МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
“ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ”

МЕТОДИЧНІ ВКАЗІВКИ

до лекційних зайнять з дисципліни
«Чисельні методи в фізиці»
для студентів спеціальності
153 «Мікро– та наносистемна техніка»

Затверджено
на засіданні кафедри фізичного
матеріалознавства для електроніки
та геліоенергетики,
протокол №7 від 30.01.2019 р.

Харків НТУ «ХПІ» 2019

Методичні вказівки до лекційних зайнять з дисципліни «Чисельні методи в фізиці» для студентів спеціальності 153 «Мікро– та наносистемна техніка» / Уклад.: К.О. Мінакова, Р.В. Зайцев, А.М. Дроздов, М.В. Кіріченко. – Харків: НТУ «ХПІ», 2019. – 73 с.

Укладачі: К.О. Мінакова,
Р.В. Зайцев,
А.М. Дроздов,
М.В. Кіріченко,

Рецензент проф. Г.С. Хрипунов

Кафедра фізичного матеріалознавства для електроніки та геліоенергетики

ВСТУП

Дисципліна «Чисельні методи в фізиці» викладається в студентам третього курсу, які навчаються за напрямом “ Мікро- та наносистемна техніка”, призначена формування знань та навиків з використання комп’ютерів у своїй професійній та повсякденній діяльності, а саме: програмування та розв’язування поширених інженерних задач автоматизації технологічних процесів з використанням чисельних методів та математичних пакетів. Методи та алгоритми розв’язування інженерних задач, наведені у методичці, також можуть використовувати студенти, які навчаються на інших напрямках, які потребують навичок розвёязування інжнерних задач за допомогою математичних пакетів.

Досвід розв'язування науково–дослідних і прикладних задач показує, що незалежно від їхньої складності кінцевої мети можна досягти або постановкою експерименту, або методом математичного моделювання. Кожен з цих методів має свої переваги і недоліки.

За допомогою експерименту можна розв'язувати навіть дуже складні задачі, при цьому достовірність результатів тим вища, чим ретельніше відпрацьована методика експерименту. Водночас здобуті результати будуть стосуватися тільки тих умов, за яких проводився експеримент, внаслідок чого узагальнення результатів на інші умови не коректне. Крім того, треба враховувати економічний бік постановки складного експерименту. Щодо цього, то більші можливості має метод математичного моделювання за допомогою електронної обчислювальної машини ЕОМ, коли аналізують не реальну задачу, а її модельне зображення.

Процес математичного моделювання зображують у такій послідовності: фізична постановка задачі; математична постановка задачі; математичне дослідження задачі; аналіз і осмислення математичного розв'язку та порівняння його з експериментом.

Розглянемо докладніше математичну постановку і математичне дослідження задачі.

Математична постановка полягає у формуванні математичної моделі досліджуваної задачі, яка звичайно е системою рівнянь математичної фізики (диференціальних, Інтегральних, інтегрально–диференціальних).

Математичне дослідження задачі власне зводиться до розв'язування системи рівнянь і аналізу здобутих результатів. Для порівняно простих задач вдається розв'язати вихід :у систему рівнянь і розв'язок подати у вигляді залежностей, виражених через елементарні та інші відомі функції. Якщо це можливо, то говорять, що знайдено аналітичний (точний) розв'язок задачі. Однак переважна більшість практично важливих задач аналітичних розв'язків не має. До таких належать, наприклад, задачі будівництва: визначення напружено–деформованого стану пластин, плит, фунда-

ментів; задачі стійкості, теплопровідності для твердих тіл; напрямленої дифузії тощо. У цих випадках використовують чисельні методи, які, оперуючи системою алгебраїчних рівнянь (аналогів рівнянь математичної фізики), дають можливість побудувати деяку послідовність арифметичних операцій, збільшення кількості яких до нескінченності дає точний розв'язок. Оскільки на практиці здійснюють скінченне число кроків (операцій), то знайдений розв'язок є наближеним. А через те що обчислювальні операції виконують над числами, то відповідні методи дістали назву чисельних. Найбільшого розвитку чисельні методи набули останнім часом завдяки застосуванню ЕОМ, що мають високу швидкість обчислень і велику ємність оперативної пам'яті. Проте основна роль при цьому відводиться, звичайно, людині, яка повинна вміти сформулювати і поставити задачу, описати її математичними залежностями (створити математичну модель об'єкта), скласти алгоритм розв'язання задачі на ЕОМ, написати програму на алгоритмічній мові, зрозумілій машині, розв'язати задачу й оцінити результати.

Щодо оцінювання результатів розрахунку, то слід зазначити, що поєднання чисельних методів і ЕОМ дає можливість зробити це ефективно й оперативно, варіюючи найсуттєвіші параметри розрахункової схеми задачі з наступним чисельним аналізом впливу їх на кінцевий результат. Фактично йдеться про чисельний експеримент, оскільки умови задачі можна змінювати багато разів.

Незважаючи на відмінності в методології, до чисельного експерименту щільно примикають фізичний експеримент і фізичне дослідження, особливо у тій частині, де потрібна оцінка достовірності здобутих результатів.

Математична модель об'єкта – це та сукупність рівнянь, за допомогою якої досліджують реальні фізичні об'єкти (процеси, явища). Математична модель не тотожна досліджуваному об'єкту, а є лише його наближеним описом, оскільки її будують з деякими спрощеннями та ідеалізацією. У моделі враховують найважливіші моменти і взаємозв'язки, найхарактерніші для досліджуваного реального об'єкта. Разом з тим внаслідок заміни реального об'єкта відповідною йому математичною моделлю стало можливим сформулювати задачу як математичну і скористатися для її розв'язання тим чи іншим математичним апаратом.

Алгоритм – це зрозумілий і точний припис (вказівка) виконавцеві здійснювати послідовність дій, спрямованих на досягнення зазначеної мети або розв'язання поставленої задачі.

Точність розв'язку – це міра близькості чисельного розв'язку до аналітичного.

Збіжність розв'язку – це поступове наближення його до точного.

Після вибору математичної моделі об'єкта і опису її на алгоритміч-

ній машинній мові здійснюють чисельну реалізацію задачі на ЕОМ. Останнім часом при реалізації практичних задач здебільшого застосовують ЕОМ, що можуть виконувати від кількох сотень до мільйонів операцій за секунду. Найбільшого застосування в інженерних розрахунках набули ЕОМ, які мають не тільки високу швидкість обчислень, сучасне програмне забезпечення, а й розвинуту сервісну частину, яка дає можливість оперативно діагностувати похибки, графічно відображати результати обчислень, здійснювати розрахунки в режимі діалогу. Великої популярності у користувачів набули також міні– та мікро–ЕОМ, персональні комп'ютери.

Розв'язування багатьох інженерних задач зводиться до обчислення коренів одного нелінійного рівняння або до розв'язання систем нелінійних рівнянь. В обох випадках нелінійні рівняння, що утворюються, можна поділити на два типи – алгебраїчні та трансцендентні.

Алгебраїчними називають рівняння, що містять лише алгебраїчні функції (цілі, раціональні, ірраціональні).

Мета дисципліни – формування у студентів знань та навичок у таких областях:

моделювання;

математичні пакети;

чисельне диференціювання та інтегрування;

розв’язок нелінійних рівнянь;

розв’язок систем алгебраїчних рівнянь;

чисельні методи розв’язку диференціальних рівнянь та систем;

інтерполяція та апроксимація функцій;

чисельні методи розв’язування задач одновимірної оптимізації функцій;

розв’язування задач багатовимірної оптимізації.

Для вивчення дисципліни необхідні знання з таких дисциплін: «Вища математика», «Фізика», «Теорія електричних кіл», «Обчислювальна техніка», «Основи програмування та мікроелелектронна техніка». Знання з дисципліни є вхідними до дисциплін «Автоматизоване проектування електронних пристроїв», «Фізика напівпровідникових приладів», «Електроніка дефектів в напівпровідниках», «Фізичні методи дослідження напівпровідникових матеріалів» тощо.

1 Основи чисельних методів

1.1 Моделювання

Моделювання є основою пізнання людиною навколишнього світу. Проводячи експерименти, теоретичні дослідження, навіть при обговоренні власних дій, намірів, висновків, ми практично займаємося моделюванням. Цілі, задачі, засоби й методи моделювання у цих випадках значно відрізняються один від одного, але загальна спрямованість залишається єдиною – одержання нових знань шляхом випробування (дослідження), деякого замінника реального об'єкта дослідження – моделі. У випадку експериментальних досліджень моделлю є реальний об'єкт, який має ту саму фізичну природу, що й досліджуваний об'єкт. При теоретичних дослідженнях модель має знакову форму – математичних формул, співвідношень, рівнянь, а задачею моделювання є встановлення нових знань про об'єкти, що описуються цими співвідношеннями. Обговорення встановлює правомірність тих припущень і висновків, які були зроблені шляхом моделювання.

Взагалі, спрощено моделювання можна розглядати як певний експеримент, об'єктом якого у першому випадку є матеріальний аналог досліджуваного об'єкта, у другому випадку об'єктом досліджень є знакова (математична) модель.

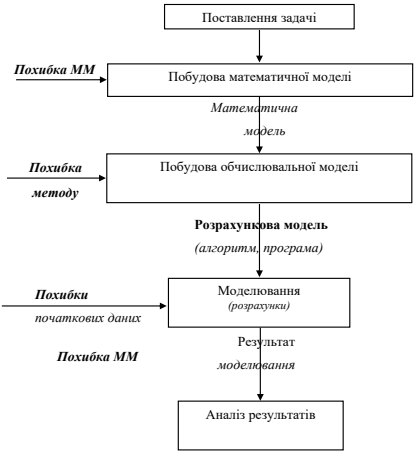

Рисунок 1.1 – Схема розв'язання інженерної задачі

Результатом розв'язування інженерних (прикладних) задач будь–якого рівня є, як правило, чисельні оцінки (параметрів пристроїв, проце-

сів, технічних характеристик тощо), які є наслідком розрахунків, що здійснюються з наближеними первісними даними. Більшість прикладних задач зводяться до математичних задач, які розв'язуються різноманітними обчислювальними методами.

Послідовність розв'язування таких задач можна подати у вигляді наступних етапів:

1) поставлення задачі;
2) створення математичної моделі (формулювання задачі); перевірка моделі на адекватність;
3) побудова розрахункової (обчислювальної) моделі, яка відповідає прийнятій математичній моделі;
4) проведення розрахунків за обраною обчислювальною моделлю при заданих (відомих) значеннях первісних даних;
5) аналіз одержаних результатів.

У цілому процес розв'язування інженерної задачі може бути поданий у вигляді схеми, наведеної на рис. 1.1.
Розглянемо докладніше кожний із цих етапів.

1.1.1 Математичне моделювання

Модель утворюється для подальшого її дослідження з метою одержання нових знань про відповідний реальний об'єкт. Таке дослідження вже готової моделі називають моделюванням. Дослідження математичної моделі називатимемо математичним моделюванням [1].

Математична задача є абстрагованою від конкретної сутності задачі. Для її розв'язання створюються спеціальні обчислювальні методи, причому до тієї самої математичної моделі можуть зводитися зовсім різні прикладні задачі.

У загальному значенні під *адекватністю* математичної моделі розуміють правильний якісний, повний і достатньо точний кількісний опис саме тих характеристик реального технічного об’єкта, які важливі в даному конкретному випадку.

Економічність математичної моделі оцінюють витратами на обчислювальні ресурси (машинний час і пам'ять), необхідні для реалізації математичної моделі на ЕОМ. Ці витрати залежать від числа арифметичних операцій при використанні моделі, від розмірності простору фазових змінних, від особливостей використовуваної ЕОМ та інших чинників. Очевидно, що вимоги економічності, високої точності і досить широкої області адекватності математичної моделі суперечливі і на практиці можуть бути задоволені лише на основі розумного компромісу. Властивість економічності математичної моделі часто пов'язують з її простотою. Більш того, кількісний аналіз деяких спрощених варіантів математичної

моделі може бути здійснений і без залучення сучасної обчислювальної техніки.

Робастність математичної моделі (від англійського слова robust – міцний, стійкий) характеризує її стійкість щодо похибок початкових даних, здатність нівелювати ці похибки і не допускати їх надмірного впливу на результат обчислювального експерименту. Причинами низької робастності математичної моделі може бути необхідність при її кількісному аналізі віднімання близьких один до одного наближених значень величин або ділення на малу за модулем величину, а також використання в математичній моделі функцій, що швидко змінюються у проміжку, де значення аргументу відоме з невисокою точністю. Іноді прагнення збільшити повноту математичної моделі призводить до зниження її робастності внаслідок введення додаткових параметрів, які відомі з невисокою точністю або входять в дуже наближені співвідношення.

Продуктивність математичної моделі пов'язана з можливістю мати у своєму розпорядженні достатньо достовірні початкові дані. Якщо вони є результатом вимірювань, то точність їх вимірювання повинна бути вищою, ніж для тих параметрів, які одержані при використанні математичної моделі.

Наочність математичної моделі є її бажаною, але необов'язковою властивістю. Проте використання математичної моделі і її класифікація спрощуються, якщо її складові (наприклад, члени рівнянь) мають зрозумілий змістовний сенс. Це звичайно дозволяє орієнтовно передбачати результати обчислювального експерименту і полегшує контроль їх правильності.

1.1.2 Побудова обчислювальної моделі

Побудова обчислювальної моделі може здійснюватися різними методами, які можна поділити на точні й наближені. Точні методи – це такі, які після скінченної кількості дій (обчислень) приводять до точного результату за умови, що обчислення здійснюються без похибок. Наближеними називають такі методи, які за тих же умов дозволяють одержати результат лише з деякою похибкою.

При використанні точних методів етап дослідження математичної моделі поділяється на такі підетапи:

1) відшукання точного розв'язку математичної моделі;

2) підставлення вихідних даних у знайдений точний розв'язок і реалізація передбачених ним обчислень.

Дослідження математичної моделі наближеними методами поділяється на такі етапи:

1) обрання обчислювального методу (як правило, наближених чисельних методів буває декілька);

2) вивчення або складання алгоритму методу;

3) реалізація алгоритму за допомогою обчислювальних засобів.

При виборі чисельного методу суттєвими є об'єм обчислень, швидкість збіжності обчислень (як швидко одержується результат) та інші чинники. Зокрема, обрання методу залежить і від вхідних даних.

Крім того, на вибір методу впливають засоби його реалізації (ручний розрахунок, наявність обчислювальної машини, наявність готової програми тощо). Так, якщо будуть використані швидкодіюча ЕОМ і готова програма, то об'єм обчислень не повинен бути визначальним фактором при обранні методу. При ручному ж розрахунку необхідно віддати перевагу методу, який, можливо, потребує деяких певних попередніх досліджень і перетворень математичної моделі, але завдяки цьому потребує й значно меншої кількості обчислень.

1.1.3 Алгоритм методу

Алгоритмом методу називається система правил, яка задає точно визначену послідовність операцій, що приводить до необхідного результату (точного або наближеного).

Алгоритм – одне із ґрунтовних понять математики. Хід розв'язання обчислювальної (і взагалі будь–якої) задачі має бути поданий через алгоритм.

Алгоритм можна записати словесно–формульно, або у вигляді схеми.

При виконанні алгоритму перехід від однієї дії до іншої здійснюється строго у порядку їхнього запису. Якщо ж потрібно перервати природний хід дій за деякої умови, необхідно вказувати на це (див. п. 3 наведеного алгоритму).

Структурною схемою алгоритму називають графічне зображення послідовності дій обчислювального процесу.

У схемі кожна дія розміщується у певному геометричному символі (фігурі). Послідовність дій зазначається на схемі напрямком стрілок на лініях, якими з'єднують ці символи. Як правило, прийнято *початок і кінець* обчислень зображувати *овалами*, *введення* даних і *виведення* результатів – у вигляді *паралелограма*. *Обчислювальні операції* розміщуються у *прямокутниках*, а операція *перевірки* деякої *умови* зображується у вигляді *ромба*. Усередині кожної фігури розміщується стислий формульний опис відповідної операції.

Символи операцій перевірки умови мають два виходи: "так" і "ні". Стрілка на лінії, що виходить із виходу "так" вказує на операцію, до виконання якої потрібно перейти, якщо умову, яка перевіряється, виконано. Стрілка з написом "ні" вказує на операцію, до виконання якої необхідно перейти у випадку, коли умову не виконано.

На рис. 1.2. подані елементи блок–схеми алгоритму обчислень. Фігури з'єднуються лініями зі стрілками, які вказують на операцію, до виконання якої необхідно перейти.

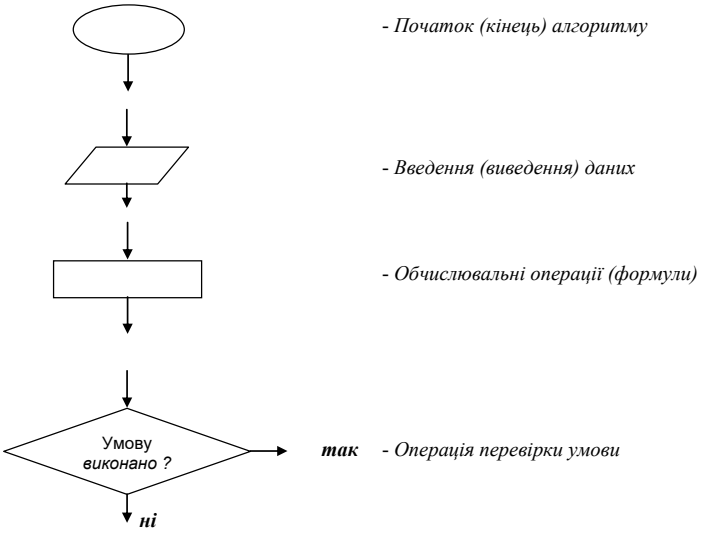

Рисунок 1.2 – Елементи блок–схеми алгоритму

1.1.4. Реалізація методу обчислень

Обчислення за алгоритмами відбувається за допомогою різних обчислювальних засобів.

Суттєвим є *контроль обчислень*, який проводять за так званим *контрольним прикладом* (тестом). Результат контрольного прикладу має бути заздалегідь відомим, тобто він або є очевидним, або його відшукують яким–небудь іншим способом. При ручному рахунку контроль рекомендується проводити поетапно. При розрахунках на ЕОМ за складеною програмою контрольний приклад заздалегідь прораховують вручну, а потім звіряють поетапно результати розрахунків із здійснюваними машиною.

1.2 Чисельне розв'язання нелінійних рівнянь

Будь–яке рівняння з одним невідомим можна записати у вигляді $f(x) = 0$. Його розв'язком називається таке значення x^* (корінь рівняння), для якого $f(x^*) = 0$. Алгоритми знаходження точного значення коре-

нів відомі тільки для вузького класу рівнянь. Тому більшість їх можливо розв'язати лише наближеними чисельними методами.

Задача знаходження наближеного значення кореня передбачає два етапи:

1) відділення коренів – визначення відрізка $[a,b]$ з області визначення функції $y = f(x)$, де знаходиться тільки один корінь;

2) уточнення наближених коренів, тобто обчислення їх із заданою точністю.

Для кожного з цих етапів розроблені свої чисельні методи [2–4].

1.2.1 Метод половинного поділу для аналітичного відділення кореня рівняння і пошуку його наближення

Нехай функція $f(x)$ – неперервна на відрізку $[a,b]$, на кінцях його набуває значень різних знаків, тобто $f(a)f(b) < 0$, похідна $f'(x)$ зберігає на цьому відрізку знак (рис. 1.3). Отже, усередині цього відрізка міститься один корінь. Ділимо відрізок $[a,b]$ навпіл, знаходимо його середину $c = (a+b)/2$. Якщо $f(c) = 0$, то корінь $x^* = c$. Інакше, позначимо через $[a_1,b_1]$ ту половину відрізка $[a,b]$, на кінцях якої функція набуває значень різних знаків. Процес послідовного поділу продовжуємо до того часу, поки на n – кроці не буде виконуватися одна з умов:

1) $f((a_n + b_n)/2) = 0$, тоді $x^* = (a_n + b_n)/2$ – шуканий корінь;

2) довжина відрізка, що містить корінь, стане менше ε, де ε – задана точність обчислень, тобто $|b_n - a_n| < 2\varepsilon$, а $x^* = (a_n + b_n)/2$.

