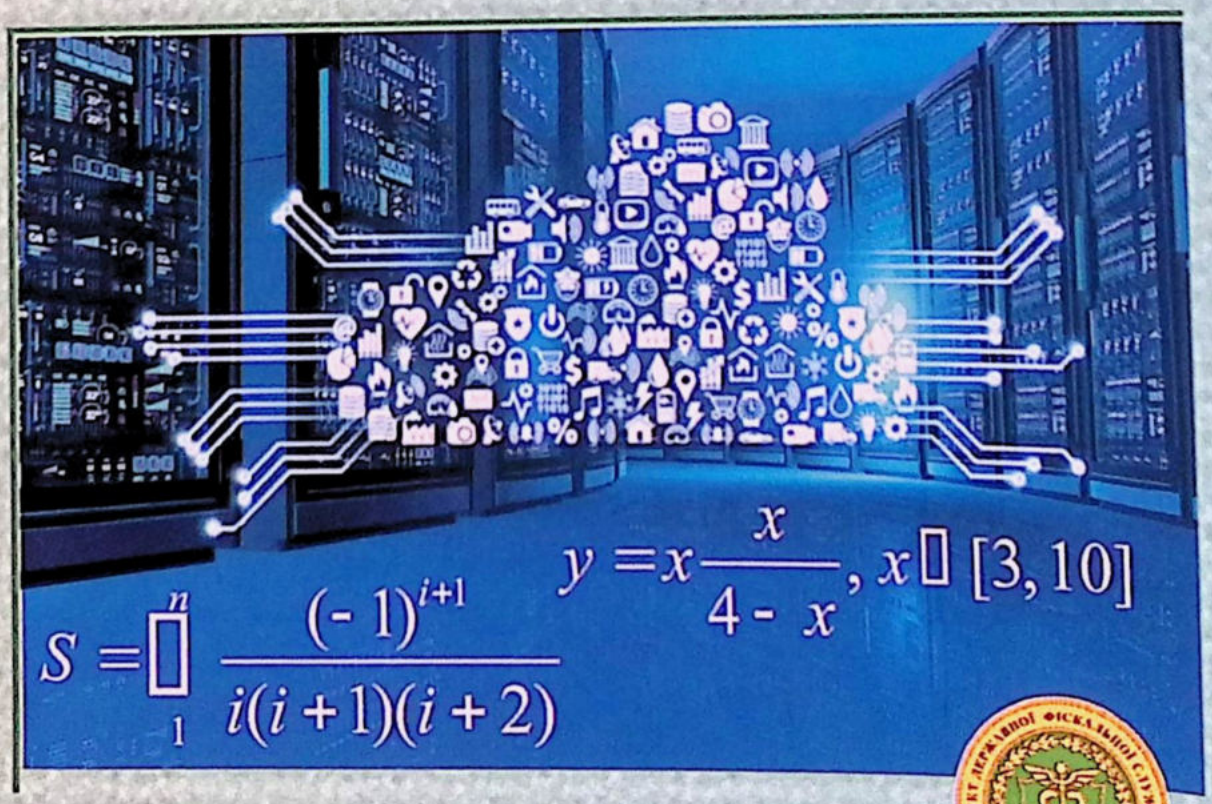
НА ДОПОМОГУ СТУДЕНТУ УДФСУ

Антонюк А. О.

ЧИСЕЛЬНІ МЕТОДИ



СЕРІЯ «НА ДОПОМОГУ СТУДЕНТУ УДФСУ»

Серію «На допомогу студенту УДФСУ» засновано 2016 року.

Редакційна колегія:
Пашко П. В., д.е.н. (голова)
Шевчук О. А., д.е.н. (заступник голови)
Топчій В. В., д.ю.н.
Мацелюх Н. П., к.е.н.
Куженєв М. О., д.е.н.
Швабій К. І., д.е.н.
Ріппа С. П., д.е.н.
Мандрагеля В. А., д.філософ.н.
Чмелюк В. В., к.ю.н.
Малинський І. Й., к.н.фіз.вих.
Шевчук В. А., к.ю.н.

У СЕРІЇ «НА ДОПОМОГУ СТУДЕНТУ УДФСУ» ВИЙШЛИ ДРУКОМ:

2016

«Методичні основи спеціальної фізичної та технічної підготовки студентів за розділом «Легка атлетика»

«Самостійна робота студента як одна з форм впливу на функціональну, фізичну та психологічну підготовленість»

«Організація роботи командира механізованого взводу»

«5.45-мм автомати Калашникова (АК-74, АКС-74, АК-74Н, АКС-74Н) та 5.45-мм ручні кулемети Калашникова (РПК-74, РПКС-74, РПК-74Н, РПКС-74Н)»

«Гранатомет підствольний ГП-25»

«Ручні гранати»

«Кулемети Калашникова – 7.62, ПК, УЖМ, ПКТ»

«Ручний протитанковий гранатомет РПГ-7»

«9-мм пістолет Макарова (ПМ)»

2017

«Вища та прикладна математика»

«Цивільний захист»

«Програмування мовою JAVA : практикум»

«Інформаційні системи і технології в юридичній практиці»

«Дослідження операцій : практикум»

«Чисельні методи»

«English for Students of Finance»

«Основи військової розвідки»

ДЕРЖАВНА ФІСКАЛЬНА СЛУЖБА УКРАЇНИ
УНІВЕРСИТЕТ ДЕРЖАВНОЇ ФІСКАЛЬНОЇ СЛУЖБИ УКРАЇНИ

СЕРІЯ «НА ДОПОМОГУ СТУДЕНТУ УДФСУ»
Заснована 2016 року

А. О. Антонюк

ЧИСЕЛЬНІ МЕТОДИ

Навчальний посібник

Ірпінь
2017

УДК 519.6(075.8)
ББК 22.19я73
А 72

Рекомендовано до друку Вченою радою
Університету державної фіскальної служби України
(протокол № 6 від 25 квітня 2017 року)

Рецензенти:
Стецюк П. І., професор, доктор фізико-математичних наук, завідувач відділу методів негладкої оптимізації Інституту кібернетики ім. В. М. Глушкова;
Семко М. М., професор, доктор фізико-математичних наук, завідувач кафедри вищої математики Університету ДФС України.

Антонюк А. О.
А 72 Чисельні методи : навчальний посібник / А. О. Антонюк ; Державна фіскальна служба України, Університет державної фіскальної служби України. – Ірпінь, 2017. – 144 с. – (Серія «На допомогу студенту УДФСУ», т. 15).

ISBN 978-966-337-453-6

У навчальному посібнику розкриваються постановки математичних задач, що виникають у ході розробки інформаційних систем і систем моделювання, та чисельні методи їх розв'язання. Його структура відповідає робочій навчальній програмі і сприяє розвитку навичок реалізації на комп'ютері чисельних методів та роботи з відомими комп'ютерними математичними пакетами.

Призначений для викладачів, студентів денної та заочної форм навчання спеціальності «Комп'ютерні науки», а також усіх бажаючих, кого цікавлять питання реалізації чисельних методів на комп'ютері.

УДК 519.6(075.8)
ББК 22.19я73

© Антонюк А. О., 2017
© Університет державної фіскальної служби України, 2017

ISBN 978-966-337-453-6

ЗМІСТ

Позначення 6
Вступ 7
1. Вступ до чисельних методів. Загальні поняття 13
1.1. Сутність чисельних методів. Загальні поняття 13
1.2. Характеристики чисельних методів 16
1.3. Похибка розв'язку 18
1.4. Похибка округлення при розрахунках на комп'ютері з плаваючою крапкою 25
1.5. Математичні пакети 27
Контрольні запитання та завдання 29
2. Прямі методи розв'язання систем лінійних рівнянь 31
2.1. Чисельні методи розв'язування СЛАР 31
2.2. Метод Гаусса. Прямий та обернений хід 32
2.3. Вибір головного елемента 34
2.4. Схема *LU*-розкладання матриці 37
2.5. Метод квадратного кореня 38
2.6. Обчислення визначника та оберненої матриці 39
2.7. Метод прогонки 40
2.8. Обумовленість систем лінійних алгебраїчних рівнянь 42
Контрольні запитання та завдання 44
3. Розв'язання систем лінійних рівнянь великої розмірності 46
3.1. СЛАР великої розмірності 46
3.2. Розріджені матриці 46
3.3. Ітераційні методи розв'язання СЛАР великої розмірності з розрідженими матрицями 49
3.4. Метод простої ітерації 50
3.5. Метод Гаусса-Зейделя 50
3.6. Метод Якобі 52
3.7. Однокрокові методи 53
Контрольні запитання та завдання 54
4. Чисельні методи розв'язування нелінійних рівнянь 55
4.1. Постановка задачі 55
4.2. Алгоритми ізолювання та уточнення коренів 56

4.3. Графічний спосіб 57
4.4. Метод ділення навпіл (дихотомії) 58
4.5. Метод простої ітерації 59
4.6. Метод Ньютона (метод дотичних) 60
4.7. Методи розв'язання систем нелінійних рівнянь 63
Контрольні запитання та завдання 66
5. Обчислення власних значень і власних векторів матриці 67
5.1. Постановка задачі 67
5.2. Степеневий метод 68
5.3. *QR*-алгоритм 70
Контрольні запитання та завдання 71
6. Чисельне диференціювання та інтегрування функції 73
6.1. Чисельне диференціювання функцій 73
6.2. Формули чисельного диференціювання на основі першої інтерполяційної формули Ньютона 74
6.3. Формули чисельного диференціювання на основі інтерполяційного полінома Лагранжа 76
6.4. Формули диференціювання для практичних обчислень 77
6.5. Чисельне інтегрування функції 78
6.6. Формули прямокутників 81
6.7. Формула трапецій 84
6.8. Формула Сімпсона 85
Контрольні запитання та завдання 86
7. Розв'язання задачі Коші для звичайних диференціальних рівнянь 88
7.1. Постановка задачі Коші 88
7.2. Метод послідовних наближень (метод Пікара) 89
7.3. Метод рядів Тейлора 89
7.4. Метод Ейлера та його модифікації 90
7.5. Метод Рунге-Кутта четвертого порядку 93
Контрольні запитання та завдання 94
8. Багатокрокові методи розв'язання диференціальних рівнянь . 96
8.1. Поняття багатокрокового методу 96
8.2. Метод Адамса-Бошфорда 96
8.3. Метод Адамса-Мултона 97
8.4. Метод прогнозу та корекції 98

Контрольні запитання та завдання 98
9. Неявні методи розв'язання жорстких задач Коші 99
9.1. Поняття жорсткої системи диференціальних рівнянь 99
9.2. Неявні методи Ейлера і Рунге-Кутта 100
Контрольні запитання та завдання 101
10. Крайові задачі для звичайних диференціальних рівнянь 102
10.1. Постановка крайової задачі 102
10.2. Методи стрільби 103
10.3. Кінцево-різницеві методи 103
10.4. Метод прогонки 105
Контрольні запитання та завдання 106
11. Розв'язання диференціальних рівнянь
з частинними похідними 107
11.1. Поняття задач математичної фізики 107
11.2. Класифікація диференціальних рівнянь
з частинними похідними 107
11.3. Метод кінцевих різниць 111
11.4. Метод кінцевих елементів 113
11.5. Розв'язання параболічних рівнянь 115
11.6. Розв'язання гіперболічних рівнянь 119
11.7. Розв'язання еліптичних рівнянь 121
Контрольні запитання та завдання 123
12. Чисельні методи наближення функцій 124
12.1. Апроксимація, інтерполяція та екстраполяція 124
12.2. Постановка задачі інтерполяції 125
12.3. Інтерполяційний поліном Лагранжа 127
12.4. Інтерполяційний поліном Ньютона 129
12.5. Кусково-лінійна та квадратична інтерполяція 133
12.6. Сплайн-інтерполяція 135
12.7. Метод найменших квадратів для апроксимації функцій ... 137
12.8. Поняття екстраполяції функцій 141
Контрольні запитання та завдання 142
Список використаних джерел 143

ПОЗНАЧЕННЯ

Позначення	Значення
R^1	простір дійсних чисел
$x \in \Omega$	x належить множині Ω (x – елемент множини Ω)
$i = \overline{1,n}$	i приймає значення від 1 до n (еквівалентно $i = 1,\ldots,n$)
$\forall x \in \Omega$	для всіх x, що належать множині Ω
$\exists x \in \Omega$	існує x, що належить множині Ω
$k >> 1$	число k набагато більше, ніж 1
R^n	n-вимірний простір дійсних вектор-стовпців
$x \in R^n$	x – n-вимірний вектор-стовпець
x^*	вектор-рядок, транспонований до x
$x^{(k)} \in R^n$	$x^{(k)}$ – n-вимірний вектор-стовпець, k -й за номером
x_i	i-та компонента вектора x
$\{x^{(k)}\}$	послідовність $x^{(k)}$, тобто $x^{(0)}, x^{(1)}, x^{(2)},\ldots$
$\{x \in \Omega: T\}$	підмножина елементів множини Ω, що задовольняє умові T
$A = (a_{ij})$	матриця з елементами a_{ij}, де i – номер рядка, j – номер стовпця
A^*	матриця, транспонована до A
A^{-1}	матриця, обернена до A
$R^{n\times m}$	множина матриць розмірності n на m ($n \times m$)
$E \in R^{n\times m}$	одинична матриця розмірності n на m
$f: R^1 \to R^1$	дійсна функція однієї змінної
$f: R^n \to R^m$	дійсна вектор-функція n змінних і m значень функції
(x,y)	скалярний добуток векторів x та y, $(x,y) = \sum_{i=1}^{n} x_i y_i$
$\|x\|$	евклідова норма вектора x, $\|x\| = \left[\sum_{i=1}^{n} x_i^2\right]^{1/2} = \sqrt{(x,x)}$
$V_r(x^*)$	окіл точки $x^* \in R^n$ радіусу r, $V_r(x^*) = \{x \in R^n : \|x - x^*\| \le r\}$
δ_{ik}	символ Кронекера, $\delta_{ik} = \begin{cases} 1, & \text{якщо } i = k, \\ 0, & \text{якщо } i \ne k. \end{cases}$

ВСТУП

Не буває зайвої інформації,
буває невміння нею розпорядитися

Неможливо уявити собі сучасну науку без широкого застосування математичного моделювання. Сутність цієї методології полягає в заміні початкового об'єкта математичною моделлю з метою подальшого вивчення моделі за допомогою реалізованих на комп'ютерах обчислювально-логічних алгоритмів.

Цей метод пізнання, конструювання, проектування містить у собі багато переваг як теорії, так і експерименту. Робота не з самим об'єктом (явищем, процесом), а з його моделлю дає можливість безболісно, швидко і без істотних витрат досліджувати його властивості і поведінку в будь-яких ситуаціях (переваги теорії). Водночас обчислювальні (комп'ютерні, симуляції, імітаційні) експерименти з моделями об'єктів дозволяють, спираючись на потужність сучасних обчислювальних методів і технічних інструментів інформатики, детально і глибоко вивчати об'єкти в достатній повноті, неприступній суто теоретичним підходам (переваги експерименту). Не дивно, що методологія математичного моделювання бурхливо розвивається, охоплюючи всі нові сфери – від розробки технічних систем і управління ними до аналізу найскладніших економічних і соціальних процесів.

Елементи математичного моделювання використовувалися з початку появи точних наук, і не випадково, що деякі методи обчислень носять імена таких корифеїв науки, як Ньютон і Ейлер, а слово «алгоритм» походить від імені арабського вченого Аль-Хорезмі. Друге «народження» цієї методології припало на кінець 40-х – початок 50-х років XX століття і було обумовлено принаймні двома причинами. Перша з них – поява ЕОМ (комп'ютерів), хоча і скромних за нинішніми мірками, позбавила вчених від величезної за обсягом рутинної обчислювальної роботи. Друга – безпрецедентне соціальне замовлення – виконання національних

програм СРСР і США зі створення ракетно-ядерного щита, які не могли бути реалізовані традиційними методами.

Математичне моделювання справилося з цією задачею: ядерні вибухи і польоти ракет і супутників були попередньо «здійснені» в надрах ЕОМ за допомогою математичних моделей і лише потім втілені на практиці. Цей успіх у цілому визначив подальші досягнення методології, без застосування якої в розвинених країнах жоден великомасштабний технологічний, екологічний або економічний проект тепер всерйоз не розглядається (сказане справедливо і щодо деяких соціально-політичних проектів).

Зараз математичне моделювання вступає в третій принципово важливий етап свого розвитку, «вбудовуючись» у структури так званого інформаційного суспільства. Вражаючий прогрес засобів переробки, передачі й зберігання інформації відповідає світовим тенденціям про ускладнення і взаємне проникнення різних сфер людської діяльності. Без володіння інформаційними «ресурсами» не можна і думати про вирішення складних і різноманітних проблем, що стоять перед світовим співтовариством.

Проте інформація як така часто мало що дає для аналізу і прогнозу, для ухвалення рішень і контролю за їх виконанням. Потрібні надійні способи переробки інформаційної «сировини» в готовий «продукт», тобто в точне знання. Історія методології математичного моделювання переконує: вона може і має бути інтелектуальним ядром інформаційних технологій усього процесу інформатизації суспільства.

Технічні, екологічні, економічні й інші системи, що вивчаються сучасною наукою, більше не піддаються дослідженню (у потрібній повноті та точності) звичайними теоретичними методами. Прямий натурний експеримент над ними – це довгий, дорогий, часто або небезпечний, або просто неможливий, оскільки багато з цих систем існують у «єдиному екземплярі». Ціна помилок і прорахунків в обігу з ними неприпустимо висока. Тому математичне (ширше – інформаційне) моделювання є обов'язковою складовою науково-технічного прогресу.

Сама постановка питання про математичне моделювання будь-якого об'єкта породжує чіткий план дій. Його можна умовно розбити на три етапи: **модель – алгоритм – програма.**

На **першому етапі** вибирається (або будується) «еквівалент» об'єкта, що відображає в математичній формі найважливіші його властивості – закони, яким він підкоряється, зв'язки, властиві складовим його частинам, і т.п. Математична модель (або її фрагменти) досліджується теоретичними методами, що дозволяє отримати важливі попередні знання про об'єкт.

Другий етап – вибір (або розробка) алгоритму для реалізації моделі на комп'ютері. Модель представляється у формі, зручній для застосування чисельних методів, визначається послідовність операцій, які потрібно провести, щоб знайти шукані величини із заданою точністю. Обчислювальні алгоритми повинні не спотворювати основні властивості моделі і початкового об'єкта, бути економічними і такими, що адаптуються до особливостей вирішуваних завдань і використовуваних комп'ютерів.

На **третьому етапі** створюються програми, що «перекладають» модель і алгоритм на доступну комп'ютеру мову. До них також пред'являються вимоги економічності й адаптивності. Їх можна назвати «електронним» еквівалентом об'єкта, що вивчається, уже придатним для безпосереднього випробування на «експериментальній установці» – комп'ютері.

Створивши **тріаду «модель-алгоритм-програма»**, дослідник отримує в руки універсальний, гнучкий і недорогий інструмент, який спочатку підладжується, тестується у «пробних» обчислювальних експериментах. Після того, як адекватність тріади початковому об'єкту засвідчена, з моделлю проводяться різноманітні й докладні «досліди», що дають усі потрібні якісні та кількісні властивості й характеристики об'єкта.

Процес моделювання супроводжується поліпшенням і уточненням, у міру необхідності, усіх ланок тріади. Будучи методологією, математичне моделювання не підміняє собою математику, фізику, біологію та інші наукові дисципліни, не конкурує з ними. Створення і застосування тріади неможливе без опори на різні

методи і підходи – від якісного аналізу нелінійних моделей до сучасних мов програмування.

Воно дає нові додаткові стимули різнонаправленням науки. Розглядаючи питання ширше, нагадаємо, що моделювання присутнє майже у всіх видах творчої активності людей різних «спеціальностей» – дослідників і підприємців, політиків і воєначальників. Звичайно ж, математичне моделювання регулятивне лише при виконанні добре відомих професійних вимог: чітке формулювання основних понять і припущень, апостеріорний аналіз адекватності моделей, гарантована точність обчислювальних алгоритмів і т.п.

Основні етапи математичного моделювання:

1. Розробка моделі – **формалізація.** Вивчається в прикладних і фундаментальних науках.

2. Розробка методу (алгоритму) розв'язання рівняння моделі – **алгоритмізація.** Вивчається в обчислювальній математиці.

3. Створення програми – **програмування.** Вивчається в інформатиці.

4. Розрахунки, аналіз результатів – **практичне використання.**

Щоб зрозуміти місце чисельних методів у процесі розв'язання практичних задач, що виникають у результаті математичного моделювання, з використанням комп'ютера, слід навести основні етапи цього процесу.

Етапи розв'язання практичних задач на комп'ютері.

1. Постановка задачі:

- словесне формулювання задачі;
- визначення кінцевої мети розв'язку.

2. Побудова математичної моделі, тобто математичне формулювання задачі.

3. Вибір чисельного методу для розв'язання математичної задачі.

4. Розроблення алгоритму.

5. Програмна реалізація алгоритму.

6. Тестування програми (налагодження на тестових задачах).

7. Проведення розрахунків на реальних даних.

8. Аналіз результатів.

Найскладнішим із перерахованих етапів є другий етап, а вивчення методів його реалізації є предметом інших навчальних дисциплін (наприклад, моделювання систем й ін.). У подальшому викладенні матеріалу цього навчального посібника передбачено, що математичне формулювання задачі вже є, потрібно тільки навчитися її розв'язувати на комп'ютері з використанням чисельних методів.

Варто зазначити, що якщо математична модель вибрана недостатньо коректно, то які б методи не застосовувалися для розрахунків з її використанням, отримані висновки будуть ненадійні, або й зовсім неправильні.

У ході розв'язання конкретної практичної задачі спеціаліст повинен, перш за все, визначити, до якого типу математичної задачі належить ця практична задача, вибрати чисельний метод для її розв'язання, розробити програмну реалізацію методу самостійно чи вміти застосувати для її розв'язання один із відомих комп'ютерних математичних пакетів.

Вивчення навчальної дисципліни «Чисельні методи» дозволяє студентам оволодіти знаннями в галузі практичних методів рішення математичних проблем, що виникають у процесі інженерної діяльності та моделювання фізичних систем, засвоїти способи розрахунків на сучасних комп'ютерах із застосуванням пакетів спеціальних прикладних програм.

У навчальному посібнику висвітлюються питання, пов'язані з розкриттям основних понять предметної області: математичної постановки задачі, чисельного методу, ітерації, характеристик чисельних методів, абсолютної та відносної похибок розв'язку, джерел погрішності та інших; розглянуті постановки основних математичних задач та найбільш відомі чисельні методи їх розв'язання.

Об'єктом вивчення навчальної дисципліни є типові математичні задачі, до яких зводиться рішення практичних проблем, що виникають у ході розробки інформаційних систем та систем

моделювання. Предметом вивчення навчальної дисципліни є чисельні методи розв'язання типових математичних задач.

Мета цього навчального посібника – ознайомлення студентів із постановками основних математичних задач і чисельних методів їх розв'язання, набуття студентами навичок реалізації на комп'ютері чисельних методів, навичок роботи з відомими комп'ютерними математичними пакетами.

Видання цієї книги було б неможливим без підтримки колег та друзів автора. Щира подяка всім, хто брав участь у її підготовці. Зокрема, Носовець Тетяні Василівні – за методичну та технічну допомогу, Шевчук Ірині Миколаївні – за технічну допомогу, Постілу Степану Дмитровичу – за методологічну та методичну підтримку і, нарешті, дружині Наталії Григорівні – за велику моральну підтримку.

1. ВСТУП ДО ЧИСЕЛЬНИХ МЕТОДІВ. ЗАГАЛЬНІ ПОНЯТТЯ

1.1. Сутність чисельних методів. Загальні поняття

Для розв'язання математичних задач в основному існує три групи методів:

1. **Аналітичні методи**, в яких розв'язок задачі подається у вигляді аналітичних виразів. Їх **перевагами** є: запис розв'язку у загальному вигляді; висока точність і малий об'єм комп'ютерної пам'яті для зберігання розв'язку. Основний **недолік** – неуніверсальність, оскільки лише невелика частина математичних задач може бути розв'язана аналітично.

2. **Графічні методи**, в яких розв'язок задачі знаходиться візуально. Їх **перевагою** є наочність. **Недоліками** графічних методів є: велика трудомісткість; низька точність (залежить від точності побудови графіків); неуніверсальність (графіки можна побудувати тільки для невеликої розмірності та ін.).

3. **Чисельні методи**, що дозволяють звести розв'язування задачі до виконання скінченного числа арифметичних і логічних дій з числами. При цьому розв'язок визначається як набір чисел, які надалі можуть бути інтерпретовані різним способом (наприклад, подані у вигляді таблиць, графіків, анімації тощо). Їх **перевагами** є: абсолютна універсальність, бо теоретично можуть бути застосовані для розв'язання будь-яких задач; добре пристосовані для реалізації на комп'ютері. **Недоліком** є велика трудомісткість у ході ручного рахунку, що зараз не є проблемою, оскільки вони призначені для використання на комп'ютері.

Таким чином, чисельні методи є основним апаратом розв'язання математичних задач, а їх значущість тільки збільшуватиметься у міру вдосконалення комп'ютерної техніки.

Чисельні методи бувають двох типів: **прямі** та **ітераційні**. У прямих методах розв'язок задачі досягається за скінченну кількість **кроків** методу після виконання останнього кроку, в ітераційних методах виконується ряд **ітерацій** методу до отримання наближеного розв'язку із заданою точністю.

В основному чисельні методи є **ітераційними**. Щоб зрозуміти сутність **ітераційного методу**, розглянемо задачу розв'язання рівняння вигляду:

$$x^2 = a.$$

З погляду звичайної математики точний розв'язок цього рівняння такий:

$$x^*_{1,2} = \pm\sqrt{a},$$

причому, якщо $a > 0$, то є два дійсних розв'язки;

якщо $a = 0$, то є тільки тривіальний розв'язок $x^*_1 = x^*_2 = 0$;

якщо $a < 0$, то дійсних розв'язків немає.

Але знак $\sqrt{}$ не вирішує задачу, оскільки не дає практичного способу (алгоритму) обчислення значення x для конкретного значення a.

Тут зазначимо, що чисельні методи не вивчають методи розв'язання конкретних задач, а пропонують алгоритми розв'язання задач у абстрактному вигляді. У цьому випадку чисельний метод пропонує такий алгоритм обчислення x^*:

1) вибрати початкове значення x, наприклад $x_0 = a$ – початкове наближення розв'язку;

2) далі обчислювати нові наближення розв'язку x_i за формулою:

$$x_i = \frac{1}{2}(x_{i-1} + \frac{a}{x_{i-1}})$$

до досягнення умови:

$$|x_i - x_{i-1}| \le \varepsilon.$$

Тут $i = 1,2\ldots$ – номер обчислення i-ої ітерації, ε – необхідна точність.

Приклад. Розв'язати рівняння $x^2 - 3 = 0$ з точністю $\varepsilon = 0{,}001$.

Розв'язання. Задамо $x_0 = a = 3$. Обчислимо перше наближення: $x_1 = \frac{1}{2}(x_0 + \frac{a}{x_0}) = \frac{1}{2}(3 + \frac{3}{3}) = 2$ та оцінимо точність:

$|x_1 - x_0| = |2-3| = 1 \geq \varepsilon$. Тобто необхідна точність не досягнута, а отже, потрібно продовжити розрахунок.

Обчислимо друге наближення: $x_2 = \frac{1}{2}(x_1 + \frac{a}{x_1}) = \frac{1}{2}(2 + \frac{3}{2}) = 1{,}75 = \frac{7}{4}$, оцінимо точність: $|1{,}75 - 2| = 0{,}25 > \varepsilon$.

Обчислимо третє наближення: $x_3 = \frac{1}{2}(x_2 + \frac{a}{x_2}) = \frac{1}{2}(\frac{7}{4} + \frac{12}{7}) = 1{,}7321$, оцінимо точність: $|1{,}7321 - 1{,}75| = 0{,}0179 > \varepsilon$.

Обчислимо четверте наближення: $x_4 = 1{,}73205$, оцінимо точність: $|1{,}73205 - 1{,}7321| = 0{,}00005 < \varepsilon$ – точність досягнута.

Відповідь: $x_{1,2} = +1{,}73205$.

Більш точне значення (до 8 значущих цифр): $\sqrt{3} = 1{,}7320508$.

Розглянутий приклад демонструє **принципи, які є загальними** для ітераційних методів розв'язання задач обчислювальної математики:

1. Вихідна задача замінюється іншою задачею – обчислювальним алгоритмом за заданими формулами, в яких використовуються лише арифметичні операції. Прийнято називати першу формулу – **формулою ітераційного процесу (ітераційним процесом)**, другу – **умовою завершення ітераційного процесу.**

2. Задача містить новий параметр i – **номер ітерації.** Очевидно, що число ітерацій впливає на точність рішення. Якщо $\lim\limits_{i \to \infty} x_i = x^* = \pm\sqrt{a}$, то ітераційний процес дозволяє отримати розв'язок початкової задачі з будь-якою точністю.

3. Розв'язок, отриманий ітераційним методом, завжди є наближеним, оскільки точне рішення отримати неможливо – потрібні нескінченні обчислення.

Як зрозуміло з попереднього, **ітерація** – це повторення сукупності операцій або процедур для покращення наявного (поточного) наближеного розв'язку задачі. Нехай x^* – розв'язок задачі, тоді ітераційний метод будує так звану **ітераційну послідовність**

$\{x^{(k)}\}$ (k = 0,1,2,...) наближень розв'язку, при цьому $x^{(k)}$ повинно наближатися до x^* зі збільшенням k.

Отже, алгоритм ітераційного методу в найзагальнішому вигляді має таку схему:

1. Задається **початкове наближення** розв'язку $x^{(0)}$ (на основі апріорних знань про задачу).

2. На k-й ітерації методу (k = 0,1,2,...) буде поточне наближення розв'язку $x^{(k)}$. Далі обчислюється таке наближення $x^{(k+1)} = \Phi(x^{(k)})$, де Φ і є сукупністю операцій або процедур для покращення наближеного розв'язку задачі (формулою ітераційного процесу), яка є суттю конкретного чисельного методу.

3. Перевіряється **критерій останову (умова завершення)**, тобто перевіряється: чи є отримане наближення $x^{(k+1)}$ розв'язку x^* достатньо близьким. Якщо цього немає, то відбувається перехід до наступної ітерації, тобто до пункту 2.

Варто зазначити, що вид критерію останову (тобто умова припинення обчислень за ітераційним методом) залежить від виду розв'язуваної математичної задачі.

1.2. Характеристики чисельних методів

Для оцінки чисельних методів, тобто порівняння між собою методів для розв'язання однієї задачі, вводять такі їх основні характеристики:

- трудомісткість;
- порядок методу;
- збіжність;
- швидкість збіжності;
- стійкість до погрішностей обчислень;
- стійкість до погрішностей у відправних даних.

Під **трудомісткістю** методу розуміють кількість і якість обчислень, необхідних для досягнення достатньо близького наближення розв'язку задачі.

Під **порядком** методу розуміють вимоги до знань про функції, що входять у математичне формулювання задачі (наприклад, використання в методі похідних цих функцій):

• метод **нульового порядку**, якщо він використовує тільки значення цих функцій;

• метод **першого порядку**, якщо він використовує значення функцій і їх перших похідних;

• метод **другого порядку**, якщо він використовує значення і функцій та їх перших і других похідних і т.д.

Чисельний метод називається **таким, що збігається**, якщо наближення x^k прямує до розв'язку зі збільшенням k. Очевидно, що методи, які не збігаються, не цікаві з прикладної точки зору. Тому одним з найважливіших етапів при введенні нового чисельного методу є теоретичне доведення його збіжності, тобто формулювання умов, за яких метод гарантовано збігається.

В основному розрізняють такі **швидкості збіжності** методів.

1. **Лінійна збіжність.** Указують, що послідовність $\{x^{(k)}\}$ (k = 0,1,2,...) лінійно збігається до розв'язку x^* (або зі швидкістю геометричної прогресії), якщо існують числа $q \in (0,1)$ і k_0 такі, що

$$\left\|x^{(k+1)} - x^*\right\| \le q\left\|x^{(k)} - x^*\right\| \text{ для всіх } k \ge k_0.$$

Тут норма $\|x - y\|$ означає відстань між x і y.

2. **Надлінійна збіжність.** Указують, що послідовність $\{x^{(k)}\}$ (k = 0,1,2,...) надлінійно збігається до розв'язку x^*, якщо існує послідовність $\{q^{(k)}\}$ (k = 0,1,2,...), $q_k \in (0,1)$ для всіх k така, що

$$\left\|x^{(k+1)} - x^*\right\| \le q_k\left\|x^{(k)} - x^*\right\| \text{ і } q_k \to 0 \text{ при } k \to \infty.$$

3. **Квадратична збіжність.** Указують, що послідовність $\{x^{(k)}\}$ (k = 0,1,2,...) квадратично збігається до розв'язку x^*, якщо існують числа $C > 0$ і $k_0 > 0$ такі, що

$$\left\|x^{(k+1)} - x^*\right\| \le C\left\|x^{(k)} - x^*\right\|^2 \text{ для всіх } k \ge k_0.$$

Під **стійкістю до погрішностей обчислень** розуміють таку властивість методу, при застосуванні якого завжди можна отримати розв'язок задачі на комп'ютері, незважаючи на помилки округлень і обчислень. Для цього в чисельних методах, якщо потрібно, передбачаються додаткові операції, які не змінюють суть методу, але забезпечують його стійкість до помилок обчислень.

Під **стійкістю до погрішностей у відправних даних** розуміють таку властивість методу, при застосуванні якого при невеликих погрішностях у відправних його даних можна отримати наближений розв'язок задачі з не дуже великою погрішністю. Стійкість до погрішностей у відправних даних досягається, як правило, шляхом модифікації чисельного методу, тобто внесенням змін до суті методу.

1.3. Похибка розв'язку

Під час розв'язування практичних задач зазвичай оперують з наближеними значеннями тих чи інших величин (наближеними числами).

Наближеним називають число, що незначно відрізняється від точного і замінює його в обчисленнях. Наближене число має практичну цінність лише тоді, коли можна вказати міру його точності, тобто оцінити його похибку. Джерелами похибок можуть бути:

1. Неточний математичний опис реальних процесів.

2. Неточне завдання вихідної інформації. Вихідні величини, як правило, – це експериментальні дані, тому міра їх точності залежить від досконалості вимірювальних приладів і надійності операцій вимірювання.

3. Застосування для розв'язування задач наближених методів. Одержання точного розв'язку задачі часто пов'язане з великими, а іноді із нездоланними труднощами, тому доводиться удаватися до наближених методів розв'язування.

4. Скінченна апроксимація нескінченних математичних процесів. Наприклад, при обчисленні e^x (або $\sin x$, $\cos x$, $\ln x$ тощо) розкладанням у ряд:

$$e^x = 1 + x + \frac{x^2}{2!} + \ldots$$

за наближене значення e^x беруть суму скінченної кількості доданків, припускаючи при цьому похибку.

5. Округлення вихідних даних, проміжних чи остаточних результатів.

6. Виконання дій над наближеними числами призводить до наближеного числа.

Похибки, що виникають у всіх зазначених випадках, можна поділити на такі групи:

1) неусувні (випадки 1, 2, 6), тобто такі, які неконтрольовані в процесі розв'язування задачі й можуть бути зменшені тільки за рахунок більш точного математичного опису задачі та вихідних даних. Дослідження неусувної похибки може бути корисним для зниження висунутих вимог до точності методу, що застосовується: недоцільно застосовувати метод розв'язування задачі з похибкою, істотно меншою, ніж величина неусувної похибки;

2) похибка методу (випадок 3);

3) похибка зрізання (випадок 4); її називають ще похибкою обмеження чи залишковою похибкою;

4) обчислювальна похибка чи похибка округлення (випадок 5).

При розв'язуванні конкретної задачі тих чи інших похибок може не бути або вони чинитимуть мізерно малий вплив. Однак у загальному випадку для повного аналізу похибок потрібно враховувати всі їхні види.

Основна задача теорії похибок – знаходження області невизначеності результату.

Для кількісного оцінювання точності чисельного методу слід ввести поняття абсолютної і відносної погрішностей величин.

