x^2} = \operatorname{arctg} 1 = \frac{\pi}{4} = 0{,}78539816\ldots$$

Отже, значення I_h має справді три правильні значущі цифри, а після екстраполювання правильними будуть усі шість значущих цифр.

На практиці часто доводиться обчислювати інтеграли з деякою наперед заданою точністю $\varepsilon > 0$, причому крок h не заданий, а його треба знайти, але так, щоб абсолютна похибка результату, обчисленого за відповідною квадратурною формулою з кроком h, не перевищувала ε.

Величину кроку h в цьому разі можна визначити, виходячи з оцінки залишкового члена відповідної квадратурної формули. При цьому абсолютна величина залишкового члена не повинна перевищувати ε. Для формули середніх прямокутників маємо (див. (34.7)):

$$\frac{b-a}{24} h^2 M_2 \leqslant \varepsilon.$$

Отже,

$$h \leqslant \sqrt{\frac{24\varepsilon}{(b-a)M_2}}. \qquad (35.7)$$

Аналогічно, врахувавши оцінки (34.16) і (34.22), для формул трапецій і Сімпсона відповідно дістанемо:

$$h \leqslant \sqrt{\frac{12\varepsilon}{(b-a)M_2}}, \qquad (35.8)$$

$$h \leqslant \sqrt[4]{\frac{180\varepsilon}{(b-a)M_4}}. \qquad (35.9)$$

Примітка. Крок h повинен не тільки задовольняти нерівності (35.7) — (35.9), а й бути таким, щоб відрізок інтегрування був поділений на рівні між собою частини, а для формули Сімпсона — на парну кількість рівних частин.

Якби значення підінтегральної функції у вузлах виражалися точними числами і всі проміжні обчислення виконувалися точно, то, визначивши інтеграл за відповідною формулою з кроком h, можна було б гарантувати, що абсолютна похибка знайденого результату не перевищує ε. Оскільки значення підінтегральної функції, як правило, виражаються наближеними числами, то для обчислення інтеграла за квадратурною формулою з кроком h значення $f(x_k)$ треба взяти з такою точністю, щоб неусувною похибкою можна було знехтувати. Таким чином, $f(x_k)$ обчислюватимемо (якщо це можна) з такою точністю, щоб неусувна похибка $\Delta_\Sigma = (b-a)\Delta_f$ була значно менша за ε (точність характеризується величиною Δ_f).

Визначивши крок і встановивши потрібну точність при обчисленні $f(x_k)$, за відповідною квадратурною формулою визначимо інтеграл.

Приклад.

За квадратурною формулою Сімпсона обчислити

$$\int_0^{\frac{\pi}{2}} \sin x dx$$ з точністю до $\varepsilon = 0{,}5 \cdot 10^{-2}$.

Величину кроку h для формули Сімпсона визначимо за формулою (35.9).

Тут $M_4 = \max\limits_{x \in \left[0;\, \frac{\pi}{2}\right]} |f^{(4)}(x)| = \max\limits_{x \in \left[0;\, \frac{\pi}{2}\right]} |\sin x| = 1$. Тоді $h \leqslant$

$\leqslant \sqrt[4]{\frac{180 \cdot 0{,}5 \cdot 10^{-2}}{\frac{\pi}{2} \cdot 1}} \approx 0{,}87$. Візьмемо $h = \frac{\pi}{4} < 0{,}87$, тобто відрізок інтегрування $\left[0;\, \frac{\pi}{2}\right]$ поділимо пополам.

Встановимо точність, з якою потрібно обчислювати значення $f(x_k)$. Щоб похибка $\Delta_\Sigma = (b-a)\Delta_f = \frac{\pi}{2}\Delta_f$ була значно меншою від $\varepsilon = 0{,}5 \cdot 10^{-2}$, достатньо обчислити значення підінтегральної функції у вузлах з точністю до $0{,}5 \cdot 10^{-3}$. Тоді $\Delta_f = 0{,}0005$ і $\Delta_\Sigma = 0{,}00079$.

Визначивши інтеграл за формулою Сімпсона, дістанемо:

$$\int_0^{\frac{\pi}{2}} \sin x\, dx \approx \frac{\pi}{12}\left(\sin 0 + 4 \sin \frac{\pi}{4} + \sin \frac{\pi}{2}\right) =$$

$$= \frac{\pi}{12}(0 + 4 \cdot 0{,}707 + 1) = 1{,}002.$$

Зауважимо, що в результаті ми спеціально зберегли сумнівну цифру. Звичайно, тут її можна було б округлити і дістати результат 1,00. Але після округлення сумнівної цифри остання збережена цифра може бути неправильною. Справді, гранична абсолютна похибка будь-якого числа після округлення сумнівної цифри може дорівнювати одиниці останнього збереження розряду (0,5 одиниці розряду останньої правильної цифри та 0,5 одиниці цього самого розряду після округлення).

Тому, коли доводиться визначати інтеграл з деякою кількістю m правильних значущих цифр і результат записувати без сумнівної цифри, треба обчислювати так, щоб у результаті можна було гарантувати правильність $m + 1$ значущих цифр, і тільки після цього округлити до m значущих цифр.

Вибирати крок інтегрування розглянутим способом доцільно, очевидно, тоді, коли легко знаходити потрібні максимуми похідних, що входять до формул (35.7) — (35.9). Проте часто такі максимуми знаходити не тільки складно, а й взагалі неможливо. У такому разі обчислити заданий інтеграл з потрібною точністю можна методом подвійного перерахунку, а саме: обчислюють заданий інтеграл за відповідною квадратурною формулою з деяким кроком h і $\frac{h}{2}$. Тоді за цими даними знаходять наближену величину залишкового члена (див. формулу (35.3)). Якщо залишковий член за модулем не більший від ε, то інтеграл, обчислений з кроком $\frac{h}{2}$, є відповіддю. У протилежному випадку зменшуємо крок вдвічі і обчислюємо значення інтеграла з кроком $\frac{h}{4}$. За двома значеннями інтеграла, що обчислені з кроками $\frac{h}{2}$ і $\frac{h}{4}$, відшукуємо значення залишкового члена, модуль якого знову порівнюємо з ε. Якщо ця величина не перевищує ε, то інтеграл, обчислений з кроком $\frac{h}{4}$, є шуканим. Якщо залишковий член за модулем більший від ε, то знову крок інтегрування зменшуємо вдвічі і так продовжуємо процес доти, поки відповідний залишковий член за модулем не буде меншим від ε або не дорівнюватиме ε. Останнє обчислене значення інтеграла і є шуканим. Оскільки величина залишкового члена останнього наближення вже обчислена, то за формулою (35.4) (див. також формули (35.5) і (35.6)) можна знайти уточнене екстраполюванням значення інтеграла, точність якого, як правило, буде більшою ніж ε.