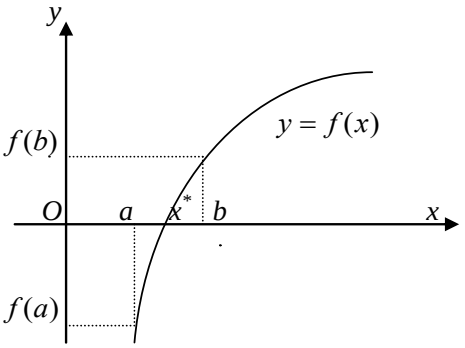

Рисунок 1.3 – Ілюстрація методу половинного поділу

1.2.2 Метод простих ітерацій

Для використання методу простих ітерацій (послідовних наближень) замінимо рівняння $f(x)=0$ еквівалентним йому рівнянням

$$x=\varphi(x)\,. \tag{1.1}$$

Виберемо деяке наближення $x_0 \in [a,b]$ кореня і підставимо його у праву частину рівняння (1.1). Одержимо $x_1=\varphi(x_0)$. Далі обчислюємо за формулою

$$x_n=\varphi(x_{n-1}), n=2,3,... \tag{1.2}$$

Отримуємо послідовність наближень $\{x_n\}$ до кореня, що у випадку її збіжності до кореня x^* може дати наближене його значення із заданою точністю ε. Необхідною і достатньою умовою існування границі послідовності є вимога: $\forall \varepsilon>0,\ N$ такий, що $\forall n>N \Rightarrow |x_n-x_{n-1}|<\varepsilon$. З цієї причини шукаємо наближення (ітерації), які б задовольняли вищезазначену умову.

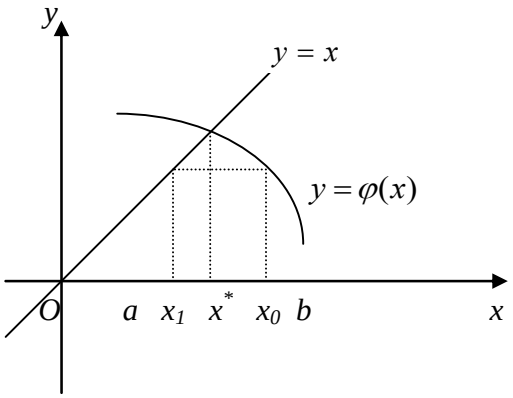

Рисунок 1.4 – Ілюстрація методу простих ітерацій

Перейти від рівняння $f(x)=0$ до еквівалентного йому $x=\varphi(x)$ можна багатьма способами. Але оптимальним є той, що задовольнить достатню умову збіжності методу простої ітерації $\forall x \in [a,b]\ |\varphi'(x)|<1$.

При виконанні умови збіжності за початкове наближення x_0 можна взяти довільне значення з інтервалу $[a,b]$.

1.2.3 Метод Ньютона

Метод Ньютона (метод дотичних), який використовується для наближеного розв'язку рівняння $f(x)=0$, полягає в побудові ітераційної послідовності $\{x_n\}, n=0,1,2,...$, що збігається до кореня рівняння на відрізку $[a,b]$ його локалізації.

На рисунку 1.5 зображено спосіб отримання першого наближення за методом дотичних: x_1 є точка перетину дотичної, проведеної до кривої в точці з координатами $(x_0, f(x_0))$. З прямокутного трикутника, гострий кут якого α, маємо

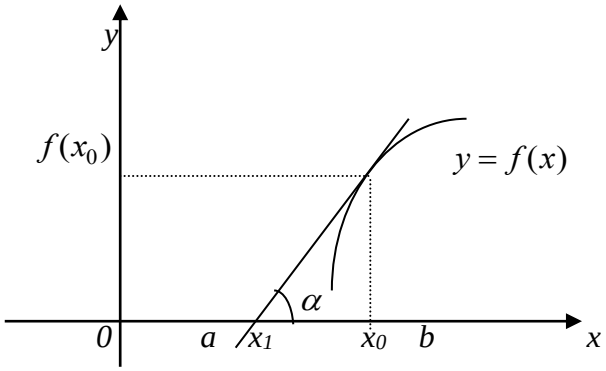

Рисунок 1.5 – Ілюстрація методу дотичних

$$tg\alpha = f'(x_0) = \frac{f(x_0)}{x_0 - x_1},\ \text{звідки}\ x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Достатні умови збіжності такі. Нехай $f(x)$ – визначена і двічі диференційована на $[a,b]$, причому похідні $f'(x), f''(x)$ зберігають знак на $[a,b]$. Тоді , виходячи з початкового наближення $x_0 \in [a,b]$, що задовольняє нерівність $f(x_0)f''(x_0) > 0$, ітераційна послідовність

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)},\ n=0,1,2,... \qquad (1.3)$$

збігається до єдиного на $[a,b]$ розв'язку ξ рівняння $f(x)=0$.

Для оцінки похибки *n*–го наближення кореня можна скористатися нерівністю

$$\left|\xi - x_n\right| \le \frac{\mathrm{M}}{2m}\left|x_n - x_{n-1}\right|^2, \tag{1.4}$$

де M – найбільше значення модуля другої похідної $\left|f''(x)\right|$ на $[a,b]$; m – найменше значення модуля першої похідної $\left|f'(x)\right|$ на $[a,b]$.

За необхідності обчислити корінь рівняння з точністю ε ітераційну послідовність переривають за умови

$$\left|x_n - x_{n-1}\right| < \sqrt{2m\varepsilon / \mathrm{M}} = \varepsilon_0 \tag{1.5}$$

і беруть x_n за наближене значення кореня ξ.

Метод Ньютона ефективний, якщо вибрано вдале початкове наближення для кореня і навколо кореня графік функції має велику кривину.

1.3 Розв'язання систем лінійних алгебраїчних рівнянь (СЛАР)

Розглянемо системи вигляду

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = c_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = c_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = c_n. \end{cases} \tag{1.6}$$

Її матричний вигляд

$$AX = C. \tag{1.7}$$

Тут A – $\{[\, a_{ij}\,], (i,j = \overline{1,n})\}$ – матриця коефіцієнтів системи,

$$C = \begin{bmatrix} c_1 \\ c_2 \\ \ldots \\ c_n \end{bmatrix}, \quad X = \begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{bmatrix} \text{ – вектори–стовпці.}$$

Методи чисельного розв'язання систем лінійних алгебраїчних рівнянь (СЛАР) поділяються на точні і наближені. Метод вважають точним, якщо, нехтуючи похибками округлення, він дає точний результат після виконання певної кількості обчислювальних операцій. Математичні пакети прикладних програм для ПЕОМ містять стандартні процедури розв'язання СЛАР такими поширеними точними методами, як метод Гаусса. До наближених методів розв'язання СЛАР відносять метод простої

ітерації та метод Зейделя. Вони дозволяють отримати послідовність $\{X^k\}$ наближень до розв'язку X^*, таку, що $\lim_{k\to\infty} X^k = X^*$.

Ітераційні методи прості, легко програмуються і мають малу похибку округлення, але вони дають збіжну послідовність наближень тільки при виконанні певної умови. Це достатня умова збіжності ітераційного процесу: модулі діагональних коефіцієнтів для кожного рівняння системи не повинні бути менше, ніж сума модулів усіх інших коефіцієнтів при невідомих

$$|a_{ii}| \geq \sum_{i \neq j} |a_{ij}|. \qquad (1.8)$$

1.3.1 Метод простої ітерації

Розглянемо СЛАР у матричному вигляді (1.7) (діагональні коефіцієнти a_{ii} не дорівнюють нулю для всіх i). Приведемо її до вигляду $X=BX+D$, де $B=[b_{ij}]$ – квадратна матриця порядку n:

$$b_{ii} = 0; \quad b_{ij} = \frac{-a_{ij}}{a_{ii}}, \quad d_i = \frac{c_i}{a_{ii}}, \quad i, j = \overline{1,n}.$$

При цьому СЛАР (1.7) набуде вигляду

$$\begin{cases} x_1 = b_{12}x_2 + b_{13}x_3 + \ldots + b_{1n}x_n + d_1, \\ x_2 = b_{21}x_1 + b_{23}x_3 + \ldots + b_{2n}x_n + d_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ x_n = b_{n1}x_1 + b_{n2}x_2 + \ldots + b_{nn-1}x_{n-1} + d_n. \end{cases} \qquad (1.9)$$

Взявши довільне початкове наближення $X^0 = col(x_1^0, \ldots x_n^0)$, будуємо ітераційний процес за формулою $X^k = BX^{k-1} + D$ ($k=1,2,\ldots$).

1.3.2 Метод Зейделя

Ітераційний процес Зейделя відрізняється від методу простої ітерації тим, що при розв'язуванні систем вигляду $X=BX+D$ обчислення наступного наближеного значення x_i при $1<i<n$ використовується при обчисленні раніше невідомих наближених значень $x_1, x_2, \ldots, x_{i-1}$.

Розглянемо тепер систему $AX=C$ (1.7) з n рівнянь із n невідомими, як і раніше припускаючи, що діагональні коефіцієнти a_{ii} не дорівнюють нулю для всіх i. Перетворимо вихідну систему вигляду $AX=C$ до вигляду $X=BX+D$, де $B=[b_{ij}]$ – квадратна матриця порядку n:

$$b_{ii}=0;\quad b_{ij}=\frac{-a_{ij}}{a_{ii}},\quad d_i=\frac{c_i}{a_{ii}},\quad i,j=\overline{1,n}.$$

У цьому випадку ітераційний процес методу Зейделя має вигляд

$$\begin{cases} x_1^{k+1}=\sum\limits_{j=2}^{n} b_{1j}\cdot x_j^k+d_1, \\ x_2^{k+1}=b_{21}\cdot x_1^{k+1}+\sum\limits_{j=3}^{n} b_{2j}\cdot x_j^k+d_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ x_n^{k+1}=\sum\limits_{j=1}^{n-1} b_{nj}\cdot x_j^{k+1}+d_n. \end{cases} \qquad (1.10)$$

Часто в практичних обчисленнях ітераційний процес припиняють, якщо два послідовних наближення відрізняються менше від наперед заданого числа Δ у змісті обраної норми:

$$\left\| X^{k+1}-X^k \right\|<\Delta,$$

тобто коли наближені розв'язки x_i^{k+1} і x_i^k $(i=\overline{1,n})$ стануть досить близькими і $\max\limits_i \left|x_i^{k+1}-x_i^k\right|<\Delta$. Величина Δ пов'язана з точністю ε розв'язання системи таким співвідношенням: $\Delta=\dfrac{1-\|B\|}{\|B\|}\cdot\varepsilon$, де $\|B\|$ – норма матриці з коефіцієнтів при невідомих у правих частинах рівнянь перетвореної системи*: X=BX+D.* Норма матриці може бути визначена по–різному, наприклад, як

$$\|B\|=\max_i \sum_{j=1}^{n}\left|b_{ij}\right|,\ \text{або}\ \|B\|=\max_j \sum_{i=1}^{n}\left|b_{ij}\right|.$$

1.4 Розв'язування систем лінійних алгебраїчних рівнянь (СЛАР)

Система лінійних алгебраїчних рівнянь (СЛАР) n –го порядку має вигляд

$$\begin{cases} a_{11}x_1+a_{12}x_2+\ldots a_{1n}x_n=b_1 \\ a_{21}x_1+a_{22}x_2+\ldots a_{2n}x_n=b_2 \\ \ldots\ldots\ldots=\ldots \\ a_{n1}x_1+a_{n2}x_2+\ldots a_{nn}x_n=b_n \end{cases} \qquad (1.11)$$

Тут позначено x_i $(i=1,2,...,n)$ – деякі змінні; $a_{ij}(i,j=1,2,...,n)$ – коефіцієнти при змінних; $b_i(i=1,2,...,n)$ – так звані "вільні" члени.

Під розв'язуванням СЛАР розуміємо відшукання таких значень змінних x_i, підстановка яких у кожне з n рівнянь перетворює їх одночасно у тотожності.

Якщо використати матричні позначення

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix};\quad B=\begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix};\quad X=\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}, \tag{1.12}$$

то система рівнянь (1.11) може бути поданою у матричній формі у такий спосіб:

$$A\cdot X=B. \tag{1.13}$$

1.4.1 Метод Крамера

Найбільш відомий у математиці метод розв'язання СЛАР – метод Крамера – дозволяє побудувати повне аналітичне розв'язання. Це розв'язання має вигляд

$$x_i=\frac{\det(A_i)}{\det(A)};\quad (i=1,2,...,n), \tag{1.14}$$

де $\det(A)$ – позначення визначника матриці A, а $\det(A_i)$ – визначник матриці

$$A_i=\begin{bmatrix} a_{11} & b_1 & \dots & a_{1n} \\ a_{21} & b_2 & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & b_n & \dots & a_{nn} \end{bmatrix}=\begin{bmatrix} a_{11} & \dots & a_{1(i-1)} & b_1 & a_{1(i+1)} & \dots & a_{1n} \\ a_{21} & \dots & a_{2(i-1)} & b_2 & a_{2(i+1)} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{n(i-1)} & b_n & a_{n(i+1)} & \dots & a_{nn} \end{bmatrix}, \tag{1.15}$$

яка виходить з матриці A коефіцієнтів шляхом заміни в ній i–го стовпця на стовпець B з вільних членів.

Як бачимо, розв'язування тісно пов'язане з обчисленням визначників. Усього для відшукання повного розв'язання потрібно обчислити $n+1$ визначників n–го порядку і додатково здійснити ще n ділень відшуканих значень визначників. Обчислення визначника за відомим правилом розкладання на суму алгебраїчних доповнень є вельми марнотратним у розумінні кількості операцій, оскільки потребує $2\cdot n!$ операцій множення. Усього ж для відшукання розв'язку потрібно, таким чином, $n+2\cdot(n+1)!$ операцій множення–ділення.

1.4.2 Метод Гаусса та його модифікації

Усі чисельні методи розв'язування СЛАР спираються на деякі перетворення вихідної системи рівнянь, які, з одного боку, не змінюють шуканого розв'язання, а, з іншого – значно спрощують подальше його відшукування. Звичайно, при цьому домагаються, щоб загальна кількість операцій разом із попередніми перетвореннями була якомога меншою у порівнянні з методом Крамера.

Розглянемо, які ж саме перетворення рівнянь (1.11) не змінюють розв'язків цієї системи. Неважко зрозуміти, що до таких перетворень належать:

1) ділення будь–якого з рівнянь на будь–яке число s, що не дорівнює нулю; така операція не змінює розв'язків СЛАР, але змінює визначник матриці, який зменшується у стільки ж разів, тобто визначник початкової матриці A дорівнюватиме значенню визначника нової матриці коефіцієнтів, помноженому на s;

2) додавання до будь–якого рівняння системи (1.11) будь–якого іншого рівняння тієї ж системи, помноженого на будь–яке число, що не дорівнює нулю;

3) переставлення місцями двох яких–небудь рівнянь; ця операція призводить до зміни знака визначника системи.

Якщо ввести у розгляд так звану "розширену матрицю" коефіцієнтів рівнянь системи

$$A^* = \begin{bmatrix} A & B \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n}b_1 \\ a_{21} & a_{22} & \dots & a_{2n}b_2 \\ \dots & \dots & \dots & \dots\dots \\ a_{n1} & a_{n2} & \dots & a_{nn}b_n \end{bmatrix}, \qquad (1.16)$$

яка матиме розміри $n*(n+1)$, то вищезгадані перетворення можна інтерпретувати як перетворення саме розширеної матриці коефіцієнтів СЛАР. До таких допустимих перетворень цієї матриці можна, як це було зазначено, віднести:

1) множення будь–якого рядка матриці на число, що не дорівнює нулю;

2) підсумовування будь–якого рядка матриці з іншим рядком цієї матриці, помноженим на число, що не дорівнює нулю;

3) переставлення місцями двох будь–яких рядків розширеної матриці; при цьому знак головного визначника матриці A змінюється на протилежний;

4) переставляння місцями двох будь–яких стовпців матриці A; ця операція є еквівалентною лише змінюванню позначень шуканих змінних; при цьому також змінюється знак визначника матриці A.

1.4.3 Схема єдиного ділення

Найпростіший варіант методу Гаусса називається схемою єдиного ділення. Розглянемо його докладніше.

Схема єдиного ділення складається із двох етапів. На першому з них (його називають прямим ходом) вихідні рівняння (1.11) перетворюються таким чином, що з наступних рівнянь вилучаються усі попередні змінні, тобто із другого і подальших рівнянь вилучається змінна x_1, з третього і подальших – змінна x_2 і так далі. У результаті таких дій випливає, що останнє рівняння міститиме лише одну змінну – x_n, передостаннє – дві змінні – x_n і x_{n-1} і так далі у порядку зростання кількості змінних.

На другому етапі (який називається зворотним ходом) визначаються шукані розв'язки СЛАР. Значення змінної x_n визначається безпосередньо з останнього одержаного рівняння, значення x_{n-1} – з передостаннього (із урахуванням відшуканого значення x_n) і так далі.

Розглянемо докладніше прямий хід.

Припускаючи, що a_{11} не дорівнює нулю, поділимо на нього перше рівняння (1.11). Одержимо перше рівняння у вигляді

$$x_1 + a_{12}^{(1)} \cdot x_2 + a_{13}^{(1)} \cdot x_3 + a_{14}^{(1)} \cdot x_4 + \ldots + a_{1n}^{(1)} \cdot x_n = b_1^{(1)},$$

де використане позначення

$$a_{1i}^{(1)} = \frac{a_{1i}}{a_{11}}, \qquad (i = 2,3,\ldots,n); \qquad b_1^{(1)} = \frac{b_1}{a_{11}},$$

а індекс угорі позначає номер кроку прямого ходу.

