НІЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

М.Г. Охріменко, О.А. Жуковська, О.О. Купка

МЕТОДИ РОЗВ'ЯЗУВАННЯ НЕКОРЕКТНО ПОСТАВЛЕНИХ ЗАДАЧ

Підручник
для студентів вищих навчальних закладів

Київ
«Центр учбової літератури»
2008

ББК 65.01я73
О-92
УДК 330.43(075.8)

Рецензенти:
Новицький В.В. – доктор фізико-математичних наук, професор;
Крум П.В. – кандидат економічних наук професор.

М.Г. Охріменко, О.А. Жуковська, О.О. Купка
О-92 Методи розв'язування некоректно поставлених задач: Навч. пос. – К.: Центр учбової літератури, 2008. – 166 с.

ISBN 978-966-364-576-6

У підручнику вперше у вітчизняній навчальній літературі викладено основні принципи розв'язування некоректно поставлених (нестійких) задач (по Ж. Адамару). Викладено і проілюстровано на конкретних прикладах методи розв'язування нестійких (по відношенню до вхідних даних) задач, пов'язаних з іменами А.Н. Тихонова та В. К. Іванова. Значну увагу в підручнику приділено розв'язку некоректних задач економічного походження: систем лінійних алгебраїчних рівнянь, задач лінійного програмування, нелінійних екстремальних задач, інтегральних рівнянь І роду тощо. Значну увагу приділено підбору та оцінкам залежностей параметрів регуляризації задач від вихідних даних (параметрів) математичних моделей. Приведено методологію алгоритмічного уточнення вихідних даних моделей, якщо ці дані одержуються обчислювальним шляхом.

Для студентів, аспірантів, викладачів вищих навчальних закладів.

ISBN 978-966-364-576-6

© М.Г. Охріменко, О.А. Жуковська, О.О. Купка, 2008
© Центр учбової літератури, 2008

Зміст

ПЕРЕДМОВА

Основою наукових досліджень є різноманітні математичні моделі процесів та явищ, які відбуваються під час практичної діяльності людини.

Побудова математичних моделей ґрунтується на вхідних даних, які одержуються або з показників технічних пристроїв, або є результатом чисельних експериментів на обчислювальних агрегатах. І в першому, і в другому випадках вихідні дані одержуються, як правило, в наближеному варіанті. Це означає, що вхідні дані, на основі яких побудовані математичні моделі, мають деякі похибки.

Математичні моделі будуються для різностороннього вивчення процесів та явищ з метою їх дослідження. Самі дослідження проводяться для одержання практичних зисків або уникнення негативних наслідків в результаті виникнення явищ, процесів. Тому на основі побудованих математичних моделей ставляться різноманітні задачі, розв'язування яких і допомагає зробити конкретні обґрунтовані висновки стосовно плину цих процесів, явищ. Розв'язок поставлених задач здійснюється, як правило, за допомогою синтезованих для цих задач алгоритмів на обчислювальних комплексах.

Серед математичних задач, що виникають з побудованих математичних моделей (на неточній вхідній інформації), існує широкий клас задач, розв'язки яких нестійкі по відношенню до малих змін вхідних даних. Вони характеризуються тим, що як завгодно малі зміни вхідних даних тягнуть за собою досить великі зміни розв'язків. Подібні задачі, по суті, є погано поставленими. Вони належать до класу некоректно (нерозумно) поставлених задач.

Якщо вхідні дані відомі наближено, то згадана нестійкість приводить до практичної неєдиності розв'язку в рамках заданої точності і до великих труднощів в інтерпретації сенсу одержаного наближеного розв'язку. Через ці особливості довгий час вважалося, що некоректно поставлені задачі не можуть мати практичного значення, бо яке б не було уточнення моделі за допомогою уточнення вихідних даних, воно не приводить до уточнення розв'язку задачі. Як правило, спосіб побудови коректно поставленої задачі невідомий.

Однак можна вказати некоректно поставлені задачі, що відносяться як до класичних розділів математики, так і до різноманітних класів практично важливих прикладних задач.

Це дозволяє робити висновки про широту класу розв'язуваних задач. Практична діяльність дослідників явищ, процесів в природі та суспільстві дозволяє зробити висновок, що потужність класу розглядуваних задач значно перевищує потужність класу коректно поставлених задач. До числа некоректно поставлених задач відносяться задачі створення систем обробки результатів експерименту, задачі оптимального керування, задачі оптимального проектування систем, задачі оптимального планування та багато інших.

Одним із суттєвих етапів обробки даних є розв'язок задач, не стійких до малих змін вхідних даних. Тому не виникає сумніву в необхідності розробки методів розв'язку таких задач. При цьому наближені розв'язки, які отримуються з наближених вхідних даним, повинні бути стійкими до малих змін останніх.