Для розв'язання задачі можна використовувати також в

парі формулу прямокутників і трапецій. І якщо $\frac{|I_h^{\text{пр}} - I_h^{\text{тр}}|}{2} < \varepsilon$, то значення $\frac{I_h^{\text{пр}} + I_h^{\text{тр}}}{2}$ і буде шуканим.

Приклад.

Обчислити за формулою середніх прямокутників інтеграл $\int_0^{1,6} e^{-x^2}dx$ з точністю до $\varepsilon = 0{,}0005$. Потрібні для обчислень значення підінтегральної функції, знайдені з точністю до $0{,}5 \cdot 10^{-4}$, подано в таблиці.

x	e^{-x^2}	x	e^{-x^2}
0,1	0,9900	0,9	0,4449
0,2	0,9608	1,0	0,3679
0,3	0,9139	1,1	0,2982
0,4	0,8521	1,2	0,2369
0,5	0,7788	1,3	0,1845
0,6	0,6977	1,4	0,1409
0,7	0,6126	1,5	0,1054
0,8	0,5273		

Неусувна похибка при будь-якому кроці h не перевищуватиме величини $1{,}6 \cdot 0{,}5 \cdot 10^{-4} = 0{,}8 \cdot 10^{-4} < \varepsilon$.

Спочатку відрізок інтегрування $[0; 1{,}6]$ поділимо надвоє, тобто візьмемо початковий крок $h = \frac{1{,}6}{2} = 0{,}8$.

Обчисливши інтеграл за формулою середніх прямокутників, з кроком $h = 0{,}8$ дістанемо:

$$I_{0,8} = 0{,}8 \sum_{i=1}^{2} f\left(\frac{2i-1}{4}\,1{,}6\right) = 0{,}8\,(0{,}8521 + 0{,}2369) = 0{,}8712.$$

Для кроку $h = 0{,}4$ матимемо:

$$I_{0,4} = 0{,}4 \sum_{i=1}^{4} f\left(\frac{2i-1}{8}\,1{,}6\right) = 0{,}4\,(0{,}9608 + 0{,}6977 + 0{,}3679 + 0{,}1409) = 0{,}8669.$$

Тепер знайдемо за формулою (35.5) наближене значення залишкового члена для $I_{0,4}$:

$$R_{0,4}(f) = \frac{I_{0,4} - I_{0,8}}{3} = \frac{1}{3}\,(0{,}8669 - 0{,}8712) = -0{,}0014.$$

Оскільки задана точність ще не досягнута, бо $|R_{0,4}(f)| = |-0{,}0014| > 0{,}0005$, знову обчислимо інтеграл за формулою середніх прямокутників з кроком $h = 0{,}2$:

$$I_{0,2} = 0{,}2\sum_{i=1}^{8} f\left(\frac{2i-1}{16}1{,}6\right) = 0{,}2\,(0{,}9900 + 0{,}9139 + 0{,}7788 + 0{,}6126 + 0{,}4449 + 0{,}2982 + 0{,}1845 + 0{,}1054) = 0{,}8657.$$

Обчислимо тепер

$$R_{0,2}(f) = \frac{I_{0,2} - I_{0,4}}{3} = \frac{1}{3}(0{,}8657 - 0{,}8669) = -0{,}0004.$$

$$|R_{0,2}(f)| = |-0{,}0004| < 0{,}0005.$$

Тому знайдене значення $I_{0,2}$ є шуканим значенням інтеграла. Отже, наближене значення інтеграла $I = 0{,}8657$, обчислене за формулою середніх прямокутників з кроком $h = 0{,}2$, відрізняється від точного значення не більш як на $\varepsilon = 0{,}0005$.

Якщо виконати ще екстраполювання за Річардсоном, то знайдемо уточнене значення $I_{0,2;0,4} = 0{,}8657 - 0{,}0004 = 0{,}8653$. Для порівняння наведемо значення шуканого інтеграла з чотирма правильними значущими цифрами:

$$\int_0^{1,6} e^{-x^2}dx = 0{,}8653.$$

Вправи.

1. Вивести квадратурну формулу

$$\int_a^b f(x)\,dx \approx (b-a)\left(\frac{1}{8}f(a) + \frac{3}{8}f\left(\frac{2a+b}{3}\right) + \frac{3}{8}f\left(\frac{a+2b}{3}\right) + \frac{1}{8}f(b)\right)$$

і знайти її залишковий член.

В к а з і в к а. Тут записана формула Ньютона — Котеса для випадку $n = 3$. Вона називається формулою трьох восьмих.

2. Знайти оцінки для залишкового члена квадратурної формули Сімпсона для функцій класів $C^{(1)}[a;\ b]$, $C^{(2)}[a;\ b]$ і $C^{(3)}[a;\ b]$.

3. Вивести квадратурну формулу

$$\int_{-a}^{b} f(x)\,dx \approx \frac{a+b}{6ab}\,(b\,(2a-b)\,f(-a) + (a+b)^2 f(0) + a\,(2b-a)\,f(b)),$$

де $a,\ b > 0$,

яку можна використати для випадку нерівних інтервалів табулювання.

В к а з і в к а. Замінити функцію $f(x)$ інтерполяційним многочленом другого степеня.

4. За допомогою формули Сімпсона вивести просте правило для обчислення об'єму бочки.

В к а з і в к а. Об'єм бочки (як і будь-якого тіла) можна виразити інтегралом $\int_a^b f(x)\,dx$, де функція $f(x)$ — площа довільного перерізу бочки площиною, яка перпендикулярна до осі Ox (вісь Ox збігається з віссю бочки).