Тепер виключимо із другого рівняння (1.11) змінну x_1. Для цього помножимо перетворене перше рівняння (1.11) на коефіцієнт a_{21} і віднімемо його від другого рівняння (1.11). Одержимо друге рівняння у вигляді

$$a_{22}^{(1)} \cdot x_2 + a_{23}^{(1)} \cdot x_3 + a_{24}^{(1)} \cdot x_4 + \ldots + a_{2n}^{(1)} \cdot x_n = b_2^{(1)},$$

де позначено

$$a_{2i}^{(1)} = a_{2i} - a_{21} \cdot a_{1i}^{(1)}, \qquad (i = 2,3,\ldots,n);$$

$$b_2^{(1)} = b_2 - a_{21} \cdot b_1^{(1)}.$$

Так само перетворюються усі подальші рівняння. Після цього вони набудуть вигляду (k –номер рівняння):

$$a_{k2}^{(1)} \cdot x_2 + a_{k3}^{(1)} \cdot x_3 + a_{k4}^{(1)} \cdot x_4 + \ldots + a_{kn}^{(1)} \cdot x_n = b_k^{(1)},$$

причому

$$a_{ki}^{(1)} = a_{ki} - a_{k1} \cdot a_{1i}^{(1)}, \ (i = 2,3,\ldots,n); \ b_k^{(1)} = b_k - a_{k1} \cdot b_1^{(1)}. \qquad (1.17)$$

У результаті першого кроку прямого ходу система рівнянь (1.11) набуває вигляду (1.18). При цьому усі рівняння системи, починаючи із другого, матимуть на одну змінну менше за вихідну систему (1.11), тобто у сукупності утворюють СЛАР ($n-1$)–го порядку.

$$\begin{cases} x_1 + a_{12}^{(1)}\ x_2 + \ldots + a_{1n}^{(1)} x_n = b_1^{(1)}, \\ a_{22}^{(1)} x_2 + \ldots + a_{2n}^{(1)} x_n = b_2^{(1)}, \\ \ldots\ldots\ldots = \ldots, \\ a_{n2}^{(1)} x_2 + \ldots + a_{nn}^{(1)} x_n = b_n^{(1)}. \end{cases} \qquad (1.18)$$

Другий крок прямого ходу методу Гаусса полягає у аналогічному перетворенні СЛАР ($n-1$)–го порядку, яку складають одержані рівняння (1.18) з другого по останнє. У результаті виходить така система рівнянь:

$$\begin{cases} x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 + \ldots a_{1n}^{(1)} x_n = b_1^{(1)}, \\ x_2 + a_{23}^{(2)} x_3 + \ldots a_{2n}^{(2)} x_n = b_2^{(2)}, \\ \ldots\ldots\ldots = \ldots, \\ a_{n3}^{(2)} x_3 + \ldots a_{nn}^{(2)} x_n = b_n^{(2)}. \end{cases} \qquad (1.19)$$

Коефіцієнти визначаються аналогічно:

$$a_{2i}^{(2)} = \frac{a_{2i}^{(1)}}{a_{22}^{(1)}}, \qquad b_2^{(2)} = \frac{b_2^{(1)}}{a_{22}^{(1)}},$$

$$a_{ki}^{(2)} = a_{ki}^{(1)} - a_{k2}^{(1)} \cdot a_{2i}^{(2)},$$

$$b_k^{(2)} = b_k^{(1)} - a_{k2}^{(1)} \cdot b_2^{(2)}; \ (k, i = 3,4,\ldots,n). \qquad (1.20)$$

У такий само спосіб здійснюються подальші кроки прямого ходу. Формули перетворення на m –му кроці визначаються співвідношеннями:

$$a_{mi}^{(m)} = \frac{a_{mi}^{(m-1)}}{a_{mm}^{(m-1)}}, \quad b_m^{(m)} = \frac{b_m^{(m-1)}}{a_{mm}^{(m-1)}}, \quad a_{ki}^{(m)} = a_{ki}^{(m-1)} - a_{k2}^{(m-1)} \cdot a_{mi}^{(m)},$$

$$b_k^{(m)} = b_k^{(m-1)} - a_{km}^{(m-1)} \cdot b_m^{(m)}; \ (k, i = m+1, m+2,\ldots,n). \qquad (1.21)$$

У підсумку за $m-1$ –м кроком утворюється така система рівнянь:

$$\begin{cases} x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 + a_{14}^{(1)} x_4 + \ldots a_{1n}^{(1)} x_n = b_1^{(1)}, \\ x_2 + a_{23}^{(2)} x_3 + a_{24}^{(2)} x_4 + \ldots a_{2n}^{(2)} x_n = b_2^{(2)}, \\ x_3 + a_{34}^{(3)} x_4 + \ldots a_{3n}^{(3)} x_n = b_2^{(3)}, \\ \ldots\ldots = \ldots, \\ x_n = b_n^{(n)}. \end{cases} \qquad (1.22)$$

Матриця $A_{(n-1)}$ коефіцієнтів цієї системи є верхньою трикутною матрицею з одиничними елементами вдовж головної діагоналі:

$$A_{(n-1)} = \begin{bmatrix} 1 & a_{12}^{(1)} & a_{13}^{(1)} & a_{14}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & 1 & a_{23}^{(2)} & a_{24}^{(2)} & \dots & a_{2n}^{(2)} \\ 0 & 0 & 1 & a_{34}^{(3)} & \dots & a_{3n}^{(3)} \\ 0 & 0 & 0 & 1 & \dots & a_{4n}^{(4)} \\ 0 & 0 & 0 & 0 & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}. \qquad (1.23)$$

Відповідна до цієї системи розширена матриця коефіцієнтів має вигляд

$$A_{(n-1)}^{*} = \begin{bmatrix} 1 & a_{12}^{(1)} & a_{13}^{(1)} & a_{14}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & 1 & a_{23}^{(2)} & a_{24}^{(2)} & \dots & a_{2n}^{(2)} & b_2^{(2)} \\ 0 & 0 & 1 & a_{34}^{(3)} & \dots & a_{3n}^{(3)} & b_3^{(3)} \\ 0 & 0 & 0 & 1 & \dots & a_{4n}^{(4)} & b_4^{(4)} \\ 0 & 0 & 0 & 0 & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 & b_n^{(n)} \end{bmatrix}. \qquad (1.24)$$

Підсумовуючи, можна сказати, що основною метою прямого ходу методу Гаусса є перетворення розширеної матриці системи до трикутної форми (14), після чого відшукання розв'язків СЛАР легко здійснюється зворотним ходом за співвідношеннями:

$$x_n = b_n^{(n)};\ x_m = b_m^{(m)} - \sum_{j=m+1}^{n} a_{mj}^{(m)} \cdot x_j\ ,\ (m = n-1, n-2, \dots, 1). \qquad (1.25)$$

Таким чином, прямий хід методу Гаусса зводиться до побудови розширеної матриці системи (1.16) і подальшого її перетворення до верхньої трикутної форми за допомогою таких операцій:

1) ділення елементів першого рядка матриці на перший елемент цього рядка (який міститься на головній діагоналі); цей елемент називається роздільним;

2) віднімання з подальших рядків матриці першого рядка, помноженого на елемент відповідного рядка, що знаходиться у тому зі стовпців, що й роздільний елемент; обнуління елементів, які містяться у стовпці роздільного елемента;

3) повторення цих дій щодо другого рядка, а потім і для усіх подальших рядків нових одержаних матриць.

1.4.4 Метод Гаусса з обранням роздільного елемента

Неважко помітити суттєвий недолік розглянутого методу єдиного ділення, який випливає з того, що однією з необхідних операцій є ділення на роздільний елемент. Якщо на головній діагоналі перед цим виявиться, що роздільний елемент дорівнює нулю, відповідна операція стає неможливою. Якщо ж роздільний елемент виявиться досить малою за абсолютним значенням величиною, то хоч операція ділення при цьому є здійсненною, але вона приводить до значної похибки у визначенні подальших коефіцієнтів – членів розширеної матриці. Тому необхідно здійснювати заходи з уникнення цього явища.

У методі Гаусса з обранням роздільного елемента цього досягають у такий спосіб. Перед діленням рядка на роздільний елемент порівнюють за модулем усі елементи відповідного стовпця матриці, відшукують той рядок (із подальших), де міститься елемент, найбільший за модулем, і змінюють місцями поточний рядок із знайденим рядком із максимальним елементом. Таким чином домагаються, щоб роздільний елемент був якомога більшим за модулем.

При обміні місцями рядків визначник матриці A коефіцієнтів змінює свій знак на протилежний. Тому якщо завданням є обчислення у подальшому визначника цієї матриці, необхідно лічити кількість p таких обмінів рядків місцями. Тоді визначник початкової матриці A може бути обчислений як значення визначника перетвореної матриці $A_{(n-1)}$, помножене на $(-1)^p$.

1.4.5 Метод Гаусса–Жордана (метод повного виключення)

Недоліком методу Гаусса є наявність зворотного ходу. Виникає питання, чи не можна позбутися цього етапу, і на першому ж етапі виключення змінних визначити шукані розв'язки СЛАР.

Виявляється, це є можливим завдяки поширенню виключення змінних не тільки на подальші рівняння, а й на попередні. Цей процес і називають повним виключенням, або методом Гаусса–Жордана.

Перший крок цього методу є таким само, як у методі Гаусса. На другому кроці необхідно ту саму операцію виключення розповсюдити і на перше рівняння, тобто з першого рівняння (1.19) відняти друге, помножене на $a_{12}^{(1)}$. Це є рівносильним до того, що обчислення за формулами (1.20) поширюються й на $k=1$:

$$a_{2i}^{(2)} = \frac{a_{2i}^{(1)}}{a_{22}^{(1)}}, \qquad b_2^{(2)} = \frac{b_2^{(1)}}{a_{22}^{(1)}},$$

$$a_{ki}^{(2)} = a_{ki}^{(1)} - a_{k2}^{(1)} \cdot a_{2i}^{(2)}, \qquad b_k^{(2)} = b_k^{(1)} - a_{k2}^{(1)} \cdot b_2^{(2)};$$
$$(i = 3,4,\ldots,n; \qquad k = 1,3,4,\ldots,n). \tag{1.26}$$

Внаслідок цього після другого кроку матимемо замість системи (1.19) іншу:

$$\begin{cases} x_1 + 0 + a_{13}^{(2)} x_3 + \ldots a_{1n}^{(2)} x_n = b_1^{(2)}, \\ x_2 + a_{23}^{(2)} x_3 + \ldots a_{2n}^{(2)} x_n = b_2^{(2)}, \\ \ldots\ldots = \ldots, \\ a_{n3}^{(2)} x_3 + \ldots a_{nn}^{(2)} x_n = b_n^{(2)}. \end{cases} \tag{1.27}$$

Аналогічно, на m–му кроці треба здійснювати обчислення за фор–мулами (1.11), поширюючи їх на усі попередні рівняння:

$$a_{mi}^{(m)} = \frac{a_{mi}^{(m-1)}}{a_{mm}^{(m-1)}}, \quad b_m^{(m)} = \frac{b_m^{(m-1)}}{a_{mm}^{(m-1)}},$$
$$a_{ki}^{(m)} = a_{ki}^{(m-1)} - a_{k2}^{(m-1)} \cdot a_{mi}^{(m)}, \qquad b_k^{(m)} = b_k^{(m-1)} - a_{km}^{(m-1)} \cdot b_m^{(m)};$$
$$(i = m+1, m+2,\ldots,n; \qquad k = 1,2,\ldots,m-1, m+1, m+2,\ldots,n). \tag{1.28}$$

У підсумку випливає "рівняння" вигляду

$$\begin{cases} x_1 = b_1^{(n)} \\ x_2 = b_2^{(n)} \\ \ldots\ldots = \ldots \\ x_n = b_n^{(n)} \end{cases} \tag{1.29}$$

які вже, очевидно, є просто розв'язками початкової СЛАР.

Неважко зрозуміти, що виграш у вигляді зникнення операцій зворотного ходу у цьому методі відбувається за рахунок збільшення кількості операцій у прямому ході.

1.5 Розв'язання систем нелінійних рівнянь

Розглянемо систему нелінійних рівнянь

$$\begin{cases} f_1(x_1, x_2,\ldots,x_n) = 0, \\ f_2(x_1, x_2,\ldots,x_n) = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots, \\ f_n(x_1, x_2,\ldots,x_n) = 0. \end{cases} \tag{1.30}$$

Для її розв'язання можна застосувати ітераційні методи. Вони дозволяють отримати послідовність наближень $X^{(k)} = (x_1^k, x_2^k,\ldots,x_n^k)$

$k = 0,1,2,\ldots$. Якщо ітераційний процес збігається, то граничне значення $X^* = \lim\limits_{k\to\infty} X^{(k)}$ є розв'язком заданої системи рівнянь.

Обов'язковою умовою збігу є виконання певних вимог до системи, що розглядається.

1.5.1 Метод простої ітерації

Систему (1.30) наведемо у вигляді

$$\begin{cases} x_1 = \varphi_1(x_1, x_2, \ldots, x_n), \\ x_2 = \varphi_2(x_1, x_2, \ldots, x_n), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ x_n = \varphi_n(x_1, x_2, \ldots, x_n) \end{cases} \tag{1.31}$$

або у векторному вигляді $X = \phi(X)$. Причому перехід від системи (1.30) до (1.31) має відбутися тільки за умови, щоб

$$\overline{\varphi(X)} = \begin{bmatrix} \varphi_1(x_1, x_2, \ldots, x_n) \\ \varphi_2(x_1, x_2, \ldots, x_n) \\ \ldots\ldots\ldots\ldots\ldots \\ \varphi_n(x_1, x_2, \ldots, x_n) \end{bmatrix}$$

виявилося стискуючим відображенням. Необхідно зазначити, що загального способу для переходу від (1.30) до (1.31) не існує.

Ітераційна послідовність будується за формулою

$$X^{(k+1)} = \phi(X^{(k)}), \; k = 0,1,2,\ldots, \tag{1.32}$$

де $X^{(0)}$ – початкове наближення, яке має бути задано.

Достатньою умовою збіжності ітераційного процесу є виконання умови

$$\|M\| \le q < 1 \quad , \tag{1.33}$$

де M – матриця з елементами $m_{ij} = \dfrac{\partial \varphi_i}{\partial x_j}$ ($\|M\| = \max\limits_j \sum\limits_{i=1}^{n} \left| \dfrac{\partial \varphi_i(x)}{\partial x_j} \right|$ – норма матриці M) для довільного X із області визначення розв'язку. Відображення $\overline{\varphi(X)}$ називають стискуючим, якщо для двох довільних елементів X_1 та X_2 справедливе

$$\left\| \overline{\varphi(X_1)} - \overline{\varphi(X_2)} \right\| \le q \| X_1 - X_2 \|,$$

де коефіцієнт стискання q задовольняє нерівність $0 < q < 1$.

1.5.2 Метод Ньютона для нелінійних систем

При його використанні для нелінійних систем припускається, що в деякій області G, яка містить розв'язок $X^* = (x_1^*, x_2^*, ..., x_n^*)$ системи (1.30), функції $f_i(X)$ мають неперервні похідні першого порядку, і в деякому околі точки X^* матриця Якобі $F(X)$ невироджена:

$$F(X) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}. \qquad (1.34)$$

Ітераційний процес будується за формулою

$$X^{(k+1)} = X^{(k)} - F^{-1}(X^{(k)})\overline{f}(X^{(k)}) \quad (k = 0,1,2,...), \qquad (1.35)$$

де $\overline{f}(X^{(k)}) = (f_1(X^{(k)}), f_2(X^{(k)}), ..., f_n(X^{(k)}))$, а $F^{-1}(X^{(k)})$ – обернена матриця до матриці Якобі. За початкове наближення необхідно брати вектор розв'язку $X^{(0)}$, достатньо близький до шуканого розв'язку X^* системи.

Через потребу на кожному кроці розраховувати обернену матрицю обчислювальна формула (1.35) виявляється досить громіздкою. Замість неї зручніше скористатися такою схемою:

$$\begin{aligned} F(X^{(k)})\overline{\varepsilon^{(k)}} &= -\overline{f}(X^{(k)}); \\ X^{(k+1)} &= X^{(k)} + \overline{\varepsilon^{(k)}}. \end{aligned} \qquad (1.36)$$

При її реалізації для кожного наближення розв'язується система лінійних рівнянь з матрицею $F(X^{(k)})$, а потім за знайденим приростом $\overline{\varepsilon^{(k)}}$ відшукується наступне наближення $X^{(k+1)}$.

Недоліком методу Ньютона є те, що при невдалому виборі наближення $X^{(0)}$ ітераційна послідовність не має границі.

1.6 Інтерполяція функцій

1.6.1 Поставлення задачі інтерполяції

На відрізку $[a,b]$ задано N точок $x_1, x_2, ..., x_N$, що називаються вузлами інтерполяції, і значення деякої функції $y = f(x)$ в цих точках: $f(x_1) = y_1; f(x_2) = y_2; ...; f(x_N) = y_N$. Потрібно побудувати функцію $F(x)$ (функцію , що інтерполює), яка б збігалася з $f(x)$ у вузлах інтерполяції і наближала її між ними, тобто таку, що

$F(x_1)=y_1; F(x_2)=y_2;...;F(x_N)=y_N$. Геометрична інтерпретація задачі інтерполяції полягає в тому, що потрібно знайти таку криву $y=F(x)$ певного типу, що проходить через задану систему точок $M_i(x_i,y_i),(i=1,2,..,N)$. За допомогою цієї кривої можна знайти наближене значення $y=f(x^*)\approx F(x^*)$, де $x^*\in[x_i,x_{i+1}], i=\overline{1,n-1}$. Задача інтерполяції стає однозначною, якщо замість довільної функції $F(x)$ шукати многочлен $P_{N-1}(x)$ степеня не вище N, що задовольняє умови

$$P_{N-1}(x_1)=y_1; P_{N-1}(x_2)=y_2;...;P_{N-1}(x_N)=y_N.$$

Інтерполяційний многочлен $P_{N-1}(x)$ завжди однозначний, оскільки існує тільки один многочлен степеня $N-1$, що в даних точках набуває заданих значень. Існує декілька способів побудови інтерполяційного многочлена.

1.6.2 Інтерполяційний многочлен Лагранжа

Інтерполяційний многочлен Лагранжа, що набуває у вузлах інтерполяції $x_1, x_2,..., x_N$ відповідно значень $y_1, y_2,..., y_N$, має вигляд

$$P_{N-1}=\frac{(x-x_2)(x-x_3)...(x-x_N)}{(x_1-x_2)(x_1-x_3)...(x_1-x_N)}y_1+\frac{(x-x_1)(x-x_3)..(x-x_N)}{(x_2-x_1)(x_2-x_3)..(x_2-x_N)}y_2+$$
$$+...+\frac{(x-x_1)(x-x_2)..(x-x_{N-1})}{(x_N-x_1)(x_N-x_2)..(x_N-x_{N-1})}y_N. \qquad (1.37)$$

З формули безпосередньо випливає , що ступінь многочлена $P_{N-1}(x)$ дорівнює $N-1$ і многочлен Лагранжа задовольняє всі умови задачі інтерполяції.

Якщо відстань між всіма сусідніми вузлами інтерполювання є однаковою, тобто $x_{i+1}-x_i=h, i=\overline{1,n-1}$, формула (1.37) суттєво спрощується. Введемо нову змінну $t=\frac{x-x_1}{h}$, тоді $x=x_1+ht,\ \ x_i=x_1+ih.$ Інтерполяційний поліном Лагранжа набуде такого вигляду:

$$P_{N-1}(x)=\sum_{i=1}^{N}(-1)^{N-i}C_N^i\frac{t(t-1)...(t-N)}{(t-i)N!}f(x_i). \qquad (1.38)$$

Тут $C_N^i=\frac{N!}{i!(N-i)!}$. Коефіцієнти, що стоять перед величинами $f(x_i), i=\overline{1,N}$ у формулі (1.38), не залежать ні від функції $f(x)$, ні від кроку h, а лише від величин i, N. Тому таблицями, що складені для різ-

них значень N, можна скористатися при розв'язуванні різноманітних задач інтерполювання для рівновіддалених вузлів.

Виникає питання, наскільки близько многочлен Лагранжа наближається до функції $f(x)$ в інших точках (не вузлових), тобто наскільки великий залишковий член. На функцію $y = f(x)$ накладають додаткові обмеження. А саме: припускають, що в розглянутій області $a \le x \le b$ зміни x, що містить вузли інтерполяції, функція $f(x)$ має усі похідні $f'(x), f''(x),..., f^{(N)}(x)$ до N –го порядку включно. Тоді оцінка для абсолютної похибки інтерполяційної формули Лагранжа має вигляд

$$\left|R_{N-1}(x)\right| = \left|f(x) - P_{N-1}(x)\right| \le \frac{M_N}{N!}\left|(x-x_1)(x-x_2)...(x-x_N)\right|, \qquad (1.39)$$

де $M_N = \max\left|f^{N}(x)\right|$.

1.6.3 Інтерполяційний поліном Ньютона

Поділеними різницями називають співвідношення вигляду:

– першого порядку

$$f(x_i, x_{i+1}) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}, \ i = 0,1,...,n-1;$$

– другого порядку

$$f(x_i, x_{i+1}, x_{i+2}) = \frac{f(x_{i+1}, x_{i+2}) - f(x_i, x_{i+1})}{x_{i+2} - x_i}, \ i = 0,1,...,n-2; \qquad (1.40)$$

– *n*– го порядку

$$f(x_0, x_1,...,x_n) = \frac{f(x_1, x_2,...,x_n) - f(x_0, x_1,...,x_{n-1})}{x_n - x_0}.$$

З їх допомогою можна побудувати многочлен

$$P_n(x) = f(x_0) + (x-x_0)\cdot f(x_0, x_1) + (x-x_0)(x-x_1)\cdot f(x_0, x_1, x_2) + ... +$$
$$+ (x-x_0)(x-x_1)\ ...\ (x-x_{n-1})\cdot f(x_0, x_1,\ ...\ , x_n). \qquad (1.41)$$

Він називається інтерполяційним поліномом Ньютона для заданої функції. Ця форма запису більш зручна для застосування, оскільки при додаванні до вузлів $x_0, x_1, ..., x_n$ нового x_{n+1} всі обчислені раніше члени залишаються без зміни, а у формулу додається тільки один доданок. При застосуванні ж формули Лагранжа треба робити всі обчислення знову.