Вхідні дані некоректно поставлених задач одержуються, як правило, в результаті вимірів або обчислень, містять випадкові похибки. З цієї причини при побудові наближених розв'язків та при оцінці їх похибок залежно від характеру вхідної інформації, можливий як детермінований, так і ймовірнісний підхід. У даному посібнику ми обмежимося, в основному, детермінованим варіантом.

У посібнику приводяться методи регуляризації побудови наближених розв'язків некоректно поставлених задач, розроблені в працях [1-4].

Автори не ставили за мету дати повний виклад проблем при розв'язку некоректно поставлених задач. Ставилося завдання ознайомити студентів, аспірантів, викладачів, інженерів, економістів і наукових співробітників з основними проблемами при розв'язуванні задач, що виникають під час розв'язування задач математичної обробки даних, прогнозування експерименту, планування в економічній діяльності та інше.

Автори вважають своїм обов'язком висловити щиру вдячність Петру Васильовичу Крушу за критичні зауваження, зроблені в процесі підготовки посібника, до друку що сприяли його суттєвому удосконаленню.

Автори: Охріменко М. Г.,
Жуковська О.А.,
Купка О. О.

Кожна аналітична проблема завжди коректно поставлена, якщо проблема має механічну чи фізичну інтерпретацію
(Жак Адамар)

Коректно поставлені задачі – це далеко не єдині задачі, які правильно відображають фізичні явища.
(Лаврентьєв М.М.)

ВСТУП

Після виникнення сучасної електронної техніки почалося бурхливе зростання її широкого використання для розв'язання широких класів задач, які виникали в різноманітних галузях людської діяльності: науці, техніці, економіці, соціології тощо. Для розв'язання назрілих задач на обчислювальних пристроях необхідні чіткі схеми досягнення розв'язку (алгоритми). Почався процес швидкого розвитку обчислювальних алгоритмів для широких класів задач. Але слід усвідомлювати, що треба розуміти під „розв'язком" задачі, яким вимогам повинні задовольняти алгоритми знаходження „розв'язків"? Класичні концепції і постановки задач не відображали багатьох особливостей тих задач, які зустрічаються в практичній діяльності, а термін „розв'язок задачі" носив розпливчастий характер. Слід зауважити, що найсучасніші обчислювальні пристрої, як правило, здійснюючи чисельні розрахунки і аналіз різноманітних явищ, оперують неточною, наближеною інформацією. Інформація для роботи алгоритмів, програм прикладних задач одержується часто або з технічних пристроїв, або є результатом інших обчислень. Для різних прикладних задач потрібна точність вихідних даних різна і зумовлена потребою одержати достовірний розв'язок задачі з наперед заданим рівнем достовірності (наперед заданою точністю).

Існує досить широкий клас практичних задач, для яких точність одержаних розв'язків залежить від точності вхідних даних так, що підвищення точності вхідних даних тягне за собою підвищення точності розв'язку. Але така ситуація має місце не завжди.

Наведемо приклад. Розглянемо систему лінійних алгебраїчних рівнянь:

$$Ax = y, \qquad (1)$$

де x – шуканий вектор, y – відомий вектор, $A = \{a_{ij}\}$ – квадратна матриця з відомими елементами a_{ij}, $i, j = 1, 2, ..., n$.

Якщо система (1) не вироджена [5], то існує $\det A \neq 0$ ($\det A$ – визначник матриці A), то вона має єдиний розв'язок, котрий можна знайти за відомими формулами Крамера або за допомогою схеми Гаусса, чи будь-яким іншим методом.

Якщо система (1) вироджена, то вона має розв'язок (до того ж не єдиний) лише за виконання умов розв'язуваності, які складаються з рівності нулю відповідних визначників [5].

Таким чином, перш ніж розв'язувати систему (1), треба перевірити, вироджена вона чи ні. Для цього необхідно обчислити визначник системи $\det A$.

Якщо n – порядок системи, то для обчислення $\det A$ треба виконати порядку n^3 арифметичних операцій. З якою б точністю ми не здійснювали обчислення, при досить великому значенні n, внаслідок накопичення похибок обчислень, ми можемо одержати значення $\det A$, яке може суттєво відрізнятись від точного [6]. З цієї причини бажано будувати такі алгоритми знаходження розв'язку системи (1), котрі не вимагають попереднього вияснення факту виродженості чи невиродженості її.

Крім того, в практичних задачах часто права частина y і елементи матриці A, коефіцієнти системи рівнянь (1), відомі нам наближено. В цих випадках, замість системи (1), ми маємо справу з деякою іншою системою:

$$\widetilde{A}x = \widetilde{y} \tag{2}$$

такою, що $\left\|\widetilde{A} - A\right\| \le h$, $\left\|\widetilde{y} - y\right\| \le \delta$, h і δ – достатньо малі числа, цей сенс норм є запозиченим з [6],[12].