Нехай x^* – точне значення деякої величини, а x – її відоме наближене значення.

Абсолютною похибкою (або погрішністю) числа x^* називається деяка величина $\Delta(x^*)$, що задовольняє умові

$$\|x - x^*\| \le \Delta(x^*).$$

Тут норма $\|x - x^*\|$ означає відстань між x і x^*. Таким чином, якщо відомі наближений розв'язок задачі x та абсолютна погрішність $\Delta(x^*)$, то невідомий точний розв'язок задачі x^* знаходиться від x на відстані не більше ніж $\Delta(x^*)$.

Також варто зазначити, що абсолютна погрішність розв'язку – це не різниця між точним і наближеним розв'язками задачі, а лише оцінка цієї різниці.

Відносною похибкою числа x^* називається деяка величина δx^*, що задовольняє умові

$$\frac{\|x - x^*\|}{\|x^*\|} \le \delta(x^*).$$

Зазначимо, що точність результату краще характеризує відносна похибка. Інформацію про абсолютну та відносну похибки можна використати для наступного представлення числа x:

$$x = x^* \pm \Delta(x^*),$$

$$x = x^*(1 \pm \delta(x^*)).$$

Варто звернути увагу на те, що наближений розв'язок задачі x, отриманий за допомогою чисельного методу, зазвичай залежить від вибраних параметрів цього методу. Без обмеження загальності можна вважати, що такий параметр у методі один (умовно можна позначити його через h) і при цьому він задовольняє умовам $0 < h < 1$. Тоді абсолютна погрішність $\Delta(x)$ також залежатиме від h, тобто можна вважати, що $\Delta(x) = \Delta(h)$.

Якщо існують числа $M > 0$ і $k > 0$ (не обов'язково цілі) такі, що виконується нерівність $\Delta(h) \le Mh^k$, то говорять, що абсолютна погрішність $\Delta(h)$ має порядок $O(h^k)$. Цим підкреслюється **якісна** (на відміну від кількісної) залежність погрішності наближеного розв'язку задачі x від параметра h чисельного методу. Наприклад, це показує, що зменшення параметра h на один порядок призводить до зменшення величини абсолютної погрішності $\Delta(x)$ на k порядків.

Розрізняють пряму та обернену задачі теорії похибок.

Пряма задача теорії похибок формулюється таким чином. У деякій області Ω n-вимірного простору розглядається неперервно-диференційована функція $y = f(x_1, x_2, \ldots, x_n)$. Припустимо, що потрібно обчислити значення цієї функції в точці $(x_1, x_2, \ldots, x_n) \in \Omega$, а відомі тільки наближені значення $x_1^*, x_2^*, \ldots, x_n^*$ такі, що точка $(x_1^*, x_2^*, \ldots, x_n^*) \in \Omega$, та відомі їх похибки. Обчислимо наближене значення $y^* = f(x_1^*, x_2^*, \ldots, x_n^*)$ та оцінимо його абсолютну похибку.

Використовуючи формулу Лагранжа, будемо мати

$$\Delta(y^*) = \left| f(x_1^*, x_2^*, ..., x_n^*) - f(x_1, x_2, ..., x_n) \right| \le \sum_{j=1}^{n} B_j \Delta(x_j^*),$$

де

$$B_j = \sup_G \left| \frac{\partial f(x_1, x_2, ..., x_n)}{\partial x_j} \right|.$$

При практичних розрахунках використовують оцінку

$$\Delta(y^*) \le \sum_{j=1}^{n} \left| \frac{\partial f(x_1^*, x_2^*, ..., x_n^*)}{\partial x_j} \right| \Delta(x_j^*),$$

яку називають лінійною оцінкою похибки. Знайдемо відносну похибку:

$$\delta(y^*) \le \sum_{j=1}^{n} \left| \frac{\dfrac{\partial f(x_1^*, x_2^*, ..., x_n^*)}{\partial x_j}}{f(x_1^*, x_2^*, ..., x_n^*)} \right|.$$

Використовуючи ці формули, визначимо похибки результатів окремих математичних операцій.

1. Похибка суми

$$y = f(x_1, x_2) = x_1 + x_2, \quad x_1, x_2 > 0..$$

Оскільки $f'_{x_j}(x^*) = 1$, то будемо мати

$$\Delta(y^*) = \Delta(x_1^*) + \Delta(x_2^*),$$

і відповідно

$$\delta(y^*) = \left| \frac{x_1^*}{x_1^* + x_2^*} \right| \delta(x_1^*) + \left| \frac{x_2^*}{x_1^* + x_2^*} \right| \delta(x_2^*).$$

Аналогічно знаходимо похибки для інших математичних операцій.

2. Похибка різниці.

$$y = f(x_1, x_2) = x_1 - x_2, \quad x_1 > x_2 > 0$$

$$\Delta(y^*) = \Delta(x_1^*) + \Delta(x_2^*),$$

$$\delta(y^*) = \frac{x_1^* \delta(x_1^*) + x_2^* \delta(x_2^*)}{x_1^* - x_2^*}.$$

3. Похибка множення.

$$y = f(x_1, x_2) = x_1 \cdot x_2, \quad x_1, x_2 > 0.$$

$$\Delta(y^*) = |x_2^*|\Delta(x_1^*) + |x_1^*|\Delta(x_2^*),$$

$$\delta(y^*) = \delta(x_1^*) + \delta(x_2^*).$$

4. Похибка ділення.

$$y = f(x_1, x_2) = x_1 / x_2, \quad x_1, x_2 > 0.$$

$$\Delta(y^*) = \frac{|x_2^*|\Delta(x_1^*) + |x_1^*|\Delta(x_2^*)}{(x_2^*)^2},$$

$$\delta(y^*) = \delta(x_1^*) + \delta(x_2^*).$$

Зазначимо, що для суми та різниці абсолютні похибки додаються, а для операцій множення та ділення складаються відносні похибки. З формули видно, що якщо віднімаються два близьких числа, то відносна похибка результату може значно зрости. А при діленні на досить мале число може значно зрости абсолютна похибка.

Розглянемо деякі приклади.

Приклад. Заокруглюючи наступні числа до трьох значущих цифр, визначити абсолютну та відносну похибки отриманих наближених чисел:

1) 0,1545; 2) 1,343; 3) –372,75.

Розв'язання

1) $x = 0{,}1545$. Заокруглення до трьох значущих цифр дає $x^* = 0{,}155$, тоді $\Delta(x^*) = 0{,}0005 = 5\cdot10^{-4}$, а відносна похибка

$\delta(x^*) = 5\cdot10^{-4}/0{,}155 \approx 0{,}32\cdot10^{-4}$.

2) $x = 1{,}343$. Тоді $x^* = 1{,}34$, $\Delta(x^*) = |x^* - x| = 0{,}003$. Відповідно відносна похибка

$\delta(x^*) = 3\cdot10^{-3}/1{,}34 = 2{,}2\cdot10^{-3}$.

3) $x = -372{,}75$. Тоді $x^* = -373$, $\Delta(x^*) = 0{,}25$, а

$\delta(x^*) = 0{,}25/373 = 6{,}7\cdot10^{-4}$.

Відзначимо досить цікавий реальний факт – поведінка (величина) обчислювальної похибки залежить від правил заокруглення та алгоритму чисельного розв'язування задачі.

Приклад. На гіпотетичній ЕОМ з мантисою довжини чотири знайти суму $S = 0{,}2764 + 0{,}3944 + 1{,}475 + 26{,}46 + 1364$:

а) сумуючи від меншого доданку до більшого;

б) сумуючи від більшого доданку до меншого.

Розв'язання

а) маємо $S_2 = 0{,}2764 + 0{,}3944 = 0{,}6708$, $S_3 = S_2 + 1{,}475$. Вирівнюючи порядки у цих двох доданків, будемо мати $S_3 = 1{,}475+0{,}671 = 2{,}146$. Аналогічно далі

$$S_4 = S_3 + 26{,}46 = 2{,}15 + 26{,}46 = 28{,}61,$$
$$S = S_5 = S_4 + 1364 = 29 + 1364 = 1393.$$

б) маємо $S_2 = 1364 + 26{,}46 = 1364 + 26 = 1390$,

$$S_3 = S_2 + 1{,}475 = 1390+1 = 1391,$$
$$S_4 = S_3 + 0{,}3944 = 191,$$
$$S = S_5 = S_4 + 0{,}2764 = 1391.$$

Враховуючи, що точне значення $S = 1392{,}6058$, бачимо, що сумування потрібно проводити починаючи з менших доданків. У протилежному випадку може мати місце значна втрата значущих цифр.

Приклад. Оцінити похибку обчислення функції:

$$f(x,y,z) = \frac{x^2 z}{y^3},$$

якщо $x = 0{,}15\pm0{,}005$, $y = 2{,}13\pm0{,}01$, $z = 1{,}14\pm0{,}007$.

Розв'язання. Для абсолютної похибки результату отримаємо

$$\Delta(f^*) = \left|\frac{2x^*z^*}{(y^*)^3}\right|\Delta(x^*) + \left|\frac{3(x^*)^2 z}{(y^*)^4}\right|\Delta(y^*) + \left|\frac{(x^*)^2}{(y^*)^3}\right|\Delta(z^*) = \frac{2\cdot0{,}15\cdot1{,}14}{(2{,}13)^3}\cdot0{,}005 +$$

$$+\frac{3\cdot(0{,}15)^2\cdot1{,}14}{(2{,}13)^4}\cdot0{,}01 + \frac{(0{,}15)^2}{(2{,}13)^3}\cdot0{,}007 = 0{,}0001769 + 0{,}0000373 +$$

$$+0{,}0000162 + 0{,}00023 = 2{,}3\cdot10^{-4}.$$

Знайдемо $f(x^*, y^*, z^*) = \dfrac{(0{,}15)^2\cdot1{,}14}{(2{,}13)^3} = 0{,}0022265429$.

Тоді $\delta(f^*) = \dfrac{2{,}3\cdot10^{-4}}{0{,}00265429} = 0{,}08665$.

Приклад. Висота та радіус основи циліндра виміряні з точністю до 0,5%. Яка відносна похибка при обчисленні об'єму циліндра, якщо $\pi^* = 3{,}14$?

Розв'язання. $V = \pi R^2 h$. Більш точне значення $\pi = 3{,}14159265$, отже $\Delta(\pi^*) = 0{,}16\cdot10^{-2}$, а $\delta(\pi^*) = 0{,}16\cdot10^{-2}/3{,}14 = 0{,}0005 = 0{,}05\%$.

Тоді, згідно з формулою про відносну похибку добутку будемо мати $\delta(V^*) = \delta(\pi^*) + 2\delta(R^*) + \delta(h) = 1{,}55\%$.

Приклад. Ребро куба виміряне з точністю до 0,02 см. дорівнює 8 см. Знайти абсолютну та відносну похибки при обчисленні об'єму куба.

Розв'язання. Позначимо сторону куба через a. Тоді $V = a^3$, $V^* = (a^*)^3 = 512$ см^3. Тоді $\Delta(V^*) = 3(a^*)^2 \cdot \Delta(a^*) = (3 \cdot 8^2 \cdot 0{,}02)$ см^3 = 3,84 см^3, а $\delta(V^*) = (3{,}84 / 512) = 0{,}0075$.

Обернена задача теорії похибок полягає в наступному: з якою точністю потрібно задати значення аргументів $x_1^*, x_2^*, ..., x_n^*$ функції $f(x_1, x_2, ..., x_n)$, щоб похибка значення функції $f(x_1^*, x_2^*, ..., x_n^*)$ не перевищувала заданої величини ε.

Для функції однієї змінної $y = f(x)$ абсолютну похибку, можна наближено обчислити за формулою

$$\Delta(x^*) = \frac{\Delta(y^*)}{|f'(x^*)|}, \quad f'(x^*) \neq 0.$$

Для функції декількох змінних $y = f(x_1, x_2, ..., x_n)$ задача розв'язується за допомогою таких рекомендацій:

а) використовуємо принцип рівних впливів, тобто вважаємо, що всі доданки $c_i = |\partial f / \partial x_i| \Delta(x_i^*)$, $i = \overline{1, n}$ рівні між собою. Тоді абсолютні похибки всіх аргументів визначаються формулою:

$$\Delta(x_i^*) = \frac{\Delta(y^*)}{n\left|\frac{\partial f^*}{\partial x_i}\right|}, \quad i = \overline{1, n};$$

б) вважаємо всі похибки рівними, причому максимально можливими, тобто покладемо

$$\Delta(x_1^*) = \Delta(x_2^*) = ... = \Delta(x_n^*) = \delta,$$

де

$$\delta = \varepsilon / (c_1 + c_2 + ... + c_n).$$

Приклад. Сторона квадрату дорівнює 2 м. З якою точністю її потрібно виміряти, щоб похибка знаходження площі не перевищувала 1 см^2?

Розв'язання. Позначимо сторону квадрату через x; $S = x^2$, $S' = 2x$. Тоді за формулою отримаємо

$$\Delta(x^*) = \frac{1}{2 \cdot 200} = \frac{1}{4} 10^{-2} \text{см}.$$

1.4. Похибка округлення при розрахунках на комп'ютері з плаваючою крапкою

Одним із основних джерел обчислювальних погрішностей є наближене подання дійсних чисел у комп'ютері, обумовлене скінченністю розрядної сітки. При розв'язанні обчислювальних задач зазвичай використовують подання чисел у **формі з плаваючою крапкою** (комою). Число a у формі з плаваючою крапкою подається у вигляді:

$$a = \text{sign}(a) M r^p,$$

де r – основа системи числення; p – порядок числа a; M – мантиса числа a; $\text{sign}(a)$ – знак числа a.

При цьому повинна виконуватися умова нормування $r^{-1} \le M \le 1$, тобто перша цифра в мантисі повинна бути відмінною від нуля.

Усі персональні комп'ютери і більшість робочих станцій підтримують міжнародний IEEE-стандарт двійкової арифметики (тобто $r = 2$). Стандарт передбачає два типи чисел із плаваючою крапкою: **числа звичайної точності** (подаються 4 байтами – 32 бітами)

1	8	23
Знак	Порядок	Мантиса

і **числа подвійної точності** (подаються 8 байтами – 64 бітами)

1	11	52
Знак	Порядок	Мантиса

Через скінченність розрядної сітки (кількість знаків у мантисі) в комп'ютері можна подати не всі числа. Число a, яке не можна подати в комп'ютері, піддається округленню, тобто замінюєть близьким числом $\bar{a}$ і саме воно подається в комп'ютері уже точно. Отже, виникає погрішність при поданні будь-якого числа в комп'ютері.

Знайдемо оцінку відносної погрішності подання числа з плаваючою крапкою [2]. Припускається, що застосовується просте округлення – відкидання всіх розрядів числа, що виходять за межі розрядної сітки. Нехай потрібно записати число, що є нескінченним двійковим дробом (система числення – двійкова):

$$a = \pm 2^p \left(\frac{a_1}{2} + \frac{a_2}{2^2} + \ldots + \frac{a_t}{2^t} + \frac{a_{t+1}}{2^{t+1}} + \ldots \right),$$

де – a_j = (0 або 1), j = 1,2,... – цифри мантиси; t – розрядність мантиси (23 або 52 залежно від типу числа).

Тоді, після відкидання зайвих розрядів, отримаємо округлене число (але точне для комп'ютера):

$$\bar{a} = \pm 2^p \left(\frac{a_1}{2} + \frac{a_2}{2^2} + \ldots + \frac{a_t}{2^t} \right).$$

Погрішність округлення в цьому випадку дорівнює:

$$\bar{a} - a = \pm 2^p \left(\frac{a_{t+1}}{2^{t+1}} + \frac{a_{t+2}}{2^{t+2}} + \ldots \right).$$

Зрозуміло, що найбільша погрішність буде у випадку, коли всі $a_{t+1} = 1$, $a_{t+2} = 1$,... Тоді абсолютна погрішність округлення дорівнюватиме:

$$|\bar{a} - a| = 2^p \frac{1}{2^{t+1}} \left(1 + \frac{1}{2} + \frac{1}{2^2} + \ldots \right) = 2^p \frac{1}{2^{t+1}} 2 = 2^{p-t}.$$

З умови нормування виходить, що $M \geq ½$, а отже завжди $a_1 = 1$. Тоді із загальної формули подання числа отримуємо, що $|a| \geq 2^p \times 2^{-1} = 2^{p-1}$ і тому $\frac{|\bar{a} - a|}{|a|} \leq 2^{-t+1}$, тобто відносна погрішність дорівнює 2^{-t+1}.

У реальності застосовують точніші методи округлення (ніж просте відкидання розрядів), тому відносна погрішність подання чисел у пам'яті комп'ютера дорівнює:

$$\frac{|\bar{a}-a|}{|a|} \le 2^{-t},$$

тобто точність подання чисел цілком визначається розрядністю мантиси *t*.

Слід зазначити, що з цієї формули випливає, що наближене подане в комп'ютері число $\bar{a}$ можна записати у вигляді $\bar{a} = a(1 \pm \varepsilon)$, де $|\varepsilon| \le 2^{-t}$. Тому відносну погрішність подання чисел 2^{-t} часто ще називають «машинним епсілоном».

Більш детально похибки у ході розрахунків на комп'ютері з плаваючою крапкою вивчаються у [1, 2, 7].

1.5. Математичні пакети

За останні 50–60 років, пов'язаних із використанням комп'ютерної техніки, в усьому світі накопичено велику кількість програмного забезпечення у вигляді бібліотек комп'ютерних підпрограм (написаних, у першу чергу, мовами Fortran та C), призначених для розв'язання типових математичних задач. Крім того, розроблено ряд універсальних математичних пакетів, за допомогою яких можна достатньо швидко як розв'язати багато відомих математичних задач, так і провести попередні розрахунки з подальшою реалізацією вже в проектованій комп'ютерній системі.

Найбільш відомими є такі математичні пакети: R, Mathematica, Maple, MatLab, MathCad [2, 10].

Пакет **Mathematica** найпопулярніший у наукових колах, особливо у теоретиків. Він надає широкі можливості під час проведення символічних (аналітичних) перетворень, дозволяє швидко й ефективно розв'язувати багато задач обчислювальної математики. В пакеті Mathematica більшість задач розв'язується в діалоговому режимі, без традиційного програмування з використовуванням стандартних операторів. Крім того, у цьому пакеті реалізована мова програмування високого рівня, яка дозволяє

створювати достатньо складні програми. Також цей пакет дозволяє конвертувати документи у формат LaTex, який є стандартним форматом переважної більшості наукових видавництв світу.

Пакет **Maple** – найдавніший серед пакетів символьної математики – також популярний у наукових колах. Крім здійснення аналітичних перетворень, з його допомогою можна розв'язувати різноманітні математичні задачі чисельно. Графічний редактор пакета дозволяє одержувати зображення тривимірних фігур із перерізами та конвертувати документи у формат LaTex.

Пакет **MatLab**, як і згадані пакети, є своєрідною мовою програмування високого рівня, що орієнтована переважно на інженерні розрахунки теорії управління, електро- і радіотехніки, а також моделювання технічних систем. Пакет MatLab став фактично міжнародним стандартом сучасного навчального програмного забезпечення і використовується в багатьох провідних університетах світу [10].

Пакет **MathCad** відрізняється від інших такою особливістю, як використання звичайних математичних позначень. Документ пакета MathCad на екрані монітора виглядає як звичайний математичний розрахунок. Крім того, MathCad є середовищем візуального програмування, тобто не вимагає знання спеціального набору команд. Хоча пакет орієнтований, у першу чергу, на проведення чисельних розрахунків, він має вбудований символьний процесор Maple, що дозволяє виконувати аналітичні перетворювання. Також у пакеті MathCad передбачена можливість створювати зв'язки його документів із документами пакета MatLab.

Система **R** відрізняється від наведених пакетів в основному тим, що є вільним програмним продуктом, який розповсюджується на умові ліцензії GNU. Вона включає в себе об'єктно-орієнтовану мову програмування R і середовище в сукупності з великим набором бібліотек (більше ніж 1 600), які доступні для усіх основних платформ – Linux, Windows і MacOS. Постійно виходять нові версії та розширення, які доступні на офіційному сайті http://cran.r-project.org [33], розробники пишуть для системи R пакети, адаптовані до конкретних галузей. Важливим показником популярності цього програмного продукту є і те, що постійно

публікуються нові книги з описом застосування R для різних видів аналізу даних [14,15]. Мова R вбудовується в усілякі системи. Наприклад, корпорація Oracle інтегрувала мову програмування R у СУБД Oracle Database 11g.

З цих причин саме пакет R використовується у цьому навчальному посібнику для програмування, виконання розрахунків та наочного подання результатів розв'язання. Математичний пакет R все частіше застосовується у ході проведення лабораторних занять із різних навчальних дисциплін, тому тим, хто бажає більш досконально опанувати цей пакет, пропонується звернутися до додаткових джерел [16–19].

Треба зазначити, що останнім часом позначилася тенденція до зближення та інтеграції різних математичних пакетів. Так, останні версії пакетів Mathematica і Maple мають потужні засоби для візуального програмування; в пакети MatLab та MathCad включена бібліотека аналітичних перстворень Maple; MathCad дозволяє працювати разом із MatLab і т.д. Тому, за великим рахунком, не має значення, який математичний пакет застосовується у навчальному процесі та засвоюється студентами. Головне – принципово навчитися використовувати будь-який математичний пакет для розв'язання практичних задач.

Контрольні запитання та завдання

1. Охарактеризуйте тріаду «модель-алгоритм-програма».
2. Подайте приклади математичних моделей.
3. Назвіть основні групи методів розв'язання математичних задач та їх характеристики.
4. Які види чисельних методів вам відомі?
5. Назвіть основні характеристики чисельних методів. Розкрийте їх суть.
6. Які є основні швидкості збіжності ітераційних методів?
7. Назвіть та охарактеризуйте основні етапи розв'язання практичних задач на комп'ютері.
8. Що називають «ітерацією»? Наведіть загальну схему ітераційного методу.

9. Які методи розв'язання математичних задач називають ітераційними?

10. Які методи розв'язання математичних задач називають чисельними?

11. Дайте визначення поняттям абсолютної та відносної погрішностей.

12. Дайте визначення лінійної і квадратичної швидкості збіжності ітераційного методу.

13. Назвіть джерела виникнення похибок.

14. Що є причинами появи погрішності при арифметичних обчисленнях на комп'ютері?

15. Як подані числа в комп'ютері? Вкажіть розмір абсолютної та відносної погрішностей округлення чисел у комп'ютері.

16. Які універсальні математичні пакети ви знаєте? Дайте їх порівняльну характеристику. Для чого вони використовуються?

2. ПРЯМІ МЕТОДИ РОЗВ'ЯЗАННЯ СИСТЕМ ЛІНІЙНИХ РІВНЯНЬ

2.1. Чисельні методи розв'язування СЛАР

Теорія розв'язку систем лінійних алгебраїчних рівнянь (СЛАР) досить проста, давно і добре відома, проте практична реалізація чисельних методів іноді наштовхується на великі труднощі. Це пов'язано перш за все з тим, що більшість з них є дуже чутливими до впливу помилок заокруглення та збурень вхідних даних. Більшість методів базується на розкладанні матриці СЛАР на множники – чим ефективніше здійснюється таке розкладання, тим ліпшими характеристиками володітиме відповідний метод. Зазвичай порівняльні характеристики розкладань включають: кількість операцій для реалізації методу, точність, додаткову пам'ять. Крім того, слід враховувати можливість розв'язку СЛАР з матрицями спеціального виду (симетричні, трикутні і т.д.) та важливий випадок з багатьма правими частинами і однією матрицею. У загальному випадку не можна надати якихось рекомендацій щодо вибору того чи іншого методу. У цілому процес розв'язку СЛАР на ЕОМ включає такі етапи:

- попереднє перетворення вихідної системи до більш простого вигляду;
- розв'язок системи простого вигляду;
- оцінка числа обумовленості та точності розв'язку;
- уточнення розв'язку;
- загальний аналіз розв'язку.

Попереднє перетворення вихідної системи до більш простого вигляду здійснюється за допомогою еквівалентних перетворень, які включають такі дії:

- переміна місцями двох рівнянь у системі;
- множення будь-якого рівняння системи на дійсне число $c \neq 0$;
- додавання до одного рівняння іншого рівняння, помноженого на будь-яке число.

Еквівалентні перетворення дозволяють перейти від заданої системи до системи, яка еквівалентна вихідній, але розв'язується «простіше», ніж вихідна. Саме на цих перетвореннях базується більшість методів розв'язку СЛАР, зокрема, і класичний метод Гаусса.

Зазвичай методи розв'язування СЛАР поділяються на прямі та ітераційні. За умови точного виконання обчислень прямі методи за скінченну кількість операцій у результаті дають точний розв'язок. Використовуються вони для невеликих та середніх розмірностей СЛАР, тобто при $n = 10^2$–10^4. Ітераційні ж методи використовуються для великих СЛАР $n > 10^5$, які, як правило, мають розріджені матриці. Вони генерують послідовність наближень, яка збігається до розв'язку.

2.2. Метод Гаусса. Прямий та обернений хід

Розглянемо задачу розв'язання СЛАР

$$A\vec{x} = \vec{b},$$

причому $A = (a_{ij})_{i,j=1}^{n}$, $\det A \neq 0$, $\vec{x} = (x_i)_{i=1}^{n}$, $\vec{b} = (b_j)_{j=1}^{n}$.

Умова $\det A \neq 0$ є необхідною і достатньою для існування розв'язку цієї системи [10]. У цьому випадку її розв'язок може бути отриманий за формулами Крамера:

$$x_i = \frac{\det A_i}{\det A}, \; i = 1,\ldots,n,$$

де матриця A_i утворюється з матриці A заміною її i-го стовпця стовпцем вільних членів b.

Але такий спосіб розв'язання системи лінійних рівнянь з n невідомими призводить до обчислення $(n+1)$-го визначника порядку n, що є трудомісткою операцією при великих n. Він (метод Крамера з обчисленням визначників) має складність $Q = O(n!\,n)$.

Отже, розглянемо більш ефективний метод Гаусса. Запишемо СЛАР у вигляді

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \equiv a_{1n+1} \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_x = b_2 \equiv a_{2n+1} \\ \ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n \equiv a_{nn+1} \end{cases}$$

Далі скористаємося еквівалентними перетвореннями, а саме: якщо $a_{11} \neq 0$, то ділимо перше рівняння на нього, множимо на число $-a_{i1}$ і додаємо його до кожного з інших рівнянь. Таким чином, виключається x_1 з усіх інших рівнянь:

$$\begin{cases} x_1 + a_{12}^{(1)}x_2 + \ldots + a_{1n}^{(1)}x_n = a_{1n+1}^{(1)} \\ a_{22}^{(1)}x_2 + \ldots + a_{2n}^{(1)}x_n = a_{2n+1}^{(1)} \\ \ldots \\ a_{n2}^{(1)}x_2 + \ldots + a_{nn}^{(1)}x_n = a_{nn+1}^{(1)} \end{cases}$$

Процес повторюємо для $x_2,\ldots,x_n$. У результаті отримуємо систему з трикутною матрицею:

$$\begin{cases} x_1 + a_{12}^{(1)}x_2 + \ldots + a_{1n}^{(1)}x_n = a_{1n+1}^{(1)} \\ x_2 + \ldots + a_{2n}^{(2)}x_n = a_{2n+1}^{(2)} \\ \ldots \\ x_n = a_{nn+1}^{(n)} \end{cases}$$

Це прямий хід методу Гаусса. Формули прямого ходу:

$$\begin{cases} k = \overline{1, n-1}: \\ a_{kj}^{(k)} = \dfrac{a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}}, j = \overline{k+1, n+1}; \\ a_{ij}^{(k)} = a_{ij}^{(k-1)} - a_{ik}^{(k-1)}a_{kj}^{(k)}, \\ i = \overline{k+1, n}; j = \overline{k+1, n+1}. \end{cases}$$

Звідси:

$$x_n = a_{nn+1}^{(n)}, x_i = a_{in+1}^{(i)} - \sum_{j=i+1}^{n} a_{ij}^{(i)}x_j, i = \overline{n-1, 1}.$$

Це формули оберненого ходу.

Складність, тобто кількість операцій, яку необхідно виконати для реалізації методу, – $Q_{пр} = \frac{2}{3}n^3 + O(n^2)$ для прямого ходу, $Q_{об} = n^2 + O(n)$ для оберненого ходу.

Слід зазначити, що з логічної точки зору кращою є модифікація методу Гаусса, яка зветься метод Гаусса-Жордана. Його суть полягає в тому, що обернений хід виконується не за наведеними вище формулами, а виключення невідомих в оберненому порядку проводиться з рівнянь, отриманих після виконання прямого ходу методу Гаусса. Тоді система приводиться до діагонального виду з одиницями на діагоналі, а розв'язок виявляється на місці вектора b. У такому випадку обернений хід дійсно є оберненим, оскільки виконуються операції, симетричні щодо прямого ходу.

2.3. Вибір головного елемента

Умова $a_{kk}^{(k-1)} \neq 0$ не суттєва, оскільки завжди знайдеться m, для якого $|a_{mk}^{(k-1)}| = \max_i |a_{ik}^{(k-1)}| \neq 0$ (оскільки $\det A \neq 0$). Тоді міняємо місцями рядки номерів k і m. Елемент $a_{kk}^{(k-1)} \neq 0$ називається ведучим.

Можна досягти зменшення обчислювальної похибки шляхом перестановки рядків і стовпців матриці так, щоб головний елемент на k-му кроці виключення був найбільшим за модулем з коефіцієнтів, що беруть участь у подальшому виключенні:

$$|a_{kk}^{(k-1)}| = \max_{k \le i,j \le n} |a_{ij}^{(k-1)}|,\ 1 \le k \le n,\ \text{де}\ a_{ij}^{(0)} = a_{ij}.$$

Проте повне впорядкування вимагає великої додаткової обчислювальної роботи. Тому часто зупиняються на методі виключення з вибором провідних елементів по рядках:

$$|a_{kk}^{(k-1)}| = \max_{k \le i \le n} |a_{ik}^{(k-1)}|.$$

У системах з діагональним переважанням елементів $|a_{ii}| > \sum_{j=1, j \neq i}^{n} |a_{ij}|, k = 1, \ldots, n$ помилки округлення не накопичуються.

Обчислювальна похибка методу виключення може бути усунена в арифметиці цілих чисел. При цьому, якщо розподіл на

провідні елементи виключення проводиться з залишком, то воно не виконується, а запам'ятовуються лише чисельник і знаменник. Компоненти вектора розв'язку визначаються у вигляді дробів. Метод Гаусса в арифметиці цілих чисел для загальних матриць A вимагає певного мистецтва масштабування, щоб уникнути переповнення після виконання арифметичних операцій.

Введемо матриці

$$M_k = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ 0 & & m_{kk} & & \\ & & \vdots & \ddots & \\ 0 & & m_{nk} & & 1 \end{pmatrix},$$

елементи якої обчислюються так: $m_{kk} = \dfrac{1}{a_{kk}^{(k-1)}}$, $m_{ik} = -\dfrac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}$.

Нехай на k–му кроці маємо систему $A_{k-1}\vec{x} = \vec{b}_{k-1}$. Множимо цю СЛАР зліва на M_k: $M_k A_{k-1}\vec{x} = M_k \vec{b}_{k-1}$. Позначимо $A_k = M_k A_{k-1}$; $A_0 = A$. Тоді прямий хід методу Гаусса можна записати у вигляді

$$M_n M_{n-1} \ldots M_1 A\vec{x} = M_n M_{n-1} \ldots M_1 \vec{b}.$$

Позначимо останню систему так:

$$U\vec{x} = \vec{c},\ U = (u_{ij})_{i,j=1}^{n},$$

причому

$$\begin{cases} u_{ii} = 1, \\ u_{ij} = 0, i > j. \end{cases}$$

Таким чином $U = M_n M_{n-1} \ldots M_1 A$. Введемо матриці

$$L_k = M_k^{-1} = \begin{pmatrix} 1 & & 0 & & 0 \\ & \ddots & & & \\ 0 & & a_{kk}^{(k-1)} & & 0 \\ & & \vdots & \ddots & \\ 0 & & a_{nk}^{(k-1)} & & 1 \end{pmatrix}.$$

Тоді

$$A = L_1 \ldots L_n U = LU; \quad L = L_1 \ldots L_n.$$

Тут L – нижня трикутна матриця, U – верхня трикутна матриця. Таким чином метод Гаусса можна трактувати як розклад матриці A в добуток двох трикутних матриць – (LU)–розклад.

Введемо матрицю перестановок на k-му кроці (це матриця, отримана з одиничної матриці перестановкою k-го і m-го рядків). Тоді при множенні на неї матриці A_{k-1} робимо ведучим елементом максимальний за модулем.

$$P_k = \begin{pmatrix} 1 & & \ldots & & 0 \\ & \ddots & & & \\ & & 0 & 1 & \\ & & 1 & 0 & \\ 0 & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \leftarrow k \\ \leftarrow m \\ \\ \\ \end{matrix}$$

$$\begin{matrix} & \uparrow & \uparrow \\ & k & m \end{matrix}$$

За допомогою цих матриць перехід до трикутної системи тепер має вигляд:

$$M_n M_{n-1} P_{n-1} \ldots M_1 P_1 A\vec{x} = M_n M_{n-1} P_{n-1} \ldots M_1 P_1 \vec{b}.$$

Теорема. Знайдеться така матриця P – перестановок, що $PA = LU$ – розклад матриці на нижню трикутну з ненульовими діагональними елементами і верхню трикутну матрицю з одиницями на діагоналі.

Висновки про переваги трикутного розкладу:

1. Розділення прямого і оберненого ходів дає змогу економно розв'язувати декілька систем з однаковою матрицею та різними правими частинами.

2. Зберігання M, або L та U на місці A.

3. Обчислюючи число l – кількість перестановок, можна встановити знак визначника.

2.4. Схема *LU*-розкладання матриці

Метод *LU*-розкладання в принципі еквівалентний методу Гаусса, відмінність полягає тільки в порядку дій.

У методі *LU*-розкладання матриці *A* системи спочатку подається у вигляді *LU*-розкладання, тобто у вигляді добутку двох матриць:

$$A = LU,$$

де *L* – нижня трикутна матриця;

U – верхня трикутна матриця з одиницями на діагоналі.

Визначимо нижню трикутну матрицю *L* з одиничною головною діагоналлю в такий спосіб:

$$L = \begin{pmatrix} 1 & & & \\ \frac{a_{21}}{a_{11}} & & & 0 \\ \frac{a_{31}}{a_{11}} & \frac{a_{32}^{(1)}}{a_{22}^{(1)}} & & \\ \cdots & \cdots & \cdots & \cdots \\ \frac{a_{n1}}{a_{11}} & \frac{a_{n2}^{(1)}}{a_{22}^{(1)}} & \cdots & \frac{a_{nn-1}^{(n-2)}}{a_{n-1n-1}^{(n-2)}} \end{pmatrix}.$$

Представлення матриці *A* у вигляді *LU*-розкладання приводить до розв'язку вихідної системи з двома трикутними матрицями:

$$L\bar{g} = \bar{b};$$

$$U\bar{x} = \bar{g}.$$

Таким чином, розглянутий варіант методу виключення Гаусса – це визначення вектора $\bar{g}$ і потім визначення вектора $\bar{x}$.

Слід зазначити, що якщо *A* – симетрична знакопостійна матриця [10], то *LU*-розкладання може бути записане у вигляді:

$$A = U^{*}DU,$$

де U^{*} – нижня трикутна матриця, транспонована до *U*;

D – діагональна матриця, всі її елементи d_{ii} мають однаковий знак.

Застосування такого розкладання для розв'язання системи має назву методу Холецького.

2.5. Метод квадратного кореня

Метод квадратного кореня використовується для розв'язку СЛАР з симетричної матрицею A

$$A\vec{x} = \vec{b}\,,\ A^* = A\,.$$

Він оснований на розкладі матриці A в добуток:

$$A = S^* DS\,,$$

де S – верхня трикутна матриця, S^T – нижня трикутна матриця, D – діагональна матрицю.