Якщо значення функції задані для рівновіддалених значень аргументу $x_0, x_1 = x_0 + h,\ ...,\ x_n = x_0 + n\cdot h$ (постійну величину $h = x_{i+1} - x_i$, $i=0,1,...,n$ називають кроком інтерполяції), то інтерполяційний поліном буде мати вигляд

$$P_n(x) = y_0 + \frac{\Delta y_0}{h}(x - x_0) + \frac{\Delta^2 y_0}{2!h^2}(x - x_0)(x - x_1) + \ldots + \\ + \frac{\Delta^n y_0}{n!h^n}(x - x_0)(x - x_1)\ \ldots\ (x - x_{n-1}). \tag{1.42}$$

Тут $\Delta^k y_0$ – скінченні різниці *к*–го порядку. Вони визначаються за формулою $\Delta^k y_0 = \sum_{i=0}^{k} (-1)^i c_k^i y_{k-i}$, де c_k^i –біноміальні коефіцієнти.

Порівнюючи цю формулу з попередньою, легко встановити, що при $x_i = x_0 + ih \quad (i = \overline{0,n})$ скінченні і поділені різниці пов'язані співвідношенням вигляду

$$f(x_0, x_1, \ldots, x_n) = \frac{\Delta^n y_0}{n!h^n}. \tag{1.43}$$

Для практичного використання формулу (1.42) записують у перетвореному вигляді. Для цього введемо нову змінну величину t, припустивши, що $x = x_0 + ih$, де $t = \frac{x - x_0}{h}$ – кількість кроків h, необхідних для досягнення точки x із точки x_0. Після внесення вказаних величин у вираз для $P_n(x)$ отримаємо першу інтерполяційну формулу Ньютона для інтерполювання вперед, тобто поблизу початку таблиці:

$$P_n(x) = y_0 + \frac{t}{1!}\Delta y_0 + \frac{t(t-1)}{2!}\Delta^2 y_0 + \ldots + \frac{t(t-1)\ldots(t-n+1)}{n!}\Delta^n y_0. \tag{1.44}$$

Припустимо, що точка інтерполяції розташована поблизу кінцевої точки x_n таблиці. У цьому випадку вузли інтерполяції необхідно брати у порядку $x_n, x_n - h, x_n - 2h, \ldots$ Формула Ньютона для інтерполювання назад тоді матиме вигляд

$$P_n(x) = f(x_n) + f(x_n, x_{n-1})(x - x_n) + f(x_n, x_{n-1}, x_{n-2})(x - x_n)(x - x_{n-1}) + \\ + \ldots + f(x_0, x_1, \ldots, x_n)(x - x_n)(x - x_{n-1})\ldots(x - x_1). \tag{1.45}$$

Поділені різниці можна виразити через скінченні різниці, якщо скористатися можливістю переставляти в них аргументи, та співвідношенням (1.43), із яких випливає:

$$f(x_n) = y_n; \qquad f(x_n, x_{n-1}) = f(x_{n-1}, x_n) = \frac{\Delta y_{n-1}}{1!h};$$

$$f(x_n, x_{n-1}, x_{n-2}) = f(x_{n-2}, x_{n-1}, x_n) = \frac{\Delta^2 y_{n-2}}{2!h^2}.$$

Введемо змінну t, припустивши, що $x = x_n + th,$ отримаємо для $f(x) = y(x)$ другу інтерполяційну формулу Ньютона для інтерполювання в кінці таблиці:

$$P_n(x_n + th) = y_n + \frac{t}{1!}\Delta y_{n-1} + \frac{t(t+1)}{2!}\Delta^2 y_{n-2} + \ldots + \frac{t(t+1)\ldots(t+n-1)}{n!}\Delta^n y_0 .$$

Як перша, так і друга інтерполяційні формули Ньютона можуть бути використані для екстраполяції функції, тобто для знаходження значень функції y, значення аргументів x якої лежать поза таблицею. Якщо $x < x_0$ і значення x близьке до x_0, то вигідно використовувати перший інтерполяційний поліном Ньютона, тоді $t = \frac{x - x_n}{h}$ і $t > 0.$ Таким чином, перша інтерполяційна формула Ньютона застосовується для інтерполювання вперед та екстраполювання назад, а друга – навпаки, для інтерполювання назад та екстраполювання вперед.

Зазначимо, що операція екстраполювання, взагалі кажучи, менш точна, ніж операція інтерполювання.

Інтерполяційні формули Ньютона вигідні, оскільки при додаванні m нових вузлів інтерполяції потрібні додаткові обчислення тільки для m нових членів, без зміни старих.

1.6.4 Многочлени Чебишева

Як видно з формули (1.39), похибка заміни функції $y = f(x)$ інтерполяційним многочленом залежить від вибору вузлів інтерполяції $x_1, x_2, \ldots, x_N$. Перш ніж перейти до питання про раціональний вибір вузлів інтерполяції, розглянемо деякі властивості одного з найважливіших й добре вивчених зараз класів спеціальних функцій – многочленів Чебишева першого роду, що часто використовуються для наближення функцій. Многочлен Чебишева $n-$ го степеня визначається за формулою

$$T_n(x) = \frac{2^n n!}{(2n)!}\sqrt{x^2 - 1}\,\frac{d^n}{dx^n}\left((x^2 - 1)^{n-\frac{1}{2}}\right). \qquad (1.46)$$

При $n = 0,1,2,3,4$ з (1.40) отримаємо перші п’ять многочленів першого роду:

$$T_0(x) = 1;\ T_1(x) = x;\ T_2(x) = 2x^2 - 1;$$

$$T_3(x) = 4x^3 - 3x;\ T_4(x) = 8x^4 - 8x^2 + 1.$$

Для визначення многочленів Чебишева часто користуються тригонометричною формою запису

$$T_n(x) = \cos(n \arccos x), |x| \le 1, \qquad (1.47)$$

що приводить до таких же виразів для $T_n(x)$, як і формула (1.40).

Із тотожності

$$\cos(n+1)\theta = 2\cos\theta\cos n\theta - \cos(n-1)\theta$$

при $\theta = \arccos x$ маємо рекурентну формулу

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)\,.$$

Многочлен $T_n(x)$ має n коренів, які можна отримати, розв'язавши рівняння $\cos(n\arccos x) = 0$ або $n\arccos x = \frac{\pi}{2}(2m+1)$;

$$x = \cos\frac{(2m+1)\pi}{2n}, m = 0,1,\ldots,n-1. \qquad (1.48)$$

Як видно з (1.42), всі n коренів, що відповідають значенням $m = 0,1,\ldots,n-1,$ знаходяться на відрізку $[-1,1]$, причому ці точки не рівновіддалені, а згущуються ближче до кінця даного відрізка. З формули (1.41) також очевидно, що на відрізку $[-1,1]$

$$\max\left|T_n(x)\right| = 1. \qquad (1.49)$$

Доведено, що серед усіх можливих n значень x на відрізку $[-1,1]$ корені $x_0^{(T)}, x_1^{(T)},\ldots,x_{n-1}^{(T)}$ многочлена $T_n(x)$ мають ту "чудову" властивість, що для них величина

$$\omega_n(x) = (x-x_0)(x-x_1)\ldots(x-x_{n-1}) = \frac{1}{2^{n-1}}T_n(x) \qquad (1.50)$$

має найменше за абсолютною величиною максимальне значення.

Беручи до уваги (1.43), запишемо

$$\max\left|\omega_n(x)\right| = \frac{1}{2^{n-1}}\,. \qquad (1.50)$$

Виходячи із властивостей коренів многочленів Чебишева першого роду і визначення многочлена Лагранжа $L_n(x)$ n–го степеня на відрізку $[-1,1]$, можна стверджувати, що якщо за n вузлів інтерполювання взяти корені многочлена $T_n(x)$, то максимальне значення похибки на цьому відрізку буде найменшим для всіх можливих варіантів вибору n вузлів інтерполювання. Інтерполяційний многочлен, що має таку властивість, називається многочленом найкращого наближення. Оцінка (6.3) при цьому набуває вигляду

$$\left|f(x) - L_n(x)\right| \le \frac{M_{n+1}}{2^n(n+1)!}\,.$$

Якщо інтерполювання проводиться на довільному відрізку $[a,b]$, то заміною змінної

$$x = \frac{1}{2}((b-a)z + (b+a)), z = \frac{1}{b-a}(2x - b - a)$$

цей відрізок можна звести до відрізка $[-1,1]$. При цьому корені многочлена $T_n(x)$ будуть знаходитися в точках

$$x_m = \frac{1}{2}(b-a)\cos\frac{2m+1}{2n}\pi + (b+a).$$

Оцінка похибки має вигляд

$$\left|f(x) - L_n(x)\right| \le \frac{M_{n+1}}{2^{2n+1}(n+1)!}(b-a)^{n+1}.$$

1.6.5 Інтерполяція за допомогою сплайнів

Підвищення точності інтерполювання вимагає збільшення вузлів інтерполяції**.** Це призведе до зростання ступеня інтерполяційних многочленів. Але в умовах відсутності додаткової інформації про задану таблично функцію останні дають досить значну похибку. В цьому випадку більш ефективним є використання сплайнів, що на проміжку між вузлами інтерполювання є поліномом невисокого ступеня. На всьому проміжку інтерполяції $[a,b]$ сплайн – це функція, що складена з різних частин поліномів. Отже, розглянемо на відрізку $[a,b]$ систему вузлів $a = x_0 < x_1 < x_2 < \ldots < x_n = b$. Сплайном $S_m(x)$ називається функція, що визначена на $[a,b]$, має на ньому неперервні похідні $m-1$ –го порядку і на кожному частковому відрізку $[x_i, x_{i+1}]$ збігається з деяким многочленом ступеня не вище m. При цьому хоча б на одному з відрізків ступінь многочлена дорівнює m. Якщо $S_m(x_i) = f(x_i)$, то це інтерполюючий сплайн.

Лінійний сплайн – це ламана, що проходить через вузли інтерполювання. Рівняння ламаної для $x \in [x_i, x_{i+1}]$

$$S_i(x) = \frac{x - x_i}{x_{i+1} - x_i}(y_{i+1} - y_i) + y_i \,. \qquad (1.51)$$

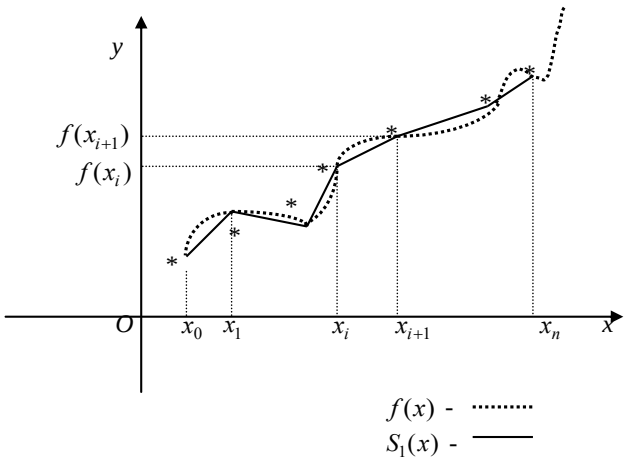

Рисунок 1.6 – Лінійний сплайн

На практиці широкого застосування набули кубічні сплайни. Доведено, що такий інтерполюючий сплайн – єдина функція з мінімальною кривиною серед усіх функцій, які інтерполюють задану функцію і мають квадратично інтегровану другу похідну. В цьому розумінні кубічний сплайн з крайовими умовами є найкращою з функцій, що інтерполюють задану функцію.

Отже, якщо $h = \dfrac{b-a}{n}$, $x_i = x_0 + ih\ (i = \overline{0, n-1})$, $x \in [x_i, x_{i+1}]$, то кубічний сплайн на цьому відрізку

$$S_3(x) = \frac{(x_{i+1}-x)^2(2(x-x_i)+h)}{h^3} f(x_i) + \frac{(x-x_i)^2(2(x_{i+1}-x)+h)}{h^3} f(x_{i+1}) +$$

$$+ \frac{(x_{i+1}-x)^2(x-x_i)}{h^2} m_i + \frac{(x-x_i)^2(x-x_{i+1})}{h^2} m_{i+1}.$$

Тут $m_i = S_3'(x_i)$; $m_{i+1} = S_3'(x_{i+1})$. Для їх визначення накладають умови неперервності другої похідної в точці x_i та обмеження на значення сплайна і його похідних на кінцях проміжка $[a,b]$ – крайові умови. Тобто потрібна додаткова інформація про функцію, для якої є потреба в інтерполюванні.

Випадки використання кубічного сплайна:

1 Якщо відомо, що $S_3'(a) = f'(a); S_3'(b) = f'(b)$, то для визначення m_i маємо систему рівнянь

$$\begin{cases} m_0 = f_0', \\ m_n = f_n', \\ m_{i-1} + 4m_i + m_{i+1} = \dfrac{3}{h}(f(x_{i+1}) - f(x_i)), i = \overline{1, n-1}. \end{cases} \quad (1.52)$$

2 Якщо відомо $S_3''(a) = f''(a), S_3''(b) = f''(b)$, то відповідна система рівнянь

$$\begin{cases} 2m_0 + m_1 = \dfrac{3}{h}(f(x_1) - f(x_0)) - \dfrac{h}{2} f''(x_0), \\ 2m_n + m_{n-1} = \dfrac{3}{h}(f(x_n) - f(x_{n-1})) + \dfrac{h}{2} f''(x_n), \\ m_{i-1} + 4m_i + m_{i+1} = \dfrac{3}{h}(f(x_{i+1}) - f(x_{i-1})), i = \overline{1, n-1}. \end{cases} \quad (1.53)$$

3 Якщо $f(x)$ – періодична функція, тобто $f(x) = f(x+T)$, то $f(x_0) = f(x_n)$, $f(x_1) = f(x_{n+1}) \Rightarrow m_0 = m_n$, $m_1 = m_{n+1}$ і система рівнянь має вигляд

$$\begin{cases} 4m_1 + m_2 + m_n = \dfrac{3}{h}(f(x_2) - f(x_0)), \\ m_{i-1} + 4m_i + m_{i+1} = \dfrac{3}{h}(f(x_{i+1}) - f(x_{i-1})), i = 2,3,\ldots,n-1, \\ m_1 + m_{n-1} + 4m_n = \dfrac{3}{h}(f(x_1) - f(x_{n-1})). \end{cases} \quad (1.54)$$

1.7 Чисельне інтегрування функції одного аргументу

Необхідно обчислити визначений інтеграл

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

за умови, що a і b – певні числа, а $f(x)$ є неперервною функцією на інтервалі інтегрування $[a,b]$. Для цього розіб'ємо відрізок $[a,b]$ на n часткових інтервалів точками $a = x_0 < x_1 < \ldots < x_n = b$ з кроком $h = (b-a)/n$, у кожному з яких виберемо довільні точки $\xi_i, x_i \le \xi_i \le x_{i+1}$.

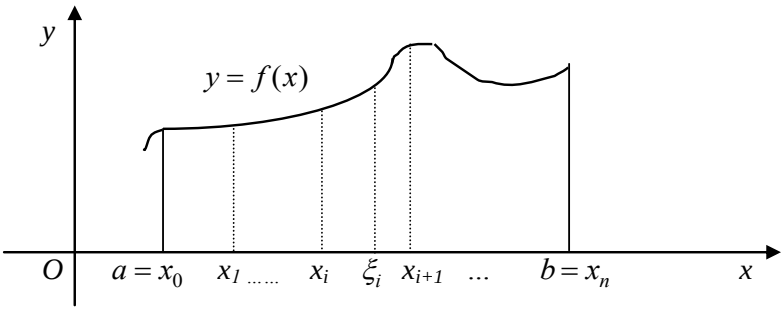

Рисунок 1.7 – Геометричний зміст визначеного інтеграла

Геометричний зміст визначеного інтеграла полягає в тому, що він чисельно дорівнює площі криволінійної трапеції, обмеженої прямими *y=0, x=a, x=b* та функцією $y = f(x)$. У самому визначенні його фактично вже закладена основна ідея чисельного інтегрування. Адже шукана площа може бути показана як границя інтегральної суми (рис. 1.7), тобто

$$S = \sum_{i=0}^{i=n-1} f(\xi_i)(x_{i+1} - x_i), \tag{1.55}$$

$$I = \lim_{\max|\Delta x_i| \to 0} S, \text{ де } \Delta x_i = x_{i+1} - x_i. \tag{1.56}$$

Обчислення суми S, якщо не визначати границю, дає найпростіший приклад чисельного інтегрування. А сам інтеграл можна подати як $I = S + R$. Тут S називають квадратурною формулою, а R – похибкою квадратурної формули.

Для деяких класів функцій можна записати квадратурні формули з похибкою $R = 0$. Такі квадратурні формули називають точними. Наприклад, для поліномів

$$P_m(x) = a_0 + a_1 x + \ldots + a_m x^m$$

квадратурна формула

$$Q = \sum q_i P_m(\xi_i), \tag{1.57}$$

де вибрано довільні точки $\xi_i \in [a,b], i = \overline{0,m}$) і

$$q_i = \int_a^b \frac{(x-\xi_0)\ldots(x-\xi_{i-1})(x-\xi_{i+1})\ldots(x-\xi_m)}{(\xi_i-\xi_0)\ldots(\xi_i-\xi_{i-1})(\xi_i-\xi_{i+1})\ldots(\xi_i-\xi_m)} dx$$

є точною.

Очевидно, що застосування формули (1.57) для інтегрування поліномів не має сенсу, адже він легко обчислюється безпосередньо. Практичний зміст точних квадратурних формул проявляється при інтегруванні таких класів функцій $f(x)$, які можуть бути вдало апроксимовані поліномами на інтервалі $[a,b]$. Замінивши ними підінтегральну функцію та скориставшись (1.57), можна сподіватись на малу похибку R.

Розглянемо деякі найбільш вживані формули наближеного обчислення інтегралів.

1.7.1 Формула прямокутників

Шукану площу замінимо сумою площ прямокутників, побудованих заміною на кожному відрізку $[x_i, x_{i+1}]$ $(i=\overline{0,n-1})$ функції $y=f(x)$ відрізком прямої $y=f(\xi_i)$, $\xi_i \in [x_i, x_{i+1}]$.

Звідси формула прямокутників має вигляд

$$\int_a^b f(x)dx \approx \sum_{i=0}^{n-1} f(\xi_i)(x_{i+1}-x_i) = h\sum_{i=0}^{n-1} f(\xi_i)\,. \qquad (1.58)$$

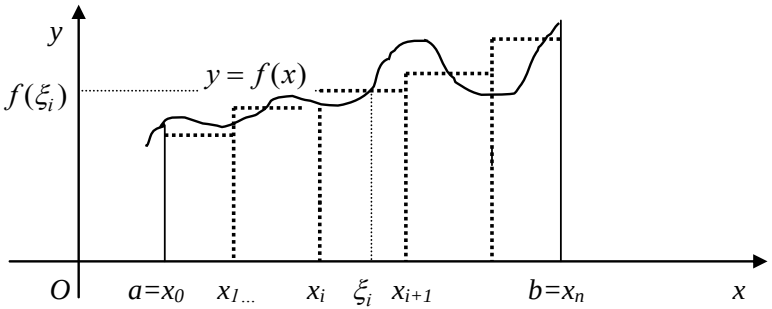

Рисунок 1.8 – Ілюстрація формули прямокутників

1.7.2 Формула трапецій

У цьому випадку підінтегральна функція $f(x)$ для $x \in [a,b]$ замінюється лінійним сплайном, для якого вузлами інтерполяції будуть точки розбиття x_i $(i=\overline{0,n})$. Загальна площа буде розглядатися як сума площ трапецій, що утворилися ланками ламаної і прямими $y=0$, $x=x_i$, $x=x_{i+1}$ на кожному з інтервалів розбиття.