Маючи замість матриці A матрицю $\widetilde{A}$, ми тим більше не можемо висловити ніякого судження про виродженість чи невиродженість системи.

У цьому випадку про точну систему $Ax = y$ нам відомо лише те, що для матриці A і правої частини y виконуються нерівності $\left\|\widetilde{A} - A\right\| \le h$ і $\left\|\widetilde{y} - y\right\| \le \delta$. Але систем з такими вхідними даними (A, y) нескінченно багато, і в рамках відомого нам рівня похибок вони не відрізняються. Серед таких „можливих точних систем" можуть бути і вироджені системи.

Оскільки замість точної системи (1) ми маємо наближену систему (2), то мова може йти лише про знаходження наближеного

розв'язку. Але наближена система (2) може бути і нерозв'язуваною. Виникає питання, що треба розуміти під наближеним розв'язком системи (1)? Він повинен також бути стійким по відношенню до малих змін вхідних даних (A, y).

Задачі, для яких малим змінам вхідних даних відповідають достатньо малі зміни вихідних даних називаються *коректно поставленими*. Вперше на такі задачі звернув увагу великий французький математик Жак Адамар. Ним і було введене поняття коректно поставленої задачі. Воно визначалося такими умовами (пояснимо на прикладі розв'язку системи (1)):

1) для всякого y існує розв'язок x у множині дійсних чисел;

2) розв'язок визначається однозначно (єдиний);

3) задача стійка (це означає, що для всякого $\varepsilon > 0$ знайдуться пара додатніх чисел $\delta(\varepsilon)$, $h(\varepsilon)$, які задовольняють умовам: із того, що $\left\|\widetilde{A} - A\right\| \le h(\varepsilon)$ і $\left\|\widetilde{y} - y\right\| \le \delta(\varepsilon)$ випливає виконання умови $\left\|\widetilde{x} - x\right\| \le \varepsilon$, де $\widetilde{x}$ – розв'язок системи (2)).

Задачі, які не задовольняють хоч одній з цих умов, називають **некоректно поставленими**.

В математичній літературі довгий час існувала думка, згідно з якою всяка математична задача повинна задовольняти цим умовам. Жак Адамар також вважав, що розглядати некоректно поставлені задачі немає сенсу.

Коректна постановка задачі часто трактувалася як умова, якій повинна задовольняти всяка математична задача, відповідно якої-небудь фізичної чи технічної задачі. Домінуюча думка серед математиків та авторитет Жака Адамара ставили під сумнів доцільність вивчення некоректно поставлених задач. Однак така думка, цілком природна для застосування до деяких явищ, що розвиваються в часі, не може бути перенесена на всі задачі. У даному посібнику будуть наведені приклади некоректно поставлених задач, що відносяться як до основного апарата математики, так і до широкого класу прикладних задач техніки, економіки, соціології і т.д. Широким класом некоректно поставлених задач є, так звані, обернені задачі.

Наприклад, застосування методу найменших квадратів (МНК) для побудови математичних моделей реальних процесів за результатами конкретних вимірів вихідних параметрів приводить до необхідності розв'язку некоректно поставлених задач [2,3,4].

До числа важливих задач відносяться задачі створення систем автоматичної математичної обробки результатів фізичного

експерименту. Одним із етапів обробки є розв'язок обернених задач виду (1) відносно x.

Велика кількість сучасних експериментальних установок для дослідження різноманітних фізичних явищ і об'єктів є складними і дороговартісними комплексами (прискорювачі елементарних часток, устаткування для одержання і дослідження високотемпературної плазми, для дослідження властивостей речовин за понаднизьких температур та ін.).

Бажання мати надійну інформацію про досліджуване явище, вивчення „рідких" і „слабких" ефектів часто приводить до необхідності багаторазового повторення однотипного експерименту. Автоматизація проведення експерименту і способів реєстрації його результатів дозволяють одержувати за короткий час досить великий обсяг необхідної інформації (десятки і сотні тисяч знімків, осціолограм, показників детекторів і т.п.). Для одержання із цієї інформації необхідних характеристик явища чи об'єкта, що вивчається, потрібна наступна обробка результатів спостережень. У багатьох випадках цю обробку треба проводити практично одночасно з проведенням експерименту або з невеликим доступним зсувом у часі. Таку обробку, яка вимагає переробки великого обсягу інформації, можна провести лише за допомогою сучасних електронних обчислювальних пристроїв (ЕОМ).

Для широкого класу експериментальних пристроїв можна виділити такі етапи обробки результатів спостережень.

Перший етап. Зняття інформації з реєструючої апаратури або з постійного носія (наприклад, з фото- чи кіноплівки), переведення її в числовий код і пересилання в пам'ять ЕОМ.