5. Обчислити інтеграл $\int_0^{10} e^{-x}dx = 0{,}999955$ за квадратурними формулами прямокутників, трапецій, Сімпсона з кроком $h = 1$. Порівняти результати обчислень.

6. Обчислити за формулами прямокутників, трапецій і Сімпсона інтеграли: а) $\int_0^1 \sqrt{1+x}\,dx$, $h = 0{,}1$; б) $\int_0^{\frac{\pi}{2}} \sin x\,dx$, $h = \frac{\pi}{20}$. Оцінити похибки результатів.

7. Обчислити за формулами трапецій і Сімпсона інтеграл $\int_0^{0,8} f(x)\,dx$, якщо функція $f(x)$ задана таблично.

x	0,0	0,1	0,2	0,3	0,4	0,5
$f(x)$	0,9298	0,9851	0,9704	0,9464	0,9139	0,8737

x	0,6	0,7	0,8	0,9	1,0
$f(x)$	0,8269	0,7749	0,7189	0,6603	0,6005

Оцінити похибку результатів.

8. За формулами середніх прямокутників, трапецій і Сімпсона обчислити з точністю $\varepsilon = 0{,}005$ інтеграли:

а) $\int_1^2 \frac{1}{x}\,dx$; б) $\int_0^1 \cos x\,dx$.

9. Обчислити за формулами середніх прямокутників, трапецій і Сімпсона з вказаними значеннями h такі інтеграли:

а) $\int_0^{\frac{\pi}{2}} \cos x dx \left(h = \frac{\pi}{4},\ h = \frac{\pi}{8}\right)$; б) $\int_1^2 \frac{dx}{\sqrt{x}}$ $(h = 0{,}5;\ h = 0{,}25)$.

Результати проекстраполювати.

10. Обчислити за формулами середніх прямокутників і трапецій з точністю $\varepsilon = 0{,}5 \cdot 10^{-2}$, а за формулою Сімпсона з точністю $0{,}5 \cdot 10^{-3}$ інтеграли: а) $\int_0^{1{,}6} \sqrt{1+x^2}\,dx$; б) $\int_0^{0{,}8} \frac{dx}{1+x^2}$; в) $\int_0^1 \frac{\lg(1+x)}{1+x^2}\,dx$;

г) $\int_0^{1{,}6} e^{-x^2}dx$.

Для контролю точності користуватися подвійним перерахунком.

11. Для переведення натуральних логарифмів у десяткові використовується формула $M \ln x = \lg x$ для всіх $x > 0$, де M — константа. Підставивши $x = 10$, для M дістанемо:

$$M = \frac{\lg 10}{\ln 10} = \frac{1}{\ln 10},$$

звідки

$$\frac{1}{M} = \ln 10 = \int_1^{10} \frac{dx}{x}.$$

Знайти наближено $\frac{1}{M}$ з точністю $\varepsilon = 0{,}0005$, користуючись квадратурною формулою Сімпсона.

§ 36. Задача Коші для звичайних диференціальних рівнянь. Класифікація методів її розв'язування

Задача Коші ставиться так — необхідно знайти на відрізку $[x_0;\ x_0 + l]$ розв'язок $y(x)$ диференціального рівняння

$$y' = f(x,\ y), \tag{36.1}$$

який задовольняє початкову умову

$$y(x_0) = y_0. \tag{36.2}$$

Має місце така теорема (про існування і єдиність розв'язку).

Якщо функція $f(x, y)$ ***неперервна в прямокутнику*** $\Delta = \{(x,\ y) : x_0 \leqslant x \leqslant x_0 + l,\ |y - y_0| \leqslant b\}$ ***і задовольняє в*** Δ ***умову Ліпшиця за змінною*** y***:***

$$|f(x,\ y_1) - f(x,\ y_2)| \leqslant K|y_1 - y_2|,$$

де K ***— стала, то на відрізку*** $[x_0;\ x_0 + h]$***, де***

$$0 < h < \min\left\{l,\ \frac{b}{M},\ \frac{1}{K}\right\},\quad M = \max_{(x;y)\in\Delta} |f(x,\ y)|,$$

існує єдиний неперервний розв'язок задачі **(36.1)—(36.2).**

Доведення. Задача знаходження розв'язку задачі (36.1) — (36.2) рівносильна задачі знаходження неперервного розв'язку інтегрального рівняння

$$y(x) = y_0 + \int_{x_0}^{x} f(t,\ y(t))\,dt,\quad x \in [x_0;\ x_0 + l], \tag{36.3}$$

$y(x)$ — невідома функція.

Для доведення теореми достатньо довести існування єдиного розв'язку рівняння (36.3). Застосуємо для доведення принцип стискуючих відображень (§ 7).

Розглянемо оператор (відображення)

$$Ty = y_0 + \int_{x_0}^{x} f(t,\ y(t))\,dt, \tag{36.4}$$

визначений на множині всіх тих функцій $y = y(x)$ з $C[x_0;\ x_0 + h]$, для яких виконуються нерівності $|y - y_0| \leqslant b$. Очевидно, задана множина є повним метричним простором, який позначимо через $C[D]$. Покажемо, що оператор Ty відображає простір $C[D]$ в самого себе і що цей оператор стискуючий. Справді, функція Ty, визначена формулою (36.4), неперервна на $[x_0;\ x_0 + h]$, тобто належить $C[x_0;\ x_0 + h]$. Крім того,

$$|Ty - y_0| = \left|\int_{x_0}^{x} f(t, y(t))\,dt\right| \leqslant M|x - x_0| \leqslant Mh < M\frac{b}{M} = b.$$

Отже, $Ty \in C[D]$.