Виникає питання: як обчислити S, D по матриці A? Маємо:

$$(DS)_{ij} = \begin{cases} d_{ii}s_{ij}, i \le j \\ 0, i > j \end{cases}$$

$$(S^*DS)_{ij} = \sum_{l=1}^{n} s_{li}^* d_{ll} s_{lj} = \sum_{l=1}^{i-1} s_{li} s_{lj} d_{ll} + s_{ii} s_{ij} d_{ii} + \underbrace{s_{ii}^* \sum_{l=i+1}^{n} s_{li} s_{lj} d_{ll}}_{=0} = a_{ij}, i, j = \overline{1,n}.$$

Якщо $i = j$, то:

$$|s_{ii}^2| d_{ii} = a_{ii} - \sum_{l=1}^{i-1} |s_{li}^2| d_{ll} \equiv p_i\,.$$

Тому:

$$d_{ii} = sign\ (p_i)\ ,\ s_{ii} = \sqrt{|p_i|}\,.$$

Якщо $i < j$, то

$$s_{ij} = \left(a_{ij} - \sum_{l=1}^{i-1} s_{li} d_{ll} s_{lj}\right) / (s_{ii} d_{ii}),\ i = \overline{1,n},\ j = \overline{i+1,n}.$$

Якщо $A > 0$ (тобто головні мінори матриці A додатні), то всі $d_{ii} = +1$.

Знайдемо розв'язок рівняння СЛАР. Отже, маємо:

$$S^* D\vec{y} = \vec{b}\,,$$

$$S\vec{x} = \vec{y}\,.$$

Оскільки S – верхня трикутна матриця, а S^*D – нижня трикутна матриця, то

$$y_i = \frac{b_i - \sum_{j=1}^{i-1} s_{ji} d_{jj} y_j}{s_{ii} d_{ii}}, \quad i = \overline{1,n},$$

$$x_n = \frac{y_n}{s_{nn}}, x_i = \frac{y_i - \sum_{j=1}^{i-1} s_{ij} x_j}{s_{ii}}, \quad i = \overline{n-1,1}.$$

Метод квадратного кореня має складність $Q = \frac{1}{3}n^3 + O(n^2)$, тобто при великих n він удвічі швидше методу Гаусса. Якщо матриця A має багато нульових елементів, то і в матриці S також. Остання властивість дає економію в пам'яті та кількості арифметичних операцій. Наприклад, якщо A має m ненульових стрічок по діагоналі, то $Q = O(m^2 n)$ і вимагає вдвічі менше оперативної пам'яті ЕОМ.

2.6. Обчислення визначника та оберненої матриці

Використовуючи розглянуті методи розв'язку СЛАР і з огляду на факторизацію матриці системи у вигляді трикутних матриць, можна обчислити значення визначника матриці A за формулою:

$$\det(A) = a_{11} a_{22}^{(1)} \ldots a_{nn}^{(n-1)} \det P.$$

Оскільки матриця P виходить перестановкою строк одиничної матриці E, то

$$\det P = \begin{cases} 1 \text{ при парній кількості перестановок;} \\ -1 \text{ при непарній кількості перестановок.} \end{cases}$$

Остаточно визначник матриці A

$$\det(A) = (-1)^k a_{11} a_{22}^{(1)} \ldots a_{nn}^{(n-1)}.$$

Якщо матриця A – симетрична, то маємо:

$$\det(A) = \det L \det L' = (\det L)^2 = (\alpha_{11} \alpha_{22} \ldots \alpha_{nn})^2.$$

Обернена матриця A^{-1} знаходиться розв'язком n систем лінійних рівнянь методом виключення Гаусса:

$$A\bar{x} = e_i, \ 1 \le i \le n,$$

де e_i – i-й стовпець одиничної матриці E. Отримані вектори розв'язків $x_1, x_2, \ldots, x_n$ утворюють n стовпців матриці A^{-1}, оскільки $AA^{-1} = E$.

2.7. Метод прогонки

Це економний метод для розв'язання СЛАР з тридіагональною матрицею:

$$\begin{cases} -c_0 y_0 + b_0 y_1 = -f_0, \\ a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \\ a_n y_{n-1} - c_n y_n = -f_n. \end{cases}$$

Тобто матриця системи:

$$A = \begin{pmatrix} -c_0 & b_1 & & 0 \\ a_0 & -c_1 & \ddots & \\ & \ddots & \ddots & b_n \\ 0 & & a_n & -c_n \end{pmatrix}$$

тридіагональна.

Розв'язок представимо у вигляді:

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}, i = \overline{0, n-1}.$$

Замінимо в цьому співвідношенні $i \to i-1$ і підставимо в друге рівняння системи, тоді

$$(a_i \alpha_i - c_i) y_i + b_i y_{i+1} = -f_i - a_i \beta_i$$

Звідси

$$y_i = \frac{b_i}{c_i - a_i \alpha_i} y_{i+1} + \frac{f_i + a_i \beta_i}{c_i - a_i \alpha_i}.$$

А також

$$\alpha_{i+1} = \frac{b_i}{c_i - a_i \alpha_i}, \beta_{i+1} = \frac{f_i + a_i \beta_i}{c_i - a_i \alpha_i}, i = \overline{1, n-1}.$$

Умова розв'язності системи $c_i - a_i \alpha_i \neq 0$.

Щоб знайти всі α_i, β_i, треба задати перші значення. З першого рівняння:

$$\alpha_1 = \frac{b_0}{c_0}, \beta_1 = \frac{f_0}{c_0}.$$

Після знаходження всіх α_i, β_i обчислюємо y_n з системи

$$\begin{cases} a_n y_n - c_n y_n = -f_n \\ y_{n-1} = \alpha_n y_n + \beta_n \end{cases}.$$

Звідси

$$y_n = \frac{f_n + a_n \beta_n}{c_n - a_n \alpha_n}.$$

Тепер алгоритм методу прогонки виглядає так:

1. Покладемо $\alpha_1 = \frac{b_0}{c_0}, \beta_1 = \frac{f_0}{c_0}$.

2. Позначимо $z_i = c_i - a_i \alpha_i$, обчислимо $\alpha_{i+1} = \frac{b_i}{z_i}, \beta_{i+1} = \frac{f_i + a_i \beta_i}{z_i}$ для $i = \overline{1, N-1}$.

3. Знайдемо $y_n = \frac{f_n + a_n \beta_n}{c_n - a_n \alpha_n}$.

4. Обчислюємо $y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}, i = \overline{n-1, 0}$.

Складність алгоритму $Q = 8n - 2$.

Метод можна застосовувати, коли $c_i - a_i \alpha_i \neq 0 \; \forall i, |\alpha_i| \leq 1$. Якщо $|\alpha_i| \geq q > 1$, то $\Delta y_0 \geq q^n \Delta y_n$ (тут Δy_i абсолютна похибка обчислення y_i), а це приводить до експоненціального накопичення похибок заокруглення, тобто нестійкості алгоритму прогонки.

Теорема. (Про достатні умови стійкості методу прогонки). Нехай

$$a_i, b_i \neq 0, \text{ та } |c_i| \geq |a_i| + |b_i|, \; \forall i, (a_0 = b_n = 0)$$

та хоча б одна нерівність строга. Тоді:

$$|\alpha_i| \leq 1, \text{ та } z_i = c_i - a_i \alpha_i \neq 0, i = \overline{1, n}.$$

2.8. Обумовленість систем лінійних алгебраїчних рівнянь

Нехай задано СЛАР

$$A\vec{x} = \vec{b}\,.$$

Припустимо, що матриця і права частина системи задані неточно, тобто фактично необхідно розв'язувати наближену систему

$$B\vec{y} = \vec{h}\,,$$

де $B = A + C$, $\vec{h} = \vec{b} + \vec{\eta}$, $\vec{y} = \vec{x} + \vec{z}$.

Звернемо увагу на те, що малість детермінанту $|\det A| << 1$ не є необхідною умовою різкого збільшення похибки. Це ілюструє такий приклад:

$$A = diag(\varepsilon), a_{ij} = \varepsilon\ \delta_{ij}\,,$$

тоді $\det A = \varepsilon^n << 1$, але $x_i = \dfrac{b_i}{\varepsilon}$. Тому $\Delta x_i = \dfrac{\Delta b_i}{\varepsilon}$.

Оцінимо похибку розв'язку. Підставивши значення $B, \vec{h}$ та $\vec{z} = \vec{y} - \vec{x}$, отримуємо:

$$(A + C)(\vec{x} + \vec{z}) = \vec{b} + \vec{\eta}\,.$$

Віднімемо від цієї рівності початкове рівняння, тоді $A\vec{z} + C\vec{x} + C\vec{z} = \vec{\eta}$. Отже,

$$A\vec{z} = \vec{\eta} - C\vec{x} - C\vec{z},\ \vec{z} = A^{-1}(\vec{\eta} - C\vec{x} - C\vec{z}).$$

Для аналізу поняття числа обумовленості зручно замість евклідової норми вектора використовувати інші норми.

Зокрема, вводять такі норми векторів $\|\vec{z}\|$:

$$\|\vec{z}\|_1 = \sum_{i=1}^{n}|z_i|,\ \|\vec{z}\|_2 = \left(\sum_{i=1}^{n}|z_i|^2\right)^{\frac{1}{2}},\ \|\vec{z}\|_\infty = \max_i|z_i|.$$

Можна довести, що всі вони еквівалентні. Вводять також норми матриці, що відповідають даним нормам вектора, тобто:

$$\|A\|_m = \sup_{\|\vec{x}\|_m \neq 0}\frac{\|A\vec{x}\|_m}{\|\vec{x}\|_m}, m = 1,2,\infty,$$

такі: $\|A\|_1 = \max_j\sum_{i=1}^{n}|a_{ij}|$, $\|A\|_2 = \max_i\sqrt{\lambda_i(A^TA)}$,

$\|A\|_\infty = \max_i\sum_{j=1}^{n}|a_{ij}|$, де $\lambda_i(B)$ – власні значення матриці B.

Позначимо $\delta(\bar{x}) = \frac{\|\bar{z}\|}{\|\bar{x}\|}$, $\delta(\bar{b}) = \frac{\|\bar{\eta}\|}{\|\bar{b}\|}$, $\delta(A) = \frac{\|C\|}{\|A\|}$ – відносні похибки $\bar{x}$, $\bar{b}$,A, де $\|\cdot\| = \|\cdot\|_k$ одна з введених вище норм.

Для характеристики зв'язку між похибками правої частини та розв'язку вводять поняття числа обумовленості матриці системи.

Число обумовленості матриці A – $cond(A) = \|A\|\|A^{-1}\|$.

Теорема. Якщо $\exists A^{-1}$ та $\|A^{-1}\|\|C\| < 1$, то

$$\delta(\bar{x}) \le \frac{condA}{1 - condA\,\delta(A)}(\delta(A) + \delta(\bar{b})),$$

де $condA$ – число обумовленості.

Властивості $condA$:

1. $condA \ge 1$;
2. $condA \ge \frac{\max|\lambda_i(A)|}{\min|\lambda_i(A)|}$;
3. $cond(AB) \le condA * condB$;
4. $A^* = A^{-1} \Rightarrow condA = 1$.

Друга властивість має місце, оскільки довільна норма матриці не менше її найбільшого за модулем власного значення. Значить $\|A\| \ge \max|\lambda_i|$. Оскільки власні значення матриць A^{-1} та A взаємно обернені, то

$$\|A^{-1}\| \ge \max\frac{1}{|\lambda_i|} = \frac{1}{\min|\lambda_i|}.$$

Якщо $condA >> 1$, то система називається погано обумовленою.

Оцінка впливу похибок заокруглення при обчисленні розв'язку СЛАР така: $\delta(A) = O(n\beta^{-t})$, $\delta(\bar{b}) = O(\beta^{-t})$, де β – розрядність ЕОМ, t – кількість розрядів, що відводиться під мантису числа. Легко отримати оцінку $\delta(\bar{x}) = condA \times O(n\beta^{-t})$. Найпростіший спосіб підвищити точність обчислення розв'язку погано обумовленої СЛАР – збільшити розрядність ЕОМ при обчисленнях.

Приклад погано обумовленої системи – системи з матрицею Гільберта

$$H_n = \left(\frac{1}{i+j-1} \right)_{i,j=1}^{n}$$, наприклад $condH_3 \approx 10^3$.

Контрольні запитання та завдання

1. Необхідність чисельних методів розв'язування систем лінійних алгебраїчних рівнянь.

2. Сформулюйте постановку задачі розв'язання системи лінійних алгебраїчних рівнянь.

3. Назвіть групи методів розв'язання систем лінійних алгебраїчних рівнянь, вкажіть їх принципові відмінності.

4. У чому полягає ідея методу Гаусса?

5. Оцінка кількості арифметичних дій методу Гаусса.

6. У чому полягає суть методу виключення Гаусса з вибором головного елемента? В яких випадках його застосовують на практиці?

7. Формули методу квадратних коренів.

8. Методи обчислення визначника та оберненої матриці.

9. Метод прогонки.

10. Поняття обумовленості систем лінійних алгебраїчних рівнянь.

11. Метод простої ітерації.

12. Метод Зейделя розв'язування систем лінійних алгебраїчних рівнянь.

13. Однокрокові (двошарові) ітераційні методи.

14. Що називається *LU*-розкладанням матриці?

15. Як застосовується *LU*-розкладання матриць для розв'язання систем лінійних алгебраїчних рівнянь?

16. Як можна використати *LU*-розкладання матриці для обчислення її оберненої матриці?

17. Опишіть схему обчислення оберненої матриці методом Гаусса-Жордана.

18. За яких умов збігається метод простої ітерації? Наведіть оцінки швидкості збіжності цього методу.

19. Який порядок арифметичних обчислень потрібен для розв'язання системи лінійних алгебраїчних рівнянь з n невідомими?

20. Знайдіть коефіцієнти параболи $y = ax^2 + bx + c = 0$, яка проходить через точки (1; 8.6), (3; 30.8), (5; 65).

21. Знайдіть обернену матрицю для матриці A:

$$A = \begin{Vmatrix} 1.8 & -3.8 & 0.7 & -3.7 \\ 0.7 & 2.1 & -2.6 & -2.8 \\ 7.3 & 8.1 & 1.7 & -4.9 \\ 1.9 & -4.3 & -4.9 & -4.7 \end{Vmatrix}.$$

22. Знайдіть точку перетину двох функцій: $y = 1 - 0.3x$ і $x = 3 - 2.2y$.

3. РОЗВ'ЯЗАННЯ СИСТЕМ ЛІНІЙНИХ РІВНЯНЬ ВЕЛИКОЇ РОЗМІРНОСТІ

3.1. СЛАР великої розмірності

Як зазначалося раніше, класичні прямі чисельні методи розв'язання задачі є сенс застосовувати при невеликих та середніх розмінностях СЛАР, тобто при $n = 10^2$–10^4. Але на практиці нерідко виникає необхідність розв'язувати СЛАР і з більшою розмірністю. Такі системи вважаються системами великої розмірності. Наприклад, такі системи виникають при проектуванні великих інтегральних схем, в економічних моделях та ін. При розв'язанні систем великої розмірності на практиці з застосуванням комп'ютерної техніки виникають такі основні проблеми:

1) великий об'єм оперативної пам'яті комп'ютера, необхідний при проведенні розрахунків із матрицею;

2) великий об'єм самих розрахунків;

3) матриця A системи може бути виродженою;

4) накопичення помилок обчислення.

Наприклад, якщо матриця має розмірність $n = 10^6$, то для її зберігання необхідно $n \times n \times 8$ байт, тобто порядку 10^{13} байт, або 10^4 Гігабайт оперативної памяті комп'ютера. Очевидно, що це можливо реалізувати тільки на суперкомп'ютерах. Великий об'єм самих розрахунків приводить до того, що для розв'язання задачі необхідно багато часу. Також при цьому відбувається накопичення помилок обчислення, що призводить до неправильного результату.

Для вирішення цих проблем застосовують деякі прийоми та спеціальні чисельні методи.

3.2. Розріджені матриці

Розріджені матриці – це матриці, кількість нульових елементів в яких є такою великою, що стає ефективним урахування структури цих матриць.

Розріджені матриці часто зустрічаються в інженерних задачах, наприклад, у задачах проектування великих інтегральних

схем, у задачах моделювання складних об'єктів та оптимального управління ними.

Для економії комп'ютерної пам'яті при зберіганні розріджених матриць застосовують спеціальне кодування, а саме зберігаються тільки ненульові елементи та їх місця розташування в матриці.

Варто розглянути приклад. Нехай є матриця розмірності 6×6.

$$A = \begin{pmatrix} a_{11} & 0 & 0 & a_{14} & 0 & 0 \\ 0 & a_{22} & 0 & 0 & 0 & 0 \\ a_{31} & 0 & a_{33} & a_{34} & 0 & a_{36} \\ 0 & 0 & a_{43} & a_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{55} & 0 \\ 0 & 0 & a_{63} & 0 & 0 & a_{66} \end{pmatrix}.$$

Для її зберігання звичайним засобом потрібно 6×6× (8 байт) = 288 байт пам'яті. Але якщо зберігати тільки ненульові елементи матриці, то буде достатньо двох лінійних масивів:

$$Ae = (a_{11}, a_{31}, a_{22}, a_{33}, a_{43}, a_{63}, a_{14}, a_{34}, a_{44}, a_{55}, a_{36}, a_{66}),$$

Am = ((1,1), (3,1), (2,2), (3,3), (4,3), (6,3), (1,4), (3,4), (4,4), (5,5), (3,6), (6,6)).

В одновимірному масиві Ae зберігаються лише самі ненульові елементи матриці A по стовпцях, при цьому потрібно 12× (8 байт) = 96 байт. У масиві Am зберігаються місця розташування ненульових елементів матриці, при цьому потрібно 12×2×(2 байта) = 48 байт. Таким чином, для зберігання матриці A у кодованому вигляді потрібно загалом лише 144 байти, тобто у два рази менше. При великій розмірності ця різниця стає дуже великою.

Описаний спосіб зберігання розріджених матриць застосовують при відсутності будь-яких закономірностей у структурі розташування нульових та ненульових елементів. Але існують матриці спеціальних видів, яку досить часто зустрічаються на практиці.

Розглянемо деякі види розріджених матриць. Якщо матриця $A \in R^{n \times n}$ має вигляд:

$$A = \begin{pmatrix} * & 0 & * & 0 & 0 & 0 \\ 0 & * & 0 & * & 0 & 0 \\ * & 0 & * & 0 & * & 0 \\ 0 & * & 0 & * & 0 & * \\ 0 & 0 & * & 0 & * & 0 \\ 0 & 0 & 0 & * & 0 & * \end{pmatrix},$$

тобто ненульові елементи знаходяться лише на головній діагоналі та деяких інших діагоналях матриці, то вона називається стрічковою [10].

У цьому випадку зберігаються значення ненульових діагоналей у вигляді матриці розмірності $n \times k$, де k – кількість цих діагоналей та їх порядковий номер щодо головної діагоналі. Діагоналі, розташовані над головною діагоналлю, нумеруються додатними значеннями, а під головною діагоналлю – від'ємними значеннями. При цьому номер головної діагоналі дорівнює нулю [10].

Якщо стрічкова матриця $A \in R^{n \times n}$ має вигляд:

$$A = \begin{pmatrix} * & * & 0 & 0 & 0 & 0 \\ * & * & * & 0 & 0 & 0 \\ 0 & * & * & * & 0 & 0 \\ 0 & 0 & * & * & * & 0 \\ 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & * & * \end{pmatrix},$$

тобто ненульові елементи знаходяться тільки на головній діагоналі та двох сусідніх діагоналях матриці, то вона називається **трьохдіагональною**.

Якщо стрічкова матриця $A \in R^{n \times n}$ має вигляд:

$$A = \begin{pmatrix} * & * & * & 0 & 0 & 0 \\ * & * & * & * & 0 & 0 \\ * & * & * & * & * & 0 \\ 0 & * & * & * & * & * \\ 0 & 0 & * & * & * & * \\ 0 & 0 & 0 & * & * & * \end{pmatrix},$$

тобто ненульові елементи знаходяться тільки на головній діагоналі та чотирьох сусідніх діагоналях матриці, то вона називається п'ятидіагональною.

Трьох- та п'ятидіагональні матриці виникають, наприклад, при розв'язанні крайових задач для звичайних диференціальних рівнянь методом кінцевих різниць.

Якщо матриця $A \in R^{n \times n}$ має вигляд

$$A = \begin{pmatrix} * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & * & * & * & * \\ 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & 0 & 0 & * \end{pmatrix},$$

тобто ненульові елементи знаходяться тільки на головній діагоналі та вище неї, то вона називається **верхньою трикутною**. Аналогічно є також і **нижні трикутні матриці**. У цьому випадку зберігаються значення ненульових елементів у вигляді одномірного масиву по стовпцях.

Приклад. Нехай є матриця розмірності 4×4

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{pmatrix}$$

Для її зберігання звичайним засобом потрібно 4×4×(8 байт) = 128 байт пам'яті. Але якщо зберігати тільки ненульові елементи матриці, то буде достатньо одномірного масива:

$Ae = (a_{11}, a_{12}, a_{22}, a_{13}, a_{23}, a_{33}, a_{14}, a_{24}, a_{34}, a_{44})$

і при цьому потрібно лише 10×(8 байт) = 80 байт. При великій розмірності ця різниця стає більш значущою.

Як неважко помітити, при проведенні розрахунків замість звернення до довільного елемента a_{ij} повної матриці A тут слід звертатися до елемента $aej*(j-1)/2+i$ вектора (масива) Ae.

3.3. Ітераційні методи розв'язання СЛАР великої розмірності з розрідженими матрицями

На відміну від прямих ітераційні методи не збільшують число ненульових елементів матриці у процесі обчислення. Ефективність ітераційних методів визначається швидкістю збіжності послідовних наближень $\bar{x}^{(k)}$ до розв'язку.

Першим кроком у ітераційному методі є перетворення вихідної системи до еквівалентного виду:

$$C\bar{x} = B\bar{x} + \bar{d},$$

де C, B і вектор $\bar{d}$ визначаються по матриці A і вектору $\bar{b}$.

Другий крок – побудова ітераційного процесу (розстановка індексів або номерів наближення) і завдання нульового наближення. Наприклад:

$$C\bar{x}^{(k+1)} = B\bar{x}^{(k)} + \bar{d},\ k = 0,\ 1,\ 2,\ldots,$$

де $\bar{x}^{(0)} = \left\{x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right\}$ – заданий вектор.

Третім кроком ітераційного методу є обґрунтування збіжності послідовних наближень $\left\{\bar{x}^{(k)}\right\}$ до точного розв'язку $\bar{x}$ системи й оцінка похибки k-го наближення:

$$\left\|\bar{x}^{(k)} - \bar{x}\right\| \le \varepsilon.$$

Ця оцінка при заданому ε дозволяє зупинити ітераційний процес.

3.4. Метод простої ітерації

У методі простої ітерації матриця C вибирається одиничною: $C = E$.

Ітераційний процес має вигляд:

$$\bar{x}^{(k+1)} = B\bar{x}^{(k)} + \bar{d},\ k = 0,\ 1,\ldots,$$

або покомпонентно,

$$x_i^{(k+1)} = \sum_{j=1}^{n} b_{ij} x_j^{(k)} + d_i,\ i = 1,\ 2,\ldots,n,\ \ k = 0,\ 1,\ldots$$

Теорема. Для того щоб СЛАР мала єдиний розв'язок $\bar{x}$ і ітераційний процес збігався до нього з точністю геометричній прогресії, достатньо виконання умови:

$$\|b\| < 1.$$

Тут і далі використовуються узгоджені норми $(\| \|_m, \| \|_l, \| \|_e)$ вектора та матриці.

Слабким місцем методу простої ітерації (та й інших ітераційних методів) є відсутність єдиного практичного підходу до пошуку перетворення в системі, щоб виконувалася умова збіжності для будь-яких невироджених матриць A.

3.5. Метод Гаусса-Зейделя

Метод Зейделя (Гаусса-Зейделя) являє собою певну модифікацію методу простої ітерації. Основна його ідея полягає в

тому, що при обчисленні $(k+1)$-го наближення невідомої x_i враховується уже $(k+1)$-е наближення невідомих $x_1, x_2, \ldots, x_{i-1}$.

Ітераційний процес у матричній формі можна представити так:

$$\bar{x}^{(k+1)} = U\bar{x}^{(k)} + L\bar{x}^{(k+1)} + \bar{d},$$

де матриці U, L отримані розкладанням B у суму:

$$B = U + L,$$

матриця U – верхня трикутна частина B, включаючи діагональ; L – нижня піддіагональна частина B;

$$U = \begin{pmatrix} b_{11} & \cdots & b_{1m} \\ & \ddots & \vdots \\ & & b_{mm} \end{pmatrix}; \quad L = \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ b_{21} & 0 & & 0 \\ \cdots & \cdots & \cdots & \cdots \\ b_{m1} & \cdots & b_{m,m-1} & 0 \end{pmatrix}.$$

Метод Зейделя можна записати в іншій формі, якщо взяти позначення $C = (E - L)$; $B = U$. Тоді маємо:

$$(E - L)\bar{x}^{(k+1)} = U\bar{x}^{(k)} + \bar{d}.$$

Побудова матриці, зворотної $(E-L)^{-1}$, не становить труднощів, тому що це нижня трикутна матриця.

Теорема. Нехай виконано умову:

$$\left\|(E-L)^{-1}U\right\| < 1,$$

тоді послідовні наближення $\{\bar{x}^{(k)}\}$ сходяться до точного розв'язання системи: $\bar{x} = B\bar{x} + \bar{d}$ при будь-якому початковому векторі $\bar{x}^{(0)}$.

Подамо матрицю A у вигляді $A = A_1 + D + A_2$, де A_1 – нижній трикутник матриці A, A_2 – верхній трикутник матриці A, D – її діагональ. Тоді систему запишемо у вигляді:

$$D\vec{x} = A_1\vec{x} + A_2\vec{x} + \vec{b}$$

або

$$\vec{x} = D^{-1}A_1\vec{x} + D^{-1}A_2\vec{x} + D^{-1}\vec{b}.$$

Тоді матричний запис методу виглядатиме так:

$$\vec{x}^{k+1} = -D^{-1}A_1\vec{x}^{k+1} - D^{-1}A_2\vec{x}^{k} + D^{-1}\vec{b}$$

або

$$(D + A_1)\vec{x}^{k+1} = -A_2\vec{x}^k + \vec{b}.$$

Узагальненням методу Зейделя є метод верхньої релаксації:

$$(D + \omega A_1)\frac{\vec{x}^{k+1} + \vec{x}^k}{\omega} + A\vec{x}^k = \vec{b},$$

де D – діагональна матриця з елементами a_{ii} по діагоналі. $\omega > 0$ – заданий числовий параметр.

Тепер $B = D + \omega A_1, \tau = \omega$. Якщо $A^* = A > 0$, то метод верхньої релаксації збігається за умови $0 < \omega < 2$. Параметр підбирається експериментально з умови мінімальної кількості ітерацій.

3.6. Метод Якобі

Припустимо $a_{ii} \neq 0, \forall i$. Зведемо систему СЛАР до вигляду:

$$x_i = -\sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ii}} \cdot x_j - \sum_{j=i+1}^{n}\frac{a_{ij}}{a_{ii}} \cdot x_j + \frac{b_i}{a_{ii}}, i = \overline{1,n}.$$

Ітераційний процес запишемо у вигляді

$$x_i^{k+1} = -\sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ii}}x_j^k - \sum_{j=i+1}^{n}\frac{a_{ij}}{a_{ii}}x_j^k + \frac{b_i}{a_{ii}} \quad k = 0,1...., i = \overline{1,n}.$$

Ітераційний процес збігається до розв'язку, якщо виконується умова

$$\sum_{\substack{j=1 \\ i \neq j}}^{n} |a_{ij}| \leq |a_{ii}|, \forall i = \overline{1,n}.$$

Це умова діагональної переваги матриці A. Якщо ж:

$$\sum_{\substack{j=1 \\ i \neq j}}^{n} |a_{ij}| \leq q \cdot |a_{ii}|, \forall i = \overline{1,n}, 0 \leq q < 1,$$

то має місце оцінка точності:

$$\|\vec{x}^n - \vec{x}\| \leq \frac{q^n}{1-q}\|\vec{x}^1 - \vec{x}^0\|.$$

Якщо подати матрицю A у вигляді $A = A_1 + D + A_2$ і $\vec{x} = D^{-1}A_1\vec{x} + D^{-1}A_2\vec{x} + D^{-1}\vec{b}$, то матричний запис методу виглядатиме так:

$$x^{k+1} = -D^{-1}(A_1 + A_2)x^k + D^{-1}Б.$$

3.7. Однокрокові методи

Канонічною формою однокрокового ітераційного методу розв'язку СЛАР є його запис у вигляді:

$$B_k \frac{\vec{x}^{k+1} - \vec{x}^{k}}{\tau_{k+1}} + A\vec{x}^{k} = \vec{b}.$$

Тут $\{B_k\}$ – послідовність матриць (переобумовлюючі матриці), що задають ітераційний метод на кожному кроці; $\{\tau_{k+1}\}$ – ітераційні параметри.

Якщо $B_k = E$, то ітераційний процес називається **явним:**

$$\vec{x}^{k+1} = \vec{x}^{k} - \tau_{k+1}\left(A\vec{x}^{k} + \vec{b}\right).$$

Якщо $B_k \neq E$, то ітераційний процес називається **неявним:**

$$B_k \vec{x}^{k+1} = \vec{F}^{k}.$$

У цьому випадку на кожній ітерації необхідно розв'язувати СЛАР.

Якщо $\tau_{k+1} \equiv \tau$, $B_k \equiv B$, то ітераційний процес називається **стаціонарним;** інакше – **нестаціонарним.**

Методам, що розглянуті вище, відповідають:

методу простої ітерації:

$$B_k = E, \tau_{k+1} = \tau;$$

методу Якобі:

$$B_k = D, \tau_{k+1} = 1;$$

методу Зейделя:

$$B_k = D + A_1, \tau_{k+1} = 1.$$

Розглянемо випадок симетричних матриць $A^* = A$ і стаціонарний ітераційний процес $B_k \equiv E, \tau_{k+1} \equiv \tau$.

Нехай для A справедливі нерівності

$$\gamma_1 E \leq A \leq \gamma_2 E,\ \gamma_1, \gamma_2 > 0.$$

Тоді при виборі $\tau = \tau_0 = \dfrac{2}{\gamma_1 + \gamma_2}$ ітераційний процес збігається. Найбільш точним значенням γ_1, γ_2 при яких виконуються да-

ні нерівності, є $\gamma_1 = \min_i \lambda_i(A), \gamma_2 = \max_i \lambda_i(A)$. Тоді $q = q_0 = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1} = \frac{1-\xi}{1+\xi}, \xi = \frac{\gamma_1}{\gamma_2}$ і справедлива оцінка:

$$\|\vec{x}^n - \vec{x}\| \le \frac{q^n}{1-q}\|\vec{x}^0 - \vec{x}\|.$$

Явний метод з багатьма параметрами $\{\tau_k\}$:

$B \equiv E, \{\tau_i\}: \min_\tau q(\tau)\ n = n(\varepsilon) \to \min$,

які обчислюються за допомогою нулів багаточлена Чебишова, називається ітераційним методом з чебишевським набором параметрів.

Аналогічно будуються методи з $B \neq E$. Матриця B називається переобумовлюючою і дозволяє підвищити швидкість збіжності ітераційного процесу. Її вибирають за умов: а) легко розв'язувати СЛАР $B\vec{x}^i = \vec{F}_i$ (діагональний, трикутній, добуток трикутних та інше); б) зменшення числа обумовленості матриці $B^{-1}A$ порівняно з A.

Контрольні запитання та завдання

1. Сформулюйте постановку задачі розв'язання системи лінійних алгебраїчних рівнянь.

2. Назвіть групи методів розв'язання систем лінійних алгебраїчних рівнянь, вкажіть їхні принципові відмінності.

3. Які матриці називаються розрідженими? Назвіть види розріджених матриць.

4. У чому полягає суть методу прогону? У яких випадках його застосовують на практиці?

5. Що називається *LU*-розкладанням матриці?

6. Коли застосовується *LU*-розкладання матриць для розв'язання систем лінійних алгебраїчних рівнянь великої розмірності?

4. ЧИСЕЛЬНІ МЕТОДИ РОЗВ'ЯЗУВАННЯ НЕЛІНІЙНИХ РІВНЯНЬ

4.1. Постановка задачі

Розв'язати рівняння – означає знайти всі його розв'язки (**корені**) або показати, що їх не існує. Нагадаємо, що розв'язком рівняння з невідомим x називають точку, при підстановці якої в рівняння замість x отримується правильна числова рівність.

Нехай маємо рівняння $f(x)=0$, f – нелінійна неперервна функція однієї змінної, тобто f: $R^1 \to R^1$. Розв'язати це рівняння означає: знайти таке $x \in R^1$, для якого $f(\bar{x}) \equiv 0$. При цьому x – розв'язок (або корінь) рівняння.

Задача розв'язання цього рівняння розпадається на декілька етапів:

1. Дослідження існування дійсних коренів.
2. Знаходження границь розташування коренів і їх числа.
3. Відділення коренів, тобто розбиття числової вісі на інтервали, де знаходиться один корінь.
4. Знаходження наближених значень коренів.
5. Уточнення наближених значень коренів із заданою точністю ε.
6. Оцінки похибок наближених значень коренів.

На перших чотирьох етапах застосовуються аналітичні і графічні методи, теореми про існування і границі коренів рівняння $f(x)=0$. Зокрема, досить корисною є така теорема.

Теорема. Якщо неперервна функція $f(x)$ приймає значення різних знаків на кінцях відрізка $[\alpha,\beta]$, тобто $f(\alpha)f(\beta)<0$, то всередині цього відрізка міститься щонайменше один корінь рівняння $f(x) = 0$, тобто знайдеться хоча б одне число $\xi \in (\alpha,\beta)$ таке, що $f(\xi)=0$.

Зауваження. Корінь ξ завідомо буде єдиним, якщо похідна $f'(x)$ існує і зберігає постійний знак всередині інтервалу (α,β).

4.2. Алгоритми ізолювання та уточнення коренів

У загальному випадку рівняння $f(x)=0$ може мати багато коренів.

Чисельні методи розв'язання нелінійних рівнянь, які розглянуто далі, дозволяють знаходити один корінь на заданому відрізку $[a, b]$. При цьому на інтервалі повинен існувати тільки один корінь. Знайти відрізок, що задовольняє цю умову, можна різними способами:

а) з фізичних міркувань, тобто на основі фізичних знань про задачу;

б) на основі досвіду розв'язання аналогічних задач;

в) за допомогою графічних методів;

г) шляхом відокремлення коренів.

Якщо функція $f(x)$ заздалегідь відома, то найбільш ефективним є графічний спосіб пошуку відрізка $[a, b]$. В інших випадках, коли відрізок $[a, b]$ треба знайти автоматично (не візуально), то застосовують алгоритм відокремлення коренів.

Процес відділення коренів починається із встановлення границь області, в якій розташовані всі корені рівняння. Для вирішення цього питання можна скористатися рядом способів відшукання даних границь.

Алгоритм відокремлення коренів

Відокремити корінь – це означає вказати на осі Ox відрізок $[a, b]$, який містить лише один корінь [5]. Алгоритм відокремлення коренів розбивається на такі кроки:

1. Розбивається деяка область на деяку кількість рівних відрізків $[x_i, x_{i+1}]$, де x_0 задано, $x_{i+1} = x_i + h$, h – крок розбиття, $i = 0, 1,\ldots,n$.

2. Визначається знак функції $f(x)$ на кінцях кожного i-го відрізка $[x_i, x_{i+1}]$.

3. Якщо добуток $f(x_i)f(x_{i+1})$ додатний і h достатньо мале, то можна сподіватися, що на відрізку $[x_i, x_{i+1}]$ коренів рівняння немає.

4. Якщо добуток $f(x_i)f(x_{i+1})$ від'ємний, то на відрізку $[x_i, x_{i+1}]$ існує розв'язок рівняння, хоча можливо, що він не один.

5. Перевіряється, чи змінює знак похідна $f'(x)$ на кінцях відрізка $[x_i, x_{i+1}]$. Якщо h достатньо мале і знак похідної $f'(x)$ на кінцях відрізка $[x_i, x_{i+1}]$ не змінюється, то можна сподіватися, що на відрізку корінь один. Таким чином, корінь рівняння відокремлений.

6. Якщо знак похідної $f'(x)$ змінюється, то впевненості, що корінь один немає.