Другий етап. Первинна обробка. Вона може включати нормування даних спостерігача, приведення до встановленої системи відліку, статистичну обробку з оцінкою ступеня довіри, фільтрацію інформації тощо. Метою її є одержання „вихідних результатів" („вихідної кривої") експерименту.

Третій етап. Інтерпретація результатів, одержаних на другому етапі обробки. Вона складається, як правило, в оцінці шуканих характеристик моделі явища чи об'єкта, що вивчається.

У фізичному експерименті не реєструються, як правило, характеристики явища x, які нас не цікавлять, а лише деякі прояви $y = Ax$. Тому задача інтерпретації зводиться до розв'язку рівняння $Ax = y$. В багатьох випадках ця задача є некоректно поставленою. У даній роботі буде приділено увагу в основному методам розв'язку некоректно поставлених задач, які відносяться до систем лінійних

алгебраїчних рівнянь, інтегральних рівнянь першого роду, методів розв'язку задач лінійного програмування, методів наближення функції, чисельного диференціювання, розв'язку екстремальних лінійних та нелінійних задач та ін.

У роботі приділено увагу основним аспектам розв'язування некоректно поставлених задач, тому посібник не слід розглядати, як довідник з розв'язку нестійких задач.

§1. СИСТЕМИ ЛІНІЙНИХ АЛГЕБРАЇЧНИХ РІВНЯНЬ (СЛАР)

Розглянемо систему m рівнянь з n невідомими

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m, \end{aligned} \qquad (1.1)$$

де для розмірностей m, n може виконуватись одне з трьох співвідношень:

$$\begin{aligned} &m < n, \\ &m > n, \\ &m = n. \end{aligned}$$

Тут і надалі будемо використовувати термінологію і позначення з [1].

Далі будемо використовувати компактний векторно-матричний запис системи (1.1)

$$Ax = b, \qquad (1.2)$$

де A – матриця коефіцієнтів, яка містить m рядків і n стовпчиків

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix},$$

x – шуканий n-компонентний (n-мірний) вектор розв'язку,

b – заданий m –компонентний вектор-стовпчик правих частин системи (1.2)

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_m \end{bmatrix}, \quad x^T = x' = (x_1, \quad x_2, \quad \ldots, \quad x_n), \qquad (1.3)$$

x^T – транспонований вектор x.

Сукупність n-компонентних векторів x виду (1.3) складають n-мірний лінійний векторний простір R^n, в якому визначена операція додавання векторів

$$x + y = \begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \ldots \\ x_n + y_n \end{bmatrix}$$

і множення вектора x на дійсне число λ

$$\lambda \cdot x = \lambda \cdot \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} \lambda \cdot x_1 \\ \lambda \cdot x_2 \\ \dots \\ \lambda \cdot x_n \end{bmatrix},$$

які задовольняють наступним властивостям:

1) $x + y = y + x,$
2) $x + (y + z) = (x + y) + z,$
3) $\lambda \cdot (\mu \cdot x) = (\lambda \cdot \mu) \cdot x,$ λ, μ - дійсні числа,
4) $\lambda \cdot (x + y) = \lambda \cdot x + \lambda \cdot y,$
5) $(\lambda + \mu) \cdot x = \lambda \cdot x + \mu \cdot x.$

Для векторів *n*-мірного простору вводимо скалярний добуток за формулою

$$(x, y) = \sum_{i=1}^{n} x_i \cdot y_i \,, \tag{1.4}$$

для якого виконуються наступні властивості:

1) $(x, x) > 0,$ якщо $x \neq 0$; $(x, x) = 0,$ якщо $x = 0$, тоді $x' = (0 \quad 0 \quad \dots \quad 0)$;
2) $(x, y) = (y, x)$;
3) $(x_1 + x_2, y) = (x_1, y) + (x_2, y),$ $(x, y_1 + y_2) = (x, y_1) + (x, y_2)$;
4) $(\lambda \cdot x, y) = \lambda \cdot (x, y),$ $(x, \lambda \cdot y) = \lambda \cdot (x, y).$

Вектори $x^{(1)}, x^{(2)}, \dots, x^{(k)}$ будемо називати *лінійно залежними*, якщо існують такі постійні числа $c_1, c_2, \dots, c_k$, не рівні нулю одночасно, що виконується рівність

$$c_1 x^{(1)} + c_2 x^{(2)} + \dots + c_k x^{(k)} = 0 .$$

Якщо ж ця рівність виконується тільки у разі, коли $c_1 = c_2 = \dots = c_k = 0$, тоді *вектори* називаються *лінійно не залежними*. *Система векторів* $\left\{x^{(1)}, x^{(2)}, \dots, x^{(k)}\right\}$ називається *базисом* простору R^n, якщо будь-який вектор *x* є лінійною комбінацією цих векторів:

$$x = c_1 x^{(1)} + c_2 x^{(2)} + \dots + c_n x^{(n)}, \ \forall i = 1, \dots, n, \ c_i \neq 0.$$

§2. НОРМИ ВЕКТОРА ТА МАТРИЦІ

Для того щоб визначити поняття границі векторів та матриці і оцінити швидкість збіжності до розв'язку ітераційного процесу при розв'язуванні систем лінійних рівнянь ітераційними методами, необхідно ввести міру близькості між векторами та матрицями. Це можна зробити за допомогою поняття норми.