Покажемо, що відображення Ty стискуюче. Дійсно, для довільних y_1 і y_2, які належать $C[D]$, маємо

$$\rho(Ty_1,\ Ty_2) = \max_{x\in[x_0;x_0+h]} \left|\int_{x_0}^{x} (f(t, y_1(t)) - f(t,\ y_2(t)))\,dt\right| \leqslant$$

$$\leqslant \max_{x\in[x_0;x_0+h]} \int_{x_0}^{x} |f(t, y_1(t)) - f(t,\ y_2(t))|\,dt \leqslant$$

$$\leqslant \max_{x\in[x_0;x_0+h]} \int_{x_0}^{x} K|y_1(t) - y_2(t)|\,dt \leqslant$$

$$\leqslant K \cdot \max_{x\in[x_0;x_0+h]} |y_1(t) - y_2(t)| \cdot \max_{x\in[x_0;x_0+h]} \int_{x_0}^{x} dx \leqslant Kh\rho(y_1,\ y_2) =$$

$$= q\rho(y_1,\ y_2),$$

де $q = Kh < 1$.

Отже, відображення Ty має єдину нерухому точку $y \in C[D]$, $y = Ty$, яка є єдиним неперервним розв'язком рівняння (36.3) на відрізку $[x_0; x_0 + h]$. Цей розв'язок можна дістати методом послідовних наближень:

$$y_m(x) = y_0 + \int_{x_0}^{x} f(t, y_{m-1}(t))\,dt, \quad y_0(t) = y_0, \quad m = 1, 2, \ldots \tag{36.5}$$

Останні формули визначають ітераційний метод Пікара. Метод Пікара збігається і визначає єдиний розв'язок рівняння (36.3) або (36.1) — (36.2).

Примітка. Певним недоліком доведеного результату є те, що розв'язок $y(x)$ задачі (36.1) — (36.2) визначено не на всьому відрізку $[x_0; x_0 + l]$, а лише на $[x_0; x_0 + h]$, де h визначається умовами теореми. Проте, взявши точку $(x_0 + h; y(x_0 + h))$ за початкову і повторивши міркування теореми можна продовжити розв'язок ще на відрізок довжини h_1, якщо, звичайно, в околі цієї точки виконуються умови теореми про існування і єдність розв'язку. Продовжуючи цей процес, часто вдається продовжити розв'язок на весь відрізок $[x_0; x_0 + l]$ і навіть на $[x_0; \infty[$.

Надалі вважатимемо, що функція $f(x, y)$ в розглядуваній області має достатню кількість похідних за обома змінними x і y. Зокрема, коли $f(x, y)$ має неперервні похідні до порядку m включно, то розв'язок $y(x)$ має $m + 1$ неперервну похідну за змінною x.

Знайти точний розв'язок задачі (36.1) — (36.2), тобто виразити його через елементарні чи спеціальні функції, або подати через квадратури від елементарних чи спеціальних функцій вдається лише в небагатьох практичних задачах. Інколи, якщо навіть і вдається знайти точний розв'язок, він має досить складний вигляд і користуватися ним практично неможливо. Тому для розв'язування задачі доводиться застосовувати наближені методи. Наближені методи можна поділити на два типи: аналітичні (які дають наближений розв'язок диференціального рівняння у вигляді аналітичного виразу) і чисельні (які дають наближений розв'язок у вигляді таблиці значень).

До аналітичних наближених методів відноситься *ітераційний метод Пікара* (36.5). Обмежуючись скінченним числом m, визначають наближений вираз для розв'язку $y(x)$. Метод Пікара зручно застосовувати, якщо інтеграли в (36.5) вдається обчислювати через елементарні функції. Якщо ж права частина рівняння (36.1) більш складна, то ці інтеграли доводиться знаходити чисельними методами, і метод Пікара не досить зручний.

Іншим аналітичним методом наближеного розв'язування задачі (36.1) — (36.2) є метод, який грунтується на розкладанні розв'язку задачі Коші в ряд Тейлора. За цим методом наближений розв'язок задачі шукають у вигляді

$$y_m(x) = \sum_{i=0}^{m} \frac{y^{(i)}(x_0)}{i!}(x - x_0)^i, \quad x \in [x_0; x_0 + l] \tag{36.6}$$

(тобто $y(x)$ замінюють відрізком ряду Тейлора, похибка наближення має порядок $O((x - x_0)^{m+1})$), де

$$y^{(0)}(x_0) = y(x_0) = y_0, \quad y^{(1)}(x_0) = f(x_0, y_0),$$

а останні $y^{(i)}(x_0)$ $(i = 2, 3, \ldots, m)$ знаходять за формулами, які дістають послідовним диференціюванням рівняння (36.1). Послідовно диференціюючи (36.1), знаходимо:

$$y^{(2)}(x) = y''(x) = f'_x(x, y) + f'_y(x, y) \cdot y'(x),$$

$$y^{(3)}(x) = f''_{xx}(x, y) + 2f''_{xy}(x, y) \cdot y'(x) + f''_{yy}(x, y) \cdot (y'(x))^2 +$$

$$+ f'_y(x, y) \cdot y''(x).$$

Підставивши $x = x_0$ і $y = y_0$, знайдемо потрібні коефіцієнти $y''(x_0)$, $y'''(x_0)$, ... у формулі (36.6).

Якщо x належить області збіжності ряду $\sum_{i=0}^{\infty} \frac{y^{(i)}(x_0)}{i!}(x - x_0)^i$, то при досить великих m похибка наближеної рівності $y(x) \approx y_m(x)$ мала. Якщо $x - x_0$ більше за радіус збіжності ряду, то користуватися наближенням $y_m(x)$, яке обчислене за формулою (36.6), не можна.

Метод розкладання розв'язку в ряд Тейлора зручний тоді, коли вдається явно знайти коефіцієнти ряду.

Відомі й інші аналітичні методи розв'язування задачі (36.1) — (36.2), але ми не розглядатимемо їх.

Для розв'язування задачі Коші широко застосовуються чисельні методи. Вони дають можливість знаходити наближені значення (а інколи й точні) шуканого розв'язку $y(x)$ в деяких фіксованих точках (вузлах) $x_0 < x_1 < x_2 < \ldots < x_N = x_0 + l$.

Слід пам'ятати, що чисельні методи можна застосовувати лише для коректно поставлених задач. Причому для успішного застосування чисельних методів задача має бути не тільки формально стійкою, а й добре обумовленою (див. § 6). Інакше незначні похибки в початкових умовах чи в проміжних обчисленнях можуть призвести до великої похибки результату. Звичайно для розв'язування задачі треба брати стійкі чисельні методи (алгоритми). Наприклад, розглянемо таку задачу

Коші:

$$y' = 2xy - 2x^2 - 2x + 1, \quad 0 \leqslant x \leqslant 10, \tag{36.7}$$

$$y(0) = 1. \tag{36.8}$$

Неважко переконатися, що загальний розв'язок рівняння (36.7) є $y(x) = ce^{x^2} + x + 1$, де c — довільна константа.