7. Відрізки, для яких немає упевненості в тому, що розв'язок рівняння лише один, продовжують розбивати вже з меншим кроком, тобто повторюється для них описана процедура відокремлення коренів.

Спосіб звуження границь

Уявімо $f(x)$ у вигляді $f(x) = F(x) + \varphi(x)$, де $F(x)$ – багаточлен, що містить усі перші старші за ступенем члени багаточлена $f(x)$, які мають позитивні коефіцієнти, і всі члени з негативними коефіцієнтами; $\varphi(x)$ – багатогочлен, утворений усіма іншими членами $f(x)$. Тоді якщо знайдемо $x = a > 0$, для якого $F(a) > 0$, то $f(x) > 0$ при всіх $x \geq a$, оскільки $\varphi(x) \geq 0$ при $x > 0$ і всі корені рівняння $f(x) = 0$ будуть менші a.

Спосіб, запропонований Ньютоном

і заснований на твердженні: якщо при $x = a > 0$ мають місце нерівності $f(a) > 0, f'(a) > 0, \ldots, f^{(n)}(a) > 0$, то рівняння $f(x) = 0$ не має коренів, більших a.

4.3. Графічний спосіб

Для визначення границь коренів можна використовувати графічні методи. Будуємо графік функції $y = f(x)$ і за кресленням визначаємо інтервали, які містять точки перетину графіка з віссю OX, тобто корені рівняння $f(x) = 0$. Іноді зручно представити рівняння $f(x) = 0$ у вигляді $\varphi(x) = \psi(x)$ та, побудувавши графіки функцій $y = \varphi(x)$ і $y = \psi(x)$, визначити інтервали, які містять точки їх перетину.

Задача відділення дійсних коренів рівняння $f(x) = 0$ полягає в тому, щоб кожен з коренів укласти в інтервал, який не містить інших коренів рівняння. Зазвичай для вирішення цієї задачі спочатку знаходять нижню й верхню границі всіх дійсних коренів одним з наведених раніше способів. Потім отриманий відрізок розбивають на більш дрібні, зазвичай рівної довжини, так, щоб у кожному з них не могло міститися більше одного кореня. Якщо виявиться, що $f(a_k)f(a_{k+1}) < 0$, то в силу наведеної теореми в інтервалі (α_k, α_{k+1}) маємо корінь рівняння $f(x) = 0$. Слід додати, що графічним способом можна досягти досить високої точності визначення не тільки границь коренів, а і визначення самого кореня.

4.4. Метод ділення навпіл (дихотомії)

Припустимо, що на відрізку $[a,b]$ знаходиться лише один корінь рівняння $f(x) = 0$ для $f(x) \in C[a,b]$, який необхідно визначити. Нехай $f(a)f(b) < 0$. Припустимо, що $f(a) > 0, f(b) < 0$. Покладемо $x_1 = \frac{a+b}{2}$ і підрахуємо $f(x_1)$. Якщо $f(x_1) < 0$, тоді шуканий корінь $\bar{x}$ знаходиться на інтервалі (a, x_1). Якщо ж $f(x_1) > 0$, то $\bar{x} \in (x_1, b)$. Далі з двох інтервалів (a, x_1) і (x_1, b) вибираємо той, на границях якого функція $f(x)$ має різні знаки. Позначимо його (a_1, b_1). Знаходимо точку x_2 – середину вибраного інтервалу, підраховуємо $f(x_2)$ і повторюємо вказаний процес.

У результаті отримаємо послідовність інтервалів (a_k, b_k), що містять шуканий корінь $\bar{x}$, причому довжина кожного наступного інтервалу вдвічі менше попереднього.

Цей процес продовжується доти, доки довжина отриманого інтервалу (a_n, b_n) не стане меншою за ε: $b_n - a_n < \varepsilon$. Тоді:

$$|x_n - \bar{x}| < \varepsilon.$$

Ця умова для деякого n виконуватиметься за теоремою Больцано-Коші. Оскільки

$$|b_{k+1} - a_{k+1}| = \frac{1}{2}|b_k - a_k|,$$

то:

$$|x_n - \bar{x}| \le \frac{1}{2^n}(b-a) < \varepsilon$$

Звідси отримаємо нерівність для обчислення кількості ітерацій n для виконання умови $|x_n - \bar{x}| < \varepsilon$:

$$n = n(\varepsilon) \ge \left[\log_2\left(\frac{b-a}{\varepsilon}\right)\right] + 1$$

Степінь збіжності – лінійна, тобто геометричної прогресії зі знаменником $q = \frac{1}{2}$.

Переваги методу: простота, надійність. **Недоліки** методу: низька швидкість збіжності; метод не узагальнюється на системи рівнянь.

4.5. Метод простої ітерації

Спочатку рівняння

$$f(x) = 0$$

замінюється еквівалентним

$$x = \varphi(x).$$

Ітераційний процес має вигляд

$$x_{n+1} = \varphi(x_n),\ n = 0,\ 1,...$$

Початкове наближення x_0 задається.

Для збіжності ітераційного процесу велике значення має вибір функції $\varphi(x)$. Перший спосіб заміни рівняння полягає у відділенні змінної з якогось члена рівняння. Більш продуктивною еквівалентною заміною є використання функції $\varphi(x) = x + \tau(x)f(x)$, де $\tau(x)$ – знакостала функція на тому відрізку, де шукаємо корінь.

Кажуть, що ітераційний метод збігається, якщо $\lim_{k\to\infty} x_k = \bar{x}$.

Далі нехай $U_r = \{x : |x - x_0| \le r\}$ – відрізок довжини $2r$ з серединою в точці x_0. З'ясуємо умови, за якими збігається метод простої ітерації.

Теорема. Якщо $\max\limits_{x \in [a,b] + U_r} |\varphi'(x)| \le q < 1$, то метод простої ітерації збігається і має місце оцінка

$$|x_n - \bar{x}| \le \frac{q^n}{1-q}|x_0 - x_1| \le \frac{q^n}{1-q}(b-a).$$

Визначимо кількість ітерацій для досягнення точності ε. З останньої оцінки отримаємо

$$|x_n - \bar{x}| \le \frac{q^n}{1-q}(b-a) < \varepsilon \Rightarrow n(\varepsilon) = n \ge \left[\frac{\ln \frac{\varepsilon(1-q)}{b-a}}{\ln q}\right] + 1.$$

Практично ітераційний процес зупиняємо при $|x_n - x_{n-1}| < \varepsilon$. Але ця умова не завжди гарантує, що $|x_n - \bar{x}| < \varepsilon$.

Зауваження. Умова збіжності методу може бути замінена на умову Ліпшиця:

$$|\varphi(x) - \varphi(y)| \le q|x - y|, \ 0 < q < 1.$$

Переваги методу: простота; при $q < \frac{1}{2}$ – швидше збігається, ніж метод ділення навпіл; метод узагальнюється на системи. Недоліки методу: 1) при $q > \frac{1}{2}$ – збігається повільніше, ніж метод ділення навпіл, 2) виникають труднощі при зведенні $f(x) = 0$ до $x = \varphi(x)$.

4.6. Метод Ньютона (метод дотичних)

Припустимо, що рівняння $f(x) = 0$ має простий дійсний корінь $\bar{x}$, тобто $f(\bar{x}) = 0$, $f'(\bar{x}) \ne 0$. Нехай виконуються умови: $f(x) \in C^1[a,b]$, $f(a) \cdot f(b) < 0$. Тоді:

$$0 = f(\bar{x}) = f(x_i + \bar{x} - x_i) = f(x_i) + f'(\xi_i)(\bar{x} - x_i),$$

де $\xi_i = x_i + \theta_i(\bar{x} - x_i)$, $0 < \theta_i < 1$, $\xi_i \approx x_i$. Тому наступне наближення виберемо з рівняння

$$f(x_k) + f'(x_k)(x_{k+1} - x_k) = 0.$$

Звідси маємо ітераційний процес:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},\ k = 0, 1, 2\ldots;\ x_0 - \text{заданс.}$$

Метод Ньютона ще називають методом лінеаризації або методом дотичних.

Метод Ньютона можна інтерпретувати як метод простої ітерації з

$$\varphi(x) = x - \frac{f(x)}{f'(x)},\ \text{тобто}\ \tau(x) = -\frac{1}{f'(x)}.$$

Тому $\varphi'(x) = 1 - \dfrac{[f'(x)f'(x) - f(x)f''(x)]}{(f'(x))^2} = \dfrac{f(x)f''(x)}{(f'(x))^2}$. Якщо x – корінь $f(x)$, то $\varphi'(\bar{x}) = 0$. Тому знайдеться окіл кореня, де

$$|\varphi'(x)| = \left|\frac{f(x)f''(x)}{(f'(x))^2}\right| < 1.$$

Це означає, що збіжність методу Ньютона залежить від вибору x_0.

Модифікований метод Ньютона має вигляд:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)},\ k=0, 1, 2,\ldots$$

Цей метод має лише лінійну збіжність: $|x_{k+1} - \bar{x}| = O(|x_k - \bar{x}|)$.

У методі Ньютона, для якого $f'(x_k)$ замінюється на $\dfrac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}$, дає метод січних:

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k),\ k = 1, 2\ldots;\ x_0, x_1 - \text{задані.}$$

Теорема. Нехай $f(x) \in C^2[a,b]$; $\bar{x}$ простий дійсний корінь рівняння $f(x) = 0$ і $f'(x) \neq 0$ при $x \in U_r = \{x : |x - \bar{x}| < r\}$. Якщо:

$$q = \frac{M_2 |x_0 - \bar{x}|}{2m_1} < 1,$$

де $m_1 = \min_{U_r} |f'(x)|, M_2 = \max_{U_r} |f''(x)|$, то для $x_0 \in U_r$ метод Ньютона:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

збігається і має місце оцінка:

$$|x_n - \bar{x}| \le q^{2^n - 1} |x_0 - \bar{x}|.$$

З останньої нерівності дістанемо оцінку кількості ітерацій n для досягнення точності ε:

$$n \ge \left[\log_2 \left(1 + \frac{\ln \dfrac{\varepsilon}{b-a}}{\ln q}\right) \right] + 1.$$

Кажуть, що ітераційний метод має **степінь збіжності** m, якщо:

$$|x_{k+1} - \bar{x}| = O\left(|x_k - \bar{x}|^m\right).$$

Для методу Ньютона

$$|x_{k+1} - \bar{x}| = \frac{|x_k - \bar{x}|^2 |f''(\xi_k)|}{2|f'(x_k)|} \Rightarrow |x_{k+1} - \bar{x}| = O\left(|x_k - \bar{x}|^2\right).$$

Отже, степінь збіжності методу Ньютона $m = 2$. Для методу простої ітерації і ділення навпіл $m = 1$.

Розглянемо геометричний зміст методу Ньютона. Оскільки в методі рівняння $f(x) = 0$ заміняється наближеним лінійним (а це рівняння прямої):

$$f(x_k) + f'(x_k)(x - x_*) = 0,$$

то точка перетину цієї прямої з віссю ox і дає значення x_{k+1}.

Якщо $f(a)f''(a) > 0$ та $f''(x)$ не змінює знак, то потрібно обирати $x_0 = a$; при цьому $\{x_k\} \uparrow \bar{x}$.

Якщо $f(b)f''(b) > 0$, то $x_0 = b$; маємо $\{x_k\} \downarrow \bar{x}$.

Зауваження. Метод Ньютона можна застосовувати і для обчислення комплексного кореня $z_{k+1} = z_k - \dfrac{f(z_k)}{f'(z_k)}$. У теоремі про збіжність $q = \dfrac{|x_0 - \bar{x}| M_2}{2m_1}$, де $m_1 = \min_{U_r} |f'(z)|$, $M_2 = \max_{U_r} |f''(z)|$. Тут $|z|$ – модуль комплексного числа.

Переваги методу Ньютона: 1) висока швидкість збіжності; 2) узагальнюється на системи рівнянь; 3) узагальнюється на комплексні корені.

Недоліки методу Ньютона: 1) на кожній ітерації обчислюється не тільки $f(x_k)$, а і похідна $f'(x_k)$; 2) збіжність залежить від початкового наближення x_0, оскільки від нього залежить умова збіжності $q = \dfrac{M_2 |x_0 - \bar{x}|}{2m_1} < 1$; 3) потрібно, щоб $f(x) \in C^2[a,b]$.

4.7. Методи розв'язання систем нелінійних рівнянь

Розглянемо систему рівнянь:

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0 \\ \ldots \quad \ldots \quad \ldots \\ f_n(x_1, \ldots, x_n) = 0 \end{cases}.$$

Перепишемо її у векторному вигляді:

$$\vec{f}(\vec{x}) = 0.$$

У методі простої ітерації це рівняння зводиться до еквівалентного вигляду:

$$\vec{x} = \vec{\Phi}(\vec{x}).$$

Ітераційний процес у вигляді:

$$\vec{x}^{k+1} = \vec{\Phi}(\vec{x}^k),$$

початкове наближення $\vec{x}^0$ – задано.

Нехай оператор Φ визначений на множині R^n. За теоремою про стискуючі відображення wtq, ітераційний процес сходиться, якщо виконується умова:

$$\left\|\vec{\Phi}(\vec{x})-\vec{\Phi}(\vec{y})\right\| \le q\|\vec{x}-\vec{y}\|, \quad 0<q<1,$$

або

$$\left\|\Phi'(\vec{x})\right\| \le q < 1,$$

$\vec{x} \in U_r$ $\Phi'(x)=\left(\dfrac{\partial \varphi_i}{\partial x_j}\right)_{i,j=1}^{n}$. Для похибки справедлива оцінка

$$\left\|\vec{x}^{m}-\vec{x}\right\| \le \frac{q^n}{1-q}\left\|\vec{x}^1-\vec{x}^0\right\|.$$

Частковим випадком методу простої ітерації є метод релаксації:

$$\vec{x}^{k+1}=\vec{x}^{k}-\tau\vec{F}(\vec{x}^{k}),$$

де $\tau < \dfrac{2}{\left\|F'(\vec{x})\right\|}$.

Основними **недоліками** методу простої ітерації є:

- складність переходу від запису системи $\vec{f}(\vec{x})=0$ до виду $\vec{x}=\vec{\Phi}(\vec{x})$, зважаючи на нелінійність функцій;
- збіжність тільки для достатньо близьких до розв'язку початкових наближень $\vec{x}^0$;
- невисока швидкість збіжності.

Основною перевагою методу простої ітерації є невисока трудомісткість методу.

Розглянемо метод Ньютона для рівняння:

$$\vec{F}(\vec{x})=0.$$

Розкладемо функцію $\vec{F}(\vec{x})$ у ряд Тейлора до перших двох членів, тобто уявимо рівняння у вигляді:

$$\vec{F}(\vec{x}^{k})+F'(\vec{\xi}^{k})(\vec{x}-\vec{x}^{k})=0,$$

де $\vec{\xi}^{k}=\vec{x}^{k}+\theta_k(\vec{x}^{k}-\vec{x}^{k}), \quad 0<\theta_k<1.$

Тут $F(x)=\left(\dfrac{\partial f_i}{\partial x_j}\right)_{i,j=1}^{n}$ – матриця Якобі для $\vec{F}(\vec{x})$. Можемо приблизно вважати, що $\vec{\xi}^{k} \approx \vec{x}^{k}$. Тоді матимемо:

$$\vec{F}(\vec{x}^k) + F'(\vec{x}^k)(\vec{x}^{k+1} - \vec{x}^k) = 0.$$

Звідси виникає такий ітераційний процес у вигляді:

$$\vec{x}^{k+1} = \vec{x}^k - \left[F'(\vec{x}^k)\right]^{-1} \vec{F}(\vec{x}^k).$$

Це і є метод Ньютона. Для його реалізації необхідно, щоб на кожній ітерації існувала обернена матриця $\left[F'(\vec{x}^k)\right]^{-1}$.

Основними **недоліками** методу Ньютона є:

- збіжність тільки для достатньо близьких до розв'язку початкових наближень $\vec{x}^0$;
- висока трудомісткість методу, оскільки на кожній ітерації необхідно обчислювати матриці $F'(\vec{x}^k)$ і $\left[F'(\vec{x}^k)\right]^{-1}$.

Основною перевагою методу Ньютона є висока швидкість збіжності.

Можна не шукати обернену матрицю, а розв'язувати на кожній ітерації СЛАР

$$\begin{cases} A_k \vec{z}^k = \vec{F}(\vec{x}^k) \\ \vec{x}^{k+1} = \vec{x}^k - \vec{z}^k \end{cases}, k = 0,\ 1,\ 2,...,$$

де $\vec{x}^0$ – задано, а матриця $A_k = F'(\vec{x}^k)$.

Метод має квадратичну збіжність, якщо добре обрано початкове наближення. Складність методу (за умови використання методу Гаусса розв'язання СЛАР на кожній ітерації $Q = \frac{2}{3}n^3 + O(n^2)$,

де n – розмірність системи.

Модифікований метод Ньютона представляється у вигляді ітераційного процесу:

$$\vec{x}^{k+1} = \vec{x}^k - \left[F'(\vec{x}^0)\right]^{-1} \vec{F}(\vec{x}^k).$$

Тепер обернена матриця обчислюється тільки на нульовій ітерації. На інших обчислення нового наближення зводиться до множення матриці $A_0 = \left[F'(\vec{x}^0)\right]^{-1}$ на вектор $\vec{F}(\vec{x}^k)$ та додавання до $\vec{x}^k$.

Запишемо метод у вигляді системи лінійних рівнянь:

$$\begin{cases} A_0 \vec{z}^k = \vec{F}(\vec{x}^k) \\ \vec{x}^{k+1} = \vec{x}^k - \vec{z}^k \end{cases}$$

Оскільки матриця A_0 розкладається на трикутні (або обертається) один раз, то складність цього методу на одній ітерації (крім нульової) $Q_n = O(n^2)$. Але цей метод має лінійну швидкість збіжності.

Можливе циклічне застосування модифікованого методу Ньютона, тобто коли обернену матрицю похідних шукаємо та обертаємо через певне число кроків ітераційного процесу.

Основними недоліками методу Ньютона є:

- збіжність тільки для достатньо близьких до розв'язку початкових наближень x^0;
- висока трудомісткість методу, оскільки на кожній ітерації необхідно обчислювати матриці;
- невисока швидкість збіжності.

Контрольні запитання та завдання

1. Етапи розв'язання нелінійного рівняння.
2. Сформулюйте постановку задачі розв'язання рівняння з одним невідомим.
3. Які умови повинен задовольняти відрізок, на якому ведеться пошук розв'язку рівняння? Як його можна знайти?
4. У чому полягає метод дихотомії для розв'язання рівняння з одним невідомим? Яку він має швидкість збіжності?
5. Швидкість збіжності методу ділення навпіл.
6. Метод простої ітерації.
7. У чому полягає метод Ньютона для розв'язання рівняння з одним невідомим? Яку він має швидкість збіжності? Яку ще назву має цей метод?
8. Геометричний зміст методу Ньютона.
9. Розв'язання систем рівнянь. Розв'яжіть методом Ньютона систему:

$$\begin{cases} x_1 + x_1^2 - 2x_2x_3 - 0.1 = 0, \\ x_2 - x_2^2 + 3x_1x_3 + 0.2 = 0, \\ x_3 + x_3^2 + 2x_1x_2 - 0.3 = 0. \end{cases}$$

10. Знайдіть відрізки, на яких рівняння $x^3 - 4x + 2 = 0$ має корені методом відокремлення коренів.
11. Розв'яжіть вказане рівняння методами половинного поділу та простої ітерації.

5. ОБЧИСЛЕННЯ ВЛАСНИХ ЗНАЧЕНЬ І ВЛАСНИХ ВЕКТОРІВ МАТРИЦІ

5.1. Постановка задачі

Нехай задано матрицю $A:(n\times n)$. Тоді задача на власні значення ставиться так: знайти число λ та вектор $\vec{x}\neq 0$, які задовольняють рівнянню:

$$A\vec{x}=\lambda\vec{x}.$$

Тут λ називається власним значенням A, а $\vec{x}$ – власним вектором.

Останнє рівня рівносильне такому:

$$\det(A-\lambda E)\equiv P_n(\lambda)\equiv(-1)^n\lambda^n+a_n\lambda^{n-1}+\ldots+a_0=0.$$

Тут $P_n(\lambda)$ – характеристичний багаточлен.

У подальшому можуть бути корисними декілька наступних простих тверджень [11].

Теорема. Для того щоб ненульовий вектор $\vec{x}\neq 0$ був власним вектором матриці A, що відповідає власному значенню λ, необхідно і достатньо, щоб він був розв'язком однорідної СЛАР $A\vec{x}=\lambda\vec{x}$ з матрицею $A-\lambda E$.

Теорема. Для будь-яких квадратних матриць A і B матриці AB і BA мають однакові характеристичні багаточлени.

Теорема. Для того щоб матриця A була невиродженою, необхідно і достатньо, щоб вона не мала нульового власного значення.

Теорема. Власні значення матриці A і транспонованої до неї співпадають.

Теорема. Нехай A – невироджена матриця і λ_i, $i=1,\ldots,n$ – її власні значення. Тоді мають місце співвідношення $trA=\sum_{i=1}^{n}\lambda_i$, $\det A=\prod_{i=1}^{n}\lambda_i$ (trA – сума діагональних елементів матриці A або слід матриці A).

Для розв'язання багатьох задач механіки, фізики, хімії потрібне знання власних значень λ, а іноді й власних векторів, що відповідають λ.

Повною проблемою власних значень матриці називається задача знаходження всіх власних значень матриці A, а також власних векторів, що відповідають цим власним значенням.

Існує багато чисельних методів розв'язання повної проблеми власних значень, які, по суті, полягають у знаходженні перетворення подібності $B = T^* AT$, що приводить матрицю A загального вигляду до матриці B спеціального виду (трикутної, майже трикутної, трьохдіагональної, діагональної тощо), для якої проблема власних значень може бути вирішена більш просто. Розрізняються ці методи вибором матриці T залежно від виду матриці A. Зручно матрицею T використовувати ортогональні матриці, тому що для них $T^{-1} = T^*$.

У багатьох випадках потрібно знайти лише частину власних значень або максимальне чи мінімальне за модулем власне значення матриці. Останні проблеми називають частковими проблемами власних значень.

5.2. Степеневий метод

Найбільш розвиненим методом пошуку власних значень та векторів є степеневий метод, за допомогою якого можна розв'язати декілька часткових проблем.

1. Знаходження λ_{max}: $|\lambda_1| \equiv \lambda_{max} > |\lambda_2| \geq |\lambda_3| \geq \ldots$.

Нехай $\vec{x}^0$ – заданий вектор, будемо послідовно обчислювати вектори за формулою:

$$\vec{x}^{k+1} = A\vec{x}^k, k = 0, 1\ldots.$$

Тоді $\vec{x}^k = A^k \vec{x}^0$. Нехай $\{\vec{e}_i\}_{i=1}^n$ – система власних векторів. Представимо $\vec{x}^0$ у вигляді:

$$\vec{x}^0 = \sum_{i=1}^{n} c_i \vec{e}_i .$$

Оскільки $A\vec{e}_i = \lambda_i \vec{e}_i$, то $\vec{x}^k = \sum_{i=1}^{n} c_i \lambda_i^k \vec{e}_i$. При великих k $\vec{x}^k \approx c_1 \lambda_1^k \vec{e}_1$. Тому

$$\mu_1^{(k)} = \frac{x_m^{k+1}}{x_m^k} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right).$$

Отже $\mu_1^{(k)} \xrightarrow[k\to\infty]{} \lambda_1$.

Якщо матриця $A = A^*$ симетрична, то існує ортонормована система векторів $(\vec{e}_i, \vec{e}_j) = \delta_{ij}$. Тому:

$$\mu_1^{(k)} = \frac{(\vec{x}^{k+1}, \vec{x}^k)}{(\vec{x}^k, \vec{x}^k)} = \frac{\left(\sum_i c_i \lambda_i^{k+1} \vec{e}_i, \sum_j c_j \lambda_j^k \vec{e}_j\right)}{\left(\sum_i c_i \lambda_i^k \vec{e}_i, \sum_j c_j \lambda_j^k \vec{e}_j\right)} = \frac{\sum_i c_i^2 \lambda_i^{2k+1}}{\sum_i c_i^2 \lambda_i^{2k}} =$$

$$= \frac{c_1^2 \lambda_1^{2k+1} + c_2^2 \lambda_2^{2k+1} + \ldots}{c_1^2 \lambda_1^{2k} + c_2^2 \lambda_2^{2k} + \ldots} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right) \xrightarrow[k\to\infty]{} \lambda_1$$

Це означає збіжність до максимального за модулем власного значення з квадратичною швидкістю.

Якщо $|\lambda_1| > 1$, то при проведенні ітерацій відбувається зріст компонента вектора $\vec{x}^k$, що приводить до «переповнення» (overflow). Якщо ж $|\lambda_1| < 1$, то це приводить до зменшення компонента (underflow). Позбутися негативу такого явища можна, нормуючи вектори $\vec{x}^k$ на кожній ітерації.

Алгоритм степеневого методу знаходження максимального за модулем власного значення з точністю ε виглядає так:

1. $\vec{x}^0 \to \vec{e}_0 = \frac{\vec{x}^0}{\|\vec{x}^0\|}$;

2. $\vec{x}^{k+1} = A\vec{e}^k, \mu_1^{(k)} = (\vec{x}^{k+1}, \vec{e}^k)$, $\vec{e}^{k+1} = \frac{\vec{x}^{k+1}}{\|\vec{x}^{k+1}\|}$, $k = 0, 1\ldots$;

3. $|\mu_1^{(k+1)} - \mu_1^{(k)}| \geq \varepsilon$ *goto* 2;

4. $\lambda_1 \approx \mu_1^{(k+1)}$.

За цим алгоритмом для симетричної матриці $A = A$ швидкість прямування $\mu_1^{(k)}$ до λ_{max} – квадратична.

2. Знаходження λ_2: $|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \ldots$. Нехай λ_1, $\vec{e}_1$ відомі.

Можна довести, що якщо $|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \ldots$, то

$$\mu_2^{(k)} = \frac{x_m^{k+1} - \lambda_1 x_m^k}{x_m^k - \lambda_1 x_m^{k-1}} \xrightarrow[k\to\infty]{} \lambda_2, \text{ де } \vec{x}^{k+1} = A\vec{x}^k,$$

x_m^k – m-та компонента вектора $\vec{x}^k$.

3. Знаходження мінімального власного числа $\lambda_{min}(A) = \min_i |\lambda_i(A)|$.

Припустимо, що $\lambda_i(A) > 0$ та відоме λ_{max}. Розглянемо матрицю $B = \lambda_{max} E - A$. Маємо:

$$\lambda_i(B) = \lambda_{max} - \lambda_i(A), \ \forall i,$$

тому $\max_i \lambda_i(B) = \lambda_{max} - \min_i \lambda_i(A)$. Звідси

$$\lambda_{min}(A) = \lambda_{max}(A) - \lambda_{max}(B).$$

Якщо $\exists i : \lambda_i(A) < 0$, то будуємо матрицю $\overline{A} = \sigma E + A, \sigma > 0$ $\overline{A} > 0$ і для неї попередній розгляд дає необхідний результат. Замість λ_{max} в матриці B можна використовувати $\|A\|$.

Ще один спосіб обчислення мінімального власного значення полягає у використанні обернених ітерацій:

$$A\vec{x}^{k+1} = \vec{x}^k, k = 0, 1...$$

Але цей метод вимагає більшої кількості арифметичних операцій, хоча збігається швидче.

5.3. *QR*-алгоритм

***QR*-алгоритм** базується на процедурі *QR*-розкладання. ***QR*-розкладанням** (або ***QR*-факторизацією**) називається подання матриці A у вигляді $A = QR$, де Q – ортогональна, а R – верхня трикутна матриці [5].

***QR*-алгоритм** дозволяє вирішити повну проблему власних значень для несиметричних матриць. Цей метод вигідний для верхніх майже трикутних матриць – під час перетворення їх структура майже не змінюється. Такий метод заснований на приведенні вихідної матриці A до клітинно-трикутної форми, для якої легко знаходяться власні значення. При цьому отримана клітинно-трикутна матриця повинна бути подібною до вихідної. Дві квадратні матриці A і B називаються подібними, якщо існує така невироджена матриця P, що $B = P^{-1}AP$.

Нехай $A_0 = A$.

Далі матриця A_0 подається у вигляді $A_0 = Q_1R_1$ і обчислюється матриця $A_1 = Q_1R_1 = Q_2^{-1}R_2Q_2$. Тобто матриці A_1 і A_0 подібні.

Знову матриця A_1 подається у вигляді $A_1 = Q_2R_2$ і обчислюється матриця $A_2 = Q_2R_2 = Q_1^{-1}R_1Q_1$. І знову матриці A_2 і A_1 подібні, а отже, вони подібні і до матриці A_0.

Таким чином, будується послідовність матриць $\{A_k\}$, $k = 0,1,\ldots$, $A_k = Q_{k+1}R_{k+1}$, $Ak + 1 = Q_{k+1}^{-1}R_{k+1}Q_{k+1}$, які є подібними до відправної матриці A.

Справедливе таке твердження: якщо всі основні мінори матриці A не вироджені, то послідовність матриць $\{A_k\}$, $k = 0, 1,\ldots$ збігається до матриці блочно-трикутного виду (матриці Хессенберга [12]), власні значення якої легко знайти:

$$H = \begin{bmatrix} xxx & \ldots & xx \\ xxx & \ldots & xx \\ 0xx & \ldots & xx \\ \ldots & \ldots & \ldots \\ 000 & \ldots & xx \\ 000 & \ldots & xx \end{bmatrix}.$$

QR-алгоритм є стійким до похибок обчислення, тому вважається найбільш ефективним для пошуку власних значень і власних векторів матриці.

Для *QR*-розкладу матриці застосовують метод обертання Гівенса або перетворення Хаусхолдера [12].

Контрольні запитання та завдання

1. Надайте визначення власного значення та власного вектора квадратної матриці.

2. Сформулюйте постановку задачі обчислення власних значень і власних векторів квадратної матриці.

3. Для пошуку яких власних значень призначені ітераційні методи?

4. На якому принципі засновані чисельні методи перетворення подібності для обчислення власних значень і власних векторів матриці?

5. Як можна знайти мінімальне за модулем власне значення матриці?

6. Знайдіть чисельно максимальне та мінімальне за модулем власне значення матриці із системи лінійних алгебраїчних рівнянь:

$$\begin{cases} 16x_1 - 8x_2 - 4x_3 = -8, \\ -8x_1 + 13x_2 - 4x_3 - 3x_4 = 7, \\ -4x_1 - 4x_2 + 9x_3 = 6, \\ -3x_2 + 3x_4 = -3. \end{cases}$$

7. Що таке форма матриці Хессенберга?

6. ЧИСЕЛЬНЕ ДИФЕРЕНЦІЮВАННЯ ТА ІНТЕГРУВАННЯ ФУНКЦІЇ

6.1. Чисельне диференціювання функцій

Задача чиссльного диференціювання полягає у знаходженні значень похідних функції $y = f(x)$ у заданих точках у випадку, коли аналітичний запис функції $f(x)$ невідомий або складний чи функція задана таблично. Привабливість чисельного підходу здебільшого пояснюється наявністю простих залежностей, за допомогою яких похідні в заданих точках можна апроксимувати декількома значеннями функції в цих і близьких до них точках.

Конструювання формул наближеного диференціювання полягає в тому, що функцію $f(x)$ на заданому відрізку $[a, b]$ замінюють відповідною апроксимуючою функцією $\varphi(x)$, а потім вважають, що похідні від функцій $f(x)$ і $\varphi(x)$ збігаються, наприклад:

$$f'(x) = \varphi'(x),$$

де $x \in [a, b]$.

Аналогічно знаходять похідні вищих порядків від функції $f(x)$. При цьому апроксимуюча функція $\varphi(x)$ найчастіше задається у вигляді полінома.

Природно, що тоді похідні $f(x)$ обчислюються з деякою похибкою. Якщо через $R(x)$ позначити залишковий член

$$R(x) = f(x) - \varphi(x),$$

то похибку для $R'(x)$ можна записати як

$$R'(x) = f'(x) - \varphi'(x).$$

Такі ж формули диференціювання можна побудувати і для знаходження похідних вищих порядків, тобто $\Phi^{(k)}(x) = f^{(k)}(x) - R^{(k)}$.

Слід зазначити, що малість залишкового члена $R(x)$ не свідчить про малість залишкових членів похідних, бо похідні від малих функцій можуть бути досить великими. У загальному випадку наближене диференціювання є менш точною операцією, ніж інтерполяція, тому що для як завгодно близьких функцій $f(x)$ і $\varphi(x)$ різниця між їхніми похідними може бути як завгодно великою. Наявність великої похибки під час обчислення значень похідної пояснюється ще й тим, що значення функцій, які входять до формул диференціювання, здебільшого мають деяку похибку.

Однак, застосовуючи виважений підхід до вибору інтерполяційної формули та її порядку, можна отримати результат із необхідною точністю.

Найбільш зручними інтерполяційними багаточленами для чисельного диференціювання є поліноми Ньютона та Лагранжа. На їхній основі отримані формули різного порядку точності залежно від кількості задіяних точок x_i [1, 10].

6.2. Формули чисельного диференціювання на основі першої інтерполяційної формули Ньютона

Нехай функція $f(x)$ задана в рівновіддалених точках $x_i = a + ih$, $i = 0, 1, ..., n$ відрізка $[a,b]$ значеннями $f_i = f(x_i)$.

Щоб обчислити похідні $f'(x), f''(x)$ та ін., замінимо $f(x)$ інтерполяційним поліномом Ньютона для інтерполяції **вперед**, тобто:

$$f(x) = f_0 + \frac{\Delta f_0}{1!} t + \frac{\Delta^2 f_0}{2!} t(t-1) + \ldots + \frac{\Delta^n f_0}{n!} t(t-1)\ldots(t-n+1),$$

де $t = \dfrac{x - x_0}{h}$, h – крок таблиці, який використовується для інтерполювання функції у точках x, близьких до початку таблиці, а також для продовження таблиці (у бік зменшення аргументу) на один-два кроки, $\Delta f_0 = f_1 - f_0$, $\Delta^2 f_0 = \Delta f_1 - \Delta f_0 = f_2 - 2f_1 + f_0$ ... – скінчені різниці.

Для отримання скінченнорізницевої формули чисельного диференціювання візьмемо похідні від обох частин тотожності, скориставшись співвідношеннями:

$$f'(x) = \frac{df}{dx} = \frac{df}{dt}\frac{dt}{dx} = \frac{1}{h}\frac{df}{dt},$$

тому що:

$$\frac{dt}{dx} = \frac{1}{h}.$$

Тоді:

$$f'(x) = \frac{1}{h}\left[\Delta f_0 + \frac{2t-1}{2!}\Delta^2 f_0 + \frac{3t^2 - 6t + 2}{3!}\Delta^3 f_0 + \ldots\right].$$

Для обчислення другої похідної використаємо співвідношення

$$f''(x) = \frac{d(f'(x))}{dx} = \frac{d(f'(x))}{dt}\frac{dt}{dx} = \frac{1}{h}\frac{d(f'(x))}{dt}.$$

Тоді формула чисельного диференціювання матиме такий вигляд:

$$f''(x) = \frac{1}{h^2}\left[\Delta^2 f_0 + (t-1)\Delta^3 f_0 + \frac{6t^2 - 18t + 11}{3!}\Delta^4 f_0 + \ldots\right].$$

У разі потреби на основі різних інтерполяційних формул можна обчислити похідні функції будь-якого порядку.

Користуючись попереднім, можна отримати формули чисельного диференціювання для інтерполяційних поліномів різних степенів.

Для $n = 1$ маємо формулу чисельного диференціювання на основі лінійного інтерполяційного полінома:

$$f'(x) = \frac{\Delta f_0}{h} \text{ для } x \in [x_0, x_1].$$

Для $n = 2$ маємо формулу чисельного диференціювання на основі квадратичного полінома

$$f'(x) = \frac{1}{h}\left[\Delta f_0 + \frac{2t-1}{2!}\Delta^2 f_0\right] \text{ для } x \in [x_0, x_2].$$

Формулу на основі кубічної інтерполяції отримуємо, приймаючи, що $n = 3$:

$$f'(x) = \frac{1}{h}\left[\Delta f_0 + \frac{2t-1}{2!}\Delta^2 f_0 + \frac{3t^2 - 6t + 2}{3!}\Delta^3 f_0\right] \text{ для } x \in [x_0, x_3]$$

тощо. Слід зазначити, що в разі обчислення похідних за цими формулами у фіксованій точці x_i як x_0 варто обирати найближче табличне значення аргументу.