Формально нормою вектора x називається зівставлення цьому векторові невід'ємне число $\|x\|$, яке задовольняє трьом умовам:

1) $\|x\| > 0$, якщо $x \neq 0$ і $\|0\| = 0$;

2) $\|cx\| = |c| \cdot \|x\|$ для будь-якого числового множника c;

3) $\|x + y\| \leq \|x\| + \|y\|$ („нерівність трикутника").

Із умов 2)-3) можна вивести також нерівність

4) $\big|\|x\| - \|y\|\big| \leq \|x - y\|$.

Наведемо конкретні способи задання норми n-мірного вектора x, $x' = (x_1 \quad x_2 \quad \ldots \quad x_n)$, $\|x'\| = \|x\|$:

I) $\|x\|_{\mathrm{I}} = \max\limits_{i=1,\ldots,n} |x_i|$;

II) $\|x\|_{\mathrm{II}} = |x_1| + |x_2| + \ldots + |x_n|$;

III) $\|x\|_{\mathrm{III}} = \sqrt{|x_1|^2 + |x_2|^2 + \ldots + |x_n|^2} = \sqrt{(x,x)}$;

IV) $\|x\|_{\mathrm{IV}} = \sqrt[p]{|x_1|^p + |x_2|^p + \ldots + |x_n|^p}$, $p > 1$.

Легко перевірити, що умови 1) – 3) для них виконуються.

Поняття норми вектора узагальнює поняття модуля (довжини) вектора. *Норма* $\|x\|_{\mathrm{III}}$ називається *евклідовою* і є *модулем вектора x*.

Будемо вважати, що послідовність векторів $\{x^{(k)}\}$ збігається до вектора x, якщо $\lim\limits_{k\to\infty} \|x^{(k)} - x\| = 0$, де $\|x^{(k)} - x\|$ – будь-яка з норм I-IV. Збіжність по кожній з цих норм означає, що для будь-якої компоненти x_i ($i = 1, 2, \ldots, n$) $\lim\limits_{k\to\infty} x_i^{(k)} = x_i^0$, де $x'^{(k)} = \left(x_1^{(k)}, \ x_2^{(k)}, \ \ldots, \ x_n^{(k)}\right)$, $x = (x_1, \ x_2, \ \ldots, \ x_n)$.

Дійсно, для норми $\|x\|_{\mathrm{I}}$ це очевидно. Для інших норм це випливає із очевидних нерівностей:

$$\|x\|_{\mathrm{I}} \leq \|x\|_{\mathrm{II}} \leq n\|x\|_{\mathrm{I}},$$
$$\|x\|_{\mathrm{I}} \leq \|x\|_{\mathrm{III}} \leq \sqrt{n}\|x\|_{\mathrm{I}},$$
$$\|x\|_{\mathrm{I}} \leq \|x\|_{\mathrm{IV}} \leq \sqrt[p]{n}\|x\|_{\mathrm{I}}.$$

Аналогічно вводиться норма матриці A, як невід'ємне число, яке задовольняє властивостям:

1) $\|A\| > 0$, якщо $A \neq 0$ і $\|0\| = 0$, якщо $A \equiv 0$;

2) $\|cA\| = |c| \cdot \|A\|$, де c – числовий множник;

3) $\|A + B\| \leq \|A\| + \|B\|$;

4) $\|A \cdot B\| \leq \|A\| \cdot \|B\|$.

Добуток матриць $A = \{a_{ij}\}_{i,j=1}^{m,p}$, $B = \{b_{ij}\}_{i,j=1}^{p,m}$ в 4) визначається звичайним способом $C = \{c_{kl}\}_{k,l=1}^{m,n}$, де коефіцієнти $C_{kl} = \sum_{i=1}^{p} a_{ki} b_{il}$ $(k = 1, 2, ..., m, \quad l = 1, 2, ..., n)$.

Будемо говорити, що норма матриці узгоджена з даною нормою векторів, якщо для будь-якої матриці A і будь-якого вектора x виконується нерівність

$$\|Ax\| \leq \|A\| \cdot \|x\|. \tag{2.1}$$

Розмірності матриці A та вектора x такі, що кількість стовпчиків в матриці A збігається з кількістю компонент вектора x [5].