Враховуючи формулу обчислення площі трапеції, маємо

$$\int_a^b f(x)dx \approx \sum_{i=0}^{n-1} \frac{(f(x_{i+1}) + f(x_i)}{2}(x_{i+1} - x_i) =$$
$$= \frac{h}{2}\sum_{i=0}^{n-1}(f(x_{i+1}) + f(x_i)) == \frac{h}{2}[f(a) + 2\sum f(x_i) + f(b)]. \quad (1.59)$$

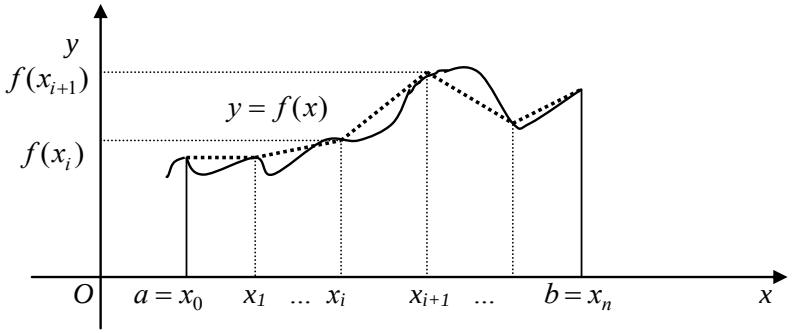

Рисунок 1.9 – Ілюстрація формули трапецій

1.7.3 Формула Сімпсона

Розглянемо один із найбільш відомих і застосовуваних методів чисельного інтегрування – метод Сімпсона. Розіб'ємо інтервал інтегрування $[a,b]$ на парне число однакових частин із кроком $h=(b-a)/(2n)$. На кожному частковому відрізку $[x_i, x_{i+1}]$ довжини h замінимо функцію $f(x)$ квадратичним сплайном, що інтерполює $f(x)$ у вузлах $[x_i, x_{i+1}, x_{i+2}, 0 \le i \le n-1]$.

Додаючи значення для інтегралів на усіх часткових відрізках (i=0,2,...,2(n–1)), одержимо квадратурну формулу Сімпсона (або формулу парабол)

$$\int_a^b f(x)dx \approx \frac{h}{3}\left\{ f(a) + 4\sum_{i=1}^{n} f(a + (2i-1)h + 2\sum_{i=1}^{n-1} f(a + 2ih) + f(b) \right\}. \quad (1.60)$$

Наведемо без доведення похибку апроксимації (обмеження) квадратурної формули Сімпсона у припущенні, що функція $f(x)$ має на відрізку $[a,b]$ неперервні похідні до четвертого порядку:

$$R = \frac{h^4(b-a)}{180} * f^{(IV)}(\xi) \ , \ \ \xi \in [a,b] \, .$$

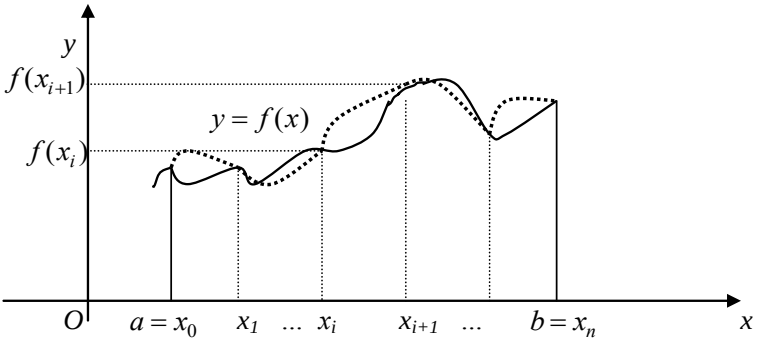

Рисунок 1.10 – Ілюстрація формули Сімпсона

Вона свідчить про те, що для будь–якої неперервної на [*a,b*] функції *f(x)* наближене значення інтеграла, отримане за формулою (1.60), наближається до точного значення при $h \to 0$.

Зауважимо, однак, що похибка обмеження методу парабол (Сімпсона) пропорційна h^4, у той час як для методу прямокутників – h^2. Це означає , що формула (1.60) відповідає ряду Тейлора з точністю до членів третього порядку включно ,а формула прямокутників відповідає цьому ряду тільки з точністю до членів першого порядку . Тому при інтегруванні полінома не вище третього степеня метод Сімпсона дає точні значення інтеграла , адже їхня четверта похідна дорівнює нулю на [*a,b*]. Методи прямокутників і трапецій, як було зазначено раніше, дають точні значення при інтегруванні поліномів першого степеня.

Формула Сімпсона (1.60) є лінійною комбінацією формул прямокутника і трапеції:

$$I_3 = \frac{2}{3} I_1 + \frac{1}{3} I_2,$$

де I_1, I_2 і I_3 – наближені значення інтеграла $\int_a^b f(x)dx$, отримані відповідно за формулами прямокутників, трапецій і Сімпсона.

Оцінка похибки квадратурних формул часто виявляється малоефективною через труднощі, пов'язані з оцінкою похідних підінтегральної функції *f(x)*. У зв'язку з цим набуло поширення практичне правило Рунге оцінки похибки, суть якого полягає в тому, щоб, організувавши обчислення двох значень інтеграла за двома множинами вузлів, потім порівняти результати обчислень і одержати оцінку похибки.

Найбільш популярне правило пов'язане з обчисленням інтеграла двічі з кроками *h* і *h/2*.

Позначимо через *I* точне значення інтеграла $\int_b^a f(x)dx$, через I_h – його наближене значення, обчислене за однією із квадратурних формул із кроком *h*, через $I_{h/2}$ – наближене значення інтеграла, обчислене за тією ж формулою з кроком *h/2*. Похибку кожної квадратурної формули з кроком *h* і *h/2* можна записати відповідно у вигляді

$$R_h = h^R M, \qquad R_{h/2} = \left(\frac{h}{2}\right)^R M,$$

де *R*–порядок точності формул, а *M*–добуток сталої на похідну $f^h(\xi)$. Для формул прямокутників і трапецій *R=2*, для формули Сімпсона *R=4*.

Обчислимо наближене значення інтеграла за однією і тією ж квадратурною формулою спочатку з кроком *h*, а потім із кроком *h/2*. Одержимо

$$I = I_h + h^R * M, \;\; I = I_{h/2} + \left(\frac{h}{2}\right)^R * M.$$

Віднімемо ці рівності: $I_{h/2} - I_h = M\left(\frac{h}{2}\right)^R \left(2^R - 1\right), M * \left(\frac{h}{2}\right)^R = \frac{I_{h/2} - I_h}{2^R - 1}$.

Одержимо оцінку похибки методом Рунге:

$$\left|R_{h/2}\right| = \left|I - I_{h/2}\right| = \left|M\right| * \left(\frac{h}{2}\right)^R, \text{ або } \left|R_{h/2}\right| = \frac{\left|I_{h/2} - I_h\right|}{2^R - 1}. \qquad (1.61)$$

Користуючись формулою, можна уточнити наближені значення інтеграла, вважаючи

$$I = I_{h/2} + R_{h/2} = I_{h/2} + \frac{I_{h/2} - I_h}{2^R - 1}. \qquad (1.62)$$

Цю формулу називають формулою екстраполяції за Річардсоном.

З огляду на порядок точності квадратурних формул випишемо наближену оцінку похибки для формул прямокутників і трапецій за методом Рунге (*R=2*):

$$R_{h/2} \approx \frac{I_{h/2} - I_h}{3} \qquad (1.63)$$

і за формулою Сімпсона (*R=4*):

$$R_{h/2} \approx \frac{I_{h/2} - I_h}{15}. \qquad (1.64)$$

1.8 Чисельні методи розв'язання звичайних диференціальних рівнянь

Нехай потрібно чисельно розв'язати задачу Коші для звичайного диференціального рівняння першого порядку, тобто знайти наближений розв'язок диференціального рівняння $y' = F(x,y)$, що задовольняє початкову умову $y(x_0)=y_0$. Чисельне розв'язання задачі полягає в побудові таблиці наближених значень $y_1, y_2, y_3, ..., y_n$ – розв'язку рівняння $y = \varphi(x)$ у точках $x_1, x_2, x_3, ..., x_n$ – вузлах сітки .

На рисунку 1.11 * позначені точки, що відповідають наближеному розв'язку задачі Коші. Треба зазначити, що частіше використовують систему рівновіддалених вузлів $x_i = x_0 + ih\ (i=1,2,..,n)$, де h – крок сітки $(h > 0)$.

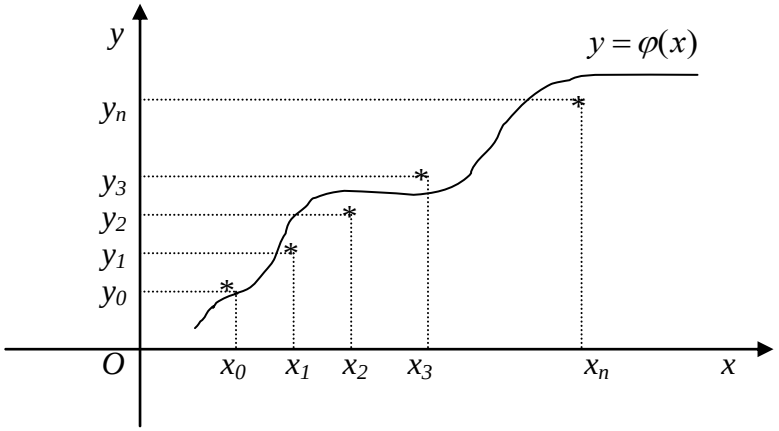

Рисунок 1.11 – Розв'язання задачі Коші

1.8.1 Методи Рунге–Кутта

Різні представники цієї категорії методів потребують більшого чи меншого об'єму обчислень і відповідно забезпечують більшу чи меншу точність. Ці методи мають ряд важливих переваг є явними, одноступінчастими, тобто значення y_{i+1} обчислюється за раніше знайденими значеннями y_i.

Допускають використання змінного кроку, що дає можливість зменшити його там, де функція швидко змінюється, і збільшити в протилежному випадку.

Є легко застосовними, оскільки що для початку розрахунку досить вибрати сітку x_i і задати значення $y_0 = f(x_0)$. Узгоджуються з рядом Тейлора включно до членів порядку h^p, де степінь p не однаковий для різних методів і називається порядком методу.

Не потребують обчислення похідних від $f(x,y)$, а вимагають тільки обчислення самої функції.

При розв'язанні конкретної задачі виникають питання, якою із формул Рунге–Кутта доцільно скористатися і як вибрати крок сітки.

Якщо $f(x,y)$ неперервна й обмежена разом із своїми четвертими похідними, то добрі результати дає метод четвертого порядку. Він описується системою таких п'яти співвідношень:

$$\begin{gathered} y_{i+1} = y_i + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4); \\ k_1 = f(x_i, y_i); \\ k_2 = f(x_i + \frac{h}{2}, y_i + \frac{h}{2}\cdot k_1). \quad (0 \le i \le n-1); \\ k_3 = f(x_i + \frac{h}{2}, y_i + \frac{h}{2}\cdot k_2); \\ k_4 = f(x_i + h, y_i + h\cdot k_3). \end{gathered} \tag{1.65}$$

Якщо функція не має зазначених похідних, порядок точності вищенаведеного методу не може бути реалізований. Тоді необхідно користуватися методами меншого порядку точності, що відповідає порядку наявних похідних.

Одним із найбільш простих і досить ефективних методів оцінки похибки й уточнення отриманих результатів є правило Рунге. Для оцінки похибки за правилом Рунге порівнюють наближені розв'язки, отримані при різних кроках сітки. При цьому використовується таке припущення: глобальна похибка методу порядку p у точці x_i подається у вигляді

$$y_i - y(x_i) = w(x_i)\cdot h^p + o(h^{p+1}); \;\; x_i = a + ih\,.$$

За формулою Рунге

$$y_{2i} - y(x_{2i}) = \frac{y_i - y_{2i}}{2^p - 1} + o(h^{p+1}) \;\; . \tag{1.66}$$

Таким чином, із точністю до $o(h^{p+1})$ (величина більш високого порядку малості) при $h \to 0$ похибка методу має вигляд

$$y_{2i} - y(x_{2i}) = \frac{y_i - y_{2i}}{2^p - 1}, \tag{1.67}$$

де y_i – наближене значення, отримане в точці x_{2i} із кроком h; y_{2i} – із кроком $h/2$; p – порядок методу; $y(x_{2i})$ – точний розв'язок задачі.

Формула Рунге для методу четвертого порядку

$$|y_{2i} - y(x_{2i})| \approx \frac{1}{15}|y_i - y_{2i}| \ . \tag{1.68}$$

1.8.2 Обчислювальна схема (алгоритм) методу Рунге–Кутта

Вибираємо початковий крок h на $[a,b]$, задаємо точність ε.

Створюємо множину рівновіддалених точок (вузлів) $x_i = a + ih$, $0 \le i \le n$, $n = \dfrac{b-a}{h}$. Знаходимо розв'язок y_{i+1} за формулами при кроку h і при кроку $h/2$, $0 \le i \le n-1$.

Перевіряємо нерівність $\dfrac{|y_h - y_{h/2}|}{15} < \varepsilon$.

Якщо ця нерівність виконується, то беремо $y(x_i) = y_{h/2}(x_i)$ і продовжуємо обчислення y_{i+2} з тим самим кроком, якщо ні, то зменшуємо початковий крок h у два рази і переходимо до пункту 3.

1.8.3 Метод прогнозу і корекції

Підправивши схему Эйлера, одержимо схему прогнозу

$$p_{n+1} = y_{n-1} + 2hy_n \ , \tag{1.69}$$

де p_{n+1} – наближене значення y_{n+1}. Цю формулу використовувати не можна, оскільки схема прогнозу нестійка. Тому використовуємо схему корекції

$$c_{n+1} = y_n + \frac{n}{2}\left(y_n^{'} + y_{n+1}^{'} \right). \tag{1.70}$$

Оцінюючи похибки прогнозу і корекції, одержимо

$$y_{n+1} - c_{n+1} = -\frac{h^3}{12} y''' \ \text{– похибка корекції;} \tag{1.71}$$

$$y_{n+1} - p_{n+1} = \frac{h^3}{3} y''' \ \text{– похибка прогнозу.} \tag{1.72}$$

На будь–якому кроці можна оцінити точність розв'язку. При заданому ε=0,0000001, наприклад, $|c_{n+1} - p_{n+1}| < \varepsilon$. Віднімаючи з (1.71) співвідношення (1.70), маємо

$$c_{n+1} - p_{n+1} = \frac{5}{12} h^3 y''' \ .$$

Уточнюємо розв'язок, виходячи з формули (1.70):

$$y_{n+1} = c_{n+1} + \frac{p_{n+1} - c_{n+1}}{5}. \qquad (1.73)$$

Ця формула завершує схеми прогнозу і корекції.

1.8.4 Задача Коші для диференціальних рівнянь вищих порядків

Необхідно знайти функцію $y = \varphi(x)$, що задовольняє диференціальне рівняння та додаткові умови

$$\begin{cases} y^{(m)} = f(x, y, y', y'', ..., y^{(m-1)}), \\ y(x_0) = y_0, \\ y'(x_0) = y_1, \\ y''(x_0) = y_2, \\ \text{.....}, \\ y^{(m-1)}(x_0) = y_{m-1}. \end{cases} \qquad (1.74)$$

Для розв'язання цієї задачі можна застосувати таку схему зведення до системи диференціальних рівнянь першого порядку:

1) вводимо нові змінні:

$$\begin{cases} u_1 = y, \\ u_2 = u_1' = y', \\ u_3 = u_2' = y'', \qquad u_m' = f(x, u_1, u_2, ..., u_m). \\ \text{.....}, \\ u_m = u_{m-1}' = y^{(m-1)}; \end{cases} \qquad (1.75)$$

2) розв'язуємо систему з m диференціальних рівнянь першого порядку.

1.8.5 Метод скінченних різниць для розв'язання лінійних крайових задач

Маємо відрізок [a,b]. Потрібно знайти розв'язок лінійного диференціального рівняння другого порядку

$$y'' + p(x) \cdot y' + q(x) \cdot y = f(x), \qquad (1.76)$$

що задовольняє такі крайові умови:

$$\begin{aligned} &a_1 y(a) + a_2 y'(a) = A, \qquad \beta_1 y(b) + \beta_2 y'(b) = B, \\ &|a_1| + |a_2| \neq 0, \qquad |\beta_1| + |\beta_2| \neq 0. \end{aligned} \qquad (1.77)$$

Виберемо рівномірну сітку: $x = a + ih$, $i = 0,1,2,\ldots,n$. Нехай $p(x_i) = p_i, q(x_i) = q_i, f(x_i) = f_i, y(x_i) = y_i, y'(x_i) = y'_i, y''(x_i) = y''_i$. Апроксимуємо $y'(x_i)$ і $y''(x_i)$ у кожному внутрішньому вузлі (i = 1, 2, …, n–1) центральними різницями $y'(x) = \dfrac{y_{i+1} - y_{i-1}}{2h}$, $y''(x_i) = \dfrac{y_{i+1} - 2y_i + y_{i-1}}{h^2}$ і на кінцях відрізка – односторонніми скінченно–різницевими апроксимаціями $y'(a) = \dfrac{(y_1 - y_0)}{h}$, $y'(b) = \dfrac{(y_n - y_{n-1})}{h}$.

Використовуючи ці формули, одержуємо різницеву апроксимацію вихідного крайового завдання:

$$\begin{cases} \dfrac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + p_i \dfrac{y_{i+1} - y_{i-1}}{2h} + q_i y_i = f_i, \ i = \overline{1, n-1}, \\ \dfrac{a_1 y_0 + a_2 (y_1 - y_0)}{h} = A, \ \dfrac{\beta_1 y_n + \beta_2 (y_n - y_{n-1})}{h} = B. \end{cases} \tag{1.78}$$

Коефіцієнти різницевих рівнянь залежать від кроку сітки. Введемо позначення:

$$\begin{aligned} b_0 &= (ha_1 - a_2), & c_0 &= a_2, \\ a_i &= \left(1 - \frac{hp_i}{2}\right), & d_0 &= hA, \\ & & b_i &= (h^2 q_i - 2), \\ c_i &= \left(1 + \frac{hp_i}{2}\right), & d_i &= h^2 f_i, \qquad i = \overline{1, n-1}, \\ & & b_n &= (h\beta_1 + \beta_2), \\ a_n &= (-\beta_2), & d_n &= hB. \end{aligned}$$

Перепишемо систему з урахуванням введених позначень

$$\begin{cases} b_0 y_0 + c_0 y_1 = d_0, \\ a_i y_{i-1} + b_i y_i + c_i y_{i+1} = d_i, \qquad i = \overline{1, n-1}. \\ a_n y_{n-1} + b_n y_n = d_n. \end{cases} \tag{1.79}$$

Маємо різницеву схему крайового завдання. Запишемо систему рівнянь у розгорнутій матричній формі:

$$\begin{bmatrix} b_0 & c_0 & & & 0 \\ a_1 & b_1 & c_1 & & \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & & & a_n & b_n \end{bmatrix} \times \begin{bmatrix} y_0 \\ y_1 \\ \ldots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ \ldots \\ d_{n-1} \\ d_n \end{bmatrix}.$$

Таким чином, завдання зводиться до розв'язання системи лінійних алгебраїчних рівнянь, що можна записати у вигляді $Ay=d$.

1.8.6 Загальна характеристика явних методів

Однокрокові і багатокрокові методи разом складають так звані *явні методи* чисельного інтегрування диференціальних рівнянь. Назву їх обумовлено тим, що усі вони з метою відшукання значень фазових змінних (змінних стану) на наступному кроці використовують інформацію лише про попередні точки.

Нагадаємо головні властивості кожної з двох груп явних методів.

Однокрокові методи.

1. Через те, що у методах Рунге–Кутта використовується інформація лише про останню обчислену точку інтегральної кривої, за допомогою цих методів можна починати інтегрування диференціальних рівнянь у випадку задачі Коші.
2. З тієї ж причини ці методи дозволяють у процесі інтегрування легко змінювати його крок.
3. Однак з тієї ж причини при використанні цих методів доводиться багаторазово обчислювати значення функцій $Z_k(t, y_1, y_2, ..., y_n)$ $(k = 1, ..., n)$ і витрачати на це багато машинного часу.
4 При застосуванні цих методів важко одержати оцінку виникаючої при інтегруванні похибки обмеження (для цього потрібні додаткові обчислення приблизно того ж обсягу, що й саме інтегрування).