Врахувавши початкову умову (36.8), знайдемо, що $c = 0$, тобто розв'язок задачі такий: $y(x) = x + 1$.

Звідси знайдемо значення розв'язку, наприклад, у точці $x = 10$: $y(10) = 11$.

Якщо замість початкової умови (36.8) взяти, наприклад, таку $y(0) = 1 + 10^{-7}$ (досить незначна зміна початкової умови), то $c = 10^{-7}$ і $y(10) = 10^{-7}e^{100} + 1 \approx 2{,}7 \cdot 10^{36}$, тобто розв'язок дуже зміниться.

§ 37. Методи типу Ейлера

Побудуємо формули, які дають можливість знайти наближене значення шуканого розв'язку $y(x)$ задачі Коші в точці x_{i+1}, якщо значення цього розв'язку відоме в попередній точці x_i ($i = 0, 1, 2, \ldots$). Оскільки розв'язок в початковій точці відомий з початкових умов задачі, то за цими формулами послідовно можна обчислити значення розв'язку в точках x_1, x_2, x_3, ...

Проінтегруємо рівняння (36.1) від x_i до x_{i+1}, де $x_i = x_0 + ih$:

$$y(x_{i+1}) = y(x_i) + \int_{x_i}^{x_{i+1}} y'(t)\,dt, \tag{37.1}$$

де

$$y'(x) = f(x,\ y(x)).$$

Для наближеного обчислення інтеграла в (37.1) можна використати квадратурні формули. Використавши ту чи іншу квадратурну формулу, можна дістати різні формули чисельного інтегрування задачі Коші.

Якщо, наприклад, інтеграл в (37.1) обчислити за формулою лівих прямокутників, тобто

$$\int_{x_i}^{x_{i+1}} y'(t)\,dt = hy'(x_i) + \frac{h^2}{2}y''(\eta), \quad \eta \in]x_i,\ x_{i+1}[,$$

то матимемо

$$y(x_{i+1}) = y(x_i) + hf(x_i,\ y(x_i)) + O(h^2). \tag{37.2}$$

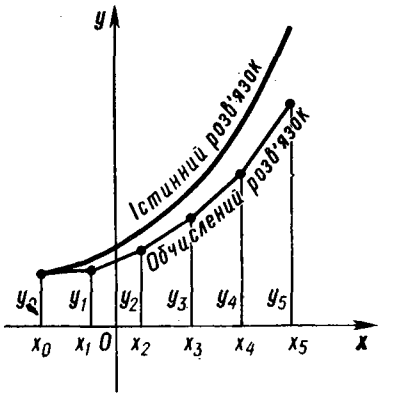
Відкинувши член порядку $O(h^2)$, з останньої рівності дістанемо:

$$y_{i+1} = y_i + hf(x_i,\ y_i), \tag{37.3}$$

де через y_i позначено наближене значення розв'язку $y(x_i)$.

Формула (37.3) *називається формулою Ейлера.* Починаючи з x_0, за формулою (37.3) можна знайти послідовно в точках $x_1 = x_0 + h$, $x_2 = x_1 + h$... наближені значення розв'язку $y(x)$ задачі (36.1) — (36.2). Похибка методу Ейлера на кожному кроці має порядок $O(h^2)$.

Мал. 43 є ілюстрацією розв'язання задачі Коші методом Ейлера. З формули (37.3) видно, що для знаходження розв'язку y_{i+1} за відомим значенням y_i треба спочатку обчислити тангенс кута нахилу дотичної до кривої $y(x)$ в точці $(x_i; y_i)$ і рухатися в напрямі дотичної до перетину з прямою $x = x_{i+1}$; точка перетину визначить y_{i+1}. Таким чином, шукану інтегральну криву $y(x)$ наближено замінюємо ламаною лінією (ламаною Ейлера).



Мал. 43

Метод Ейлера не застосовується широко в обчислювальній практиці, оскільки має невелику точність, і, крім того, досить часто виявляється нестійким (призводить до систематичного нагромадження похибок). На мал. 43 видно, як в деяких випадках ламана Ейлера може істотно відхилитися від шуканої інтегральної кривої.

Зауважимо, що формулу (37.3) можна було б дістати з формули (36.6), якщо $m = 1$ і $x = x_0 + h$, тобто розв'язок задачі (36.1) — (36.2) на кожному з відрізків $[x_i; x_{i+1}]$ подається двома членами ряду Тейлора.

Якщо для обчислення інтеграла в (37.1) використати формулу середніх прямокутників, то для чисельного інтегрування задачі Коші можна побудувати формули, похибка яких на кожному кроці матиме порядок $O(h^3)$.

За формулою середніх прямокутників

$$\int_{x_i}^{x_{i+1}} y'(t)\,dt = hy'\left(x_i + \frac{h}{2}\right) + \frac{h^3}{24}y'''(\eta), \quad \eta \in]x_i;\ x_{i+1}[.$$

Тоді з (37.1) дістанемо:

$$y(x_{i+1}) = y(x_i) + hf\left(x_i + \frac{h}{2},\ y\left(x_i + \frac{h}{2}\right)\right) + O(h^3).$$

За останньою формулою ще не можна записати формули чисельного інтегрування задачі Коші, оскільки в правій частині міститься невідоме значення $y\left(x_i + \frac{h}{2}\right)$ (значення розв'язку відомі в точці x_i і в попередніх вузлах).