Наведемо способи побудови норм, узгоджених з заданою векторною нормою:

$$\|A\| = \max_{\|x\|=1} \|Ax\|,$$

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}, \tag{2.2}$$

$$\|A\| = \min\{M : \|Ax\| \leq M \cdot \|x\|\}.$$

Безпосередньою перевіркою можна переконатися, що кожний з трьох способів (2.2) визначає одну і ту ж норму, для якої виконані властивості 1) – 4) і нерівність (2.1). Побудовану таким чином *норму матриць* називають *підпорядкованою даній нормі векторів*.

Випишемо норми матриць (розмірністю $m \times n$), які підпорядковані векторним нормам I – III, відповідно

$$\|A\|_{\text{I}} = \max_{i = \overline{1,m}} \sum_{k=1}^{n} |a_{ik}|,$$

$$\|A\|_{\text{II}} = \max_{k = \overline{1,n}} \sum_{i=1}^{m} |a_{ik}|,$$

$$\|A\|_{\text{III}} = \sqrt{\lambda_{\max}},$$

де $\lambda_{\max}$ – найбільше власне число матриці $A^* A$ (A^* – транспонована до A матриця).

Нагадаємо, що власним числом матриці A називається число λ (взагалі кажучи комплексне), для якого система однорідних рівнянь

$$A \cdot x - \lambda \cdot x = 0,$$
$$\det(A - \lambda \cdot E) = 0, \qquad (2.3)$$

$E = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$ – одинична матриця,

має ненульовий розв'язок x_λ, а x_λ називається власним вектором матриці A. Зауважимо, що власні числа матриці $A^* A$ невід'ємні [12].

У деяких випадках зручно використовувати ще й такі матричні норми, як:

$$\|A\|_{\mathrm{V}} = \sqrt{\sum_{i,j=1}^{m,n} |a_{ij}|^2},$$
$$\|A\|_{\mathrm{VI}} = n \max_{i,j=\overline{1;m,n}} |a_{ij}|.$$

Перша із записаних норм не підпорядкована ніякій векторній нормі, але узгоджена, наприклад, з евклідовою нормою III, а друга узгоджена з нормами I – III. Помітимо, що для всякої норми матриці, підпорядкованої нормі векторів, $\|E\| = 1$. Надалі, як правило, будемо використовувати евклідову норму та її підпорядковану матричну норму.

У подальшому нам буде потрібне поняття додатної напіввизначеної матриці. Це симетрична матриця ($A^* = A$) і $(Ax, x) \geq 0$ для будь-якого $x \in R^n$ або еквівалентно: всі власні числа λ – невід'ємні.

На відміну від елементарної алгебри ми будемо вивчати системи лінійних алгебраїчних рівнянь (СЛАР) з довільним числом рівнянь і невідомих. Ми не будемо вважати, що число рівнянь зберігається з числом невідомих.

Нехай нам дана система m рівнянь з n невідомими $m \neq n$. Домовимось використовувати символіку (1.1).

Розв'язком СЛАР (1.1) називається така *система чисел* $k_1, k_2, ..., k_n$, яка кожне з рівнянь (1.1) перетворює в тотожність після

заміни в ньому невідомих x_i відповідними числами k_i $(x_i=k_i)$ $\forall i = \overline{1,n}$.

СЛАР будемо називати *сумісною*, якщо вона має хоча б один розв'язок (хоч один набір чисел k_i).

СЛАР називається *визначеною*, якщо вона має єдиний розв'язок. *СЛАР* будемо називати *невизначеною*, якщо вона має більше ніж один розв'язок. Пізніше ми встановимо, що сумісні системи мають або один розв'язок або нескінченно багато.

Приклад 2.1.

Система з двох рівнянь

$$\begin{cases} x_1 + 2x_2 = 7, \\ x_1 + x_2 = 4 \end{cases}$$

визначена: вона має єдиний розв'язок $x_1 = 1,\ x_2 = 3$.

Приклад 2.2.

Система

$$\begin{cases} 3x_1 - x_2 = 1, \\ 6x_1 - 2x_2 = 2 \end{cases} \tag{1.5}$$

невизначена, бо вона має нескінченно багато розв'язків

$$x_1 = k,\ \ x_2 = 3k - 1, \tag{1.6}$$

де число k довільне дійсне число (таких чисел безліч).

Формулами (1.6) вичерпуються всі розв'язки системи (1.5).

Задача теорії СЛАР являє собою розробку методів, які дозволяють встановити, чи є сумісною дана система рівнянь; у випадку сумісності, встановити число розв'язків, а також вказати спосіб знайти ці розв'язки.

Ми почнемо з методу, найбільш зручного для практичного пошуку розв'язків СЛАР з числовими коефіцієнтами – методу Гаусса. Крім зручності у застосуванні, метод Гаусса має переваги перед іншими методами ще й у тому, що при обчисленні на ЕОМ має найменшу похибку заокруглень і є скінченним (забезпечує знаходження розв'язку за кінцеву кількість обчислювальних кроків, у випадку сумісності системи, або за таку ж кількість кроків встановлює відсутність розв'язків).