Методи прогнозу–корекції

1 Через те, що у цих методах використовується інформація про декілька попередні точки інтегральної кривої, розділені за аргументом однаковими інтервалами, за їх допомогою неможливо розпочати процес інтегрування у випадку задачі Коші.
2 Оскільки у цих методах замість обчислення $Z_k(t, y_1, y_2, ..., y_n)$ $(k = 1, ..., n)$ у попередніх точках використовуються значення, одержані при попередніх обчисленнях і збережені у пам'яті ЕОМ, вони є більш економічними у сенсі витрат машинного часу.
3 При будь–якій зміні величини кроку інтегрування доводиться тимчасово звертатися до методів Рунге–Кутта.
4 У цих методах легко одержати позитивну оцінку похибки обмеження.

Загальним для явних методів є така, ще не обговорювана обставина: *явні методи при кроках інтегрування, більших за деякий крок* $H_{\max}$, *стають нестійкими у обчислювальному відношенні*. Тобто якщо

$$h \geq H_{\max}, \tag{1.80}$$

похибка обмеження методу починає катастрофічно збільшуватися з кожним кроком, внаслідок чого одержувані результати інтегрування перестають хоч якоюсь мірою відображати якісно і кількісно дійсну інтегральну криву.

Величина $H_{\max}$ залежить від властивостей самої системи диференціальних рівнянь, яку потрібно чисельно інтегрувати, і тому не може бути визначена заздалегідь, якщо не відома ця система диференціальних рівнянь. Для систем лінійних деференціальних рівнянь з постійними коефіцієнтами цю величину можна наближено обчислити за формулою

$$H_{\max} = \frac{T_{\min}}{N}, \tag{1.81}$$

де $T_{\min}$ означає мінімальну за величиною сталу часу системи деференціальних рівнянь, що інтегрується; N – деяке число, більше за одиницю, різне для різних методів інтегрування (наприклад, для методу Рунге–Кутта 4–го порядку N =5). У свою чергу, сталі часу T_i $(i = 1,2,...,n)$ можна визначити як величини, зворотні до значень дійсних та уявних частин (точніше – їхніх модулів) коренів характеристичного рівняння системи деференціальних рівнянь.

Таким чином, щоб організувати процес чисельного інтегрування якоїсь системи диференціальних рівнянь одним із явних методів, потрібно здійснити такі дії:

1) визначити (обчислити за заданими значеннями параметрів диференціальних рівнянь) сталі часу заданої системи диференціальних рівнянь;

2) відкинувши сталі часу, що дорівнюють нескінченності (які відповідають нульовим кореням характеристичного рівняння), знайти серед них максимальну $T_{\max}$ за величиною і мінімальну $T_{\min}$;

3) задатися кількістю n_{int} (>1) кроків інтегрування на інтервалі часу, що дорівнює мінімальній сталій часу; при цьому необхідно, щоб n_{int} було більше за сталу N для заданого методу; при цьому необхідно пам'ятати також, що чим більшу за величиною n_{int} ми задаємо, тим з більшою точністю здійснюватиметься процес чисельного інтегрування, але й тим більший час він займе;

4) розрахувати величину кроку інтегрування за формулою

$$h = \frac{T_{\min}}{n_{\text{int}}}; \tag{1.82}$$

5) задатися бажаною тривалістю процесу, що виходить внаслідок інтегрування, у вигляді кількості n_T максимальних сталих часу системи; розрахувати бажану тривалість процесу за формулою

$$T = n_T \cdot T_{\max}; \tag{1.83}$$

6) здійснити присвоювання змінним стану $y_1, y_2, ..., y_n$ початкових значень;

7) організувати цикл за часом t, який би змінювався від 0 до T з кроком h, усередині якого за заданими значеннями вектора $y = [y_1, y_2, ..., y_n]$ у попередній точці обчислюється значення того ж вектора у наступній точці, а також на основі цих розрахунків формуються масиви, які у подальшому будуть використовуватися як результати інтегрування.

1.8.7 Жорсткі системи. Неявні методи

В інженерній практиці трапляються випадки, коли відношення максимальної сталої часу $T_{\max}$ системи до мінімальної $T_{\min}$ є дуже великою величиною (наприклад, більше, ніж 10^5). Такі системи диференціальних рівнянь називаються *жорсткими*.

Чисельне інтегрування жорстких систем за допомогою явних методів натикається на суттєві труднощі. У цьому випадку загальна кількість кроків інтегрування, потрібна для повного інтегрування

$$k_{\text{int}} = \frac{T_{\max}}{T_{\min}} \cdot n_{\text{int}}, \tag{1.84}$$

буде теж дуже великою. А через те, що похибки округлення накопичуються зі збільшенням кількості обчислень приблизно пропорційно цій кількості, на практиці вже після мільйона операцій підсумовування–множення похибки округлення можуть невпізнанно спотворити інтегральну криву. Якщо врахувати, що кожний крок інтегрування може вміщувати тисячі операцій підсумовування–множення, то вже через тисячі кроків інтегрування відносна похибка внаслідок округлення збільшується до 100%. Подальше інтегрування стає недоцільним.

Подолати цю перешкоду можна тільки у випадку, якщо збільшити крок інтегрування до величини, у десятки разів більшої за мінімальну сталу часу системи. Для розглянутих явних методів це не є можливим, оскільки за цих умов ці методи є нестійкими.

Тому для інтегрування жорстких систем використовують так звані *неявні методи*.

Найпростішим шляхом розв'язування "жорстких" задач є так званий *неявний метод Ейлера*, за яким інтегрування на одному кроці здійснюється за неявною формулою

$$y_{m+1} = y_m + h \cdot Z(t_m, y_{m+1}) . \qquad (1.85)$$

На відміну від явних методів цей метод є стійким при величинах кроку, набагато більших за мінімальну сталу часу, хоча точність його є помірною (порядку 10%).

2 Методи оптимізації технічних систем

2.1 Оптимальний і раціональний розв'язок

При проектуванні виробів, як правило, виходять із необхідності урахування різних вимог. Стосовно механічних конструкцій, які деформуються, такими є міцність, жорсткість, стійкість і т.д.

Розв'язок, який задовольняє всі задані обмеження, називається допустимим. З допустимих у процесі розв'язання екстремальних задач вибираються оптимальні, або раціональні розв'язки.

Під оптимальним розумітимемо такий допустимий розв'язок, який є якнайкращим з погляду вибраного критерію оптимальності. Критерій оптимальності формується на основі одного або декількох критеріїв ефективності. Критерієм ефективності може бути: маса, вартість, приведені витрати, термін виготовлення і т.п.

Але не всі критерії ефективності і обмеження, що виражають вимоги до проектованих виробів, можуть бути формалізовані, наприклад, естетичність, компактність, технологічність виготовлення та ін. У цьому випадку доцільно ввести поняття раціонального розв'язку.

Під раціональним розумітимемо такий розв'язок, який виходить неформальним шляхом, тобто з урахуванням експертних (або інших неформальних) оцінок. При пошуку оптимальної конструкції за прийнятим критерієм оптимальності і обмеження можна одержати оптимальну, але нераціональну конструкцію.

Наприклад, якщо вимога технологічності не знайшла кількісної оцінки і не увійшла до числа обмежень, то можна одержати проект мінімальної ваги (оптимальний за вагою), але нераціональний проект через нетехнологічність. Важча конструкція, але зручна з технологічної точки зору може виявитися раціональнішою. Стосовно механічних конструкцій, що деформуються, раціональне проектування може бути визначене як проектування, що грунтується засноване на принципах механіки деформованого твердого тіла і має на меті отримання оптимальної конструкції на базі вибраного проектантом критерію. Оптимізаційні задачі (і їх різні поставлення) є окремими підзадачами у процесі розроблення в цілому раціонального проекту [5,6].

2.2 Розрахункова модель

Задача проектування силової конструкції починається з вибору розрахункової моделі, яка виходить шляхом ідеалізації реальної конструкції.

При цьому необхідно розробити або прийняти ряд допоміжних моделей:

1 *Моделі форми.* Моделі форми елементів конструкцій є схематичним описом геометрії елемента за допомогою стандартних, типових елементів:

стрижень – тіло, поперечні розміри h (висота) і b (ширина), якого малі в порівнянні з його довжиною L. Для стрижня характерне співвідношення $h/L \le 1/5$ при $h \ge b$; якщо $h/L > 1/5$, то стрижень починає працювати як пластина;

пластина – тіло, форма якого визначається серединною площиною, причому товщина h набагато менша від двох інших габаритних розмірів a,b, тобто $h << a$, $h << b$;

оболонка – тіло, обмежене двома близькими криволінійними поверхнями з відстанню між ними (товщина h), значно меншого за інші габаритні розміри a,b;

просторове тіло – тіло довільної форми, всі габаритні розміри якого сумірні.

2 *Конструктивні схеми.* Класифікація конструктивних схем може бути здійснена за різними ознаками. Наприклад, залежно від використовуваних типових елементів конструктивні розрахункові схеми можуть бути: просторовими, оболонковими, пластинчастими, стрижневими і комбінованими.

До конструктивної схеми *просторового* тіла можна навести конструкції фундаменту, корпусної деталі машини, складного вузла з'єднання окремих елементів і т.д.

До *конструктивної оболонкової* схеми можуть бути віднесені конструкції купольного покриття будівлі, відсіку літака, судна і т.д.

До *конструктивної пластинчастої* схеми можуть бути наведені окремі підконструкції палуби судна, плоскої кришки посудини і т.д.

Прикладами *конструктивних стрижневих* схем є ферми мостів, зуб шестерні, корпус ракети і т.д.

До *комбінованих* конструктивних схем відносять схеми, що включають різні типи елементів. Наприклад, конструкція літального апарата є комбінованою (складовою) конструкцією, що складається з оболонок, стрижнів (стрингерів і шпангоутів), пластин.

3 *В'язі.* Всі тіла, як правило, взаємодіють із зовнішнім середовищем або іншими тілами. Якщо внаслідок яких–небудь обмежень (умов) дане тіло не може зайняти довільне положення в просторі і має довільні

швидкості, то таке тіло називається *невільним*. У цьому випадку обмеження (умови) називають в'язями.

4 *Моделі навантаження.* Зовнішні сили, що діють на конструкцію, поділяються на три групи:

зосереджені сили – сили, які діють на невеликих ділянках поверхні деталі (наприклад, тиск кульки шарикопідшипника на вал);

розподілені сили – сили, прикладені до значних ділянок поверхні (наприклад, тиск рідини або газу на стінки посудини);

об'ємні або *масові сили – сили, прикладені до кожної частинки матеріалу (наприклад, сили тяжіння, інерційні наватаження).*

Навантаження поділяють на *стаціонарні* (сталі, статичні) і *нестаціонарні* (змінні). Вони можуть мати *випадковий,* або *детермінований,* характер.

Моделі навантаження повинні враховувати дію полів і середовищ: температурного поля, електромагнітного поля, корозійних середовищ та ін. Зокрема, *температурні* дії можуть викликати зміни геометричної форми тіла і зміну сил взаємодії між його точками, зміну фізико–механічних властивостей матеріалу.

5 *Механічні властивості матеріалів*. У розрахункову модель необхідно закладати такі механічні властивості матеріалів, які відповідають вимогам експлуатації. До них відносяться: модуль пружності, модуль зсуву, коефіцієнт Пуассона, межа текучості, межа міцності.

6 *Урахування статистичної природи початкових даних у моделі.* Розрахункова модель повинна бути достатньо простою, яка допускає побудову математичної моделі і в той же час достатньо адекватно відображає стан реальної конструкції. Найчастіше в моделі передбачається детермінованість форми і структури об'єкта, який проектується, геометричних розмірів елементів, зовнішніх дій і властивостей матеріалів. Насправді всі початкові дані перебувають під впливом великої кількості неврахованих факторів і тому тією чи іншою мірою мають випадковий характер. Наприклад, змінність геометричних розмірів і форми елементів конструкції виконує істотну роль при втраті стійкості форм рівноваги конструкції. Майже всі зовнішні навантаження є випадковими. Випадковий характер мають і механічні характеристики матеріалу. Випадковий характер початкових даних у розрахунковій моделі може бути врахований при виборі коефіцієнтів запасу згідно з діючими нормами проектування або застосування статистичних методів аналізу стану конструкції з урахуванням статистичних початкових даних, що входять до розрахункової моделі.

2.3 Поставлення задач параметричної оптимізації

Кінцевою метою проектування є створення раціональних виробів, конструкцій і т.п., виходячи з наявних ресурсів і можливостей. Щоб доби-

тися якнайкращого результату, необхідно оптимізувати на всіх етапах проектування.

У процесі проектування виникає різноманітність різних поставлень задач оптимізації:

– вибір оптимальних геометричних форм;

– вибір оптимальних структур конструкцій;

– оптимальний розподіл внутрішніх зусиль за рахунок попередніх напружень;

– підбір матеріалів;

– створення конструкцій мінімальної ваги із заданою надійністю;

– оптимальне армування конструкцій і т.д.

При проектуванні розрізняють задачі визначення структури і визначення значень внутрішніх параметрів (параметрів елемента). Якщо серед варіантів структури відшукується найкращий, то таку задачу називають *структурною оптимізацією.* Розрахунок внутрішніх параметрів, оптимальних з позиції деякого критерію при заданій структурі об'єкта, називають *параметричною оптимізацією.*

2.4 Критерій ефективності. Цільова функція

Для оцінки проектних рішень у процесі їх пошуку використовуються різні показники. Ними можуть бути показники ваги, вартості, надійності, технологічності, естетичності і т.п. Показники можна розбити на два класи: кількісні і якісні.

Кількісні показники дозволяють виконати оцінку проектного рішення кількісно на основі аналізу математичної моделі ефективності. Наприклад, наявність функції ваги виробу дозволяє кількісно порівнювати варіанти, які розробляються.

Якісні показники не дозволяють виконати оцінку проектного рішення кількісно, оскільки немає математичній моделі ефективності. Наприклад, оцінка механічної системи без розробленої математичної моделі технології виготовлення здійснюється якісно експертами – технологами.

Під *критерієм ефективності* розумітимемо показник, який дозволяє кількісно виконувати оцінку ефективності проектних рішень.

2.5 Обмеження

Працездатною (допустимою) вважається така система, для якої виконуватимуться всі умови, записані у формі обмежень типу рівностей або нерівностей:

$$g_k(x) \le 0 \ \ (k = 1,2,\ldots). \qquad (2.1)$$

Наведемо основні види обмежень на поведінку пружних технічних систем.

1 *Обмеження на міцність.* Вимогу виконання умови міцності можна сформулювати в такій розмірній формі:

$$\max_{V_i} \sigma_{экв}^{i,j}(x) \le [\sigma]_{i,j} \quad (i = \overline{1,N},\ j = \overline{1,J}), \tag{2.2}$$

де N – число підконструкцій у заданій конструкції; j – номер варіанта дії на систему; $\max_{V_i} \sigma_{экв}^{i,j}$ – максимальне еквівалентне напруження в об'ємі i–ї підконструкції, яке визначається за прийнятою гіпотезою або теорією міцності при j–му варіанті дії; $[\sigma]_{i,j}$ – допустиме напруження для матеріалу i–ї підконструкції при j–му варіанті дії. Якщо ввести функцію обмежень

$$g_{i,j}^{n}(x) = \max_{V_i} \sigma_{экв}^{i,j}(x) / [\sigma]_{i,j} - 1 \quad (i = \overline{1,N},\ j = \overline{1,J}),$$

то обмеження (2.2) можна привести до канонічної форми (2.1):

$$g_{i,j}^{n}(x) \le 0. \tag{2.3}$$

2 *Обмеження на жорсткість.* Обмеження на пружні переміщення можна записати у формі

$$\max_{S_i} \left| u^{i,j}(x) \right| \le [u]_{i,j}, \tag{2.4}$$

де $\max_{S_i} \left| u^{i,j}(x) \right|$ – максимальне узагальнене переміщення поверхні i–ї підконструкції при j–му варіанті дії; $[u]_{i,j}$ – допустиме узагальнене переміщення. Під узагальненим переміщенням розуміємо різні переміщення (лінійні, кутові, відносні, абсолютні).

Обмеження (2.4) приводиться до канонічного вигляду

$$g_{i,j}^{ж}(x) \le 0, \tag{2.5}$$

якщо взяти функцію обмежень за жорсткістю у вигляді

$$g_{i,j}^{ж}(x) = \max_{S_i} \left| u^{i,j}(x) \right| / [u]_{i,j} - 1 \quad (i = \overline{1,N},\ j = \overline{1,J}).$$

3 *Умова місцевої стійкості.* Під місцевою втратою стійкості розуміємо явище втрати стійкості на локальній ділянці системи, що деформується, типу випинання. Умову місцевої стійкості можна записати в розмірній формі

$$P_{i,j} \le \min \left\{ P_{кр}^{i,j}(x) \right\} \ (i = \overline{1,N},\ j = \overline{1,J}), \tag{2.6}$$

де $P_{i,j}$ – параметр зовнішньої *j*–ї дії на *i*–ту підконструкцію; $\min\left\{P_{кр}^{i,j}(x)\right\}$ – мінімальне значення критичного параметра *j*–ї зовнішньої дії на *i*–ту підконструкцію.

Обмеження (2.6) можна записати в канонічній формі

$$g_{i,j}^{V}(x) \le 0\,, \tag{2.7}$$

якщо взяти функцію обмежень на місцеву стійкість у вигляді

$$g_{i,j}^{V}(x) = P_{i,j} / \min\left\{P_{кр}^{i,j}(x)\right\} - 1\,.$$

4 *Обмеження на частоти власних коливань.* Частоти власних коливань системи є одними з найважливіших динамічних характеристик системи. Обмеження на частоти власних коливань мають вигляд

$$\min_{i}\left\{\omega_i^j(x)\right\} \ge [\omega]^j \;\; (i = 1,2,\ldots), \tag{2.8}$$

де $\min_{i}\left\{\omega_i^j(x)\right\}$ – нижча частота власних *j*–х коливань пружної системи; $[\omega]^j$ – допустиме найменше значення частоти власних *j*–х коливань системи, що призначається як розрахункове значення частоти вимушених *j*–х коливань.

Умова (2.8) в канонічній формі має вигляд

$$g_{i,j}^{ж}(x) \le 0\,, \tag{2.9}$$

де $g_i^K(x) = [\omega]^j / \min_{i}\left\{\omega_i^j(x)\right\} - 1$.

5 *Обмеження на параметри.* При пошуку оптимальних рішень в рамках прийнятих фізичних і математичних моделей повинне виконуватися певне співвідношення керованих (ті, що змінні у процесі оптимізації) x_i і некерованих (ті, що незмінні у процесі оптимізації) параметрів у формі ($d_{i,j}$ – задане число)

$$x_i / c_j \le d_{i,j} \;\; (i = \overline{1,n},\;\; j = \overline{1,m})\,. \tag{2.10}$$

Обмеження на геометричні параметри (2.10) можна записати в канонічній формі

$$g_{i,j}^{Г}(x_i) \le 0\,, \tag{2.11}$$

де $g_{i,j}^{Г}(x_i) = x_i / (c_j d_{i,j}) - 1$.

2.6 Задачі оптимізації

У процесі розв'язання задачі оптимізації звичайно необхідно знайти оптимальні значення деяких параметрів, що визначають дану задачу [5,6]. При розв'язанні інженерних задач їх прийнято називати проект-

ними параметрами. Як проектні параметри можуть бути, зокрема, значення лінійних розмірів об'єкта, маси, температури і т.п. Число n проектних параметрів $x_1, x_2, ..., x_n$ характеризує розмірність (і ступінь складності) задачі оптимізації.

Вибір оптимального рішення або порівняння двох альтернативних рішень проводиться за допомогою деякої залежної величини (функції), яка визначається проектними параметрами. Ця величина називається цільовою функцією (або критерієм якості). У процесі розв'язання задачі оптимізації повинні бути знайдені такі значення проектних параметрів, при яких цільова функція має мінімум (або максимум). Таким чином, цільова функція – це глобальний критерій оптимальності в математичних моделях, за допомогою яких описуються інженерні задачі.