Інколи виникає потреба обчислити значення похідної від функції $f(x)$ безпосередньо у вузлах інтерполяції, тобто для $t = 0$. У цьому випадку формули чисельного диференціювання спрощуються, наприклад, маємо:

$$f'(x) = \frac{1}{h}\left[\Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \frac{1}{4}\Delta^4 f_0 + \ldots\right].$$

Формула ще більш спроститься, якщо йтиметься про окремі випадки інтерполяції – лінійної, квадратичної, кубічної. Необхідні скінченні різниці обчислюються за схемою побудови скінченних різниць.

6.3. Формули чисельного диференціювання на основі інтерполяційного полінома Лагранжа

У багатьох випадках для чисельного диференціювання досить ефективним є застосування інтерполяційного полінома Лагранжа. Дійсно, часом у формулах чисельного диференціювання зручніше застосувати не скінченні різниці функції, а її значення. Саме тоді зручніше для отримання таких формул використовувати поліном Лагранжа.

Розглянемо випадок, коли функція задана таблично в рівновіддалених точках $x_0, x_1, x_2, ..., x_n$, тобто відомі $x_i = x_0 + ih$ ($i = 0, 1,..., n$) і значення $f(x_i)$, $i = 0, 1,..., n$. Для заданої системи вузлів x_i побудуємо інтерполяційний поліном Лагранжа у вигляді:

$$L_n(x) = \sum_{i=0}^{n} \frac{\omega_{n+1}(x) f(x_i)}{(x - x_i)\omega'_{n+1}(x_i)},$$

де $\omega_{n+1}(x) = (x - x_0)(x - x_1)...(x - x_n)$.

Скористаємось заміною: $t = (x - x_0)/h$. Тоді отримаємо:

$$\omega_{n+1}(x) = (x - x_0)(x - x_1)...(x - x_n) = h^{n+1} t(t-1)...(t-n) = h^{n+1} t^{[n+1]},$$

де $t^{[n+1]} = t(t-1)...(t-n)$.

Багаточлен $\omega'_{n+1}(x_i)$ перетворимо в такий спосіб:

$$\omega'_{n+1}(x_i) = (x_i - x_0)(x_i - x_1)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n) =$$
$$= h^n i(i-1)...1(-1)...[-(n-i)] = (-1)^{n-i} h^n i!(n-i)!.$$

Тобто для випадку рівновіддалених вузлів поліном Лагранжа можемо записати в такому вигляді:

$$L_n(x) = \sum_{i=0}^{n} \frac{(-1)^{n-i} y_i t^{[n+1]}}{i!(n-i)!\, t - i}.$$

Беручи до уваги, що $dx/dt = h$, отримаємо:

$$f'(x) = L'(x) = \frac{1}{h} \sum_{i=0}^{n} \frac{(-1)^{n-i} f(x_i)}{i!(n-i)!} \frac{d}{dt}\left(\frac{t^{[n+1]}}{t-i}\right).$$

Похідні вищих порядків знаходять аналогічно.

6.4. Формули диференціювання для практичних обчислень

На основі попереднього розгляду можна записати співвідношення, які на практиці застосовуються для апроксимації похідних перших трьох порядків від функцій, заданих таблично. Для кожного співвідношення надається оцінка порядку точності.

Таблиця 6.1

Формули для оцінок перших похідних

Тип формули	Формула	Порядок точності
Несиметричні обернені, чи формули диференціювання назад	$f'(x_i) = \frac{f(x_i) - f(x_{i-1})}{h}$	$O(h)$
	$f'(x_i) = \frac{3f(x_i) - 4f(x_{i-1}) + f(x_{i-2})}{h}$	$O(h^2)$
Несиметричні прямі, чи формули диференціювання уперед	$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{h}$	$O(h)$
	$f'(x_i) = \frac{-f(x_{i+2}) + 4f(x_{i+1}) - 3f(x_i)}{2h}$	$O(h^2)$
Симетричні	$f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h}$	$O(h^2)$
	$f'(x_i) = \frac{-f(x_{i+2}) + 8f(x_{i+1}) - 8f(x_{i-1}) + f(x_{i-2})}{12h}$	$O(h^4)$

Таблиця 6.2

Формули для оцінок других похідних

Тип формули	Формула	Порядок точності
Несиметричні обернені	$f''(x_i) = \frac{f(x_i) - 2f(x_{i-1}) + f(x_{i-2})}{h^2}$	$O(h)$
	$f''(x_i) = \frac{2f(x_i) - 5f(x_{i-1}) + 4f(x_{i-2}) - f(x_{i-3})}{h^2}$	$O(h^2)$
Несиметричні прямі	$f''(x_i) = \frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{h^2}$	$O(h)$
	$f''(x_i) = \frac{f(x_{i+3}) + 4f(x_{i+2}) - 5f(x_{i+1}) + f(x_i)}{h^2}$	$O(h^2)$
Симетричні	$f''(x_i) = \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{h^2}$	$O(h^2)$
	$f''(x_i) = \frac{-f(x_{i+2}) + 16f(x_{i+1}) - 30f(x_i) + 16f(x_{i-1}) - f(x_{i-2})}{12h^2}$	$O(h^4)$

Таблиця 6.3

Формули для оцінок третіх похідних

Тип формули	Формула	Порядок точності
Несиметричні обернені	$f'''(x_i) = \frac{f(x_i) - 3f(x_{i-1}) + 3f(x_{i-2}) - f(x_{i-3})}{h^3}$	$O(h)$
	$f'''(x_i) = \frac{5f(x_i) - 18f(x_{i-1}) + 24f(x_{i-2}) - 14f(x_{i-3}) + 3f(x_{i-4})}{2h^3}$	$O(h^2)$
Несиметричні прямі	$f'''(x_i) = \frac{f(x_{i+3}) - 3f(x_{i+2}) + 3f(x_{i+1}) - f(x_i)}{h^3}$	$O(h)$
	$f'''(x_i) = \frac{-3f(x_{i+4}) + 14f(x_{i+3}) - 24f(x_{i+2}) + 18f(x_{i+1}) - 5f(x_i)}{2h^3}$	$O(h^2)$
Симетричні	$f'''(x_i) = \frac{f(x_{i+2}) - 2f(x_{i+1}) + 3f(x_{i-1}) - f(x_{i-2})}{2h^3}$	$O(h^2)$
	$f'''(x_i) = \frac{f(x_{i+3}) + 8f(x_{i+2}) - 13f(x_{i+1}) + 13f(x_{i-1}) - 8f(x_{i-2}) - f$	$O(h^4)$

Різниця між окремими формулами найбільш чітко проявляється у разі їхнього практичного застосування для оцінювання значень похідних.

Формули чисельного диференціювання застосовують досить широко. Наприклад, симетричні формули для оцінок других похідних використовують під час розв'язання крайових задач для звичайних диференціальних рівнянь і диференціальних рівнянь із частинними похідними. Несиметричні обернені формули застосовують для апроксимації перших похідних у граничних умовах для крайових задач і розв'язання жорстких звичайних диференціальних рівнянь із початковими умовами та ін.

6.5. Чисельне інтегрування функції

Розв'язати задачу інтегрування означає обчислити інтеграл для деякої функції $f(x)$ на заданому інтервалі $[a, b]$:

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

Якщо функція $f(x)$ неперервна на інтервалі $[a, b]$ і відома її первісна функція $F(x)$, то можна аналітично знайти інтеграл J за формулою Ньютона-Лейбніца:

$$J = F(b) - F(a).$$

Однак у багатьох випадках первісну функцію $F(x)$ не можна знайти аналітично чи $f(x)$ є занадто складною, що ускладнює обчислення інтеграла за формулою Ньютона-Лейбніца, або воно взагалі стає неможливим. Часто на практиці підінтегральна функція $f(x)$ задається таблично, що також унеможливлює використання аналітичних методів.

У всіх згаданих випадках для обчислення інтеграла використовують чисельні методи. Традиційний підхід полягає в тому, що функцію $f(x)$ на відрізку $[a, b]$ замінюють інтерполяційною функцією $\varphi(x)$, наприклад поліномом Лагранжа або Ньютона, а потім приймають

$$J = \int_a^b f(x)dx = \int_a^b \varphi(x)dx + R(x),$$

де $R(x)$ – деяка похибка формули інтегрування.

У цьому випадку функція $\varphi(x)$ має бути такою, щоб інтеграл можна було обчислити безпосередньо. Якщо функція $f(x)$ задана аналітично, то наближено обчислити визначений інтеграл можна заміною інтеграла скінченною сумою. При цьому відрізок інтегрування $[a, b]$ розбивається на n однакових частин з кроком $h = (b - a)/n$.

Розглянемо загальну постановку задачі чисельного інтегрування функції та підходи до її розв'язку.

Отже, нехай потрібно знайти

$$I(f) = \int_a^b \rho(x)f(x)dx,$$

де $f(x)$ – задана функція, $\rho(x) > 0$ – деякий ваговий множник. Ця задача часто вимагає чисельного розв'язку, оскільки:

- значна кількість інтегралів такого типу не можуть бути обчислені аналітично;
- інформація про функцію $f(x)$ може бути задана у вигляді таблиці.

Нагадаємо, що за означенням:

$$I(f) = \lim_{\Lambda \to 0} \sum_{i=1}^{n} \rho(\xi_i)f(\xi_i)\Delta x_i,$$

де $\Delta x_i = x_i - x_{i-1}$, а $\{x_i\}_{i=0}^{n}$ – розбиття проміжку $[a,b]$, $x_i \in [a,b]$, $\xi_i \in [x_{i-1}, x_i]$. Тому візьмемо як наближення таку суму, яка називається **квадратурною формулою:**

$$I_n(f) = \sum_{k=0}^{n} c_k f(x_k),$$

де x_k – вузли квадратурної формули, а c_k – її вагові множники. Тепер задача полягає в тому, щоб обрати $\{x_k, c_k\}_{k=0}^{n}$, так щоб похибка була найменша:

$$R_n(f) = I(f) - I_n(f) \to \min.$$

Квадратурну формулу називають **квадратурною формулою замкненого типу,** якщо $x_0 = a$ та $x_n = b$, і **відкритого типу,** якщо $x_0 > a$ та $x_n < b$.

Квадратурна формула має m-ий **степінь алгебраїчної точності,** якщо

$$R_n(f) = 0, \ \forall f \in \pi_m,$$

Де π_m – множина поліномів m-го степеня і $\exists P_{m+1}(x) \in \pi_{m+1}$, такий що $R_n(P_{m+1}) \neq 0$.

Цю умову можна замінити умовою $R_n(x^\alpha) = 0$, $\alpha = \overline{0,m}$, $R_n(x^{m+1}) \neq 0$, яка більш зручна для перевірки.

Розглянемо деякі загальні підходи до побудови квадратурних формул.

1. Інтерполяційний. Він приводить до квадратурних формул інтерполяційного типу. В інтегралі покладають $f(x) \approx L_n(x)$ по деяких вузлах $\{x_k\}_{k=0}^{n}$ (вузли фіксовані). Тоді:

$$I_n(f) = I(L_n(x)) = \int_a^b \rho(x) \sum_{k=0}^{n} \frac{f(x_k)\omega_n(x)}{(x-x_k)\omega_n'(x_k)} dx = \sum_{k=0}^{n} f(x_k) \int_a^b \rho(x) \frac{\omega_n(x)}{(x-x_k)\omega_n'(x_k)} dx.$$

Отже, вузлами цієї квадратурної формули є вузли інтерполяційного багаточлена, а вагові множники

$$c_k = \int_a^b \rho(x) \frac{\omega_n(x)}{(x-x_k)\omega_n'(x_k)} dx.$$

2. Найвищого алгебраїчного степеня точності. Обираємо одночасно x_k і c_k за умови $R_n(x^\alpha) = 0$, $\alpha = \overline{0,m}$ так, щоб m бу-

ло максимальним. Отримуємо систему нелінійних алгебраїчних рівнянь, розв'язавши яку, отримуємо квадратурні формули найвищого алгебраїчного степеня точності.

3. Складені квадратурні формули. Проміжок $[a,b]$ розбиваємо на підпроміжки (наприклад, однакової довжини), а потім на кожному проміжку використовуємо з невеликим степенем формули з пункту 1 або 2. Наприклад, для формул інтерполяційного типу:

$$I(f)=\sum_{i=1}^{N}\int_{x_{i-1}}^{x_i}\rho(x)f(x)dx \approx \sum_{i=1}^{N}\sum_{k=0}^{n}c_k' f(x_k')=I_h(f).$$

Квадратурна формула складеного типу I_h має порядок (степінь) точності p по кроку h, якщо $R_h(f)=I(f)-I_h(f)=O(h^p)$.

Далі розглянемо найбільш розповсюджені квадратурні формули прямокутників та трапецій.

6.6. Формули прямокутників

Як відомо, геометрична інтерпретація значення інтеграла така: його значення дорівнює площі під кривою, тобто площі криволінійної трапеції. Тоді найпростіший підхід до обчислення значення інтеграла полягає у заміні площі криволінійної трапеції, обмеженої графіком функції, сумою площ прямокутників, що формально означає, що функція $f(x)$ апроксимується поліномом нульового степеня.

Для побудови формули чисельного інтегрування на всьому відрізку $[a, b]$ спочатку необхідно побудувати квадратурну формулу для інтеграла на частковому відрізку, а потім скористатися властивістю інтеграла (цю властивість називають адитивністю)

$$J=\int_a^b f(x)dx=\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} f(x)dx.$$

Спочатку розглянемо метод центральних прямокутників. Замінимо значення інтеграла на частковому відрізку $[x_0, x_1]$ площею прямокутника:

$$\int_{x_0}^{x_1} f(x)dx=(x_1-x_0)f(\bar{x})+R_1(f),$$

де $\bar{x} = (x_1 + x_0)/2$ – центральна точка відрізка.

Отримана формула називається формулою центральних прямокутників на частковому відрізку. Це означає, що площа криволінійної трапеції заміняється площею прямокутника зі сторонами $[x_0, x_1]$ та $f(x_0+h/2)$. Тоді похибку на частковому відрізку визначимо як:

$$R_1(f) = \int_{x_0}^{x_1} f(x)dx - (x_1 - x_0) f(\bar{x}).$$

Припустимо, що функція $f(x)$ неперервна разом зі своїми похідними до другого порядку включно. Розкладемо функцію $f(x)$, $x \in [x_0, x_1]$ в околі точки $\bar{x}$ у ряд Тейлора.

Проінтегруємо останній вираз на інтервалі $[x_0, x_1]$:

$$\int_{x_0}^{x_1} f(x)dx = \int_{x_0}^{x_1} f(\bar{x})dx + \int_{x_0}^{x_1} f'(\bar{x})(x-\bar{x})dx + \int_{x_0}^{x_1} f''(\xi)\frac{(x-\bar{x})^2}{2!}dx =$$

$$= f(\bar{x})(x_1 - x_0) + f'(\bar{x})\left[\frac{x^2}{2}\Big|_{x_0}^{x_1} - x\bar{x}\Big|_{x_0}^{x_1}\right] + f''(\xi)\int_{x_0}^{x_1}\frac{(x-\bar{x})^2}{2}dx.$$

Виразимо похибку:

$$R_1(f) = \frac{1}{2} f''(\xi)\int_{x_0}^{x_1}(x-\bar{x})^2 dx,$$

$$\min f''(x) \le f''(\xi) \le \max f''(x),\ x_0 \le x \le x_1,$$

$$R_1(f) = \frac{1}{2} f''(\xi)\frac{(x-\bar{x})^3}{3}\Big|_{x_0}^{x_1} = \frac{1}{2} f''(\xi)\frac{1}{3}\left(\frac{h^3}{8} + \frac{h^3}{8}\right).$$

Виходячи з того, що $h = x_1 - x_0$, визначимо:

$$R_1(f) = \frac{f''(\xi)}{24} h^3.$$

Приймемо, що $M_2 = \max|f''(\xi)|$. Тоді для похибки формули центральних прямокутників на частковому відрізку справедлива оцінка:

$$R_1(f) \le \frac{h^3}{24} M_2.$$

Підсумовуючи праві частини рівняння, отримаємо узагальнену формулу центральних прямокутників:

$$\int_a^b f(x)dx = h\left[f(x_0 + \frac{h}{2}) + f(x_1 + \frac{h}{2}) + \ldots + f(x_{n-1} + \frac{h}{2})\right] + R(f) =$$

$$= h\sum_{k=1}^{n} f\left(x + \frac{(2k-1)h}{2}\right) + R(f).$$

Загальна похибка цієї формули дорівнює сумі похибок на всіх часткових відрізках:

$$R(f) = \frac{h^3}{24}\left(f''(\xi_1) + f''(\xi_2) + \ldots + f''(\xi_n)\right),$$

де $x_{i-1} < \xi_i < x_i$, $i = 1, \ldots, n$, або

$$R(f) = \frac{h^3}{24} n f''(\eta) = \frac{(b-a)^2}{24} h^2 f''(\eta),$$

$$R(f) = \frac{h^2 (b-a)^2}{24} M_2,$$

де $M_2 = \max\left|f''(\xi)\right|$, $\xi \in [a, b]$.

Порядок точності отриманої формули – $O(h^2)$.

Для оцінювання точності потрібно знати найбільше значення другої похідної. Щоб похибка не перевищувала задане значення ε, крок інтегрування необхідно обирати за умови:

$$h \le \sqrt{\frac{24\varepsilon}{(b-a)M_2}}.$$

Формули інтегрування на основі прямокутників можуть бути побудовані й за іншого розташування вузлів. У загальному випадку формулу прямокутників можна записати в такому вигляді:

$$\int_{x_0}^{x_n} f(x)dx = h\sum_{i=0}^{n-1} f(q + ih) + R(f).$$

Формулу центральних прямокутників можна отримати, якщо за q прийняти значення $q = x_0 + h/2$. Коли $q = x_0$, отримаємо формулу лівих прямокутників, а для $q = x_0 + h$ – формулу правих прямокутників.

У формулі лівих прямокутників залишковий член має такий вигляд:

$$R(f) = \frac{x_n - x_0}{2} h f'(\xi).$$

Залишковий член формули правих прямокутників має такий вигляд:

$$R(f) = -\frac{x_n - x_0}{2} hf'(\xi), x_0 \le \xi \le x_n.$$

Похибка обчислень двох останніх формул дорівнює $O(h)$; як бачимо, вона більша, ніж для формули центральних прямокутників через порушення симетрії.

6.7. Формула трапецій

Тепер розглянемо підхід, що полягає в заміні функції $f(x)$ інтерполяційним поліномам першого степеня. Спочатку визначимо варіант заміни на частковому відрізку $[x_0, x_1]$ площі криволінійної трапеції площею прямокутної трапеції, побудованої за тими ж точками. Ця заміна може бути зроблена у такий спосіб:

$$\int_{x_0}^{x_1} f(x)dx = (x_1 - x_0)\left[\frac{1}{2}f(x_0) + \frac{1}{2}f(x_1)\right] + R_1(f) = \frac{x_1 - x_0}{2}[f(x_0) + f(x_1)] + R_1(f).$$

Оцінимо похибку $R_1(f)$ у разі заміни $f(x)$ інтерполяційним поліномом першого степеня на частковому відрізку $[x_0, x_1]$:

$$f(x) = f(x_0) + (x - x_0)f(x_0, x_1) + r_1(x),$$

де $r_1(x) = \dfrac{(x - x_0)(x - x_1)f''(\xi)}{2!}$.

Виконаємо такі перетворення:

$$\int_{x_0}^{x_1} f(x)dx = \int_{x_0}^{x_1}[f(x_0) + (x - x_0)f(x_0, x_1) + r_1(x)]dx =$$

$$= f(x_0)(x_1 - x_0) + \frac{(x - x_0)^2}{2} f(x_0, x_1)\Big|_{x_0}^{x_1} + \int_{x_0}^{x_1} r_1(x)dx =$$

$$= f(x_0)(x_1 - x_0) + \frac{(x - x_0)^2}{2}\frac{f(x_1) - f(x_0)}{x_1 - x_0} + \int_{x_0}^{x_1}\frac{(x - x_0)(x - x_1)f''(\xi)}{2}dx =$$

$$= (x_1 - x_0) + \left[\frac{2f(x_0) + f(x_1) - f(x_0)}{2}\right] + \frac{f''(\xi)}{2}\int_{x_0}^{x_1}(x - x_0)(x - x_1)dx,$$

$$\int_{x_0}^{x_1}(x - x_0)(x - x_1)dx = -\frac{1}{6}(x_1 - x_0)^3,$$

$$|R_1(f)| \le \frac{(x_1 - x_0)^3}{12} M_2 = \frac{h^3}{12} M_2,$$

де $M_2 = \max|f''(\xi)|$.

Складена формула трапецій для всього інтервалу має такий вигляд:

$$\int_a^b f(x)dx = \frac{h}{2}\left[f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n) \right] + R(f).$$

Похибка цього методу обчислюється як сума похибок на всіх часткових інтервалах:

$$R(f) = \sum_{i=1}^{n} -\frac{h^3}{12} f''(\xi_i) = -\frac{h^3}{12} n f''(\eta) = -\frac{(b-a)^2}{12} h^2 f''(\eta)$$

і визначається як

$$|R(f)| \le \frac{h^2(b-a)}{12} f''(\eta) = \frac{h^2(b-a)}{12} M_2.$$

6.8. Формула Сімпсона

Відрізок $[a, b]$ розбивається на n частин з кроком $h = (b - a)/n$, при цьому точки розбиття x_i визначаються за формулою $x_i = a + ih$, $i = 1,\ldots, n$, тобто $x_0 = a$, $x_n = b$.

Метод Сімпсона заснований на кусково-квадратичній інтерполяції функції $f(x)$, побудованій за точками $(x_i, f(x_i))$, $i = 0,\ldots n$.

Оскільки на кожному інтервалі $[x_i, x_{i+1}]$ функція $f(x)$ інтерполюється параболою (тобто функцією виду $g_i(x) = a_i x^2 + b_i x + c$), то площу S_i криволінійної трапеції на цьому інтервалі нескладно обчислити аналітично:

$$S_i = \int_{x_i}^{x_{i+1}} g_i(x)dx.$$

Тоді наближене значення всієї площини дорівнюватиме:

$$S_n = \sum_{i=0}^{n-1} S_i.$$

Тоді неважко отримати формулу Сімпсона в загальному випадку [10]:

$$\int_a^b f(x)dx \approx \frac{h}{3}\left[f(x_0) + f(x_n) + 2\left(f(x_2) + f(x_4) + \ldots + f(x_{n-2})\right) + 4\left(f(x_1) + f(x_3) + \ldots + f(x_{n-1})\right)\right]$$

або коротше:

$$S_n = \frac{h}{3}\left[f(x_0) + f(x_n) + 2\sum_{k=1}^{n/2-1} f(x_{2k}) + 4\sum_{k=0}^{n/2-1} f(x_{2k+1}) \right].$$

Як видно з цієї формули, число n має бути обов'язково парним.

Оцінка погрішності формули Сімпсона така [6]:

$$|I - S_n| \le \frac{b-a}{180} M_4 h^4,$$

де $M_4 = \max\left|f^{(4)}(\xi)\right|$, $\xi \in [a,b]$.

З цієї оцінки виходить, що якщо функція $f(x)$ є багаточленом 3-ї степені, то загальна формула Сімпсона для обчислення інтеграла є теоретично точною.

Контрольні запитання та завдання

1. Сформулюйте постановку задачі диференціювання функцій.
2. У яких випадках виникає задача чисельного диференціювання?
3. Що називається інтерполяційним багаточленом Ньютона? Як його застосовують при виведенні формул чисельного диференціювання функцій?
4. На якому принципі заснований чисельний розрахунок похідних функцій у точці?
5. Який інтерполяційний багаточлен є найбільш зручним для чисельного диференціювання?
6. У яких випадках краще застосовувати формули диференціювання вперед, назад та симетричні формули?
7. Знайдіть чисельно першу та другу похідну в точці $x = 3$ функції, заданої таблицею

x	2	2.5	3	3.5	4	4.5
y	5.5	4.5	3	2	4.5	7

8. Сформулюйте постановку задачі чисельного інтегрування функції.
9. У чому полягає ідея методу прямокутників?
10. У чому полягає ідея методу трапецій?
11. У чому полягає ідея методу Сімпсона?

12. Який з відомих вам методів чисельного інтегрування має більшу точність?

13. Вкажіть алгоритм, за яким можна розрахувати значення визначеного інтеграла із заданою точністю.

14. Обчисліть визначений інтеграл функції $f(x) = \frac{5 - x}{x^2 + 2}$ на відрізку [–2, 3] за формулами трапецій і Сімпсона. Порівняйте результати.

7. РОЗВ'ЯЗАННЯ ЗАДАЧІ КОШІ ДЛЯ ЗВИЧАЙНИХ ДИФЕРЕНЦІАЛЬНИХ РІВНЯНЬ

7.1. Постановка задачі Коші

Задача Коші для звичайних диференціальних рівнянь (далі – ЗДР) використовується як математична модель при розв'язанні багатьох задач природознавства. Наприклад, задачі динаміки системи взаємодіючих тіл (у моделі руху матеріальних точок), задачі хімічної кінетики, електричних ланцюгів та ін. Ряд важливих рівнянь у частинних похідних у випадках, що допускають розділення змінних, приводить до задач для ЗДР. Це, як правило, крайові задачі (задачі про власні коливання пружних балок і пластин, визначення спектра власних значень енергії частинки у сферично-симетричних полях і багато інших).

Постановка задачі: нехай потрібно знайти розв'язок диференціального рівняння з початковими умовами

$$\frac{du}{dt} = f(t,u),\ t > t_0,\ u(t_0) = u_0,$$

$$u = (u_1,...,u_m)^*,\ f = (f_1,...,f_m)^*,\ u_k = u_k(t), f_k = f_k(t, u_1,...,u_m).$$

Якщо довільна функція f_k неперервна за кожного своєю змінного та за u_j вони Ліпшиць-неперервні, тобто:

$$\left|f_k(t,...,u_j,...) - f_k(t,...,v_j,...)\right| \le L_j\left|u_j - v_j\right|,\ \forall j,k,$$

(L_j – стала Ліпшиця), то розв'язок цієї задачі існує єдиний.

Нехай задано рівняння m-ого порядку та початкові умови:

$$\begin{cases} v^{(m)}(t) = F(t, v, v',..., v^{(m-1)}), \\ t = t_0 : v = v_1, v' = v_2,..., v^{(m-1)} = v_m. \end{cases}$$

Введемо компоненти вектора $\vec{u}$: $u_j(t) = v^{(j-1)}(t)$. Тоді ця задача записується у вигляді попередньої системи:

$$\begin{cases} \frac{du_1}{dt} = u_2, & u_1(t_0) = v_1, \\ ... \\ \frac{du_{m-1}}{dt} = u_m, & u_{m-1}(t_0) = v_{m-1}, \\ \frac{du_m}{dt} = F(t, u_1,..., u_m), & u_m(t_0) = v_m. \end{cases}$$

Тому далі в основному розглядаються методи розв'язання попередньої задачі.

Чисельні методи, що будуть розглянуті далі, мають місце для усіх постановок загальної задачі.

Чисельні методи розв'язання задачі знаходять розв'язок (тобто функцію $y(x)$ на відрізку $[a, b]$ у табличному вигляді, а саме у вигляді набору точок (x_i, y_i), $i = 0,\ldots, n$, де $x_0 = a$, $x_i = x_0 + ih$, $i = 1,\ldots, n$, $h = (b - a)/n$, n – задане число розбиття відрізка $[a, b]$, y_i, $i = 1,\ldots, n$ – знайдені наближені значення функції $y(x)$ в точках x_i (слід нагадати, що y_0 задане спочатку).

Розглянемо декілька методів розв'язання задачі Коші для диференціального рівняння на відрізку $[a, b]$.

7.2. Метод послідовних наближень (метод Пікара)

Потрібно розв'язати диференціальне рівняння з відповідними початковими умовами:

$$\frac{du}{dt} = f(t,u), \quad u(t_0) = u_0.$$

Проінтегруємо його

$$u(t) = u(t_0) + \int_{t_0}^{t} f(\xi, u(\xi))d\xi.$$

Задаємо $u^{(0)}(t)$ і запишемо ітераційний процес:

$$u^{(k+1)}(t) = u_0 + \int_{t_0}^{t} f(\xi, u^{(k)}(\xi))d\xi, \; k = 0, 1,\ldots.$$

Існує $T = T(u_0, L)$ ($L = \max_j L_j$ – стала Ліпшиця) така, що:

$$u^{(n)}(t) \underset{n\to\infty}{\to} u(t), \; t \in [t_0, T].$$

Тому $u(t) \approx u^{(n)}(t)$.

Недолік методу: необхідно проведення аналітичного інтегрування.

7.3. Метод рядів Тейлора

Нехай розв'язок початкової задачі можна представити у вигляді ряду $u(t) = \sum_{k=0}^{r} \frac{u^{(k)}(t_0)}{k!}(t - t_0)^k$.

Будемо шукати наближення у вигляді скінченої суми:

$$u(t) \approx u^N(t) = \sum_{k=0}^{N} \frac{u^{(k)}(t_0)}{k!}(t-t_0)^k,\ t \in [t_0, t_1].$$

Для визначення $u^{(k)}(t_0)$ диференціюємо початкове рівняння по t:

$$u^{(0)}(t_0) = u_0,\ u^{(1)}(t_0) = f(t_0,\ u_0),$$

$$u^{(2)}(t_0) = f_{t0} + f_{u0}f_0,\ u^{(3)}(t_0) = f_{tt0} + 2f_{tu0}f_0 + f_{uu0}f_0^2, \ldots .$$

Якщо $\tau = t_1 - t_0$ малий параметр, то:

$$\left|u(t) - u^N(t)\right| = O(\tau^{N+1}).$$

Недоліки методу:

- зростання кількості доданків при обчисленні $u^{(n)}(t_0)$;
- необхідність аналітичного диференціювання.

7.4. Метод Ейлера та його модифікації

Розглянемо задачу Коші:

$$\frac{du}{dt} = f(t,u),\ u(t_0) = u_0.$$

Використаємо перше наближення за допомогою рядів Тейлора на проміжку $t_0 \le t \le t_1$:

$$u(t) \approx u^1(t) = u(t_0) + (t-t_0)f(t_0,\ u_0).$$

Обчислимо наближене значення в точці t_1:

$$u(t_1) = u_1 \approx u^1(t_1) = u_0 + hf(t_0,\ u_0),$$

де $h = t_1 - t_0$ – деякий крок. Якщо позначити $y_1 = u(t_1),\ y_0 = u(t_0)$, то маємо формулу

$$y_1 = y_0 + hf(t_0, y_0).$$

Застосовуючи такий підхід для $t_n \le t \le t_{n+1}$, отримаємо рекурентну формулу

$$y_{n+1} = y_n + hf(t_n, y_n), n = 0,\ 1, \ldots, y_0 = u_0.$$

Це формула методу Ейлера. Крок інтегрування може змінюватися:

$$h = h_n = t_{n+1} - t_n.$$

Цей метод належить до однокрокових, тобто розв'язок на наступному кроці обчислюється тільки по одному значенню на попередньому кроці. Загальна формула однокрокових методів

$$y_{n+1} = y_n + \Phi(t_n, y_n, f).$$

Для методу Ейлера $\Phi = hf(t_n, y_n)$.

Ідея методу Ейлера [5] заснована на тому, що шукана інтегральна крива $y = y(x)$, яка проходить через точку $M_0(x_0, y_0)$, відновлюється у вигляді кусочно-лінійної ламаної $M_0M_1M_2...M_n$ з вершинами $M_i(x_i, y_i)$, $(i = 0,..., n)$ (рис. 7.1). Кожен відрізок M_iM_{i+1} цієї ламаної має напрям, що співпадає з напрямом тієї інтегральної кривої, яка проходить через точку M_i.

Тоді $y_{i+1} - y_i = h\text{tg}(\alpha)$. Тангенс кута нахилу дотичної до $y(x)$ в точці x_i дорівнює $y'(x_i)$, а отже дорівнює $f(x_i, y_i)$, звідки і з'являється формула методу Ейлера.

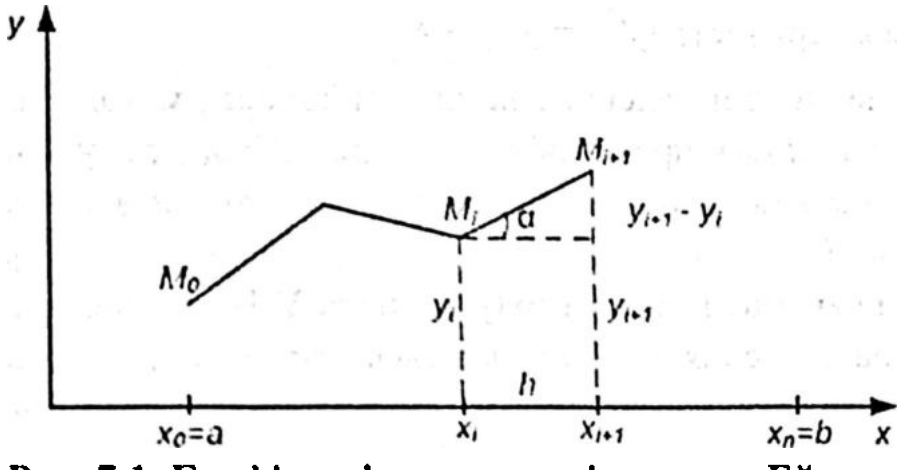


Рис. 7.1. Графічна інтерпретація методу Ейлера

Таким чином, у методі Ейлера значення y_i обчислюються рекурентно за формулою:

$$y_{i+1} = y_i + hf(x_i, y_i), i = 0,..., n.$$

Погрішність методу Ейлера [2, 3, 4, 8], що оцінюється для величини $|y_n - y(x_n)|$, має порядок $O(h)$, тобто існує деяка константа $M > 0$ така, що:

$$|y_n - y(x_n)| \le Mh.$$

Тут $y(x)$ – точний розв'язок задачі Коші, а порівняння з наближеним розв'язком (x_i, y_i), $i = 0,\ldots, n$, йде в крайній (правій) точці $x_n = b$ відрізка $[a, b]$, оскільки саме в ній погрішність теоретично буде максимальною, відносно проміжних точок x_i.

Варто зазначити, що остання оцінка має лише теоретичний характер, на практиці ж для оцінки отриманого розв'язку можна скористатися подвійним прорахунком. Для цього проводять повторні обчислення, але з кроком $h/2$, отримують вже точніший наближений розв'язок (x^*_i, y^*_i), $i = 0,\ldots, n^*$, $(n^* = 2n)$ і погрішність наближення оцінюють як:

$$\left|y^*_{n^*} - y(x_{n^*})\right| \approx \left|y^*_{n^*} - y_n\right|.$$

Таким чином, для того щоб отримати розв'язок задачі Коші із заданою точністю $\varepsilon > 0$ (застосовуючи метод Ейлера), можна, починаючи з деякого початкового значення кроку h, проводити подвійний прорахунок (тобто зменшуючи вдвічі h) допоки не виконається нерівність $\left|y^*_{n^*} - y_n\right| \le \varepsilon$.

Як видно з наведеної оцінки погрішності, метод Ейлера дає невисоку точність при розв'язанні задачі Коші, тому його рідко застосовують на практиці. Проте слід розглянути дві цікаві його модифікації.

1-ша модифікація методу Ейлера. У 1-й модифікації методу Ейлера значення y_i обчислюються рекурентно за формулами $(i = 0,\ldots, n\text{-}1)$:

$$y_{i+1/2} = y_i + \frac{h}{2} f(x_i, y_i),$$

$$y_{i+1} = y_i + hf(x_i + \frac{h}{2}, y_{i+1/2}).$$

Ідея цієї модифікації методу Ейлера полягає в тому, що спочатку обчислюється значення $y_{i+1/2}$ за першою формулою у проміжній точці $x_{i+1/2} = x + h/2$ (середині інтервалу $[x_i, x_{i+1}]$), а потім отримують y_{i+1} в точці x_{i+1} за другою формулою.