§3. МЕТОД ГАУССА

Спочатку зробимо зауваження. Нам в майбутньому доведеться робити такі перетворення СЛАР: обидві частини деякого рівняння системи, помноженого на одне і те ж число, віднімати від відповідних частин деякого іншого рівняння системи. Нехай для визначеності ми обидві частини першого рівняння системи (1.1), помножені на число *c*, віднімемо від відповідних частин другого рівняння. Ми одержимо нову систему лінійних рівнянь:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1, \\ a'_{21}x_1 + a'_{22}x_2 + ... + a'_{2n}x_n = b_2, \\ a_{31}x_1 + a_{32}x_2 + ... + a_{3n}x_n = b_3, \\ ..... \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = b_m, \end{cases} \qquad (3.1)$$

де $a'_{2j} = a_{2j} - c \cdot a_{1j}$, $\forall\ j = \overline{1,n}$, $b'_2 = b_2 - c \cdot b_1$.

Системи рівнянь (1.1) і (3.1) будемо називати еквівалентними, тобто вони є або обидві несумісні, або обидві сумісні і мають одні й ті ж розв'язки. І справді, нехай $k_1, k_2, ..., k_n$ – це довільний розв'язок системи (1.1). Ці числа задовольняють всім рівнянням (3.1), крім другого, бо всі рівняння (3.1), крім другого, збігаються з відповідними рівняннями (1.1). Але числа $k_1, k_2, ..., k_n$ задовольняють і другому рівнянню системи (3.1), так як воно виражається через перше та друге рівняння (1.1), яким ці числа задовольняють. І навпаки, будь-який розв'язок системи (3.1) буде задовольняти систему (1.1). І справді, друге рівняння системи (1.1) можна одержати, віднявши від обох частин другого рівняння (3.1) відповідні частини першого рівняння цієї системи, помножені на число *c*.

Зрозуміло, що якщо до системи (1.1) будуть декілька разів застосовані перетворення подібного виду, то нова одержана система залишиться еквівалентною до вихідної системи (1.1).

Може статися, що після виконання таких перетворень в нашій системі з'явиться рівняння, всі коефіцієнти лівої частини якого дорівнюють нулю. Якщо і вільний член цього рівняння буде рівний нулю, тоді рівняння задовольняється при всіх значеннях невідомих. Вивівши таке рівняння з системи рівнянь, ми отримаємо систему рівнянь, еквівалентну вихідній. Якщо ж вільний член рівняння, яке розглядається, буде відмінний від нуля, то це рівняння не може бути задоволене (перетворене в тотожність) ні за яких значень

невідомих. Одержана нами система рівнянь, рівно як і еквівалентна їй вихідна система, буде несумісна.

Перейдемо тепер до викладу суті методу Гаусса. Нехай задана СЛАР (1.1). Для визначеності покладемо $a_{11} \neq 0$. Насправді коефіцієнт a_{11} може бути рівний нулю, але тоді ми повинні вибрати рівняння в (1.1), яке має коефіцієнт при x_1 відмінний від нуля, переставити ці рівняння місцями, і відповідно змінити індексацію в матриці $A = \{a_{ij}\}_{i,j=1}^{m,n}$. Тоді умова $a_{11} \neq 0$ буде виконана.

Перетворюємо тепер систему (1.1), виключаючи невідоме x_1 з усіх рівнянь, крім першого. Для цього обидві частини першого рівняння помножимо на число $\frac{a_{21}}{a_{11}}$ і віднімемо від відповідних частин другого рівняння. Далі обидві частини першого рівняння, помножені на число $\frac{a_{31}}{a_{11}}$ вирахуємо з відповідних частин третього рівняння і так далі.

Ми прийдемо, таким шляхом, до нової системи з *m* рівнянь та *n* невідомими

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1, \\ a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n = b'_2, \\ a'_{32}x_2 + a'_{33}x_3 + \ldots + a'_{3n}x_n = b'_3, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a'_{m2}x_2 + a'_{m3}x_3 + \ldots + a'_{mn}x_n = b'_m, \end{cases} \qquad (3.2)$$

Система (3.2) еквівалентна системі (3.1).

Будемо перетворювати тепер систему (3.2). Перше рівняння цієї системи залишимо без змін. Перетворювати будемо далі всі рівняння системи, крім першого. При цьому ми будемо вважати, що серед цих рівнянь немає таких, всі коефіцієнти лівих частин яких рівні нулю. Якщо б такі рівняння з'явилися, то ми б їх вилучили з системи, якби праві їх частини теж дорівнювали нулю, або оголосили, що дана СЛАР немає розв'язків, якщо праві частини цих рівнянь відмінні від нуля. Таким чином, серед коефіцієнтів a_{ij} є відмінні від нуля; для визначеності приймемо, що $a_{22} \neq 0$. Перетворимо тепер систему (3.2), відраховуючи з обох частин третього і кожного з наступних рівнянь обидві частини другого рівняння, помножені відповідно на числа $\frac{a'_{32}}{a'_{22}}, \frac{a'_{42}}{a'_{22}}, \ldots, \frac{a'_{m2}}{a'_{22}}$.