Цільову функцію можна записати у вигляді

$$u = F(x_1, x_2, ..., x_n) \qquad (2.12)$$

Прикладами цільової функції, що трапляються в інженерних розрахунках, є міцність або маса конструкції, потужність установки і т.п.

У разі одного проектного параметра ($n = 1$) цільова функція (2.12) є функцією однієї змінної, і її графік – деяка крива на площині. При $n = 2$ цільова функція є функцією двох змінних, і її графіком є поверхня.

Необхідно зазначити, що цільова функція не завжди може бути представлена у вигляді формули. Іноді вона може набувати тільки деякі дискретні значення, задаватися у вигляді таблиці і т.п. У всіх випадках вона повинна бути однозначною функцією проектних параметрів.

Цільових функцій може бути декілька. Наприклад, при проектуванні виробів машинобудування одночасно вимагається забезпечити максимальну надійність, мінімальну матеріаломісткість, максимальний корисний об'єм (або вантажопідйомність). Деякі цільові функції можуть виявитися несумісними. У таких випадках необхідно вводити пріоритет тієї або іншої цільової функції.

Виділяють два типи задач оптимізації: безумовні і умовні. Безумовна задача оптимізації полягає у відшуканні максимуму або мінімуму дійсної функції (2.12) від n дійсних змінних і визначенні відповідних значень аргументів на деякій множині σ n –вимірного простору.

Умовні задачі оптимізації, або задачі з обмеженнями, – це такі задачі, при формулюванні яких задаються деякі умови (обмеження) на множині σ. Ці обмеження задаються сукупністю деяких функцій, що задовольняють рівняння або нерівності.

Обмеження–рівності виражають залежність між проектними параметрами, яка повинна враховуватися при знаходженні рішення. Ці обмеження відображають закони природи, наявність ресурсів і т.п.

У результаті обмежень область проектування σ, яка визначається всіма n проектними параметрами, може бути істотно зменшена відповідно до фізичної суті задачі.

Число обмежень–рівностей може бути довільним. Їх можна записати у вигляді

$$\left.\begin{aligned} &g_1(x_1, x_2, ..., x_n) = 0, \\ &g_2(x_1, x_2, ..., x_n) = 0, \\ &\text{.....}, \\ &g_j(x_1, x_2, ..., x_n) = 0 \end{aligned}\right\}. \qquad (2.13)$$

У ряді випадків із цих співвідношень можна виразити одні проектні параметри через інші. Це дозволяє виключити деякі параметри з процесу оптимізації, що приводить до зменшення розмірності задачі і полегшує її рішення.

Аналогічно можуть вводитися також обмеження–нерівності, що мають вигляд

$$\left.\begin{aligned} &a_1 \le g_1(x_1, x_2, ..., x_n) \le b_1, \\ &a_2 \le g_2(x_1, x_2, ..., x_n) \le b_2, \\ &\text{.....}, \\ &a_k \le g_k(x_1, x_2, ..., x_n) \le b_k. \end{aligned}\right\} \qquad (2.14)$$

Необхідно зазначити особливість у відшуканні рішення за наявності обмежень. Оптимальне рішення тут може відповідати або локальному екстремуму (максимуму або мінімуму) усередині області проектування, або значенню цільової функції на межі області. Якщо ж обмеження відсутні, то шукається оптимальне рішення на всій області проектування, тобто знаходиться глобальний екстремум.

2.7 Чисельний пошук екстремуму функції однієї змінної

Більшість задач оптимізації зводиться до пошуку найбільшого (або найменшого) значення деякої функції. Методи математичного аналізу зручні для розв'язання цієї проблеми, коли функція задається явно і є при цьому диференційованою. Коли ж функція задається табличним способом або аналітично громіздкою формулою, ефективними є чисельні методи розв'язання.

Функція однієї змінної $y = f(x)$ називається унімодальною на відрізку $[a,b]$, якщо на ньому знаходиться єдина точка $x^* \in [a,b]$, в якій функція набуває мінімального значення.

2.7.1 Метод золотого перетину

Золотим перетином відрізка $[a,b]$ називається розподіл його точкою c на дві нерівні частини так, щоб відношення усього відрізка до більшої частини дорівнювало відношенню більшої частини до меншої

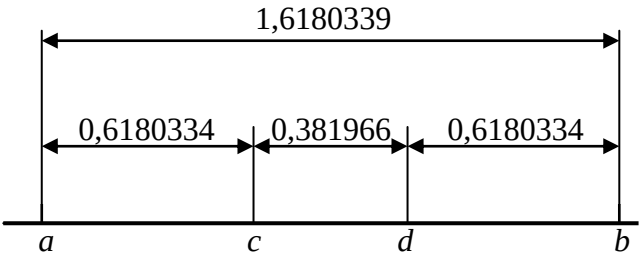

Рисунок 2.1 – Ілюстрація золотого відношення

$$\frac{b-a}{b-c}=\frac{b-c}{c-a}=r\,. \tag{2.15}$$

Число r називають золотим відношенням. Його значення відоме: $r=\frac{1+\sqrt{5}}{2}\approx 1{,}6180339$ (рис. 2.1). На відрізку $[a,b]$ можна визначити дві симетрично розміщені точки c і d, що реалізують золотий перетин. Їх знаходимо за формулами

$$d=a+\frac{1}{r}(b-a)=b-\frac{3-\sqrt{5}}{2}(b-a)\,,$$

$$c=a+(1-\frac{2}{r})(b-a)=a+\frac{3-\sqrt{5}}{2}(b-a)\,.$$

Опишемо алгоритм мінімізації функції однієї змінної $f(x)$, вважаючи її унімодальною на відрізку $[a,b]$. Початковий відрізок $[a,b]$ ділимо точками c (перша точка) і d (друга точка) за правилом золотого перетину і у цих точках обчислюємо значення функцій $f(c)$ і $f(d)$. Порівняння цих значень дозволяє відкинути або інтервал $[a,c]$, якщо $f(c)>f(d)$, або інтервал $[d,b]$, якщо $f(d)>f(c)$. Довжина відрізка, що залишився, зменшиться у r разів:

$$\frac{c-b}{b-a}=r,\frac{d-a}{b-a}=r\,.$$

Після цього процес повторюємо.

На інтервалі, що залишився, уже є одна точка, що робить його золотий перетин: c є друга точка золотого перетину відрізка $[c,b]$,а d – перша точка золотого перетину відрізка $[c,b]$. Знаючи одну з точок золотого перетину, іншу можна знайти за однією із вищезгаданих формул та обчислити значення $f(x)$ у знову знайденій точці (значення в іншій точці вже обчислено на попередньому кроці).

Таким чином, на кожному кроці, починаючи з другого, потрібно лише одне обчислення функції $f(x)$, і інтервал невизначеності зменшується в r разів: $(b_1 - a_1) = (b-a)/r$.

Точка мінімуму $x_{\min} \in [a_1, b_1]$.

Після n кроків маємо довжину інтервала невизначеності

$$(b_n - a_n) = (b-a)/r^n = \left(\frac{\sqrt{5}-1}{2}\right)^n (b-a). \qquad (2.16)$$

Процес золотого перетину продовжуємо до того часу, поки інтервал $(b_n - a_n)$ стане менше деякого заданого числа ε, названого точністю.

З формули випливає, що при $n \to \infty$ довжина відрізка, який залишився, наближається до нуля як геометрична прогресія зі знаменником $1/r$, тобто метод золотого перетину завжди збігається, причому лінійно. Число кроків n, що забезпечують задану точність ε знаходження точки мінімуму, повинно задовольняти нерівність

$$n \geq \ln\left(\frac{\varepsilon}{b-a}\right) / \ln\left(\frac{\sqrt{5}-1}{2}\right) \approx -2{,}1 \ln\left(\frac{\varepsilon}{b-a}\right). \qquad (2.17)$$

Використовуючи цю формулу, можна знайти необхідне число кроків n для забезпечення необхідної точності ε. Однак на практиці часто роблять інакше: визначають межу відрізка $[a_n, b_n]$ за формулою (2.16) і потім порівнюють із заданою точністю ε.

Метод золотого перетину гарантує знаходження мінімуму у найнесприятливіших умовах, коли функція є не тільки недиференційованою, але і навіть має розриви першого роду.

2.7.2 Алгоритм мінімізації функції «метод золотого перетину»

1 Знаходимо точки c і d за формулами

$$c = a + t(b-a), \;\; d = a + b - c, \;\; \text{де } t = (3-\sqrt{5})/2.$$

2 Обчислюємо значення функції $y = f(c)$ та $y = f(d)$ і порівнюємо їх:

a) якщо $f(c) \geq f(d)$, то за інтервал невизначеності беремо відрізок $[c,b]$ і знаходимо його золотий перетин, вважаючи, що $a_1 = c$, $b_1 = b$, $c_1 = d$, $d_1 = b_1 - t(b_1 - a_1)$, та обчислюємо $f(d)$;

b) якщо $f(c) < f(d)$, за інтервал невизначеності беремо відрізок $[a,d]$ і знаходимо його золотий перетин, вважаючи, що $a_1 = a$, $b_1 = d$, $d_1 = c$, $c_1 = a_1 + t(b_1 - a_1) = a_1 + b_1 - d_1$, та обчислюємо $f(c)$.

3 Перевіряємо умову $|b_1 - a_1| < \varepsilon$. Якщо ця нерівність виконується, то за точку мінімуму беремо величину $x_{\min} = \dfrac{a_1 + b_1}{2}$, якщо ні, то повертаємося до пункту 1.

4 Процес повторюємо до того часу, поки для деякого n не справдиться $|a_n - b_n| < \varepsilon$.

2.7.3 Метод Ньютона

Для функцій однієї змінної класичний підхід при пошуку значень *x* в точках перегину функції $F(x)$ полягає в розв'язанні рівняння

$$F'(x) = 0. \tag{2.18}$$

Робота алгоритму починається у точці x_1, яка являє собою початкове наближення координати стаціонарної точки або корінь рівняння (2.18). Потім будується лінійна апроксимація функції $F'(x)$ в точці x_1, і точка, в якій апроксимуюча лінійна функція обертається в нуль, береться як наступне наближення. Якщо точка x_k взята як поточне наближення до стаціонарної точки, то лінійна функція, яка апроксимує функцію $F'(x)$ у точці x_k, записується у вигляді

$$F'(x, x_k) = F'(x_k) + F''(x_k)(x - x_k). \tag{2.19}$$

Прирівнявши праву частину рівняння (2.19) до нуля, одержимо таке наближення:

$$x_{k+1} = x_k - \left[F'(x_k) / F''(x_k)\right].$$

На жаль, залежно від вибору початкової точки і виду функції алгоритм може як збігатися до істиної стаціонарної точки, так і розходитися.

2.7.4 Апроксимація кривими. Кубічна апроксимація

Раніше була зроблена спроба знайти малий інтервал, в якому знаходиться мінімум функції. Тут розглянемо інший підхід, згідно з яким використовується декілька значень функції у певних точках для апрокси-

мації функції звичним поліномом, принаймні, в невеликій області значень. Потім положення мінімуму функції апроксимується положенням мінімуму полінома, оскільки останній обчислити простіше. Найбільшого поширення набули методи оцінювання з використанням квадратичної і кубічної апроксимацій, оскільки побудова апроксимуючого полінома вище третього порядку стає дуже складною процедурою. У даному розділі розглядається метод оцінювання з використанням кубічної апроксимації, що забезпечує велику точність [5]. Для кубічної апроксимації в цьому методі використовуються значення функції і її похідної, обчислені у двох точках (*p, q*).

Розглянемо задачу мінімізації $F(x)$ на прямій, тобто мінімізацію функції

$$\varphi(h) = F(x_0 + hd) = F(x_{01} + hd_1,...,x_{on} + hd_n)\ ; \tag{2.20}$$

$$\frac{d\varphi}{dh} = \frac{\partial F}{\partial x_1}(x_0 + hd)d_1 + \frac{\partial F}{\partial x_2}(x_0 + hd)d_2 + ... + \frac{\partial F}{\partial x_n}(x_0 + hd)d_n\ ,$$

де x_0 – задана точка; d – заданий напрям; h – крок. Отже,

$$d\varphi / dh = \nabla F(x_0 + hd)^T d = g(x_0 + hd)^T d\ . \tag{2.21}$$

Припускаємо, що відомі такі значення:

$$\varphi(p) = \varphi_p,\ \varphi(q) = \varphi_q,\ \frac{d\varphi}{dh}(p) = G_p,\ \frac{d\varphi}{dh}(q) = G_q\ . \tag{2.22}$$

Цю інформацію можна використовувати для побудови кубічного полінома

$$a + bh + ch^2 + dh^3\ , \tag{2.23}$$

який апроксимуватиме функцію $\varphi(h)$. Якщо так, то рівняння, які визначають *a,b,c,d,* мають такий вигляд:

$$\begin{aligned} &a = \varphi_p, \\ &a + bq + cq^2 + dq^3 = \varphi_q, \\ &b = G_p, \\ &b + 2cq + 3dq^2 = G_q. \end{aligned} \tag{2.24}$$

Ці рівняння мають таке розв'язання:

$$a = \varphi_p,\ b = G_p,\ c = -\frac{(G_p + z)}{q},\ d = \frac{G_p + G_q + 2z}{3q^2}, \tag{2.25}$$

$$\text{де } z = \frac{3(\varphi_p - \varphi_q)}{q} + G_p + G_q.$$

Точки перегину кубічного полінома є розв'язком рівняння

$$G_p - 2(G_p + z)h/q + (G_p + G_q + 2z)(h/q)^2 = 0.$$

Отже, якщо r є точкою мінімуму кубічного полінома, то

$$\frac{r}{q} = \frac{(G_p + z) \pm \left[(G_p + z)^2 - G_p(G_p + G_q + 2z)\right]^{1/2}}{G_p + G_q + 2z} = \tag{2.26}$$
$$(G_p + z \pm w)/G_p + G_q + 2z,$$

де $w = \left(z^2 - G_p G_q\right)^{1/2}$. Одне із значень (2.26) відповідає мінімуму. Друга похідна дорівнює

$$2c + 6dh\,. \tag{2.27}$$

Якщо ми вибираємо додатний знак, то при $h/q = (G_p + z + w)/(G_p + G_q + 2z)$ друга похідна буде $2w/q > 0$. Тоді:

$$r/q = (G_p + z + w)/(G_p + G_q + 2z). \tag{2.28}$$

Кращі чисельні результати можуть бути одержані при використанні такої еквівалентної формули:

$$\frac{r}{q} = 1 - \frac{G_q + w - z}{G_q - G_p + 2w} = \frac{z + w - G_p}{G_q - G_p + 2w}.$$

Якщо $G_p < 0$, то необхідно вибрати значення q додатним, тобто зробити крок у напрямку спадання функції $\varphi(h)$, інакше значення q необхідно вибрати від'ємним. Значення q повинно бути таким, щоб інтервал $(0, q)$ містив мінімум. Це буде справедливо, якщо $\varphi_q > \varphi_p$ або якщо $G_q > 0$.

Якщо жодна з цих умов не виконана, то значення q подвоюється, це повторюється до того часу, поки даний інтервал не міститиме мінімум. Для визначення початкового значення q Давідон, Флетчер і Пауелл запропонували вибирати q таким чином:

$$q = \min\{\eta, -2(\varphi_p - \varphi_m)/G_p\}, \tag{2.29}$$

де φ_m – оцінка найменшого значення істинного мінімуму $\varphi(h)$; η – константа, значення якої вибирається таким, що дорівнює 2 або 1.

Ця ітераційна процедура має такі кроки:

1 Знайти $\varphi_p = F(x_0)$ і $G_p = [g(x_0)]^T d$.

2 Перевірити, чи виконується умова $G_p < 0$, і якщо вона не виконується, то необхідно здійснювати пошук уздовж напрямку d. Вибрати q з виразу (2.29). При цьому необхідно «вгадати» φ_m.

3 Обчислити $\varphi_q = F(x_0 + qd)$ і $G_q = g(x_0 + qd)^T d$.

4 Якщо $G_q > 0$ або $\varphi_q > \varphi_p$, то інтервал, що містить мінімум, знайдений. Інакше необхідно замінити *q* на *2q* і повернутися до кроку 3.

5 Використання рівняння (2.28) для апроксимації точки мінімуму на інтервалі (0, *q*) значенням *r*.

6 Якщо $|d\varphi/dh| = |[g(x_0 + rd)]^T d| = |G_r| < \varepsilon$, де ε – задана точність, то необхіжно зупинитися.

7 Повернутися на крок 5, використовуючи інтервал (0, *r),* якщо $G_r > 0$, або використовуючи інтервал (*r*, *q*), якщо $G_r \le 0$.

На кроці 6 здійснюється перевірка значення похідної. Попередні перевірки приводять до зупинення тоді, коли положення мінімуму не змінюється.

2.8 Чисельні методи пошуку екстремуму функції декількох змінних

Методи, орієнтовані на розв'язання задач безумовної оптимізації, можна розділити на три широкі класи відповідно до типу інформації, яка використовується при реалізації того або іншого методу [5,6]:

1 Методи прямого пошуку, які грунтуються на обчисленні тільки значень цільової функції.

2 Градієнтні методи, в яких використовуються значення перших похідних цільовій функції.

3 Методи другого порядку, в яких використовуються значення перших і других похідних цільовій функції.

Из методів *прямого пошуку* звичайно відзначають три методи, які є ефективними і можуть буди використані для широкого числа практичних застосувань:

1) пошук за симплексом, або S^2–метод;

2) метод пошуку Хука–Джівса;

3) метод зв'язаних напрямів Пауелла.

В основу перших двох методів прямого пошуку покладено ідею, що полягає у виборі *базової точки* і оцінюванні значень цільової функції в точках, що оточують базову точку. Наприклад, при розв'язанні задачі з двома змінними можна скористатися квадратним зразком. Потім *найкраща* точка з п'яти досліджуваних точок вибирається як наступна базова точка, навколо якої будується аналогічний зразок. Якщо жодна з кутових точок не має переваги перед базовою, розміри зразка необхідно зменшити, після чого продовжити пошук.

У процесі пошуку за S^2–методом послідовно оперують регулярними симплексами (множина (n+1)–ї рівновіддаленої точки в n–вимірному просторі) у просторі керованих змінних. При реалізації методу Хука–Джівса використовується фіксована множина (координатних) напрямів, які вибираються рекурсивним способом. Метод Пауелла орієнтований на розв'язання задач з квадратичними цільовими функціями і збігається за кінцеве число ітерацій. До загальних особливостей всіх трьох методів необхідно віднести відносну простоту відповідних обчислювальних процедур, які легко реалізуються і швидко коригуються. З іншого боку, реалізація вказаних методів часто вимагає дуже великої кількості обчислень значень функції. Ця обставина призводить до необхідності розгляду методів, що грунтуються на використанні градієнта цільової функції.

До градієнтних методів належать такі:

1) метод найшвидшого спуску; 2) метод Флетчера–Рівса (метод зв'язаних градієнтів); 3) методи змінної метрики (метод Давідона–Флетчера–Пауелла (ДФП), метод Бройдена–Флетчера–Шенно (БФШ) та ін.

У градієнтних методах для визначення *найкращого* напрямку пошуку використовується відома властивість напрямку градієнта – напрямок градієнта є найшвидшим зростанням функції. Отже, протилежний напрямок є напрямком найшвидшого спадання функції. Напрямок градієнта перпендикулярний в будь–якій точці лінії постійного рівня, оскільки уздовж цієї лінії функція постійна. Тому якщо ми знаходимося в точці x_i *на деякому кроці процесу оптимізації, то пошук мінімуму функції здійснюється уздовж напрямку* $\nabla F(x_i)$. На кроці *i* точка мінімуму апроксимується точкою x_i. Наступною апроксимацією є точка

$$x_{i+1} = x_i + \lambda_i d_i, \tag{2.30}$$

де x_i – поточне наближення до розв'язку x^*; λ_i – параметр, який характеризує довжину кроку; $d_i = \nabla F(x_i)$ – напрямок пошуку. Значення λ_i може бути знайдене за допомогою одного з методів одновимірного пошуку.

До методів другого порядку відносять: 1) метод Ньютона; 2) модифікований метод Ньютона; 3) метод Марквардта, що є комбінацією методів Коші і Ньютона.