Врахувавши, що

$$y\left(x_i + \frac{h}{2}\right) = y(x_i) + \frac{h}{2} f(x_i,\ y(x_i)) + O(h^2)$$

(див. (37.2)), матимемо:

$$y(x_{i+1}) = y(x_i) + hf\left(x_i + \frac{h}{2},\ y(x_i) + \frac{h}{2} f(x_i,\ y(x_i))\right) +$$

$$+ O(h^2)) + O(h^3) = y(x_i) + hf\left(x_i + \frac{h}{2},\ y(x_i) +\right.$$

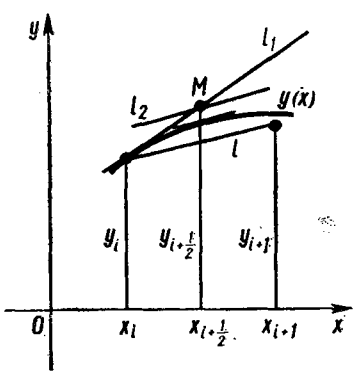
$$\left. + \frac{h}{2} f(x_i,\ y(x_i))\right) + O(h^3).$$

Звідси дістанемо такі формули:

$$y_{i+\frac{1}{2}} = y_i + \frac{h}{2} f(x_i,\ y_i), \tag{37.4}$$

$$y_{i+1} = y_i + hf\left(x_i + \frac{h}{2},\ y_{i+\frac{1}{2}}\right). \tag{37.5}$$

Формули (37.4), (37.5) визначають *удосконалений метод Ейлера*. За цим методом, як бачимо, значення розв'язку в точці x_{i+1} знаходять двома кроками: спочатку методом Ейлера з кроком $\frac{h}{2}$ обчислюють проміжне значення $y_{i+\frac{1}{2}}$, а потім уже знаходять значення y_{i+1}, враховуючи при цьому напрям інтегральної кривої в середній точці $\left(x_i + \frac{h}{2};\ y_{i+\frac{1}{2}}\right)$. На мал. 44 дано геометричну ілюстрацію удосконаленого методу Ейлера. Пряма l_1 є дотичною до кривої $y(x)$ в точці $(x_i;\ y_i)$, тобто тангенс кута нахилу цієї прямої до осі Ox дорівнює $f(x_i, y_i)$. На першому кроці удосконаленого методу Ейлера (формула (37.4)) знаходимо точку перетину цієї



Мал. 44

прямої з прямою $x = x_i + \frac{h}{2}$. Цю точку на мал. 44 позначено через M, причому $y_{i+\frac{1}{2}} = y_i + \frac{h}{2} f(x_i, y_i)$. Після цього обчислюємо тангенс кута нахилу дотичної до кривої $y(x)$ в точці M: $f\left(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}\right)$. Ця дотична позначена через l_2. Якщо тепер через точку $(x_i;\ y_i)$ провести пряму l, паралельно прямій l_2, то внаслідок перетину з прямою $x = x_{i+1}$ дістанемо шукану точку $(x_{i+1};\ y_{i+1})$. Зауважимо, що при $i \neq 0$ точка $(x_i;\ y_i)$ не обов'язково лежить на інтегральній кривій, оскільки y_i є наближеним значенням розв'язку.

Якщо для обчислення інтеграла в (37.1) використати формулу трапецій, то можна дістати ще один метод для чисельного розв'язування задачі Коші, похибка якого на кожному кроці має також порядок $O(h^3)$.

Оскільки

$$\int_{x_i}^{x_{i+1}} y'(t)\,dt = \frac{h}{2}(y'(x_i) + y'(x_{i+1})) - \frac{h^3}{12} y'''(\eta),\quad \eta \in]x_i;\ x_{i+1}[,$$

то

$$y(x_{i+1}) = y(x_i) + \frac{h}{2}(f(x_i,\ y(x_i)) + f(x_{i+1},\ y(x_{i+1})) + O(h^3). \tag{37.6}$$

У правій частині формули (37.6) міститься невідоме значення $y(x_{i+1})$. Щоб дістати потрібні для обчислення формули, як і при виведенні формули удосконаленого методу Ейлера, замінимо $y(x_{i+1})$ у правій частині (37.6) за формулою (37.2). Тоді

$$y(x_{i+1}) = y(x_i) + \frac{h}{2}(f(x_i,\ y(x_i)) + f(x_{i+1},\ y(x_i)) +$$

$$+ hf(x_i,\ y(x_i)) + O(h^2)) + O(h^3) = y(x_i) + \frac{h}{2}(f(x_i,\ y(x_i)) +$$

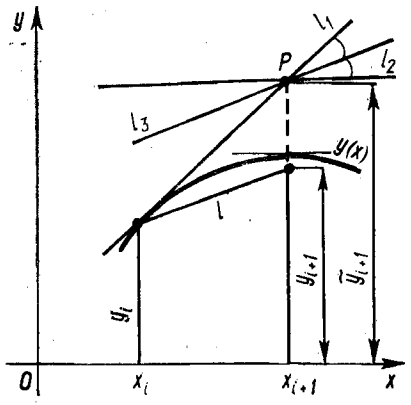
$$+ f(x_{i+1},\ y(x_i) + hf(x_i,\ y(x_i))) + O(h^3).$$

Звідси дістанемо пару таких формул:

$$\tilde{y}_{i+1} = y_i + hf(x_i,\ y_i), \tag{37.7}$$

$$y_{i+1} = y_i + \frac{h}{2}(f(x_i,\ y_i) + f(x_{i+1},\ \tilde{y}_{i+1})). \tag{37.8}$$

Ці формули визначають метод, який називається *удосконаленим методом Ейлера — Коші*.



Мал. 45

Геометричну ілюстрацію методу дано на мал. 45, де через l_1 позначено дотичну до кривої $y = y(x)$, у точці $(x_i;\ y_i)$, а через l_2 — пряму, тангенс кута нахилу якої до осі Ox такий самий, як і дотичної до $y(x)$ в точці $x = x_{i+1}$. Шукану точку $(x_{i+1};\ y_{i+1})$ дістанемо в результаті перетину прямої l, яка проходить через точку $(x_i;\ y_i)$, з прямою $x = x_{i+1}$. Причому пряма l паралельна прямій l_3, тангенс кута нахилу якої до осі Ox дорівнює середньому арифметичному кутових коефіцієнтів прямих l_1 і l_2.

Приклад.

Методами Ейлера з кроком $h = 0{,}1$ знайти чисельний розв'язок задачі Коші $y' = x + y$, $y_0 = 1$ на відрізку $[0;\ 0{,}5]$.

Результати обчислень подано в таблиці.