Погрішність 1-ї модифікації методу Ейлера [2, 3, 4, 8], що оцінюється для величини $|y_n - y(x_n)|$, має порядок $O(h^3)$, тобто існує деяка константа $M > 0$ така, що $|y_n - y(x_n)| \le Mh^3$.

Як видно з наведеної оцінки, погрішність 1-ї модифікації метода Ейлера на два порядки менша, ніж у звичайного методу Ейлера, хоча за це доводиться платити додатковими обчисленнями. А саме: функція $f(x, y)$ з правої частини диференціального рівняння обчислюється на кожному кроці 1-ї модифікації двічі. Варто зазначити, що при розв'язанні практичних задач основні обчислювальні витрати припадатимуть саме на обчислення значень функції $f(x,y)$.

2-га модифікація методу Ейлера. У 2-й модифікації методу Ейлера значення y_i обчислюються рекурентно за формулами ($i = 0,\ldots, n\text{-}1$):

$$y^*_{i+1} = y_i + hf(x_i, y_i),$$

$$y_{i+1} = y_i + \frac{h}{2} f(x_i, y_i) + f(x_{i+1}, y^*_{i+1})).$$

Ідея цієї модифікації методу Ейлера полягає в тому, що спочатку обчислюється «грубе наближення» y^*_{i+1} за першою формулою і за ним обчислюють значення $f(x_{i+1}, y^*_{i+1})$, а потім «остаточне наближення» y_{i+1} отримують за другою формулою, в якому фігурує середнє значення $\frac{1}{2}((f(x_i, y_i) + f(x_{i+1}, y^*_{i+1}))$.

Погрішність 2-ї модифікації методу Ейлера, що оцінюється для величини $|y_n - y(x_n)|$, також має порядок $O(h^3)$, тобто існує деяка константа $M > 0$ така, що $|y_n - y(x_n)| \le Mh^3$. Значення функції $f(x, y)$ з правої частини диференціального рівняння також обчислюються двічі на кожному кроці цієї модифікації.

7.5. Метод Рунге-Кутта четвертого порядку

Методи типу Ейлера мають низьку точність. Рунге запропонував метод третього, а Кутта розвинув його ідею та отримав методи четвертого порядку точності [5].

Розглянемо задачу Коші

$$\frac{du}{dt} = f(t,u),\ u(t_0) = u_0.$$

Ідея методу Рунге-Кутта заснована на застосуванні, на відміну від методу Ейлера (кусочно-лінійна ламана), кривих вищого порядку для відновлення значень шуканої функції $y(x)$ на відрізку $[a, b]$.

Найчастіше при розв'язанні практичних задач використовується метод Рунге-Кутта четвертого порядку. Згідно з цим методом значення y_i обчислюються рекурентно за формулою

$$y_{i+1} = y_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4),\ i = 0,\ldots,n-1,$$

де

$$k_1 = hf(x_i, y_i), k_2 = hf(x_i + \frac{h}{2}, y_i + \frac{k_1}{2}), k_3 = hf(x_i + \frac{h}{2}, y_i + \frac{k_2}{2}),$$
$$k_4 = hf(x_i + h, y_i + k_3).$$

Погрішність методу Рунге-Кутта 4-го порядку, що оцінюється для величини $|y_n - y(x_n)|$, має порядок $O(h^4)$. Як видно, погрішність при використанні цього методу нижча, ніж при використанні методу Ейлера і його модифікацій, але за це доводиться платити додатковими обчисленнями. А саме: функція $f(x, y)$ з правої частини диференціального рівняння обчислюється на кожному кроці методу Рунге-Кутта 4-го порядку чотири рази, що у випадках, коли обчислення значення функції $f(x, y)$ є трудомістким, може істотно уповільнити пошук розв'язку.

Контрольні запитання та завдання

1. Що називається ЗДР?

2. Сформулюйте постановку задачі Коші для ЗДР. Що є її розв'язком? У якому вигляді подається розв'язок чисельним методом?

3. Метод послідовних наближень Пікара.

4. Метод рядів Тейлора.

5. Методи типу Ейлера.

6. Похибка методу Ейлера.

7. Модифікований метод Ейлера 1-ша модифікація.

8. Модифікований метод Ейлера 2-га модифікація.

9. Поняття методів типу Рунге-Кутта.

10. Який з відомих методів розв'язання задачі Коші для ЗДР треба застосовувати в тих чи інших випадках?

11. Розв'яжіть задачу Коші для диференціального рівняння $y' = \frac{x+y-3}{x-y-1}$ на відрізку [2, 25] (початкова умова $y(2) = 1$) чисельно методами Ейлера і його модифікаціями й методом Рунге-Кутта із числом розбиття відрізка $n = 20$. Побудуйте графіки отриманих розв'язків.

8. БАГАТОКРОКОВІ МЕТОДИ РОЗВ'ЯЗАННЯ ДИФЕРЕНЦІАЛЬНИХ РІВНЯНЬ

8.1. Поняття багатокрокового методу

Розглянуті раніше чисельні методи розв'язання задачі Коші для ЗДР належать до однокрокових методів тому, що при розрахунку поточного значення y_{i+1} на i-му кроці використовується тільки інформація на останньому відрізку $[x_i, x_{i+1}]$.

Усе, що робилось на попередніх кроках методу, явно не використовується. Навпаки, розроблені багатокрокові методи розв'язання задачі Коші, які використовують те, що було отримано на попередніх кроках методу, явно. Такими є, наприклад, методи: Адамса-Бошфорда, Адамса-Мултона, Гіра-Брайтона [5, 10].

8.2. Метод Адамса-Бошфорда

В основі методу Адамса – Бошфорда лежить формула Ньютона-Лейбніца:

$$\int_a^b f(x)dx = \Phi(b) - \Phi(a),$$

де $\Phi(x)$ – будь-яка первісна для підінтегральної функції $f(x)$ на відрізку $[a, b]$.

Застосування формули Ньютона-Лейбніца на будь-якому відрізку $[x_i, x_{i+1}]$ приводить до рівняння:

$$y(x_{i+1}) - y(x_i) = \int_{x_i}^{x_{i+1}} f(x, y(x))dx.$$

Тоді, якщо вже були отримані значення $y_{i-m+1}, y_{i-m+2}, \ldots, y_{i-1}, y_i$ в m попередніх точках (вузлах) $x_{i-m+1}, x_{i-m+2}, \ldots, x_{i-1}, x_i$, то підінтегральну функцію $f(x,y(x))$ можна замінити інтерполяційним поліномом Ньютона $H_{m-1}(x)$, побудованим за цими m вузлами. Таким чином, отримуємо загальну рекурентну формулу методу Адамса-Бошфорда m-го порядку:

$$y_{i+1} - y_i = \int_{x_i}^{x_{i+1}} H_{m-1}(x)dx.$$

Оскільки $H_{m-1}(x)$ – поліном, то інтеграл від нього береться аналітично і таким чином отримують різницеву схему розв'язання

задачі Коші для будь-якого m. Наприклад, якщо $m = 1$, то $H_{m-1}(x) = f(x_i, y_i)$ і:

$$y_{i+1} = y_i + hf(x_i, y_i),$$

де $h = x_{i+1} - x_i$.

Таким чином, при $m = 1$ метод Адамса-Бошфорда співпадає з методом Ейлера.

Найчастіше застосовують метод Адамса-Бошфорда 4-го порядку, який має таку різницеву схему [10]:

$$y_{i+1} = y_i + \frac{h}{24}(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}),$$

де $f_i = f(x_i, y_i)$.

Очевидно, що за цією схемою можна проводити обчислення тільки при $i \geq 3$, тому що на i-му кроці методу треба знати значення $f_{i-3}, f_{i-2}, f_{i-1}, f_i$. Тому перш ніж застосовувати метод Адамса-Бошфорда 4-го порядку, необхідно спочатку обчислити значення f_0, f_1, f_2 для перших трьох кроків. Зазвичай для цього застосовують метод Рунге-Кутта, бо він має достатньо велику точність.

Погрішність методу Адамса-Бошфорда 4-го порядку, що оцінюється для величини $|y_n - y(x_n)|$, має порядок $O(h^4)$ [10].

8.3. Метод Адамса-Мултона

Слід зазначити, що в методі Адамса-Бошфорда інтерполяційний поліномом Ньютона $H_{m-1}(x)$ застосовується для екстраполяції функції $f(x, y(x))$ на відрізку $[x_i, x_{i+1}]$, оскільки він будується за вузлами $x_{i-m+1}, x_{i-m+2}, \ldots, x_{i-1}, x_i$. У методі Адамса-Мултона також застосовується інтерполяційний поліномом Ньютона $H_{m-1}(x)$, але не для екстраполяції, а для інтерполяції функції $f(x, y(x))$ на відрізку $[x_i, x_{i+1}]$, оскільки він будується за вузлами $x_{i-m+2}, x_{i-m+3}, \ldots, x_i, x_{i+1}$.

Тому для методу Адамса-Мултона 4-го порядку отримано таку різницеву схему [10]:

$$y_{i+1} = y_i + \frac{h}{24}(9f_{i+1} + 19f_i - 5f_{i-1} + f_{i-2}),$$

де $f_{i+1} = f(x_{i+1}, y_{i+1})$.

На відміну від попередньої схеми, в цій схемі у правій частині фігурує ще невідоме значення y_{i+1}. Такі методи носять назву неявних методів.

Тому для пошуку y_{i+1} треба розв'язати це нелінійне рівняння відносно y_{i+1}. Слід зазначити, що в загальній постановці задачі Коші при застосуванні методу Адамса-Мултона доведеться на кожному кроці розв'язувати систему нелінійних рівнянь.

Погрішність методу Адамса-Мултона 4-го порядку, що оцінюється для величини $|y_n - y(x_n)|$, має порядок $O(h^5)$ [10].

8.4. Метод прогнозу та корекції

Останню різницеву схему застосовують і для уточнення значення y_{i+1}, що було розраховане за попередньою різницевою схемою. Така комбінація методів Адамса-Бошфорда і Адамса-Мултона 4-го порядку має назву методу прогнозу та корекції [10].

У багатокроковому методі прогнозу та корекції 4-го порядку у ході розв'язання задачі Коші значення y_i у вузлах сітки x_i, $i = 1,\ldots, n$ обчислюються рекурентно за формулами:

$$y_{i+1}^{(pred)} = y_i + \frac{h}{24}(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3},$$

$$f_{i+1}^{(pred)} = f(x_{i+1}, y_{i+1}^{(pred)}),$$

$$y_{i+1} = y_i + \frac{h}{24}(9f_{i+1}^{(pred)} + 19f_i - 5f_{i-1} + f_{i-2}),$$

де $f_i = f(x_i, y_i)$.

Похибка методу прогнозу та корекції 4-го порядку, оцінювана для величини $|y_n - y(x_n)|$, має порядок $O(h^5)$ [4, 10].

Якщо порівнювати однокрокові та багатокрокові методи розв'язання задачі Коші для ЗДР, треба підкреслити, що в багатокрокових методах крок h можна обирати більшим, ніж в однокрокових методах. Це надає можливість значно зменшити кількість кроків, а отже і трудомісткість розв'язання задачі в цілому.

Контрольні запитання та завдання

1. Які чисельні методи розв'язання задачі Коші для ЗДР називаються багатокроковими методами?
2. Які особливості мають багатокрокові методи відносно однокрокових методів?
3. Наведіть загальну схему всіх методів розв'язання задачі Коші для системи ЗДР.

9. НЕЯВНІ МЕТОДИ РОЗВ'ЯЗАННЯ ЖОРСТКИХ ЗАДАЧ КОШІ

9.1. Поняття жорсткої системи диференціальних рівнянь

Поняття жорсткої задачі Коші вводиться для системи диференціальних рівнянь [5].

Для підвищення точності та адекватності математичних моделей складних об'єктів і процесів при побудові моделей доводиться враховувати велику кількість факторів та параметрів. При цьому в математичній моделі, що описується системою диференціальних рівнянь, опиняються складові з великими і малими значеннями похідних від шуканих функцій $y_i(x)$. Це і призводить до так званої жорсткої системи диференціальних рівнянь. Тут треба зазначити, що жорсткість є властивістю самої математичної задачі, а не чисельного методу її розв'язання. Застосування описаних вище явних чисельних методів для жорстких систем диференціальних рівнянь дає велику похибку у розв'язку, тому для них розроблені так звані неявні методи [10].

Розглянемо лінійну систему диференціальних рівнянь з функціями $Y \in R^m$ та постійною (незалежною від x) матрицею $A \in R^{m \times m}$:

$$Y' = A \times Y(x).$$

Нехай λ_i, $i = 1,\ldots,m$ – множина власних чисел матриці A і $K = \dfrac{\max|\mathrm{Re}(\lambda_i)|}{\min|\mathrm{Re}(\lambda_i)|}$, де $\mathrm{Re}(\lambda_i)$ – дійсна частина i-го власного числа, а операції max та min здійснюються по всіх власних числах.

Така система диференціальних рівнянь називається жорсткою, якщо [10]:

$$\mathrm{Re}(\lambda_i) < 0,\ i = 1,\ldots,m,\ K >> 1.$$

При цьому число K називається числом жорсткості лінійної системи.

Це поняття жорсткої системи узагальнюється і на нелінійну систему:

$$Y' = F(x,Y).$$

Роль матриці A при цьому відіграє матриця Якобі:

$$A(x) = F'_Y(x,Y) = \begin{pmatrix} \dfrac{\partial f_1(x,Y)}{\partial y_1} & \cdots & \dfrac{\partial f_1(x,Y)}{\partial y_m} \\ \cdots & \cdots & \cdots \\ \dfrac{\partial f_m(x,Y)}{\partial y_1} & \cdots & \dfrac{\partial f_m(x,Y)}{\partial y_m} \end{pmatrix},$$

яка в загальному випадку вже буде залежати від x, тобто змінюватись у часі буде і число жорсткості системи.

Як уже згадувалося раніше, застосування явних чисельних методів для жорстких систем диференціальних рівнянь дає велику похибку у розв'язку, якщо крок обчислень не є досить малим, а саме він повинен задовольняти обмеженню [10]:

$$h < \frac{c}{|\lambda_{\max}|},$$

де c – константа, яка залежить від умов задачі Коші;

λ_{max} – максимальне за модулем власне число матриці Якобі.

9.2. Неявні методи Ейлера і Рунге-Кутта

Поняття неявного методу було введено раніше. Це методи, що застосовують схему (наприклад, як у схемі методу Адамса-Мултона 4-го порядку), в якій у лівій і правій частині фігурує ще невідоме значення Y_{i+1}. Розглянемо деякі неявні схеми методів, що базуються на явних схемах методів, які вивчалися раніше.

Неявний метод Ейлера. Явна схема методу Ейлера має вид:

$$y_{i+1} = y_i + h \times f(x_i, y_i),\ i = 0,\ldots,n\text{-}1.$$

Раніше було зазначено, що цей метод Ейлера співпадає з методом Адамса-Бошфорда 1-го порядку. У той же час метод Адамса-Мултона дає алгоритм переходу від явного методу до неявного шляхом заміни виду інтерполяційного поліному Ньютона. Так, при $m = 1$, якщо побудувати інтерполяційний поліном Ньютона за вузлом x_{i+1}, то він буде мати вигляд $H_{m-1}(x) = F(x_{i+1}, Y_{i+1})$. Таким чином, неважко отримати схему уже неявного методу Ейлера:

$$Y_{i+1} = Y_i + h \times f(x_{i+1}, Y_{i+1}).$$

Але тепер бачимо, що для визначення Y_{i+1} треба спочатку розв'язувати систему нелінійних рівнянь. Система ж нелінійна, оскільки в загальному випадку векторна функція $F(x,Y)$ нелінійна відносно Y. Розв'язати таку систему нелінійних рівнянь можна, наприклад, методом Ньютона.

Неявний метод Рунге-Кутта. У неявному методі Рунге-Кутта 4-го порядку для розв'язання жорстких систем звичайних диференціальних рівнянь значення Y_i у вузлах сітки x_i, $i = 1,\ldots,n$, рекуррентно обчислюються як розв'язання системи рівнянь [10]:

$$Y_{i+1} = Y_i + \frac{1}{6}(K_1^* + 2K_2^* + 2K_3^* + K_4^*),\ i = 1,\ldots,n,$$

$$\text{де } K_1^* = hF(x_{i+1}, Y_{i+1}),\ K_2^* = hF(x_{i+1} - \frac{h}{2}, Y_{i+1} - \frac{h}{2}K_1^*),$$

$$K_3^* = hF(x_{i+1} - \frac{h}{2}, Y_{i+1} - \frac{h}{2}K_2^*),\ K_4^* = hF(x_{i+1}, Y_{i+1} - hK_3^*).$$

Глобальна похибка неявного методу Рунге-Кутта 4-го порядку, оцінювана для величини $\| Y_n - Y(x_n) \|$, має порядок $O(h^4)$, а локальна – $O(h^5)$ [10].

Контрольні запитання та завдання

1. У чому полягає суть жорсткої задачі Коші для системи ЗДР?

2. Які методи розв'язання задачі Коші для ЗДР називаються неявними методами?

3. Які особливості мають неявні методи відносно явних методів?

4. Сформулюйте постановку задачі Коші для системи ЗДР. Що є її розв'язком? У якому вигляді подається розв'язок чисельним методом?

5. Який з відомих вам методів розв'язання задачі Коші для системи ЗДР треба застосовувати в тих чи інших випадках?

10. КРАЙОВІ ЗАДАЧІ ДЛЯ ЗВИЧАЙНИХ ДИФЕРЕНЦІАЛЬНИХ РІВНЯНЬ

10.1. Постановка крайової задачі

Крайова задача – задача теорії диференціальних рівнянь, в якій граничні умови задаються в різних точках. Наприклад, при коливаннях струни із закріпленеми кінцями зміщення на кожному з кінців дорівнює нулю.

Крайові задачі важче розв'язувати, ніж задачі Коші, особливо чисельно. Вони виникають як у теорії ЗДР, так і в теорії диференційних рівнянь із частковими похідними, особливо рівнянь еліптичного типу [5].

Особливий вид крайової задачі – вимога певної поведінки функції (скінченності) при прямуванні аргументу до нескінченності або в околі особливих точок.

У загальному вигляді **крайова задача** для ЗДР 2-го порядку виду:

$$F(x,y,y',y'') = 0$$

полягає в такому: знайти функцію $y = y(x)$ на заданому відрізку $[a,b]$, що задовольняє даному рівнянню і крайовим умовам:

$$\begin{cases} \Phi_1(y(a), y'(a)) = 0, \\ \Phi_2(y(b), y'(b)) = 0, \end{cases}$$

де F, Φ_1, Φ_2 – задані неперервні функції відповідної кількості аргументів.

Крайова задача називається **лінійною**, якщо всі функції F, Φ_1, Φ_2 лінійні відносно y, y', y''.

Розглянемо звичайне диференціальне рівняння другого порядку:

$$\frac{d^2y}{dx^2} = f(x, y, y')$$

при граничних умовах $y(a) = A$, $y(b) = B$.

Рівняння більш високих порядків можна розв'язувати тими ж методами, які поділимо на дві групи:

1. Методи, основані на заміні розв'язку крайової задачі розв'язком декількох задач Коші.

2. Методи, в яких використовується кінцево-різницева форма диференціального рівняння.

10.2. Методи стрільби

Якщо ЗДР 2-го порядку є лінійним, тобто має вигляд

$y'' = f_1(x)y' + f_2(x)y + f_3(x)$

при додаткових умовах $y(a) = A$, $y(b) = B$, то крайову задачу можна звести до задачі Коші за допомогою початкових умов $y(a) = A_1$, $y'(a) = \alpha_1$.

Знайшовши розв'язок $y_1(x)$, можна поставити другі граничні умови:

$$y(a) = A_1,\ y'(a) = \alpha_2.$$

і отримати другий розв'язок $y_2(x)$.

Якщо $y(b) = \beta_1$, а $y'(b) = \beta_2$, причому $\beta_1 \neq \beta_2$, то розв'язок

$$y(x) = \frac{1}{\beta_1 - \beta_2}((B - \beta_2)y_1(x) + (\beta_1 - B)y_2(x))$$

задовольняє обом початковим граничним умовам.

Якщо розв'язується нелінійне ЗДР, то розв'язок крайової задачі можна звести до розв'язку декількох задач Коші, послідовно вводячи в початкові умови значення α:

$$y(a) = A,\ y'(a) = \alpha$$

і прагнучи найти розв'язок, який задовольняє умову $y(b) = B$. При цьому може допомогти інтерполяція, що дозволяє побудувати впорядковану послідовність α і звести до мінімуму об'єм обчислень.

10.3. Кінцево-різницеві методи

Перевага кінцево-різницевих методів у тому, що вони дозволяють звести розв'язок крайової задачі до розв'язку системи алгебраїчних рівнянь. При розв'язку двоточкової крайової задачі

$$y'' = f(x,y,y'),$$

при $y(a) = A$ і $y(b) = B$ інтервал $[a, b]$ можна розділити на n рівних частин:

$$x_i = x_0 + ih,\ i = 1, 2,\ldots, n,$$

де $x_0 = a$, $x_n = b$, а $h = (b - a)/n$. У точках x_i, що називаються вузлами, намагаються найти значення розв'язку y_i. Знаючи координати вузлів і користуючись кінцево-різницевими виразами для похідних

$$y'(x_i) = \frac{1}{2h}(y_{i+1} - y_{i-1}),$$

$$y''(x_i) = \frac{1}{h^2}(y_{i+1} - 2y_i + y_{i-1}),$$

можна представити диференціальне рівняння у вигляді різницевого рівняння:

а) якщо початкове ЗДР лінійне, то задача буде складатися з розв'язання системи лінійних алгебраїчних рівнянь;

б) якщо початкове ЗДР нелінійне, то задача зводиться до розв'язування системи нелінійних алгебраїчних або трансцендентних рівнянь;

в) методи, алгоритми та програми розв'язування таких лінійних та нелінійних рівнянь відомі, але звести розв'язування крайової задачі методом кінцевих різниць до стандартної програми важко, оскільки формулювання кожної задачі залежить від вигляду рівняння, що розглядається.

Приклад. Нехай потрібно вирішити диференціальне рівняння $y'' = 2x + y$ при умовах $y(0) = 0$ і $y(1) = 1$ і кроці $h = 0.2$. У різницевій формі це рівняння має вид:

$$y''(x_i) = \frac{1}{0.04}(y_{i+1} - 2y_i + y_{i-1}) = 2x_i + y_i.$$

Використовуючи цю формулу і граничні умови, можна виписати наступну систему чотирьох лінійних рівнянь з чотирма невідомими:

$$\begin{cases} -2.12y_1 + y_2 = 0.016, \\ y_1 - 2.12y_2 + y_3 = 0.032, \\ y_2 - 2.12y_3 + y_4 = -0.064, \\ -2.12y_4 + y_5 = 0.936. \end{cases}$$

$$y(x) = \frac{5}{3}\frac{sh\sqrt{3}x}{sh\sqrt{3}} - \frac{2}{3}.$$

10.4. Метод прогонки

Якщо ЗДР 2-го порядку розв'язати методом скінченних різниць, то початкове ЗДР зводиться до системи лінійних алгебраїчних рівнянь з тридіагональною матрицею коефіцієнтів, тобто такою, що кожне рівняння системи має три сусідні невідомі:

$$\begin{cases} a_{11}y_0 + a_{12}y_1 + 0\cdot y_2 + \ldots + 0\cdot y_n = f_1, \\ a_{21}y_0 + a_{22}y_1 + a_{23}y_2 + 0\cdot y_3 + \ldots + 0\cdot y_n = f_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ 0\cdot y_0 + 0\cdot y_1 + 0\cdot y_2 + \ldots + a_{n-2}y_{n-2} + a_{n-1}y_{n-1} + a_{nn}y_n = f_n. \end{cases}$$

Для розв'язання такої системи розроблено спеціальний метод – **прогонки**. Розглянемо лінійне диференціальне рівняння другого порядку:

$$y'' + p(x)y' + q(x)y = f(x)$$

і лінійні крайові двоточкові умови:

$$\begin{cases} \alpha_0 y(a) + \alpha_1 y'(a) = A, \\ \beta_0 y(b) + \beta_1 y'(b) = B, \end{cases}$$

де $p(x)$, $q(x)$, $f(x)$ – задані на $[a,b]$ неперервні функції від x;

$\alpha_0, \alpha_1, \beta_0, \beta_1, A, B$ – задані константи, причому $|\alpha_0| + |\alpha_1 \neq 0|$, $|\beta_0| + |\beta_1| \neq 0$.

Прийдемо до скінченнорізницевого рівняння:

$$h = (B - A)/N;\ x_i = x_0 + ih;\ x_0 = A;\ x_n = B;$$

$$q(x_i) = q_i;\ p(x_i) = p_i;\ f(x_i) = f_i;\ y(x_i) = y_i.$$

Підставляючи цей вираз у початкове ЗДР, отримаємо для внутрішніх точок x_i ($i = 1,2,\ldots,n$-1) систему скінченнорізницевих рівнянь:

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + p_i \frac{y_{i+1} - y_{i-1}}{2h} + q_i y_i = f_i.$$

Після деяких перетворень матимемо

$$y_{i+1} + m_i y_i + n_i y_{i-1} = F_i h^2,\ i = 1,2,\ldots,n\text{-}1,$$

де $m_i = \dfrac{2 - q_i h^2}{1 + \frac{p_i}{2}h}$, $n_i = \dfrac{1 - \frac{p_i}{2}h}{1 + \frac{p_i}{2}h}$, $F_i = \dfrac{f_i}{1 + \frac{p_i}{2}h}$.

Для похідних на кінцях відрізка інтегрування $x_0 = A$ і $x_n = B$ знаходимо скінченнорізницеві рівняння виду:

$$y_0' = (y_1 - y_0)/h,\ y_n' = (y_{n-1} - y_n)/(-h)$$

і, підставляючи їх у граничні умови, отримаємо два рівняння:

$$\alpha_0 y_0 + \frac{\alpha_1}{h}(y_1 - y_0) = A,$$

$$\beta_0 y_n + \frac{\beta_1}{-h}(y_{n-1} - y_n) = B.$$

Система $N+1$ рівнянь відносно невідомих $y_0, y_1, \ldots, y_n$ являє собою систему лінійних алгебраїчних рівнянь з тридіагональною матрицею коефіцієнтів. Цю систему зручно розв'язати методом прогонки.

Метод прогонки достатньо простий, легко алгоритмізується та включає прямий та зворотній хід. Причому:

1) прямий хід полягає в знаходженні коефіцієнтів за рекурентними формулами;

2) зворотній хід полягає у знаходженні розв'язку початкового ЗДР $y_{n-1}, y_{n-2}, \ldots, y_1$.

Контрольні запитання та завдання

1. Що називається ЗДР?

2. Сформулюйте постановку крайової задачі для ЗДР. Що є її розв'язком? У якому вигляді подається розв'язок чисельним методом?

3. Метод стрільби.

4. Метод прогонки.

5. У чому суть методу кінцевих різниць для лінійної крайової задачі? Які він має оцінки погрішності?

6. Розв'яжіть методом кінцевих різниць крайову задачу $y'' + x^2y' + (1 - x)y = x/(x^2 + 3)$, $y(0) = 0$, $y(1) = 0$, з числом розбиття відрізка $n = 25$. Побудуйте графік отриманого розв'язку.

7. Розв'яжіть методом кінцевих різниць нелінійне диференціальне рівняння другого порядку $y'' - 2xy' - 2y^2 = 0$, $y(0) = 0$, $y(1) = 3.7$ з числом розбиття відрізка $n = 20$. Побудуйте графік отриманого розв'язку.

11. РОЗВ'ЯЗАННЯ ДИФЕРЕНЦІАЛЬНИХ РІВНЯНЬ З ЧАСТИННИМИ ПОХІДНИМИ

11.1. Поняття задач математичної фізики

Інженеру часто доводиться стикатися із задачами, в яких шукана величина залежить від декількох змінних, наприклад, $u = u(t,x,y,z)$. У цьому випадку розв'язувані диференціальні рівняння містять частинні похідні і називаються диференціальними рівняннями з частинними похідними [2, 6, 10, 12, 13].

Рівняннями з частинними похідними описується багато фізичних процесів у таких областях, як механіка суцільних середовищ, термодинаміка, квантова механіка, електродинаміка, теорія пружності та ін. Тому розділ математики, що вивчає можливість розв'язання диференціальних рівнянь із частинними похідними, називається математичною фізикою, а рівняння – рівняннями математичної фізики.

На жаль, багато з таких рівнянь не мають аналітичного розв'язку, і щоб їх розв'язати, доводиться вдаватися до чисельних методів. Якщо для розв'язання звичайних диференціальних рівнянь існує багато різних методів, то для розв'язання диференціальних рівнянь із частинними похідними доводиться вибирати в основному між методом кінцевих різниць і методом кінцевих елементів.

У цьому розділі питання про чисельне інтегрування диференціальних рівнянь із частинними похідними розглядається з точки зору застосування цих методів для вирішення різних технічних завдань.

Дасться також класифікація диференціальних рівнянь з частинними похідними, що часто зустрічаються, і вказуються раціональні шляхи їх чисельного розв'язання.

11.2. Класифікація диференціальних рівнянь з частинними похідними

Диференціальні рівняння з частинними похідними класифікують або залежно від їх математичної природи (еліптичні, параболічні тощо), або від фізичного змісту розв'язуваних за їх допо-

могою задач (рівняння дифузії, хвильове рівняння тощо). Для того, щоб користуватися математичною літературою та літературою з прикладних дисциплін, фахівець повинен бути знайомий з обома цими класифікаціями [13].

Для спрощення викладання слід розглянути випадок, коли шукана величина залежить від двох змінних, тобто $u = u(x,y)$.

Лінійні рівняння в частинних похідних у загальному вигляді записуються у вигляді:

$$A(x,y)\frac{\partial^2 u}{\partial x^2} + B(x,y)\frac{\partial^2 u}{\partial x \partial y} + C(x,y)\frac{\partial^2 u}{\partial y^2} + a(x,y)\frac{\partial u}{\partial x} +$$

$$+ b(x,y)\frac{\partial u}{\partial y} + c(x,y)u = F(x,y),$$

де A, B, C, a, b, c – задані неперервні функції двох змінних, що мають неперервні частинні похідні, u – шукана функція. Для скорочення запису введемо позначення:

$$u_{xx} = \frac{\partial u}{\partial x^2};\quad u_{xy} = \frac{\partial^2 u}{\partial x \partial y};\quad u_{yy} = \frac{\partial^2 u}{\partial y^2};\quad u_x = \frac{\partial u}{\partial x};\quad u_y = \frac{\partial u}{\partial y}.$$

Тоді наведена загальна форма отримає спрощену форму запису вигляду

$$A(x,y)u_{xx} + B(x,y)u_{xy} + C(x,y)u_{yy} + a(x,y)u_x +$$

$$+ b(x,y)u_y + c(x,y)u = F(x,y)$$

і розглянемо окремий випадок, коли $a = b = c = F \equiv 0$, тобто

$$A(x,y)u_{xx} + B(x,y)u_{xy} + C(x,y)u_{yy} = 0$$

Приклади рівнянь з частинними похідними:

$u_t = u_{xx}$ (одновимірне рівняння теплопровідності);

$u_t = u_{xx} + u_{yy}$ (двовимірне рівняння теплопровідності);

$u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta} = 0$ (рівняння Лапласа в полярних координатах);

$u_{tt} = u_{xx} + u_{yy} + u_{zz}$ (тривимірне хвилеве рівняння);

$u_{tt} = u_{xx} + \alpha u_t + \beta u$ (телеграфне рівняння).

Зазначимо, що в наведених раніше прикладах невідома функція u залежить більше ніж від однієї змінної. Змінні t, x, по яких

проводиться диференціювання, називаються незалежними змінними.

Методи розв'язання рівнянь у частинних похідних:

- метод розділення змінних;
- метод інтегральних перетворень;
- метод перетворення координат;
- чисельні методи;
- метод теорії обурень;
- метод функцій Гріна;
- метод інтегральних рівнянь;
- варіаційні методи;
- метод розкладання по власних функціях.

Для кожного класу рівнянь існує своя загальна теорія і методи розв'язання. Рівняння в частинних похідних можна класифікувати за багатьма ознаками:

1. За порядком рівняння (порядком рівняння називають найвищий порядок частинних похідних, що входять у рівняння):

- $u_t = u_x$ (рівняння першого порядку);
- $u_t = u_{xx}$ (рівняння другого порядку);
- $u_t = uu_{xxx} + \sin x$ (рівняння третього порядку).

2. За числом змінних (числом змінних називають число незалежних змінних).

- $u_t = u_{xx}$ (рівняння з двома змінними);
- $u_t = u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta}$ (рівняння з трьома змінними r, θ, t).

3. За критерієм лінійне/нелінійне.

Лінійним рівнянням другого порядку з двома незалежними змінними називається рівняння:

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = G,$$

де A, B, C, D, E, F і G константи або задані функції змінних x і y.

- $u_{xx} + yu_{yy} = 0, u_{tt} = e^{-t}u_{xx} + \sin t$ (лінійні рівняння);
- $uu_{xx} + u_t = 0, xu_x + yu_y + u^2 = 0$ (нелінійні рівняння).

4. За критерієм однорідне/неоднорідне.

Рівняння називається однорідним, якщо права частина $G(x,y)$ тотожно дорівнює нулю для всіх x і y. Якщо $G(x,y)$ не дорівнює нулю тотожно, то рівняння називається неоднорідним.

5. За виглядом коефіцієнтів.

Якщо коефіцієнти A, B, C, D, E, F і G рівняння – константи, то рівняння називається рівнянням з постійними коефіцієнтами, інакше рівнянням із змінними коефіцієнтами.

З математичної точки зору диференціальні рівняння другого порядку з частинними похідними з двома незалежними змінними: класифікуються залежно від співвідношення коефіцієнтів A, B, C. Вони можуть залежати від змінних x і y. Залежність функцій A, B, C від x і y ускладнює ситуацію, оскільки робить можливим зміну типу рівняння при переході з однієї частини розглянутої області в іншу.

Виділяють також такі типи лінійних рівнянь.

Параболічний тип. Рівняння параболічного типу описують процеси теплопровідності і дифузії і визначаються умовою:

$$B^2 - 4AC = 0.$$

Гіперболічний тип. Рівняння гіперболічного типу описують коливальні системи і хвилеві рухи і визначаються умовою:

$$B^2 - 4AC > 0.$$

Еліптичний тип. Рівняння еліптичного типу описують сталі процеси і визначаються умовою:

$$B^2 - 4AC < 0.$$

Приклади лінійних рівнянь різних типів:

$u_t = u_{xx}$, $B^2 - 4AC = 0$ (параболічне);

$u_{tt} = u_{xx}$, $B^2 - 4AC = 4$ (гіперболічне);

$u_{xy} = 0$, $B^2 - 4AC = 1$ (гіперболічне);

$u_{xx} + u_{yy} = 0$, $B^2 - 4AC = -4$ (еліптичне);

$yu_{xx} + u_{yy} = 0$, $B^2 - 4AC = -4y$ (еліптичне при $y > 0$, параболічне при $y = 0$, гіперболічне при $y < 0$).

Еліптичні рівняння описують усталені (стаціонарні) процеси; задача ставиться в замкнутій області, і в кожній точці межі цієї області задаються граничні умови.

Параболічними і гіперболічними рівняннями описуються еволюційні (нестаціонарні) процеси (процеси «поширення»). У таких задачах на одній частині границі ставляться граничні умови, на іншій – початкові; можливі також відкриті області, в які «поширюється розв'язок».

Додатковими умовами для диференціальних рівнянь другого порядку з частинними похідними можуть слугувати граничні або початкові умови, а також комбінація тих та інших.

Серед чисельних методів розв'язку диференціальних рівнянь з частинними похідними слід виділити метод кінцевих різниць та метод кінцевих елементів.