Зважаючи на великий обсяг інформації, у цьому розділі розглянемо два градієнтних методи: 1) метод найскорішого спуску, на прикладі якого можна продемонструвати окремі прийоми, що використовуються при реалізації різних градієнтних алгоритмів; 2) метод ДФП, який відрізняється найширшим використанням при розв'язанні різноманітніших задач.

2.8.1 Метод найшвидшого спуску

Припустимо, що в деякій точці $\bar{x}$ простору керованих змінних вимагається визначити напрямок найшвидшого локального спуску, тобто найбільшого локального зменшення цільової функції. Розкладемо цільову функцію навколо точки $\bar{x}$ в ряд Тейлора:

$$F(x) = F(\bar{x}) + \nabla F(\bar{x})^T \Delta x + \ldots$$

і відкинемо члени другого порядку і вище. Очевидно, що локальне зменшення цільової функції визначається другим доданком, оскільки значення $F(\bar{x})$ фіксоване. Найбільше зменшення F асоціюється з вибором такого напрямку в (4.1), якому відповідає найбільша від'ємна величина скалярного добутку другої складової розкладу. З властивості скалярного добутку випливає, що названий вибір забезпечується при $d(\bar{x}) = -\nabla F(\bar{x})$ і другий доданок набуває вигляду $\lambda \nabla F(\bar{x})^T \nabla F(\bar{x})$. Розглянутий випадок відповідає найшвидшому локальному спуску. Визначення λ доцільно проводити на кожній ітерації $x_{i+1} = x_i - \lambda_i \nabla F(x_i)$. Блок–схема методу найшвидшого спуску наведена на рис. 2.2. Для пошуку мінімуму функції $\varphi(\lambda) = F(x_i + \lambda_i d_i)$ у напрямку d_i з точки x_i використовується метод квадратичної апроксимації.

У точці x_i $\lambda = 0$, і ми вибираємо довжину кроку такою, щоб крок «перекрив» мінімум функції $\varphi(\lambda)$. Похідна $d\varphi/d\lambda = \nabla F(x_i)(x_i + \lambda_i d_i)^T d_i$. У алгоритмі перевіряється умова «перекриття» мінімуму, яка виконується, якщо або $\varphi(\lambda) \geq \varphi(0)$, або $d\varphi(\lambda)/d\lambda \geq 0$. Якщо мінімум не потрапив у відрізок $(0, \lambda)$, то λ подвоюється, і це повторюється стільки разів, скільки необхідно для виконання умови «перекриття».

Упевнившись, що відрізок $(0, \lambda)$ містить мінімум, як третю точку візьмемо точку $\lambda/2$, і потім проводиться квадратична апроксимація з метою пошуку мінімальної точки. У процесі пошуку припускається збіжність до екстремуму, тому для ефективності процедури зменшується довжина кроку. При цьому поділ кроку саме вибирається довільно.

Критерій завершення кожної ітерації повинен лише призводити до зменшення значення функції, в порівнянні із значенням $F(x_i)$ це дозволить скоротити об'єм обчислень порівняно з об'ємом обчислень при точному пошуку мінімуму, який може виявитися дуже значним.

Метод найшвидшого спуску не рекомендується як «серйозна» оптимізаційна процедура, оскільки для практичного застосування «працює» дуже повільно. Це пояснюється тим, що властивість найшвидшого спуску

є лише локальною властивістю і тому необхідна часта зміна напряму, що і приводить у результаті до неефективної обчислювальної процедури.

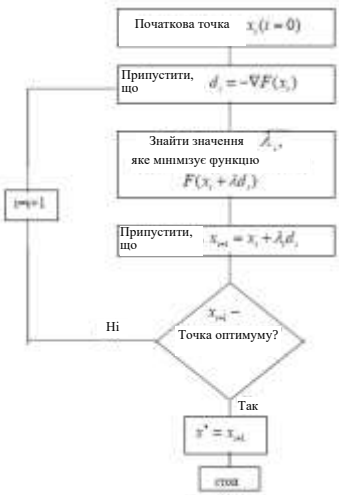

Рисунок 2.2 – Блок–схема методу найшвидшого спуску

2.8.2 Метод Давідона–Флетчера–Пауелла

У цьому методі напрямок пошуку на кроці i є напрямком $H_i \nabla F(x_i)$, де H_i – додатно певна симетрична матриця, яка відновлюється на кожному кроці.

Почнемо пошук з початкової точки x_0, взявши як початкову матрицю H_0 (як правило, одиничну матрицю). Ітераційна процедура може бути представлена таким чином (замість $\nabla F(x_i)$ далі пишеться g_i):

1. На кроці i є точка x_i і додатно визначена симетрична матриця H_i.
2. Як напрямок пошуку береться напрямок $d_i = -H_i g_i$.
3. Щоб знайти функцію λ_i, що мінімізує функцію $F(x_i + \lambda_i d_i)$, необхідно виконати одновимірний пошук уздовж прямої $x_i + \lambda_i d_i$.
4. Припустити, що $v_i = \lambda_i d_i$.
5. Припустити, що $x_{i+1} = x_i + v_i$.
6. Знайти $F(x_{i+1})$ і g_{i+1}. Завершити процедуру, якщо величини $|g_{i+1}|$ або $|v_i|$ достатньо малі. Інакше необхідно продовжити.
7. Припустити, що $u_i = g_{i+1} - g_i$.
8. Обновити матрицю H таким чином: $H_{i+1} = H_i + A_i + B_i$,

де $A_i = v_i v_i^T / \left(v_i^T u_i\right)$; $B_i = -H_i u_i u_i^T H_i / u_i^T H_i u_i$.

9. Збільшити *i* на одиницю і повернутися на крок 2.

Одновимірний пошук необхідно здійснювати кубічною апроксимацією, але пошук закінчується, не досягнувши граничної збіжності, а після того, як буде одержано покращення значення функції.

2.9 Методи оптимізації за наявності обмежень

2.9.1 Поняття штрафної функції

Розглянемо задачу нелінійного програмування такого вигляду:

$$\text{мінімізувати } F(x) \tag{2.31}$$

при обмеженнях

$$g(x) \geq 0,\ j = 1,2,...,m. \tag{2.32}$$

Початкова задача (2.31), (2.32) умовної оптимізації перетвориться в задачу пошуку мінімуму без обмежень (у задачу безумовної оптимізації) функції

$$Z = F(x) + P(r, g_j(x)), \tag{2.33}$$

де $P(r, g_j(x))$ – штрафна функція; r – штрафний параметр.

Необхідно, щоб при порушенні обмежень $P(r, g_j(x))$ «штрафувала» функцію *Z*, тобто збільшувала її значення. У цьому випадку мінімум *Z* знаходитиметься усередині області обмежень. Функція $P(r, g_j(x))$, що задовольняє цю умову, може бути не єдиною. Розглянемо основні типи штрафів.

1 Нескінченний бар'єр. Відповідний вираз набуває нескінченно великих значень в недопустимих точках і нульові значення в допустимих точках. На практиці у якості «нескінченність» використовується велике додатне число, що допускає запис в ЕОМ, наприклад:

$$P = 10^{20} \sum_{j=1}^{m} \left| g_j(x) \right|. \tag{2.34}$$

2 Логарифмічний штраф

$$P = r \sum_{j=1}^{m} \ln\left[g_j(x) \right]. \tag{2.35}$$

Цей штраф додатний при всіх *x*, таких, що $0 < g_j(x) < 1$, і від'ємний при $g_j(x) > 1$.

3 Штраф, заданий зворотною функцією

$$P = r\sum_{j=1}^{m}\left(1/g_j(x)\right) \tag{2.36}$$

не має від'ємних значень в допустимій області.

4 Штраф типу квадрата зрізання

$$P = r\left\langle g_j(x)\right\rangle^2, \quad \left\langle g_j(x)\right\rangle = \begin{cases}\alpha, \alpha \le 0\\ 0, \alpha > 0\end{cases}. \tag{2.37}$$

Розглянемо розв'язок задачі на основі, наприклад, штрафу, заданого зворотною функцією (2.36). Тоді функція (2.33) набуде вигляду

$$Z = \varphi(x,r) = F(x) + r\sum_{j=1}^{m}\left(1/g_j(x)\right). \tag{2.38}$$

Якщо x набуває допустимих значень, тобто значення, для яких $g_j(x) \ge 0$, то Z набуває значення, які більші за відповідні значення $F(x)$ (істинної цільової функції задачі), і різницю можна зменшити за рахунок того, що r може бути дуже малою величиною. Але якщо x набуває значень, які хоча і є допустимими, але близькі до границі області обмежень і, принаймні, одна з функцій $g_j(x)$ близька до нуля, тоді значення функції $P = \left(r,\ g_j(x)\right)$, і, отже, значення функції Z стануть дуже великими. Таким чином, вплив функції $P = \left(r,\ g_j(x)\right)$ полягає у створенні «гребеня з крутими краями» уздовж кожної границі області обмежень. Отже, якщо пошук починається з допустимої точки, і здійснюється пошук мінімуму функції $\varphi(x,r)$ без обмежень, то мінімум буде досягнутий усередині допустимої області для задачі з обмеженнями. Вважаючи r достатньо малою величиною, для того, щоб вплив $P = \left(r,\ g_j(x)\right)$ був малим у точці мінімуму, можна зробити точку мінімуму функції $\varphi(x,r)$ без обмежень такою, що збігається з точкою мінімуму функції $F(x)$ з обмеженнями.

2.9.2 Метод SUMT

Результати попереднього розділу показують, що можна розв'язати задачу мінімізації з обмеженнями, розв'язуючи, для послідовності значень r, яка прагне до нуля, задачу без обмежень наступного вигляду:

мінімізувати $\varphi(x,r) = F(x) + r\sum_{j=1}^{m}\left(1/g_j(x)\right).$

Метод SUMT (sequential unconstrained minimisation technique) вперше запропонований Керролом в 1961 р. Його ідеї були досліджені Фіакко і Маккорміком, які розглянули теоретичні питання і збіжність методу, а також створили практичну систему для його реалізації. На практиці необхідно побудувати обчислювальний метод, що використовує теоретичну властивість збіжності, яка розглянута в попередньому розділі. Для заданих функції $F(x)$ і обмежень $g(x) \geq 0,\ j = 1,2,...,m$ необхідно вибрати початкове значення $r = r_0$, щоб сформувати функцію, яка мінімізується без обмежень методом ДФП, який був наведений раніше. Визначивши мінімум функції $\varphi(x, r_0)$, необхідно зменшити значення r шляхом $r_1 = r_0 / c$, де константа $c > 1$. Потім необхідно мінімізувати функцію $\varphi(x, r_1)$, знову використовуючи метод ДФП. Таким чином, буде розроблена ітераційна процедура. На *к*–му кроці мінімізується функція $\varphi(x, r_k)$, мінімум якої знаходиться у точці x_k^*. Важливо, що її можна використовувати надалі як першу точку в ітераційній процедурі мінімізації функції $\varphi(x, r_{k+1})$, де $r_{k+1} = r_k / c$. Тепер зрозуміло, що послідовність r_k спадає і прямує до нуля, отже, послідовність точок мінімуму сходитиметься до розв'язання задачі з обмеженнями.

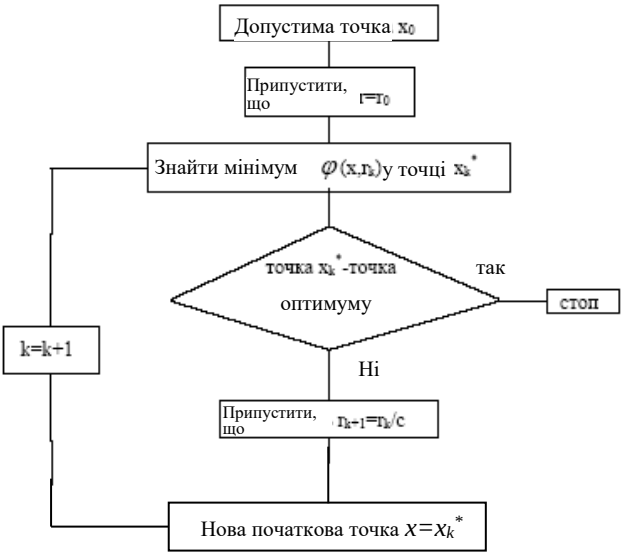

Рисунок 2.3 – Блок–схема методу SUMT

Блок–схема методу наведена на рис.2.3.

Передбачається, що на початку процедури є допустима точка. Важливо, щоб у процесі подальших обчислень одержані точки належали допустимій області. Метод ДФП є градієнтним методом мінімізації, який використовує при одномірному пошуку кубічну апроксимацію. Тоді у міру наближення точки *x* до границі усередині допустимої області $\varphi(x,r) \to \infty$, а у міру наближення точки *x* до границі зовні допустимої області $\varphi(x,r) \to -\infty$. Таким чином, якщо пошук здійснюється уздовж прямої, що з'єднує дві точки, одна з яких лежить всередині, а інша поза областю обмежень, то кубічна апроксимація виявляється неприйнятною, оскільки функція розривна уздовж даної прямої.

Необхідно ретельно досліджувати такі питання при використанні методу ДФП в даній задачі. Цілком ефективним є наступний прийом. Нехай є точка *p* і напрям пошуку $d = -Hg$. Наступна точка $q = p + \lambda d$ необхідна для здійснення кубічної апроксимації. Почнемо із значення $\lambda = 2$ (подвоєний крок у методі Ньютона) і перевіримо, чи є точка *q* допустимою, тобто чи виконується нерівність $g_j(q) > 0$ для всіх *j*. Якщо вона виконується, то λ не змінюється, але якщо нерівність не виконується, то λ замінюється на λ/a, знаходиться нова точка *q* і знову виконується перевірка, поки допустима точка *q* не буде знайдена. Вибір значення *a* не цілком очевидний. Вибір $a = 2$ може бути успішним, при $a = 1{,}05$ довжина кроку стає близькою до відстані до найближчої границі області обмежень і тому є безпечною для інтерполяційної процедури.

Вибір початкового значення *r* важливий з погляду скорочення числа ітерацій при мінімізації функції $\varphi(x,r)$. Якщо спочатку *r* вибрано дуже малим, то метод збігатиметься дуже швидко. Проте такий вибір може призвести до серйозних ускладнень при обчисленнях. Для малих *r* функція $\varphi(x,r)$ швидко змінюватиметься довкола мінімуму, що може викликати ускладнення при використанні градієнтного методу. Дуже велике значення *r* може привести до того, що штрафна функція $P = \left(r,\, g_j(x)\right)$ в рівнянні (2.37) стане домінуючою. Тому «розумний» вибір початкового значення *r* дуже важливий.

Для багатьох задач можна припустити, що $r_0 = 1$. Більш раціональніший підхід грунтується на тому, щоб зрозуміти, що якщо початкова точка *x* лежатиме поблизу мінімуму функції

$\varphi(x,r) = F(x) + r\sum_{j=1}^{m}\left(1/g_j(x)\right) = F(x) + rP(r, g_j(x))$, то градієнт функції $\varphi(x,r)$ буде малий: $\nabla\varphi(x,r) = \nabla F(x) + r\nabla P(r, g_j(x))$. Квадрат норми цього вектора

$$\nabla F(x)^T \nabla F(x) + 2r\nabla F(x)^T \nabla P(r, g_j(x)) + r^2 \nabla P(r, g_j(x))^T \nabla P(r, g_j(x))$$

і мінімум буде досягнутий при

$$r = -\nabla F(x)^T \nabla P(r, g_j(x)) / [\nabla P(r, g_j(x))^T \nabla P(r, g_j(x))] .$$

Таке початкове значення *r,* яке припускають Фіакко і Маккормік, повинне давати позитивні результати в загальному випадку. Функція $\varphi(x,r)$ мінімізується до того часу, поки два послідовні значення F_1 і F_2 не стануть такими, що $\left|(F_1 - F_2)/F_1\right| < 0{,}00001$. Ця умова, звичайно, може бути змінена. Програма закінчує роботу, коли $r\sum_{j=1}^{m}\left(1/g_j(x^*)\right) < 0{,}00001$.

Зазначимо, що записані вище елементні і глобальні вектори вузлових сил статично еквівалентні об'ємним і поверхневим силам, заданим на даному елементі або у всьому об'ємі тіла.

Нарешті, використовуючи основне формулювання принципу можливих переміщень (3.105), згідно з яким віртуальна робота зовнішніх сил дорівнює варіації потенційної енергії деформації тіла:

$$\delta A = \delta\pi , \qquad (3.107)$$

і підставляючи потім в дану рівність отримані вище вирази елементарної роботи і варіації енергії деформації та виконавши порівняння виразів зліва і справа, одержимо таке матричне рівняння:

$$KU = F_V + F_S . \qquad (3.108)$$

Даний вираз є основним розв'язним співвідношенням методу скінченних елементів і є системою лінійних алгебраїчних рівнянь стосовно вузлових переміщень. Після коректного накладення заданих кінематичних граничних умов рівняння може бути розв'язане одним із відомих чисельних методів, які використовуються в теорії розв'язання великих систем з розрідженими матрицями.

4 Після визначення глобального вектора вузлових переміщень

$$U = K^{-1}(F_V + F_S)$$

можуть бути обчислені деформації і напруження в будь–якій точці кожного елемента за записаними вище формулами, а отже, будуть отримані значення деформацій і напружень у довільних точках тіла:

$$e = B^e U^e = B^e a_f^e U = \varepsilon(x,y,z) , \ \sigma = D^e\varepsilon = D^e B^e a_f^e U = \sigma(x,y,z) .$$

Список літератури

1. Зарубин В.С. Математическое моделирование в технике: Учеб. для вузов / Под ред. В.С. Зарубина, А.П. Крищенко. – М.: Изд–во МГТУ им. Н.Э. Баумана, 2001. – 496 с.
2. Лященко М.Я., Головань М.С. Чисельні методи: Підручник. – К.: Либідь, 1996. – 288 с.
3. Турчак Л.И. Основы численных методов: Учебное пособие. – М.: Наука. Гл. ред. физ.– мат. лит., 1987. – 320 с.
4. Калиткин Н.Н. Численные методы. – М.: Наука, 1978. – 512с.
5. Банди Б. Методы оптимизации. Вводный курс / Пер. с англ. – М.: Радио и связь, 1988. – 128 с.
6. Реклейтис Г., Рейвиндран А., Рэгсдел К. Оптимизация в технике: В 2 кн. / Пер. с англ. – М.: Мир, 1986. – 349 с.
7. Маслов Л.Б. Численные методы механики: Курс лекций – Иваново: ИГЭУ, 2001.
8. Зенкевич О. Метод конечных элементов в технике / Перевод с английского под редакцией Б.Е. Победри. –М.: Издательство «Мир», 1975. – 541 с.
9. Норри Д., де Фриз Ж. Введение в метод конечных элементов / Пер. с англ.— М.: Мир, 1981. — 304 с.
10. Felippa C., Introduction to Finite Element Methods, University of Colorado Press, 2002.

ЗМІСТ

ДЛЯ НОТАТОК

Навчальне видання

МЕТОДИЧНІ ВКАЗІВКИ
до лекційних зайнять з дисципліни
«Чисельні методи в фізиці»
для студентів спеціальності
153 «Мікро– та наносистемна техніка»

Укладачі: МІНАКОВА Ксенія Олександрівна
ЗАЙЦЕВ Роман Валентинович
ДРОЗДОВ Антон Миколайович
КІРІЧЕНКО Михайло Валерійович

Відповідальний за випуск Р.В. Зайцев

План 2019 р.

Підписано до друку . Формат 60×84 1/16. Папір друк. №2.
Друк – ризографія. Гарнітура Times New Roman. Ум. друк. арк. 2,9.
Обл.–вид. 4,1. Тираж 50 прим.

Видавничий центр НТУ «ХПІ». 61002, Харків, вул. Кирипичева, 2.
Свідоцтво про державну реєстрацію ДК № 116 від 10.07.2000 р.

Надруковано у типографії ТОВ «Цифра Принт»
на цифровому комплексі Xerox DocuTech 6135.
Свідоцтво про Державну реєстрацію А01 № 432705 від 3.08.2009 г.