	Методи			
x_i	Ейлера y_i	Удосконалений Ейлера y_i	Ейлера — Коші y_i	Точний розв'язок $2e^x - x - 1$
0	1	1	1	1
0,1	1,1	1,11	1,11	1,1103
0,2	1,22	1,2421	1,2421	1,2428
0,3	1,362	1,3985	1,3985	1,3997
0,4	1,5282	1,5819	1,5819	1,5836
0,5	1,7210	1,7950	1,7950	1,7974

Як бачимо, за удосконаленим методом Ейлера і удосконаленим методом Ейлера — Коші дістали однакові результати.

З формули (37.6) для визначення шуканого значення розв'язку в точці x_{i+1} можна дістати формулу

$$y_{i+1} = y_i + \frac{h}{2}(f(x_i,\ y_i) + f(x_{i+1},\ y_{i+1})), \qquad (37.9)$$

похибка якої має порядок $O(h^3)$.

Ми маємо, таким чином, ще один метод чисельного розв'язування задачі Коші. Він принципово відрізняється від раніше розглянутих методів, які відносять до так званих *явних методів*: у них значення y_{i+1} визначається через значення розв'язку в попередній точці явними формулами. Метод, що визначається формулою (37.9), належить до *неявних методів*: щоб знайти значення y_{i+1}, користуються алгебраїчним рівнянням (37.9). Тому виникає питання про розв'язування цього рівняння. Рівняння (37.9) можна розв'язати, наприклад, методом послідовних наближень:

$$y_{i+1}^{(k)} = y_i + \frac{h}{2}(f(x_i,\ y_i) + f(x_{i+1},\ y_{i+1}^{(k-1)})), \quad k = 1,\ 2,\ \ldots \qquad (37.10)$$

За початкове наближення $y_{i+1}^{(0)}$ візьмемо значення, обчислене методом Ейлера, тобто

$$y_{i+1}^{(0)} = y_i + hf(x_i,\ y_i). \qquad (37.11)$$

Якщо ітераційний процес (37.10) збігається, то можна знайти шукане значення y_{i+1}. Ітерації припиняють, коли

$$|y_{i+1}^{(k)} - y_i^{(k-1)}| < \varepsilon, \qquad (37.12)$$

де ε — деяка задана точність. Якщо умова (37.12) виконується, то покладають $y_{i+1} = y_{i+1}^{(k)}$. Визначимо умови, при яких ітераційний процес (37.10) збігається. Віднявши від рівності (37.9) рівність (37.10), матимемо:

$$|y_{i+1} - y_{i+1}^{(k)}| = \frac{h}{2}|f(x_{i+1},\ y_{i+1}) - f(x_{i+1},\ y_{i+1}^{(k-1)})|. \qquad (37.13)$$

Якщо функція $f(x, y)$ за змінною y задовольняє умову Ліпшиця, тобто

$$|f(x_{i+1},\ y_{i+1}) - f(x_{i+1},\ y_{i+1}^{(k-1)})| \leqslant K_i |y_{i+1} - y_{i+1}^{(k-1)}|,$$

де K_i — константа, то з (37.13) дістанемо:

$$|y_{i+1} - y_{i+1}^{(k)}| \leqslant \frac{h}{2} K_i |y_{i+1} - y_{i+1}^{(k-1)}|.$$

Звідси випливає, що

$$|y_{i+1} - y_{i+1}^{(k)}| \leqslant \left(\frac{h}{2} K_i\right)^k |y_{i+1} - y_{i+1}^{(0)}|.$$

Отже, якщо $\frac{h}{2} K_i < 1$, то $y_{i+1}^{(k)} \to y_{i+1}$ при $k \to \infty$. Ітераційний процес збігається тим швидше, чим менша величина $\frac{h}{2} K_i$.

Тому $\frac{h}{2}K_l$ розглядається тут як коефіцієнт збіжності. Очевидно, при $\frac{h}{2}K_l < 1$ відображення, що задається правою частиною (37.9), стискуюче (див. § 7).

Розглянутий метод називається *удосконаленим методом Ейлера з ітераційною обробкою.*

Таким чином, завжди можна підібрати такий малий крок h, щоб останній метод Ейлера збігався. Під час обчислень користуються правилом: якщо після трьох-чотирьох ітерацій умова (37.12) не справджується, то потрібно зменшити крок h.

Приклад.

Методом Ейлера з ітераційною обробкою з точністю до $0{,}5 \cdot 10^{-4}$ знайти на відрізку $[0;\ 0{,}5]$ розв'язок задачі Коші $y' = x + y$, $y(0) = 1$. Візьмемо крок $h = 0{,}1$.

Спочатку за формулою (37.11) знайдемо наближення $y_1^{(0)}$

$$y_1^{(0)} = y_0 + hf(x_0,\ y_0) = 1 + 0{,}1\,(0 + 1) = 1{,}1.$$

Тоді за формулою (37.10) знаходимо послідовні наближення для y_1:

$$y_1^{(1)} = 1 + \frac{0{,}1}{2}(1 + 0{,}1 + 1{,}1) = 1{,}11;$$

$$y_1^{(2)} = 1 + \frac{0{,}1}{2}(1 + 0{,}1 + 1{,}11) = 1{,}1105;$$

$$y_1^{(3)} = 1 + \frac{0{,}1}{2}(1 + 0{,}1 + 1{,}1105) = 1{,}1105.$$

Оскільки $y_1^{(2)}$ і $y_1^{(3)}$ збігаються до чотирьох десяткових знаків, то покладемо $y_1 = 1{,}1105$.

Аналогічно за значеннями $x_1 = 0{,}1$ та $y_1 = 1{,}1105$ знайдемо значення y_2:

$$y_2^{(0)} = 1{,}1105 + 0{,}1\,(0{,}1 + 1{,}1105) = 1{,}2316;$$

$$y_2^{(1)} = 1{,}1105 + (0{,}1/2)\,(0{,}1 + 1{,}1105 + 0{,}2 + 1{,}2316) = 1{,}2426;$$

$$y_2^{(2)} = 1{,}1105 + (0{,}1/2)\,(1{,}2105 + 0{,}2 + 1{,}2426) = 1{,}2432;$$

$$y_2^{(3)} = 1{,}1105 + (0{,}1/2)\,(1{,}2105 + 0{,}2 + 1{,}2432) = 1{,}2432.$$

Отже, $y_2 = 1{,}2432$. Аналогічно обчислюють решту значень. Результати обчислень подано в таблиці.

x	0	0,1	0,2	0,3	0,4	0,5
y	1	1,1105	1,2432	1,4004	1,5847	1,7989

Вправи.