11.3. Метод кінцевих різниць

В основі розв'язання рівнянь з частинними похідними методом кінцевих різниць лежить [5], природно, – різницева апроксимація похідних, яка багато в чому нагадує апроксимацію похідних, присвячених розв'язанню крайової задачі для звичайних диференціальних рівнянь. Розв'язання здійснюється в три етапи. Спочатку в області розв'язку вводять рівномірну сітку «вузлових точок», що відповідає характеру задачі і граничним умовам. Потім розв'язуване рівняння з частинними похідними записують у найбільш зручній системі координат і, представляючи похідні в кінцеворізницевій формі, приводять його до виду різницевого рівняння.

Отримане різницеве рівняння використовують надалі для опису функціонального зв'язку між сусідніми вузлами сітки. Різницеве рівняння записують для всіх вузлів сітки і отримують у результаті систему n алгебраїчних рівнянь з n невідомими. На останньому етапі отриману систему алгебраїчних рівнянь розв'язують одним із чисельних методів.

На перший погляд, ця процедура, яка складається з трьох етапів, може здатися простою і такою, що прямо приводить до розв'язку, однак насправді це не так – широке розмаїття типів і

розмірів сіток, видів рівнянь із частинними похідними, можливих кінцево-різницевих апроксимацій цих рівнянь і методів розв'язання отриманих систем алгебраїчних рівнянь роблять задачу чисельного розв'язання рівнянь із частинними похідними виключно багатогранним і цікавим дослідженням.

Варто розглянути тепер усі три етапи рішення детальніше.

Етап 1. Сітки, що застосовуються при поданні диференціальних рівнянь з частинними похідними в кінцево-різницевій формі.

Усі раніше наведені рівняння з частинними похідними були записані в декартовій системі координат, однак іноді буває зручніше користуватися іншими системами координат, що володіють спеціальними геометричними властивостями і враховують форму фізичного тіла [13].

Найчастіше в інженерній практиці застосовується декартова, циліндрична та сферична системи координат.

Конфігурація сітки має відповідати характеру задачі і граничним умовам використання диференціального рівняння з частинними похідними для отримання різницевого виразу, що описує функціональні зв'язки між сусідніми вузлами сітки розв'язання отриманої системи n рівнянь з n невідомими з метою отримання наближеного розв'язку у вузлах сітки, які найчастіше застосовують при розв'язанні рівнянь з частинними похідними – прямокутна, полярна, трикутна та скошена. Нерідко доводиться мати справу з областями неправильної форми.

Етап 2. Подання частинних похідних у кінцево-різницевому вигляді.

Для подання частинних похідних у кінцево-різницевому вигляді застосовують формули, аналогічні формулам чисельного диференціювання. Вони тільки переписуються для випадку декількох змінних. Для двох змінних на практиці найчастіше застосовують симетричні формули:

$$\frac{\partial u}{\partial x} \approx \frac{u_{i+1,j} - u_{i-1,j}}{2h},$$

$$\frac{\partial u}{\partial y} \approx \frac{u_{i+1,j} - u_{i-1,j}}{2h},$$

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2},$$

$$\frac{\partial^2 u}{\partial y^2} \approx \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2},$$

$$\frac{\partial^2 u}{\partial x \partial y} \approx \frac{u_{i+1,j+1} - u_{i-1,j+1} - u_{i+1,j-1} + u_{i-1,j-1}}{4h^2}.$$

Тут $u_{k,m}$ – значення функції $u = (x,y)$ у вузлах, розташованих в околі центральної точки (x_i, y_j), якій відповідає значення $u_{i,j}$.

Етап 3. Чисельні методи розв'язання систем лінійних алгебраїчних рівнянь.

Для розв'язання систем лінійних алгебраїчних рівнянь на 3-му етапі методу кінцевих різниць застосовують чисельні методи, що були описані в розділах 2 та 3. Зазвичай матриці такої системи виявляються розрідженими, оскільки в більшій частині розрахункових схем застосовуються лише сусідні вузли, а не всі вузли сітки.

11.4. Метод кінцевих елементів

Метод кінцевих елементів для опису суцільних середовищ уперше було застосовано в середині 50-х років XX століття. Він широко застосовується в гідродинаміці, теорії поля, при розрахунку складних напружених станів та в інших областях науки і техніки. Хоча метод кінцевих елементів застосовується для розв'язання тих же самих задач, що і метод кінцевих різниць, але базуються вони на різних ідеях. У методі кінцевих різниць виконується різницева апроксимація похідних, що входять у диференціальне рівняння. У той час як у методі кінцевих елементів фізична задача заміноється кусочно-гладкою моделлю, приблизно так, як це робиться при інтерполяції функції сплайнами тільки в об'ємному вигляді.

Для повного розуміння методу кінцевих елементів необхідні досить глибокі знання вищої математики (зокрема, знання варіаційного зчислення), тому обмежимось тільки коротким описом його основ. Щоб було більш зрозуміло, слід розглянути метод

кінцевих елементів на прикладі дослідження напруженого стану тіла [5,13].

Основні етапи застосування методу кінцевих елементів наступні:

1. Розділення тіла на елементи.
2. Вибір схеми інтерполяції переміщень у середині елементів.
3. Визначення залежності між напругою і деформацією у вузлах.
4. Виведення рівнянь для системи у цілому.
5. Розв'язання системи лінійних рівнянь.
6. Обчислення значень інших величин.

Перший етап полягає в поділі тіла на малі елементи простої форми, що стикаються в точках, які називаються вузлами. Поділ на елементи можна виконати багатьма різними способами, оскільки вибір розмірів, форми й орієнтації елементів цілком визначається уявленнями інженера про те, як простіше розв'язати дану задачу. Елементи плоского тіла мають зазвичай трикутну (у цьому випадку цей процес називається тріангуляцією – розбиття геометричного об'єкта на симплекси) або чотирикутну форму, а елементи тривимірних тіл – форму тетраедрів або гексаедрів. Ті ділянки тіла, для яких із фізичних міркувань потрібно отримати більш детальну інформацію, розбиваються на більшу кількість дрібних елементів. Якщо фізичні властивості тіла змінюються в точці або вздовж лінії, то можна змінювати форму, розміри або орієнтацію елементів на цій ділянці тіла.

Другий етап застосування методу скінченних елементів полягає у виборі будь-якої простої схеми інтерполяції, що дозволяє виразити переміщення в будь-якій точці всередині елемента через його значення у вузлах. Зазвичай переміщення задається будь-яким простим поліномом. У межах кожного елемента для інтерполяції значень переміщення використовуються поліноми з коефіцієнтами, обумовленими у процесі розв'язання.

На третьому етапі виписуються залежності між напруженнями і деформаціями у вузлах усіх елементів. На цій стадії з великою користю може бути використана концепція матриць і коефіцієнтів впливу. Знаючи співвідношення між напругою і

деформаціями елементів, можна побудувати такі ж співвідношення для системи в цілому. При цьому деформації дотичних елементів повинні бути рівні, а сили, що діють у вузлах, складати в сумі зовнішню силу, прикладену в тій же точці. У результаті отримується система лінійних алгебраїчних рівнянь виду:

$$Ad = b,$$

де A – відома матриця жорсткості системи;

d – шуканий вектор переміщень системи;

b – відомий вектор навантаження.

Отримана система рівнянь містить досить багато нульових елементів, оскільки не кожен вузол належить кожному елементу. У разі довільної деформації кожен з m вузлів може мати n незалежних переміщень (наприклад, по x і y в плоскому випадку). Тому матриця жорсткості A буде мати розмірність $(nm)\times(nm)$, а вектори деформації і сили – розмірність (nm). Деякі значення переміщень безпосередньо визначаються граничними умовами. Відомі значення переміщень можна виключити із системи рівнянь і тим самим знизити її порядок. Оскільки в цьому випадку виходить розріджена система лінійних алгебраїчних рівнянь великої розмірності, для її розв'язання зручно користуватися методом Гаусса-Зейделя. У результаті отримують значення переміщень для всіх вузлів. Отримавши розподіл переміщень, можна за допомогою звичайних рівнянь теорії пружності знайти розподіл напружень і деформацій.

11.5. Розв'язання параболічних рівнянь

Як приклад рівняння параболічного типу, зупинимося на рівнянні теплопровідності для однорідного стрижня завдовжки $0 \le x \le l$:

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2},$$

де $u = u(x,t)$ – температура, x – координата, t – час, a – коефіцієнт теплопровідності. Припускатимемо, що $a=1$. Тобто отримаємо рівняння

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}.$$

Нехай заданий розподіл температури $u(x,0) = \varphi_1(x)$ у початковий момент часу $t = 0$ і закони зміни температури залежно від часу на кінцях стрижня $x = 0$ і $x = l$: $u(0,t) = \varphi_2(t)$; $u(l,t) = \varphi_3(t)$. Потрібно знайти розподіл температури $u = u(x,t)$ уздовж стрижня завдовжки $0 \le x \le l$ у будь-який момент часу t. Функція $u(x,t)$ має бути неперервна і двічі неперервно диференційована по своїх змінних в області $\Omega = \{t, x : t \in [0,T], x \in [0,l]\}$.

Неперервну область Ω замінимо сітковою, розбиваючи її за допомогою кроку h по x і за допомогою кроку k по t. У результаті заміни неперервної області Ω дискретною множиною вузлових точок $\Omega_h = \{x_i, t_j\}, i = \overline{0,n}; j = \overline{0,m}$, вихідна задача змінюється. Тепер шукатимемо рішення $u(x,t)$ тільки на дискретній множині Ω. Тобто $u_{ij} = u(x_i, t_j)$ – двовимірна таблиця значень шуканої функції у вузлових точках.

Представимо основне рівняння у скінченнорізницевій формі, замінюючи $\frac{\partial u}{\partial t}$ і $\frac{\partial^2 u}{\partial x^2}$ скінченнорізницевим аналогом у вузлових точках x_i, t_j:

$$\frac{\partial u(x_i,t_j)}{\partial t} = \frac{u(x_i,t_j+k) - u(x_i,t_j-k)}{2k} = \frac{u_{ij+1} - u_{ij-1}}{2k}$$

$$\frac{\partial^2 u(x_i,t_j)}{\partial x^2} = \frac{u(x_i+h,t_j) - 2u(x_i,t_j) + u(x_i-h,t_j)}{h^2} = \frac{u_{i+1j} - 2u_{ij} + u_{i-1j}}{h^2}$$

Отримаємо скінченнорізницевий аналог початкового завдання: потрібно знайти значення функції $u(x_i,t_j)$, що задовольняє скінченнорізницевому рівнянню вигляду:

$$\frac{u_{ij+1} - u_{ij-1}}{2k} = \frac{u_{i+1j} - 2u_{ij} + u_{i-1j}}{h^2}, \quad i = \overline{1,n-1}; j = \overline{1,m-1}.$$

і додатковим умовам:

$$u(x_i,0) = \varphi_1(x_i), i = \overline{0,n}, u(0,t_j) = \varphi_2(t_j), j = \overline{0,m}, u(l,t_j) = \varphi_3(t_j), j = \overline{0,m}$$

Отримаємо систему лінійних рівнянь алгебри, яку можна розв'язати будь-яким відомим методом. Дослідження показали,

що значення h і k мають бути зв'язані між собою таким чином: $k=\delta h^2$, де $\delta = Const$, $0<\delta<1/2$. Апроксимуємо основне рівняння скінченнорізницевим

$$\frac{u_{ij+1}-u_{ij-1}}{2\delta h^2}=\frac{2\delta u_{i+1j}-4\delta u_{ij}+2\delta u_{i-1j}}{2\delta h^2}.$$

Розв'язуючи таку систему з урахуванням додаткових умов, отримаємо $u(x_i,t_j)$ – шукану функцію в точках $x_i,t_j, i=\overline{0,n}; j=\overline{0,m}$.

Другий варіант скінченнорізницевого аналога основного диференціального рівняння, це т.з. явна схема, виходить за рахунок того, що перші похідні $\frac{\partial u}{\partial t}$ в узлових точках x_i,t_j представлені у вигляді:

$$\frac{\partial u(x_i,t_j)}{\partial t}=\frac{u(x_i,t_j+h)-u(x_i,t_j)}{k},$$

а друга похідна залишається попередньою. Отримаємо двошарову схему в скінченнорізницевій формі:

$$\frac{u_{ij+1}-u_{ij}}{k}=\frac{u_{i+1j}-2u_{ij}+u_{i-1j}}{h^2}.$$

Покладаючи $k=\delta h^2$, отримуємо $u_{ij+1}-u_{ij}=\delta u_{i+1j}-2\delta u_{ij}+\delta u_{i-1j}$, або $u_{ij+1}=\delta u_{i+1j}+(1-2\delta)u_{ij}+\delta u_{i-1j}$, $i=\overline{1,n-1}; j=\overline{0,m-1}$. По цій формулі для кожного значення u_{ij+1} для шара $(j+1)$ по осі t використовуються три значення u_{ij},u_{i-1j},u_{i+1j} на попередньому шарі з номером j. Для початку обчислень використовуємо додаткові умови.

У результаті розв'язання задачі в скінченнорізницевій формі набуваємо значень шуканої функції в точках x_i,t_j, які є наближеним розв'язком вихідної задачі. На практиці вважають $\delta=1/6$, тоді розрахункова формула спрощується і набирає такого вигляду:

$$u_{ij+1}=\frac{1}{6}\left(u_{i+1j}+4u_{ij}+u_{i-1j}\right), i=\overline{1,n-1}; j=\overline{0,m}.$$

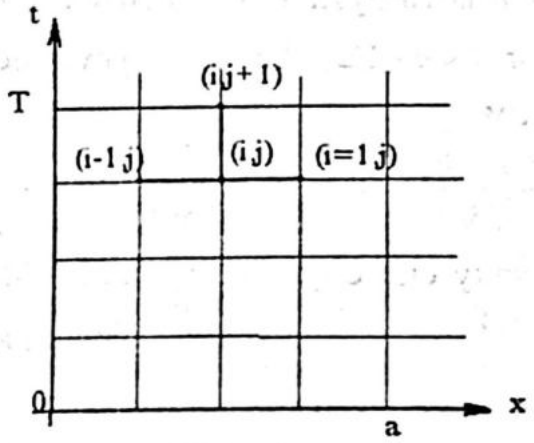


Рис. 11.1. Явна схема

Така розрахункова формула дає якнайкраще наближення до шуканого розв'язку, забезпечуючи стійкість скінченнорізницевої схеми і якнайкращу апроксимацію диференціального рівняння скінченнорізницевим.

У разі неявної схеми (рис. 11.2) використовується інший вид апроксимації і нове співвідношення між кроками h і k у вигляді $h^2 = kS$.

Вихідне диференціальне рівняння апроксимується скінченнорізницевим рівнянням вигляду:

$$\frac{u_{ij+1} - u_{ij}}{h^2/S} = \frac{u_{i+1j+1} - 2u_{ij+1} - u_{i-1j+1}}{h^2}.$$

Початкові і граничні умови залишаються тими ж, що у попередньому випадку. Для розв'язання цієї системи лінійних рівнянь застосовується метод прогону.

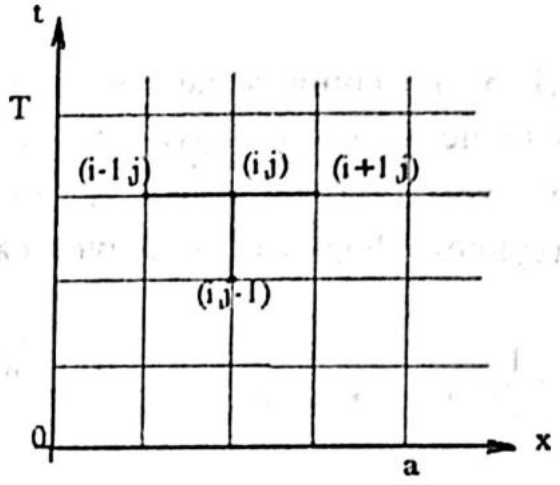


Рис. 11.2. Неявна схема

Суть методу прогону полягає в тому, що спочатку обчислюються значення $u_{i0} = \varphi_{1i}$, вибирається значення S з метою отримання необхідної швидкості просування осі t.

Позначимо $u_{ij} = u(x_i, t_j)$, $\varphi_1(x_i) = \varphi_{1i}$, $\varphi_2(t_j) = \varphi_{2j}$, $\varphi_3(t_j) = \varphi_{3j}$. У прямому ході на черговому $(j+1)$ тимчасовому шарі обчислюються допоміжні функції:

$$a_{1j+1} = \frac{1}{2+S}, \quad b_{1j+1} = \varphi_{2j+1} + Su_{1j}.$$

У зворотному ході обчислюються значення шуканої функції на $(j+1)$ шарі за формулою $u_{ij+1} = a_{ij+1}(b_{ij+1} + u_{i+1j+1})$. Величина $u_{nj+1} = \varphi_{3j+1}$ є значенням шуканої функції в точці (x_n, t_{j+1}), а $u_{0j+1} = \varphi_{2j+1}$ – в точці (x_0, t_{j+1}). Похибка методу $O(h^2 + k)$. З аналізу стійкості неявної схеми випливає, що слід призначати $S > 0$.

11.6. Розв'язання гіперболічних рівнянь

Розглянемо гіперболічне рівняння коливання однорідної і обмеженої струни. Завдання полягає у відшуканні функції $u(x,t)$ при $t > 0$, що задовольняє рівнянню гіперболічного типу:

$$\frac{\partial^2 u}{\partial t^2} = C \cdot \frac{\partial^2 u}{\partial x^2},$$

де: $0 < x < a$; $0 < t < T$, початковим умовам

$$u(x, 0) = f(x);$$

$$\frac{\partial u}{\partial t}(x, 0) = g(x);$$

$$0 \le x \le a$$

і краєвим умовам

$$u(0,t) = u_1(t);$$

$$u(a,t) = u_2(t);$$

$$0 \le t \le T.$$

Замінимо t на $\sqrt{C}t$ і отримаємо рівняння

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2};$$

надалі вважатимемо, що $C = 1$.

Для побудови різницевої схеми розв'язку даної побудуємо в області $D=\{(x,t),0\le x\le a,0\le t\le T\}$ сітку $x_i=ih$; i=0,1,…,n; $a=h\cdot n$; $t_j=j\tau$; j=0,1,…,m; $\tau m=T$.

Апроксимуємо рівняння різницевими похідними другого порядку точності по кроках h,τ:

$$\frac{u_{ij-1}-2\cdot u_{ij}+u_{ij+1}}{\phi^2}=\frac{u_{i+1j}-2\cdot u_{ij}+u_{i-1j}}{h^2}.$$

Вважаючи $\lambda=\tau/h$, перепишемо, виразив U_{ij+1}. Таким чином отримаємо трьохшарову різницеву схему:

$$u_{ij+1}=\lambda^2(u_{i+1j}+u_{i-1j})+2\cdot(1-\lambda^2)\cdot u_{ij}-u_{ij-1},$$

де i = 1,…,n; j = 1,…,m. Задаємо нульові граничні умови $\mu_1(t)=0$, $\mu_2(t)=0$. Тоді маємо $u_{0j}=0$, $u_{nj}=0$ для всіх j.

Остання схема називається трьохшаровою, оскільки вона зв'язує значення шуканої функції на трьох тимчасових шарах j-1, j, j+1.

Чисельне рішення задачі полягає в обчисленні наближених значень u_{ij} розв'язку $u(x,t)$ у вузлах (x_i,t_j) при i = 1,…,n; j = 1,…,m. Алгоритм розв'язання заснований на тому, що розв'язок на кожному наступному шарі (j = 2,3..,n) можна отримати за допомогою перерахунку розв'язків з двох попередніх шарів (j = 0,1..,n-1). При j = 0 розв'язок відомий з початкової умови $u_{i0}=f(x_i)$. Для обчислення розв'язку на першому шарі (j = 1) покладемо

$$\frac{\partial u}{\partial t}(x,0)\approx\frac{u(x,\tau)-u(x,0)}{\phi},$$

тоді $u_{i1}=u_{i0}-\tau g(x_i)$, i = 1,2,…,n.

Описана схема апроксимує задачу з точністю $O(\tau+h^2)$. Невисокий порядок апроксимації по τ пояснюється грубістю апроксимації.

Схема буде стійкою, якщо виконана умова $\tau<h$.

11.7. Розв'язання еліптичних рівнянь

Як еліптичне рівняння розглянемо рівняння Пуассона

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=f(x,y),\ (x,y)\in\Omega,$$

де $u(x,y)$ – шукана функція двох змінних (x, y – координати);

Ω – задана область (тіло);

$f(x,y)$ – функція, яка задає внутрішні джерела енергії (чи відтоків).

Якщо $f(x,y)=0$ в області Ω, то наведене рівняння називають рівнянням Лапласа.

Рівняння Лапласа і Пуассона описують: потоки ідеальних рідин у стаціонарних потоках, стаціонарний розподіл температури або напруженості електричних і магнітних полів. Рівняння Лапласа описує ці процеси за відсутності внутрішніх джерел енергії (чи відтоків).

Розглянемо рівняння Лапласа:

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0\ .$$

Рівняння описує розповсюдження електромагнітних хвиль (полів). Розглядатимемо рівняння Лапласа в прямокутній області $\Omega=\{(x,y),0\le x\le a,0\le y\le b\}$ з краєвими умовами

$$u(0,y)=f_1(y);\ u(a,y)=f_2(y);\ u(x,0)=f_3(x);\ u(x,b)=f_4(x),$$

де f_1, f_2, f_3, f_4 – задані функції. Зазначимо, що найчастіше область буває не прямокутною.

Введемо позначення $u_{ij}=u(x_i,y_j)$. Накладаємо на прямокутну область сітку $x_i=h\cdot i$; $i=0,1,\ldots,n$; $y_j=1\cdot j$; $j=0,1,\ldots,m$. Тоді $x_n=h\cdot n$, $y_m=l\cdot m$.

Частинні похідні апроксимуємо по формулах:

$$\frac{\partial^2 u}{\partial x^2}=\frac{u_{i+1}-2\cdot u_{i,j}+u_{i-1,j}}{h^2}+O(h^2);$$

$$\frac{\partial^2 u}{\partial y^2}=\frac{u_{i,j+1}-2\cdot u_{i,j}+u_{i,j-1}}{l^2}+O(l^2),$$

і замінимо рівняння Лапласа кінцеворізницевим рівнянням:

$$\frac{u_{i+1,j}-2\cdot u_{i,j}+u_{i-1,j}}{h^2}+\frac{u_{i,j+1}-2\cdot u_{i,j}+u_{i,j-1}}{l^2}=0,$$

де $i = 1,\ldots, n\text{-}1, j=1,\ldots, m\text{-}1$ (тобто для внутрішніх вузлів).

Похибка заміни диференціального рівняння різницевим складає величину $O(h^2 + l^2)$. Виразимо $u_{i,j}$ при $h = 1$ і отримуємо систему

$$u_{ij} = (u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1})/4;$$
$$u_{i0} = f_3(x_i);$$
$$u_{im} = f_4(x_i);$$
$$u_{0j} = f_1(y_j);$$
$$u_{nj} = f_2(y_j).$$

Систему лінійних алгебраїчних рівнянь можна розв'язати будь-яким ітераційним методом (Зейделя, простих ітерацій і так далі). При цій побудові системи використовувалася схема типу «хрест» (рис. 11.3).

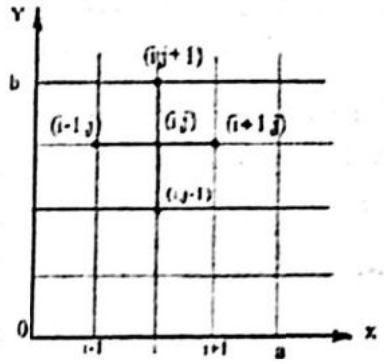


Рис. 11.3. Схема «хрест»

Будуємо послідовність ітерацій за методом Гаусса-Зейделя

$$u_{ij}^{(s+1)} = \frac{1}{4}\left(u_{i-1,j}^{(s+1)} + u_{i+1,j}^{(s)} + u_{ij+1}^{(s)} + u_{ij-1}^{(s+1)} \right),$$

де s-поточна ітерація.

Умова закінчення ітераційного процесу:

$$\max_{i,j}\left|u_{ij}^{(S+1)} - u_{ij}^{(S)}\right| < \varepsilon.$$

Однак остання умова не завжди надійна і на практиці використовують інший критерій

$$\max_{ij}\left|u_{ij}^{(S+1)}-u_{ij}^{(S)}\right|\le\varepsilon(1-\nu),$$

де $$\nu=\frac{\max\left|u_{ij}^{(S+1)}-u_{ij}^{(S)}\right|}{\max\left|u_{ij}^{(S)}-u_{ij}^{(S-1)}\right|}.$$

Схема «хрест» – явна стійка схема (мала зміна вхідних даних веде до малої зміни вихідних даних).

Контрольні запитання та завдання

1. Запишіть загальний вигляд диференціального рівняння в частинних похідних другого порядку з двома змінними. Які типи класифікації використовуються для цих рівнянь?

2. Наведіть класифікацію диференціальних рівнянь у частинних похідних другого порядку з двома змінними залежно від фізичного сенсу розв'язуваних з їх допомогою задач.

3. Наведіть класифікацію диференціальних рівнянь у частинних похідних другого порядку з двома змінними залежно від їх математичної природи.

4. Сформулюйте постановку задачі Діріхле для рівняння Пуассона.

5. Сформулюйте постановку задачі розв'язання диференціального рівняння в частинних похідних другого порядку з двома змінними для рівняння параболічного типу.

6. Апроксимація частинних похідних.

7. Метод сіток розв'язання диференціальних рівнянь еліптичного типу.

8. Метод сіток розв'язання диференціальних рівнянь гіперболічного типу.

9. Метод сіток розв'язання диференціальних рівнянь параболічного типу.

10. Різницева апроксимація диференціального рівняння та граничних умов. Поняття коректності та стійкості різницевої схеми.

11. Метод прогонки розв'язання граничних задач для рівнянь у частинних похідних.

12. ЧИСЕЛЬНІ МЕТОДИ НАБЛИЖЕННЯ ФУНКЦІЙ

12.1. Апроксимація, інтерполяція та екстраполяція

Задача наближення функцій виникає при розв'язанні багатьох практичних і теоретичних задач, а інколи і як самостійна. Так, наближення функцій є важливим допоміжним апаратом при розв'язанні інших задач чисельного аналізу: чисельного інтегрування і диференціювання, розв'язання диференціальних рівнянь, розв'язання систем нелінійних рівнянь, задач оптимізації та ін.

Зазвичай розглядається декілька постановок задачі наближення функцій.

Постановка 1. Проста задача, що приводить до наближення функцій, полягає в наступному. Нехай є деяка функція $f(x)$, $f: R^1 \to R^1$, визначена на відрізку $[a,b]$ і про яку відомо, що в n точках x_1, $x_2,\ldots,x_n$ вона приймає, відповідно, значення y_1, $y_2,\ldots,y_n$, тобто $y_i = f(x_i)$, $i = \overline{1,n}$. Потрібно відновити її значення при інших значеннях $\forall x \in [a,b]$. Функція $y = f(x)$ може бути як невідомою, тобто її значення тільки вимірюються (так званий «чорний ящик»), а може бути і просто складною для обчислень. Наприклад, вона може використовуватися в будь-яких фізико-технічних або суто математичних розрахунках, де її потрібно багато разів обчислювати.

Слід зазначити, що якщо функція $y = f(x)$ невідома, то набір точок (x_i,y_i), $i = \overline{1,n}$ називають **табличним завданням функції**, оскільки ці значення подають у вигляді таблиці (таку функцію ще називають табульованою).

У цих випадках функцію $y = f(x)$ намагаються замінити на простішу функцію $g(x;a)$, близьку до $f(x)$, тобто:

$$f(x) \approx g(x;a),$$

де $a \in R^k$ – деякі параметри функції g, вид якої відомий. Процес наближення однієї функції іншою ще називають апроксимацією, а функцію g при цьому – апроксимуючою для f.

Саме параметри $a \in R^k$ інструментом для підгонки (наближення) функції g до функції f.

Якщо параметри a_1, a_2,..., a_k визначаються з умови збігу значень функції $f(x)$ і апроксимуючої функції в точках x_1, x_2,...,x_n, тобто:

$$g(x_i;a) = f(x_i),\ \forall\ i = \overline{1,n},$$

то такий спосіб наближення називають **інтерполяцією** або **інтерполюванням.**

Зазвичай точки x_i, $i = \overline{1,n}$ називають **вузлами інтерполяції** або ще полюсами, тобто «інтерполяція» – це відновлення функції між полюсами. Множина точок x_i, $i = \overline{1,n}$ може виникати також внаслідок певного розбиття проміжку $[a,b]$, причому $x_0<x_1<...<x_n$. Тоді множину точок x_i, $i = \overline{0,n}$ називають **сіткою вузлів** на відрізку $[a, b]$. Сітка може бути як рівномірною, так і нерівномірною.

Постановка 2. У задачах планування експериментів виникає така проблема. Відомо вид гарного наближення функції, наприклад функція до $f(x)$ добре наближається поліномом 2-го степеня. У той же час вимірювані значення функції y_i мають великі помилки. Потрібно отримати найкраще в певному значенні наближення при мінімальній кількості вимірювань. Така задача виникає при плануванні експериментів у біології, хімії, фізиці, географії, військовій справі та ін.

Постановка 3. Задача наближення функцій розв'язується і при складанні стандартних процедур обчислення елементарних і спеціальних функцій, наприклад $\cos(x)$, $\sin(x)$ та ін. Так:

$$e^x \approx \sum_{k=0}^{n} \frac{x^k}{k!},$$

$$\sin(x) \approx x + \sum_{k=1}^{n} (-1)^k \frac{x^{2k-1}}{(2k-1)!},$$

де n підбирається так, щоб забезпечити достатню точність обчислень.

12.2. Постановка задачі інтерполяції

Нехай функція $f(x) \in C[a,b]$ задана своїми значеннями $y_i = f(x_i), x_i \in [a,b]$, $i = \overline{0,n}$, причому при $x_i \neq x_j$ для $i \neq j$.

Функція $\Phi(x)$ називається **інтерполюючою** для $f(x)$ на сітці $\{x_i\}_{i=0}^{n}$, якщо $\Phi(x_i)=y_i$, $i=\overline{0,n}$.

Задача інтерполювання функції має не єдиний розв'язок. Для побудови інтерполюючих функцій використовуються різноманітні системи так званих лінійно незалежних функцій. Виберемо деяку систему лінійно незалежних функцій $\{\varphi_i(x)\}_{i=0}^{n}$, $\varphi_i(x)\in C[a,b]$ і побудуємо їх лінійну комбінацію

$$\Phi(x)=\Phi_n(x)=\sum_{i=0}^{n}c_i\varphi_i(x),$$

яка називається **узагальненим багаточленом**. Умови інтерполювання дають СЛАР:

$$\sum_{i=0}^{n}c_i\varphi_i(x_i)=y_i, i=\overline{0,n},$$

розв'язком якої є $\vec{c}=(c_0,...,c_n)$. Якщо

$$D(x_0,...,x_n)=\begin{vmatrix}\varphi_0(x_0) & ... & \varphi_n(x_0)\\ ... & ... & ...\\ \varphi_0(x_n) & ... & \varphi_n(x_n)\end{vmatrix}\neq 0,$$

то СЛАР має єдиний розв'язок. Останній визначник називають визначником Вронського.

Система функцій $\{\varphi_k(x)\}$, $k=\overline{0,n}$ називається **системою Чебишова**, якщо, $D(x_0,...x_n)\neq 0, \forall\{x_i\}_{i=0}^{n}, x_i\in[a,b] i=\overline{0,n}, x_i\neq x_j, \forall i\neq j$.

Приклади систем Чебишова.

1. $\varphi_i(x)=x^i$ – алгебраїчна система.

Визначник $D(x_0,...,x_n)\neq 0$ є визначником Вандермонда:

$$D(x_0,..,x_n)=\begin{vmatrix}1 & x_0 & ... & x_0^{n-1}\\ ... & ... & ... & ...\\ 1 & x_n & ... & x_n^{n-1}\end{vmatrix}=\prod_{0\le i<m\le n}(x_i-x_m)\neq 0.$$

2. $\varphi_i(x)=L_i(x)$ – ортогональні багаточлени Лежандра; $\varphi_i(x)=T_i(x)$ – ортогональні багаточлени Чебишова.

3. $\varphi_i(x)$: $1,\cos x,\sin x,...,\cos nx,\sin nx$.

Тоді $\Phi_n(x)=T_n(x)=a_0+\sum_{k=1}^{n}(a_k\cos kx+b_k\sin kx)$ – тригонометричний багаточлен.

12.3. Інтерполяційний поліном Лагранжа

Якщо $\varphi_k(x)=x^k$, то $\Phi_n(x)=P_n(x)=\sum_{k=0}^{n}c_kx^k$. Задача інтерполювання функції $f(x)$ алгебраїчним багаточленом полягає в знаходженні коефіцієнтів c_k, $k=\overline{0,n}$, для яких виконується умова $f(x_i)=\varphi(x_i)$ $i=\overline{0,n}$. Представимо інтерполяційний багаточлен у вигляді:

$$P_n(x)=L_n(x)=\sum_{k=0}^{n}f(x_k)\Phi_k^{(n)}(x).$$

Тут $L_n(x)$ – **інтерполяційний поліном**; $\Phi_k^{(n)}(x)$ – поліноми n-го степеня, які називають **множниками Лагранжа.** З умови $L_n(x_i)=f(x_i)$ випливає, що множник Лагранжа повинен задовольняти умови

$$\Phi_k^{(n)}(x_i)=\delta_{ik}.$$

Так як $\Phi_k^{(n)}(x)$ – багаточлен степеня n, то він має вигляд:

$\Phi_k^{(n)}(x)=A_k(x-x_0)\ldots(x-x_{k-1})(x-x_{k+1})\ldots(x-x_n)$,

де A_k – число. Знайдемо його з умови $\Phi_k^{(n)}(x_k)=1$:

$$A_k=\frac{1}{(x_k-x_0)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_n)}.$$

Таким чином, багаточлени $\Phi_k^{(n)}(x)$ мають вигляд:

$$\Phi_k^{(n)}(x)=\frac{(x-x_0)\ldots(x-x_{k-1})(x-x_{k+1})\ldots(x-x_n)}{(x_k-x_0)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_n)}.$$

Позначивши $\omega_n(x)=\prod_{i=0}^{n}(x-x_i)$, маємо $\Phi_k^{(n)}(x)=\frac{\omega_n(x)}{(x-x_k)\omega_n'(x_k)}$.

Остаточно інтерполяційний багаточлен Лагранжа має вигляд:

$$L_n(x)=\sum_{k=0}^{n}f(x_k)\frac{\omega_n(x)}{(x-x_k)\omega_n'(x_k)}$$

або

$$L_n(x) = \sum_{k=0}^{n} f(x_k) \left[\prod_{j=1, j \neq i}^{n} \left(\frac{x - x_j}{x_i - x_j} \right) \right]$$

При цьому, як бачимо, в точках $x_0, x_1, \ldots, x_n$ значення багаточлена $L_n(x)$ і функції $f(x)$ збігаються. При інших значеннях x різниця $R_n(x) = f(x) - L_n(x)$ в загальному випадку відмінна від нуля і являє собою справжню помилку методу. Величина $R_n(x)$ є залишковим членом інтерполяції. Якщо функція $f(x)$ у проміжку $[a,b]$ має неперервні похідні до $(n+1)$-го порядку, залишковий член можна представити у вигляді:

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)(x - x_1)\ldots(x - x_n),$$

де (ξ) залежить від x і лежить усередині відрізка $[a,b]$.

Позначивши $M_{n+1} = \max_{[a,b]} |f^{(n+1)}(x)|$, одержимо оцінку абсолютної похибки інтерполяційної формули Лагранжа:

$$|R_n(x)| = |f(x) - L_n(x)| \leq \frac{M_{n+1}}{(n+1)!} |\Pi(x)|,$$

де $\Pi(x) = (x - x_0)(x - x_1)\ldots(x - x_n)$.

Отже, погрішність інтерполяції за допомогою багаточлена Лагранжа має порядок $O(h^n)$, де $h = \max|x_{i+1} - x_i|$. Однак між вузлами інтерполяції x_i поліном Лагранжа, як правило, нестійкий до погрішностей обчислень, причому нестійкість зростає зі збільшенням n. Крім того, навіть невеликі погрішності значень $f(x_i)$ можуть сильно змінити поведінку полінома між вузлами. Зазначені ефекти нестійкості виявляються вже при $n > 15$, а в практичних задачах, коли n близьке до 100, нестійкість полінома Лагранжа досить велика.

Недоліком формули Лагранжа також є те, що при додаванні нових вузлів інтерполяції необхідно перераховувати всі коефіцієнти Лагранжа, тому що кожний з них залежить від повної сукупності вузлів інтерполяції.

При обчисленні $f(x)$ для конкретних значень x за допомогою інтерполяційного полінома Лагранжа для нерівномірно віддалених

значень аргументу зручно користуватися схемою Ейткена. Відповідно до цієї схеми послідовно обчислюються багаточлени $L_{i,i+1}(x)$, $L_{i,i+1,i+2}(x)$ і т.д. Інтерполяційний поліном n-го ступеня, що приймає в точках $x_0, x_1, \ldots, x_n$ значення $y_0, y_1, \ldots, y_n$, запишеться таким чином:

$$L_{0,1,\ldots,n}(x) = \frac{1}{x_n - x_0}\begin{vmatrix} L_{0\,1,\ldots,n-1}(x) & x_0 - x \\ L_{1,2\ldots,n}(x) & x_n - x \end{vmatrix}.$$

Обчислення за схемою Ейткена зазвичай ведуть до тих пір, поки послідовні значення $L_{0,1,\ldots,n}$, $L_{0,1,\ldots,n,n+1}$ не співпадуть у межах заданої точності. Обчислення слід припинити, якщо в межах заданої точності збільшення ступеня многочлена не веде до уточнення $L_n(x)$.

12.4. Інтерполяційний поліном Ньютона

Введемо поняття розділених різниць. Розділені різниці є аналогом похідної для функції, що задана таблицею.

Розділеною різницею 1-го порядку для функції $f(x)$ називатимемо:

$$f(x_i; x_j) = \frac{f(x_i) - f(x_j)}{x_i - x_j}.$$

Розділеною різницею 2-го порядку для функції $f(x)$ називатимемо:

$$f(x_i; x_j; x_k) = \frac{f(x_i; x_j) - f(x_j; x_k)}{x_i - x_k}.$$

Аналогічно визначаються розділені різниці довільного порядку.

Наведемо деякі властивості розділених різниць:

1. $f(x_0; \ldots; x_n) = \sum_{i=0}^{n} \frac{f(x_i)}{\prod_{i \neq j} (x_i - x_j)}$.

2. Розділена різниця – лінійна функція:

$(\alpha_1 f_1 + \alpha_2 f_2)(x_0; x_1) = \alpha_1 f_1(x_0; x_1) + \alpha_2 f_2(x_0; x_1)$

3. Розділена різниця – симетрична функція:

$$f(x_1,...,x_i,...,x_j,...,x_n) = f(x_1,...,x_j,...,x_i,...,x_n)$$

4. Для $f(x) \in C^n([a,b]), \exists \xi \in (a,b): f(x_0, x_1,..., x_n) = \dfrac{f^{(n)}(\xi)}{n!}$.

Запишемо формулу Лагранжа інтерполяційного багаточлена:

$$L_n(x) = \sum_{i=0}^{n} f(x_i) \frac{\omega_n(x)}{(x - x_i)\omega_n'(x_i)},$$

де $\omega_n(x) = \prod_{j=0}^{n} (x - x_j)$.

Позначимо $\Phi_j(x) = L_j(x) - L_{j-1}(x)$. Тоді, оскільки

$$L_n(x) = L_0(x) + (L_1(x) - L_0(x)) + ... + (L_n(x) - L_{n-1}(x)),$$

$$L_j(x_i) = L_{j-1}(x_i) = f(x_i) \ \forall i \in \overline{0, j-1},$$

то

$$\Phi_j(x_i) = A_j(x - x_0)...(x - x_{j-1}),$$

де A_j визначається з умови $\Phi_j(x_j) = L_j(x_j) - L_{j-1}(x_j) = f(x_j) - L_{j-1}(x_j)$. Звідси

$$\Phi_j(x) = \frac{f(x_j) - L_{j-1}(x_j)}{(x_j - x_0)...(x_j - x_{j-1})}(x - x_0)...(x - x_j).$$

Тоді

$$A_j = \frac{f(x_j) - L_{j-1}(x_j)}{(x_j - x_0)...(x_j - x_{j-1})} = \frac{f(x_j)}{(x_j - x_0)...(x_j - x_{j-1})} -$$

$$- \sum_{i=0}^{j-1} \frac{f(x_i)}{(x_i - x_0).(x_i - x_{i-1})(x_i - x_{i+1}).(x_i - x_{j-1})(x_j - x_i)} =$$

$$= \frac{f(x_j)}{(x_j - x_0)...(x_j - x_{j-1})} + \sum_{i=0}^{j-1} \frac{f(x_i)}{(x_i - x_0).(x_i - x_{i-1})(x_i - x_{i+1}).(x_i - x_{j-1})(x_i - x_j)} =$$

$$= \sum_{i=0}^{j} \frac{f(x_i)}{(x_i - x_0)...(x_i - x_{i-1})(x_i - x_{i+1})..(x_i - x_j)} = f(x_0,...x_j).$$

Звідси маємо **інтерполяційну (першу) формулу Ньютона вперед** ($x_0 \to x_n$):

$$L_n(x) = f(x_0) + f(x_0, x_1)(x - x_0) + ... + f(x_0,..., x_n)(x - x_0)...(x - x_{n-1}).$$

Аналогічно, **інтерполяційна (друга) формула Ньютона назад** ($x_n \to x_0$):

$$L_n(x) = f(x_n) + f(x_n, x_{n-1})(x - x_n) + \ldots + f(x_0, \ldots, x_n)(x - x_n)\ldots(x - x_1).$$

Також маємо рекурсію за степенем багаточлена

$$L_n(x) = L_{n-1}(x) + f(x_0; \ldots; x_n)(x - x_0)\ldots(x - x_{n-1}).$$

Звідси:

$$L_n(x) = f(x_0) + (x - x_0)\{f(x_0, x_1) + (x - x_1)\{\ldots + (x - x_{n-1})f(x_0, x_1, \ldots, x_n)\}\}.$$

Виведемо нову формулу для похибки інтерполювання. Для $x \neq x_i, i = \overline{0,n}$ розглянемо розділену різницю

$$f(x; x_0; \ldots; x_n) = \frac{f(x)}{(x - x_0)\ldots(x - x_n)} + \sum_{k=0}^{n} \frac{f(x_k)}{(x_k - x)\prod_{i \neq k}(x_k - x_i)}.$$

Звідси

$$f(x) = f(x_0)\frac{(x - x_1)\ldots(x - x_n)}{(x_0 - x_1)\ldots(x_0 - x_n)} + \ldots + f(x_n)\frac{(x - x_0)\ldots(x - x_{n-1})}{(x_n - x_0)\ldots(x_n - x_{n-1})} +$$

$$+ f(x; x_0; \ldots; x_n)(x - x_0)\ldots(x - x_n) = L_n(x) + f(x; x_0; \ldots; x_n)\omega_n(x)$$

Тоді похибка має вигляд:

$$r_n(x) = f(x) - L_n(x) = f(x; x_0; \ldots; x_n)\omega_n(x)$$

або

$$f(x, x_0, \ldots, x_n) = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Оскільки оцінити абсолютну похідну високого порядку часто важко, а часом неможливо, при використанні інтерполяційних формул надходять практично так: ступінь інтерполяційного многочлена $P_n(x)$ повинна збігатися з порядком практично постійних (у межах заданої точності) різниць таблиці $f(x)$. При цьому похибка методу інтерполяції не перевищує одиниці молодшого розряду табличних значень (це правило не застосовується в тих випадках, коли похідні функції сильно змінюються).

Інтерполяційні формули Ньютона можна представити і в інших термінах. Нехай маємо сітку рівновіддалених вузлів: $x_i = a + ih,\ i = \overline{0,n},\ h = \frac{b-a}{n}$. Позначимо $\Delta f_0 = f_1 - f_0$, $\Delta^2 f_0 = \Delta f_1 - \Delta f_0 = f_2 - 2f_1 + f_0$ … – скінчені різниці.

Тоді в цих термінах **перша інтерполяційна формула Ньютона:**

$$L_n(x_0+th)=f_0+\frac{\Delta f_0}{1!}t+\frac{\Delta^2 f_0}{2!}t(t-1)+\ldots+\frac{\Delta^n f_0}{n!}t(t-1)\ldots(t-n+1),$$

де $t=\dfrac{x-x_0}{h}$, h – крок таблиці, що використовується для інтерполювання і екстраполювання функції $y=f(x)$ у точках x, близьких до початку таблиці, а також для продовження таблиці (у бік зменшення аргументу) на один-два кроки.

Також друга **інтерполяційна формула Ньютона:**

$$L_n(x_0+th)=f_n+\frac{\Delta f_{n-1}}{1!}t+\frac{\Delta^2 f_{n-2}}{2!}t(t+1)+\ldots+\frac{\Delta^n f_0}{n!}t(t+1)\ldots(t+n-1),$$

де $t=\dfrac{x-x_0}{h}$, h – крок таблиці, що використовується для інтерполювання і екстраполювання функції $y=f(x)$ у точках, близьких до кінця таблиці (x_0– кінець таблиці, вузли $x_{-1}=x_0-h$, $x_{-2}=x_0-2h,\ldots$, розташовані лівіше x_0), і для продовження таблиці на один-два кроки в бік збільшення аргументу.

При побудові інтерполяційних формул Ньютона використовуються лише значення, що лежать по один бік від обраного початкового значення аргументу, тобто ці формули мають односторонній характер. У багатьох випадках виявляються корисними інтерполяційні формули, що містять як подальші, так і попередні значення функцій відносно початкового її значення. Так, вибір вузлів по черзі ліворуч і праворуч від початкового вузла x_0 призводить до формули Гаусса, з якої, у свою чергу, можна отримати формули Стірлінга і Бесселя, що використовуються для інтерполювання в середині таблиці. Деякі результати щодо інтерполяції функцій використовувалися в попередніх розділах.

12.5. Кусково-лінійна та квадратична інтерполяція

Інтерполяція багаточленом Лагранжа або Ньютона на відрізку $[a, b]$ з використанням великої кількості вузлів інтерполяції часто приводить до поганого наближення через розбіжність процесу інтерполювання. Для того щоб уникнути великої похибки, весь відрізок $[a, b]$ розбивають на частинні відрізки $[x_{i-1}, x_i]$ і на кожному з частинних відрізків замінюють функцію $f(x)$ багаточленом невисокого степеню. У цьому і полягає кусково-поліноміальна інтерполяція.

Розглянемо найпростішу таку інтерполяцію – кусково-лінійну. Нехай задана $f(x)$ значеннями $f(x_i)$ $i = \overline{0,n}$. Побудуємо функцію $\Phi_1(x)$ – лінійну на $x \in [x_{i-1}, x_i]$, що інтерполює ці значення:

$$\Phi_1(x) = L_1^i(x) = f(x_{i-1})\frac{x - x_{i-1}}{x_i - x_{i-1}} + f(x_i)\frac{x_i - x}{x_i - x_{i-1}},\ x \in [x_{i-1}, x_i]$$

Представимо її у вигляді:

$$\Phi_1(x) = \sum_{i=0}^{n} f(x_i)\varphi_i(x).$$

З умов інтерполювання маємо:

$$\Phi_1(x_j) = \sum_{i=0}^{n} f(x_i)\varphi_i(x_j) = f(x_j).$$

Звідси:

$$\varphi_i(x_j) = \delta_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}.$$

Значить:

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, x_{i-1} \le x \le x_i \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i}, x_i \le x \le x_{i+1} \\ 0 \qquad\quad , x \le x_{i-1},\ x \ge x_{i+1} \end{cases}$$

Теорема. Для $f(x) \in C^2[a,b]$ правильна оцінка:

$$\|f(x) - \Phi_1(x)\|_{C[a,b]} \le \frac{M_2}{8}|h|^2,$$

де $\Phi_1(x)$ – кусково-лінійна функція побудована по значеннях $f(x_i), i = \overline{0,n}$, $|h| = \max_i h_i, h_i = x_i - x_{i-1}$, $M_2 = \max_{x\in[a,b]}\left|f^{(2)}(x)\right|$.

Отже маємо збіжність процесу інтерполювання за допомогою кусково-лінійної функції $\|f(x) - \Phi_1(x)\|_{C[a,b]} \xrightarrow[h\to 0 (n\to\infty)]{} 0$.

Зауважимо, що кусково-лінійна інтерполяція негладка, тому на практиці застосовують квадратичні, а найчастіше – кубічні поліноми на кожному підпроміжку $x \in [x_{i-1}, x_i]$.

Подібно до кусково-лінійної інтерполяції можна використовувати для наближення на кожному окремому інтервалі $[x_i, x_{i+1}]$ поліноми вищого порядку, наприклад квадратичну функцію $g(x;a,b,c) = ax^2 + bx + c$.

Неважко перевірити, що квадратична функція, що проходить через 3 точки (x_i, y_i), (x_{i+1}, y_{i+1}), (x_{i+2}, y_{i+2}), може бути записана у вигляді:

$$g_i(x) = \frac{(x-x_{i+1})(x-x_{i+2})}{(x_i-x_{i+1})(x_i-x_{i+2})}y_i + \frac{(x-x_i)(x-x_{i+2})}{(x_{i+1}-x_i)(x_{i-1}-x_{i+2})}y_{i+1} +$$

$$+\frac{(x-x_i)(x-x_{i+1})}{(x_{i+2}-x_i)(x_{i+2})(x_{i-2}-x_{i+1})}y_{i+2}.$$

Теорема. Нехай функція $f(x)$ тричі неперервно-диференційована на відрізку $[x_1, x_n]$. Тоді для всіх $x \in [x_1, x_n]$:

$$|f(x) - g(x)| \le M_3 h^3,$$

де $M_3 = \max_{x\in[a,b]}\left|f^{(3)}(x)\right|$, $|h| = \max_i h_i, h_i = x_i - x_{i-1}$.

Таким чином, погрішність кусково-квадратичної інтерполяції має порядок $O(h^3)$.

12.6. Сплайн-інтерполяція

Інтерполяція за допомогою поліномів Лагранжа або Ньютона з використанням великої кількості вузлів часто призводить до поганого наближення, що пояснюється накопиченням великої погрішності у процесі обчислень. З іншого боку, кусково-лінійна і кусково-квадратична інтерполяція не дозволяють досягати гарного наближення, зважаючи на їх теоретичну неточність. Одним із способів досягати гарного наближення в практичних задачах є інтерполяція за допомогою сплайн-функцій.

Сплайном (з англ. *spline* – рейка, лінійка) називається кусково-поліноміальна функція, визначена на відрізку $[x_1, x_n]$ і яка має на цьому відрізку не менше двох неперервних похідних.

Сплайн буде позначено через $S(x)$. Інтерполяція за допомогою сплайнів називається сплайн-інтерполяцією.

Розглянемо кубічну сплайн-інтерполяцію [10], яка найчастіше застосовується на практиці, тобто на кожному інтервалі $[x_i, x_{i+1}]$, $i = \overline{1, n-1}$ функція $S(x)$ задається поліномом третьої степені:

$$S(x_i) = y_i + c_{1i}(x - x_i) + c_{2i}(x - x_{i+1})^2 + c_{3i}(x - x_{i+2})^3.$$

Очевидно, $S(x_i) = S_i(x_i) = y_i$, $i = \overline{1, n-1}$.

Слід зазначити, що у випадку застосування кубічного сплайна він повинен мати дві неперервні похідні на відрізку $[x_i, x_{i+1}]$.

Для того щоб побудувати кубічний сплайн $S(x)$, необхідно знайти всі коефіцієнти c_{1i}, c_{2i}, c_{3i} поліномів $S_i(x)$, $i = \overline{1, n-1}$.

Кількість цих невідомих коефіцієнтів $3(n-1)$.

Коефіцієнти c_{1i}, c_{2i}, c_{3i}, $i = \overline{1, n-1}$ підбираються так, щоб на межах інтервалів $[x_i, x_{i+1}]$ забезпечити неперервність як функції $S(x)$, так і її першої $S'(x)$ і другої $S''(x)$ похідних. Варто зазначити, що для всіх $i = \overline{1, n-1}$

$$S'_i(x_i) = c_{1i} + 2c_{2i}(x - x_{i+1}) + 3c_{3i}(x - x_{i+2})^2.$$

$$S''_i(x_i) = 2c_{2i} + 6c_{3i}(x - x_{i+2}).$$

Таким чином, повинні виконуватися умови:

$$S_i(x_{i+1}) = S_{i-1}(x_{i+1}) = y_{i+1},\ i = \overline{1, n-2},\ S_{n-1}(x_n) = y_n,$$

$$S'_i(x_{i+1}) = S'_{i+1}(x_{i+1}) = c_{1i+1},\ i = \overline{1, n-2},$$

$$S''_i(x_{i+1}) = S''_{i+1}(x_{i-1}) = c_{2i+1},\ i = \overline{1, n-2}.$$

Можна переписати останні співвідношення в більш наочному вигляді (у звичному вигляді – системи рівнянь):

$$\begin{cases} y_i + c_{1i}(x_{i-1} - x_i) + c_{2i}(x_{i+1} - x_i)^2 + c_{3i}(x_{i+1} - x_i)^3 = y_{i+1},\ i = 1,\ldots, n-1, \\ c_{1i} + 2c_{2i}(x_{i+1} - x_i) + 3c_{3i}(x_{i+1} - x_i)^2 = c_{1,i+1},\ i = 1,\ldots, n-2, \\ 2c_{2i} + 6c_{3i}(x_{i+1} - x_i) = 2c_{2,i+1},\ i = 1,\ldots, n-2. \end{cases}$$

Як видно, така система визначає $3(n-1) - 2$ лінійних рівняння для пошуку $3(n-1)$ невідомих c_{1i}, c_{2i}, c_{3i}, $i = \overline{1, n-1}$. Тому необхідно довизначити ще два рівняння. Для цього задають так звані граничні умови – значення першої або другої похідної функції $f(x)$ на межах інтервалу $[x_1, x_n]$. Якщо значення однієї з похідних на межах відомі, тобто задавши їх, можна отримати точну інтерполяційну схему.

Якщо ж ці значення невідомі, то можна задати другу похідну на межах рівною нулю і отримати також достатньо гарні результати. Необхідно розглянути простий випадок, коли:

$$f''(x_1) = 0, f''(x_n) = 0,$$

тобто

$$\begin{cases} S''_1(x_1) = 2c_{21} = 0, \\ S''_{n-1}(x_n) = 2c_{2,n-1} + 6c_{3,n-1}(x_n - x_{n-1}) = 0. \end{cases}$$

Таким чином, для обчислення коефіцієнтів c_{1i}, c_{2i}, c_{3i}, $i = \overline{1, n-1}$ необхідно розв'язати систему лінійних алгебраїчних рівнянь.

Теорема. Нехай функція $f(x)$ 4 рази неперервно-диференційована на відрізку $[x_1, x_n]$. Тоді для всіх $x \in [x_1, x_n]$:

$$|f(x) - g(x)| \le \frac{1}{8} M_4 h^4,$$

де $M_4 = \max_{x \in [a,b]} \left| f^{(4)}(x) \right|$, $|h| = \max_i h_i$, $h_i = x_i - x_{i-1}$.

Таким чином, погрішність інтерполяції за допомогою кубічного сплайна має порядок $O(h^4)$. Перевагою кубічних сплайнів перед іншими наведеними способами інтерполяції є: по-перше, їх стійкість до процесу обчислень, і, по-друге, досить висока точність.

При використовуванні сплайн-інтерполяції для таблично заданої функції діють таким чином. Спочатку один раз визначають усі коефіцієнти c_{1i}, c_{2i}, c_{3i}, $i = \overline{1, n-1}$, а потім вже для обчислення наближеного значення функції $f(x)$ у деякій точці $x \in [x_1, x_n]$ спочатку знаходять інтервал $[x_i, x_{i+1}]$, якому точка x належить, а потім обчислюють наближене значення функції $f(x)$ в точці x як $S_i(x)$.

12.7. Метод найменших квадратів для апроксимації функцій

Наближення функцій застосовують у випадках, якщо

– функція складна (трансцендентна або є розв'язком складної задачі) і її замінюють функцією, яка легко обчислюється (найчастіше поліномом);

– необхідно побудувати функцію неперервного аргументу для функції, яка задана своїми значеннями (таблична або табульована);

– таблична функція наближається табличною ж функцією (згладжування).

Інтерполювання не кращий спосіб наближення функцій через розбіжність цього процесу для поліномів. Тим більше доцільність застосування інтерполювання сумнівна, якщо функція таблична, а її значення неточні. Потрібно будувати апроксимуючу функцію з інших міркувань.

Найбільш загальний принцип: наблизити $f(x)$ функцією $\Phi(x)$ так, щоб досягалася деяка задана точність ε:

$$\|f(x) - \Phi(x)\| < \varepsilon.$$

Але розв'язок у такій постановці може не існувати або бути не єдиним.

Загальна постановка задачі наближення така. Нехай маємо елемент f лінійного нормованого простору R. Побудуємо під-

простір M_n, в якому елементи є лінійною комбінацією $\Phi = \sum_{i=0}^{n} c_i \varphi_i \in M_n \subset R$ по елементах лінійно незалежної системи:

$$\{\varphi_i\}_{i=0}^{n},\ \varphi_i \in R.$$

Відхилення $\Phi \in M_n$ від $f \in R$ є число:

$$\Delta(f, \Phi) = \|f - \Phi\|.$$

Найбільш відомим і ефективним із методів розв'язання задачі апроксимації функцій є **метод найменших квадратів**. Він призначений безпосередньо для знаходження параметрів a апроксимуючої функції $\Phi(x; a_0, a_2, \ldots, a_n)$, вид якої вже відомий або обраний. Суть методу полягає в тому, що нехай у результаті вимірювань функції $f(x)$ маємо таблицю значень:

$$y_i \approx f(x_i),\ i = \overline{1, N},\ x_i \in [a, b].$$

За даними цієї таблиці треба побудувати аналітичну формулу $\Phi(x; a_0, a_2, \ldots, a_n)$ таку, що

$$\Phi(x_i; a_0, a_2, \ldots, a_n) \approx y_i,\ i = \overline{1, N}.$$

Виконувати це інтерполюванням, тобто задавати

$$\Phi(x_i; a_0, a_2, \ldots, a_n) = y_i,\ i = \overline{1, N}$$

нераціонально, бо $N >> n$ і система перевизначена; її розв'язки, як правило, не існують. Вигляд функції $\Phi(x; a_0, a_2, \ldots, a_n)$ і число параметрів a_i у деяких випадках відомі. В інших випадках вони визначаються за графіком, побудованим за відомими значеннями $f(x_i)$ так, щоб залежність була досить простою і добре відображала результати спостережень. Але такі міркування не дають змогу побудувати єдиний елемент та й ще найкращого наближення.

Тому визначають параметри $a_0, \ldots, a_n$ так, щоб у деякому розумінні всі рівняння системи одночасно задовольнялись з найменшою похибкою (нев'язкою), наприклад, щоб виконувалося:

$$I(a_0, \ldots, a_n) = \sum_{i=1}^{N} [y_i - \Phi(x_i; a_0, \ldots, a_n)]^2 \to \min.$$

Такий метод розв'язання системи і називають **методом найменших квадратів**, оскільки мінімізується сума квадратів відхилення $\Phi(x; a_0, a_2, \ldots, a_n)$ від значень $f(x_i)$ (сума нев'язок).

Для реалізації мінімуму необхідно та достатньо виконання умов:

$$\frac{\partial I}{\partial a_i} = 0,\ i = \overline{0,n}.$$

Наприклад, якщо $\Phi(x_i; a_0, \ldots, a_n)$ лінійно залежить від параметрів $a_0, \ldots, a_n$, тобто

$$\Phi(x_i; a_0, \ldots, a_n) = \sum_{k=0}^{n} a_k \varphi_k(x),$$

то маємо СЛАР

$$\sum_{k=0}^{n} a_k \varphi_k(x_i) = y_i, \quad i = \overline{1,N},$$

яку називають **системою умовних рівнянь**. Позначивши:

$$C = (\varphi_k(x_i))_{k=\overline{0,n}}^{i=\overline{1,N}},\ \vec{a} = (a_0, \ldots, a_n)^*,\ \vec{y} = (y_1, \ldots, y_N)^*,$$

маємо матричний запис СЛАР

$$C\vec{a} = \vec{y}.$$

Помноживши систему умовних рівнянь зліва на транспоновану до C матрицю C^*, отримаємо систему **нормальних рівнянь**

$$C^* C\vec{a} = C^* \vec{y},$$

$$G = A = C^* C,\ \dim G = n+1,\ G = [g_{ik}]_{i,k=0}^{n},$$

$$g_{ik} = \sum_{j=1}^{N} c_{ij} c_{jk} = \sum_{j=1}^{N} c_{ji} c_{jk} = \sum_{j=1}^{N} \varphi_i(x_i)\varphi_j(x_i),$$

$$C^* \vec{y} = \left(\sum_{i=1}^{N} c_{ik} y_i \right)_{k=0}^{n},$$

з якої власно і обчислюють невідомі коефіцієнти.

$$\Delta(f, \Phi) = \|f - \Phi\|^2 = (f - \Phi, f - \Phi) = \sum_{i=1}^{N} (y_i - \Phi(x_i, \vec{a}))^2 \to \inf.$$

Нехай тепер $\Phi(x, a_0, \ldots, a_n)$ – нелінійна функція параметрів $\vec{a} = (a_0, \ldots, a_n)$, наприклад:

$$\Phi = a_0 e^{a_1 x} + a_2 e^{a_3 x} + \ldots,$$

або

$$\Phi = a_0 \cos a_1 x + a_2 \sin a_3 x + \ldots.$$

Оскільки тепер $\Phi(x; a_0, \ldots, a_n)$ нелінійна, то застосуємо метод лінеаризації.

Нехай відомі наближені значення $\vec{a}^0 = (a_0^0, ..., a_n^0)$. Розкладемо в ряд $\Phi(x, \vec{a})$ в околі $\vec{a}^0$. Тоді отримаємо лінійне наближення до $\Phi(x, \vec{a})$:

$$\Phi(x; \vec{a}) \approx \Phi(x; \vec{a}^0) + \sum_{k=0}^{n} \frac{\partial \Phi}{\partial a_k}(x, \vec{a}^0)(a_k - a_k^0).$$

Якщо ввести позначення:

$$\vec{z} = \vec{a} - \vec{a}^0,\ y_i^* = y_i - \Phi(x, \vec{a}^0),\ c_{ik} = \Phi'_{a_k}(x_i, \vec{a}^0),$$

то отримаємо систему умовних рівнянь відносно поправок до $\vec{a}^0$:

$$C\vec{z} = \vec{y}^*.$$

Замінимо її на систему нормальних рівнянь:

$$C^* C\vec{z} = C^* \vec{y}^*.$$

Знайшовши $\vec{z}$, обчислюємо наступне наближення: $\vec{a}^1 = \vec{a}^0 + \vec{z}$. Цей процес можна продовжувати: на кожній ітерації знаходимо $\vec{z}^m, m = 0,1,...$ і уточнюємо наближення до $\vec{a}$: $\vec{a}^m = \vec{a}^{m-1} + \vec{z}^{m-1}$.

Умова припинення ітерацій

$$\|\vec{z}^m\| = \left(\sum_{k=0}^{n} (z_k^m)^2\right)^{\frac{1}{2}} < \varepsilon.$$

Важливим є вибір початкового наближення $\vec{a}^0$. З системи умовних рівнянь (нелінійної) виберемо деякі $n+1$. Розв'язок цієї системи і дасть початкове наближення.

Для деяких простих нелінійних залежностей від невеликої кількості параметрів задачу можна лінеаризувати аналітично. Наприклад, розглянемо наближення даних за таким законом:

$$y_i \approx f(x_i),\ \Phi(x, A, \alpha) = Ax^{\alpha}.$$

Система умовних рівнянь має вигляд:

$$\Phi(x_i) = Ax_i^{\alpha} = y_i,\ i = \overline{1, N}.$$

Прологарифмуємо її:

$$\psi(x_i) = \ln \Phi(x_i) = \ln A + \alpha \ln x_i = \ln y_i, i = \overline{1, N}$$

Введемо $a = \ln A$. Тепер функція $\psi(x, a, \alpha)$ лінійна. Система умовних рівнянь відносно параметрів a та a має вигляд.

$$C\vec{z} = \vec{b},\ \vec{z} = (a, \alpha),\ \vec{b} = (\ln y_i)_{i=1}^{N},$$

$$C = \begin{pmatrix} 1 & \ln x_1 \\ \ldots\ldots\ldots\ldots & \\ 1 & \ln x_N \end{pmatrix}.$$

Запишемо систему нормальних рівнянь для методу найменших квадратів

$$C^{\cdot}C\vec{z} = C^{\cdot}\vec{b},$$

$$G = C^{\cdot}C = \begin{pmatrix} N & \sum_{i=1}^{N} \ln x_i \\ \sum_{i=1}^{N} \ln x_i & \sum_{i=1}^{N} (\ln x_i)^2 \end{pmatrix},\ C^{\cdot}\vec{b} = \begin{pmatrix} \sum_{i=1}^{N} \ln y_i \\ \sum_{i=1}^{N} \ln x_i \ln y_i \end{pmatrix}.$$

Розв'язавши систему, знаходимо і a та $A = \exp(a)$.

12.8. Поняття екстраполяції функцій

На відміну від задачі інтерполяції (з англ. *inter* – між), тобто відновлення функції між полюсами, задача екстраполяції (з англ. *extra* –додатково) полягає у відновленні функції за межами полюсів. Тобто якщо для деякої функції $f(x)$, f: $R^1 \to R^1$ відомо, що в n точках x_1, $x_2,\ldots,$ x_n вона приймає, відповідно, значення y_1, $y_2,\ldots,$ y_n, тобто $y_i = f(x_i)$, $i = 1,\ldots,n$, то потрібно відновити її значення при $x > x_n$ або $x < x_1$.

Така задача виникає, наприклад, при прогнозуванні в часі значень деякого показника (економічного або фізичного) на основі відомих його значень у вже минулих моментах часу. У цьому випадку x – це час, y – значення показника. Відомий набір значень $y_1, y_2,\ldots, y_n$ називається часовим рядом. Прикладами часових рядів є: курси деяких акцій, поквартальний обсяг виробництва, річний ВВП країни, добовий обсяг продажів, погодинний обсяг водоспоживання міста, середньорічна температура на планеті Земля та ін.

У процесі розв'язання задачі екстраполяції застосовують методи апроксимації, оскільки значення y_i, як правило, містять у собі значні випадкові шуми і тому методи інтерполяції дають велику похибку наближення. Для прогнозування часових рядів, окрім апроксимації (прямі методи), застосовують і спеціальні статистичні методи (адаптивні або стохастичні) [15].

Контрольні запитання та завдання

1. Сформулюйте постановку задачі наближення функцій. Які постановки задачі наближення функцій ви ще знаєте?

2. Чим відрізняються задачі апроксимації, інтерполяції та екстраполяції функцій?

3. У чому полягає метод найменших квадратів для апроксимації функцій?

4. У чому полягає лінійна інтерполяція функцій? Коли її слід застосовувати?

5. У чому полягає квадратична інтерполяція функцій? Коли її слід застосовувати?

6. Що називається інтерполяційним поліномом Лагранжа? Коли його слід застосовувати?

7. Інтерполяційна формула Ньютона.

8. Кусково-лінійна інтерполяція.

9. Що називають сплайном?

10. У чому полягає лінійна сплайн-інтерполяція функції? Коли її слід застосовувати?

11. Використовуючи метод найменших квадратів, підберіть найкращу апроксимацію для функції, заданої таблично

x	1.6	1.8	2	2.2	2.4	2.6
y	6.6	6.7	6.2	5.8	4.3	3.4

серед функцій виду:

1) $y = a_0 + a_1 x$;

2) $y = a_0 + a_1 x + a_2 x$;

3) $y = a_0 + a_1 sin(a_2 x)$.

Побудуйте графіки апроксимуючих функцій і графіки залишків.

12. Побудуйте кусково-лінійну та кусково-квадратичну інтерполяцію для функції, заданої таблицею в завданні 8.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Амосов А. А. Вычислительные методы для инженеров: учебн. пособ. / А. А. Амосов, Ю. А. Дубинский, Н. В. Копченова. – М. : Высшая школа, 1994. – 544 с.

2. Бахвалов Н. С. Численные методы / Н. С. Бахвалов, Н. П. Жидков, Г.М. Кобельков. – М. : Бином, 2007. – 636 с.

3. Волков Е.А. Численные методы / Е.А. Волков. – М. : Высшая школа, 1987. – 312 с.

4. Демидович Б. П. Основы вычислительной математики / Б. П. Демидович. – М.: Наука, 1994. – 664 с.

5. Задачин В.М. Чисельні методи: навчальний посібник / В.М. Задачин, І.Г. Конюшенко. – Х.: Вид. ХНЕУ ім. С.Кузнеця, 2016. – 180 с. (Укр. Мов.).

6. Калиткин Н. Н. Численные методы / Н. Н. Калиткин. – М. : Наука, 1978. – 512 с.

7. Тейлор Дж. Введение в теорию ошибок / Дж. Тейлор. – М.: Мир, 1985. – 16 с.

8. Турчак Л.И. Основы численных методов / Л.И. Турчак. – М.: Наука, 1997. – 320 с.

9. Федоренко Р. П. Введение в вычислительную физику / Р. П. Федоренко. – М. : Изд. МФТИ, 1994. – 528 с.

10. Фельдман Л. П. Чисельні методи в інформатиці / Л. П. Фельдман, А. І. Петренко, О. А. Дмитрієва. – К. : Видавнича група BHV. – 2006. – 480 с.

11. Шикин Е. В. Кривые и поверхности на экране компьютера : руководство по сплайнам для пользователей / Е. В. Шикин, А. И. Плис. – М. : ДИАЛОГ–МИФИ, 1996. – 50 с.

12. Шуп Т. Решение инженерных задач на ЭВМ: практическое руководство: учебн. изд. / Т. Шуп ; пер. с англ. – М. : Мир, 1982. – 237 с.

13. Everitt B. A handbook of statistical analyses using R / B. Everitt, T. Hothorn. – 2-nd ed. – Chapman and HALL/CRC, 2009. – 376 p.

14. Shumway R.H. Time series analyses and its applications: With R examples / R.H. Shumway, D.S. Stoffer. – 3-rd ed. – New York: Springer, 2011. – 596 p.

15. Zadachyn V. Calculation of optimal path for parallel car parking / V. Zadachyn, O. Dorokhov // Transport and Telecommunication. – Volume 13. – 2012. – pp. 303–309.

Ресурси мережі Інтернет

16. Сайт персональних навчальних систем ХНЕУ ім. С. Кузнеця. – http://www.ikt.hneu.edu.ua/.

17. Quick-R [Electronic resource]. – Access mode: http://www.stat-methods.net/index.html.

18. R Site Search [Electronic resource]. – Access mode: http://finzi.psych. upenn.edu/nmz.html.

19. Rtips. Revival 2014! [Electronic resource]. – Access mode: http://pj.freefaculty.org/R/Rtips.html.

20. Statistics with R [Electronic resource]. – Access mode: http://zoonek2. free.fr/UNIX/48_R/all.html.

21. The Comprehensive R Archive Network [Electronic resource]. – Access mode: http://cran.r-project.org.

Навчальне видання

СЕРІЯ «НА ДОПОМОГУ СТУДЕНТУ УДФСУ»

Анатолій Олександрович Антонюк

ЧИСЕЛЬНІ МЕТОДИ

Навчальний посібник

Редактор — *Н. І. Грицюк, Л. Б. Дьомена*

Форматування та комп'ютерна верстка — *Н. М. Шамардак*

Здано до друку 23.11.2017. Формат 60×84/16.
Папір офсетний № 1. Гарнітура «Times New Roman».
Друк. арк. 7.6.
Тираж 300 примірників. Замовлення № 644.

Підготовлено до друку Видавничо-поліграфічним центром Університету ДФС України
08200, вул. Університетська, 31, м. Ірпінь, Київська область, Україна

Свідоцтво про внесення суб'єкта видавничої справи до державного реєстру видавців, виготовлювачів і розповсюджувачів видавничої продукції
Серія ДК № 5104 від 20.05.2016