1. За допомогою розкладу в ряд Тейлора знайти розв'язок задачі Коші

$$y' = x + y,$$
$$y(0) = 1$$

для $x = 0{,}1$ і $x = 0{,}2$. Взяти перші п'ять членів ряду.

2. Тіло з початковою масою 200 кг рухається під дією постійної сили, що становить 2000 Н, причому величина сили опору повітря дорівнює подвійному значенню швидкості. Маса тіла зменшується за кожну секунду на 1 кг. Потрібно знайти швидкість тіла через 50 сек, якщо в початковий момент $t = 0$ тіло перебувало в спокої.

Якщо через v позначимо швидкість тіла, яка є функцією від t, то для швидкості маємо таке диференціальне рівняння:

$$\frac{dv}{dt} = \frac{2000 - 2v}{200 - t}, \quad \text{причому} \quad v(0) = 0.$$

Розв'язати задачу чисельно з кроком $h = 10$: а) методом Ейлера; б) удосконаленим методом Ейлера; в) удосконаленим методом Ейлера — Коші.

Результати порівняти із значенням $v(50)$, яке дістанемо з точного розв'язку рівняння $v(t) = 10t - \frac{1}{40}t^2$.

СПИСОК ВИКОРИСТАНОЇ ТА РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

1. Швец М. Н. О приближенных числах. К. : Рад. шк., 1968. 124 с.
2. Тесленко І. Ф., Слєдзінський І. Ф. Метричні простори в шкільному курсі математики. К. : Рад. шк., 1978. 110 с.
3. Вайнберг М. М. Функциональный анализ. Специальный курс для пединститутов. М. : Просвещение, 1979. 128 с.
4. Жалдак М. І. Про лінійне програмування.— У кн.: У світі математики. Вип. 2. К. : Рад. шк., 1970, с. 40—65.
5. Жалдак М. І., Ковбасенко Б. С., Рамський Ю. С. Обчислювальна математика. К. : Рад. шк., 1973. 184 с.
6. Пулькин С. П. Вычислительная математика. М. : Просвещение, 1972. 272 с.
7. Демидович Б. П., Марон И. А. Основы вычислительной математики. М. : Наука, 1970. 664 с.
8. Демидович Б. П., Марон И. А., Шувалова Э. З. Численные методы анализа. М. : Физматгиз, 1963. 400 с.

ЗМІСТ

Мирослав Иванович Жалдак,
Юрий Савиянович Рамский

Численные методы математики

Пособие для самообразования учителей

(На украинском языке)

Киев, «Радянська школа»

Зав. редакцією математики *О. П. Бондаренко.*
Редактор *Л. П. Слабошицька.* Літредактор *Л. О. Поповченко.* Художній редактор *П. В. Кузь.* Обкладинка художника *М. Ю. Севрюгіної.* Технічний редактор *Н. М. Горбунова.* Коректор *Н. П. Васеніна.*

Інформ. бланк № 2386

Здано до набору 13.01.84. Підписано до друку 30.07.84. Формат 84×108/32. Папір № 1 друкарськ. Гарнітура літерат. Спосіб друку високий. Умовн. арк. 10,92. Умовн. фарбо.-відб. 11,13. Обл.-видавн. арк. 10,05. Тираж 10000 пр.. Видавн. № 27367. Зам. № 30. Ціна 50 к.

Видавництво «Радянська школа»
252053. Київ, Ю. Коцюбинського, 5

Надруковано з матриць Головного підприємства РВО «Поліграфкнига» з Дніпропетровській обласній друкарні. 320091, Дніпропетровськ, вул. Горького, 20.

КНИЖКОВА ПОЛИЦЯ

У видавництві «Радянська школа» вийшли і вийдуть у світ у 1984 році такі методичні посібники для вчителів математики:

Авраменко М. І. *Уроки алгебри і початків аналізу в 10 класі:* К.: Рад. шк., 1984.— 10 арк.— Мова укр.— 45 к.

У посібнику подано розробки уроків алгебри й початків аналізу з усіх програмних тем 10 класу. Чітко визначено структурні елементи й дидактичну мету кожного уроку. Значну увагу приділено світоглядним проблемам, організації самостійної творчої діяльності учнів, проведенню уроків-семінарів, технічному оснащенню уроків.

Бурда М. І. *Вивчення геометрії в 7 класі.*— К.: Рад. шк., 1984.— 8 арк.— Мова укр.— 25 к.

У посібнику розглядається методика вивчення геометрії в 7 класі за підручником О. В. Погорелова.

Подаються поради щодо погодинного планування, викладу теоретичного матеріалу і розв'язування задач. Наводяться розробки уроків за всіма програмними темами. До кожного уроку пропонується система додаткових вправ. Значна увага приділяється зразкам записів розв'язань, міркуванням учнів.

Возняк Г. М., Маланюк К. П. *Прикладна спрямованість шкільного курсу математики.*— К.: Рад. шк., 1984.— 8 арк.— Мова укр.— 30 к.

У посібнику подаються методичні рекомендації щодо добору і способів розв'язування екстремальних задач прикладного характеру в 4—8 класах. Особлива увага приділяється тому, як навчити учнів складати математичні моделі реальних ситуацій, висувати, перевіряти гіпотези, робити висновки й узагальнення. Наведено понад 200 алгебраїчних і геометричних задач з докладними розв'язаннями.

Середа В. Ю. *Математична логіка в шкільному курсі математики.*— К.: Рад. шк., 1984.— 8 арк.— Мова укр.— 25 к.

У першій частині посібника в довідковій формі подаються теоретичні відомості з математичної логіки в обсязі, достатньому для характеристики логічних основ шкільного курсу математики. У другій частині розкриваються зв'язки найважливіших понять традиційної логіки з відповідними