Таким чином, із рівнянь системи, крім першого та другого, буде вилучено невідоме x_2. Ми прийдемо до наступної системи рівнянь, еквівалентної до систем (1.1) та (3.2):

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1, \\ a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n = b'_2, \\ a''_{33}x_3 + \ldots + a''_{3n}x_n = b''_3, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a''_{t3}x_3 + \ldots + a''_{tn}x_n = b''_t. \end{cases}$$

Наша система має тепер t рівнянь $t \le m$, бо деякі рівняння могли бути відкинуті. Зауважимо, що число рівнянь могло зменшитися вже після першого кроку Гаусса в системі (3.2).

У подальшому перетворенням підлягає лише частина одержаної системи, яка містить всі рівняння, крім двох перших. І так далі.

Виконуючи послідовно кроки Гаусса (послідовного виключення), ми можемо прийти до такої системи, в якій є рівняння, вільний член якого відмінний від нуля, а всі коефіцієнти лівої частини рівні нулю. Тоді вихідна система несумісна. Якщо ж такий випадок місця не має, то ми одержимо систему рівнянь еквівалентну (1.1):

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1, \\ a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n = b'_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a^{(k-2)}_{(k-1)(k-1)}x_{k-1} + a^{(k-2)}_{(k-1)k}x_k + \ldots + a^{(k-2)}_{(k-1)n}x_n = b^{(k-2)}_{k-1}, \\ a^{(k-1)}_{kk}x_k + \ldots + a^{(k-1)}_{kn}x_n = b^{(k-1)}_k \end{cases} \qquad (3.3)$$

Тут $a_{11} \neq 0$, $a'_{22} \neq 0, \ldots, a^{(k-2)}_{(k-1)(k-1)} \neq 0$, $a^{(k-1)}_{kk} \neq 0$. Відмітимо, що $k \le s$ і, очевидно, $k \le n$.

У цьому випадку система (1.1) сумісна. Вона буде визначеною при $k = n$ і не визначеною при $k < n$.

І справді, якщо $k = n$, то система (3.3) матиме вигляд

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1, \\ a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n = b'_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a^{(n-1)}_{nn}x_n = b^{(n-1)}_n \end{cases} . \qquad (3.4)$$

Із останнього рівняння ми одержимо значення $x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}$. Підставляючи його в передостаннє рівняння і розв'язуючи одне рівняння з одним невідомим, одержимо x_{n-1}. Продовжуючи так далі (здійснюємо обернені кроки Гаусса), ми знайдемо всі значення $x_n, x_{n-1}, ..., x_1$.

Якщо ж $k < n$, то в (3.3) невідомі $x_{k+1}, ..., x_n$ можна покласти рівними довільним числам числової прямої. Їх будемо називати вільними змінними. Всі інші змінні можуть бути визначені здійснюючи обернений хід Гаусса, а саме $x_k = b_k^{(k-1)} - \frac{a_{k\,k+1}^{(k-1)} x_{k+1} + ... + a_{kn}^{(k-1)} x_n}{a_{kk}^{(k-1)}}$. Підставляючи значення x_k в $k-1$ рівняння і розв'язуючи його відносно x_{k-1}, знайдемо вираз для цієї змінної, як функції відносно вільних змінних $x_{k+1}, ..., x_n$ і т. д. Звідси випливає нескінченність розв'язків системи (3.3), а значить невизначеність системи (1.1).

Спостереження показують, що трикутна форма (3.4) СЛАР або „трапецієподібна" форма (3.3) (при $k < n$) одержана при допущенні, що коефіцієнти a_{11}, a'_{22} і т. д. відмінні від нуля. В загальному випадку система рівнянь, до якої ми прийдемо, здійснюючи прямий хід Гаусса (виключення невідомих), набуває трикутної чи трапецієподібної форми лише після належної зміни нумерації невідомих.

Узагальнюючи вищевикладене, ми можемо стверджувати, що метод Гаусса можна застосовувати до довільної СЛАР. При цьому система буде несумісною, якщо в процесі прямих перетворень Гаусса ми одержимо рівняння, в якому коефіцієнти при невідомих рівні нулю, а вільний член відмінний від нуля. Якщо ж ми такого рівняння не одержимо, то система буде сумісною. У свою чергу, сумісна система рівнянь буде визначена, якщо вона приводиться до трикутної форми вигляду (3.4), і не визначена, якщо приводиться до трапецієподібного вигляду (3.3) при $k < n$.

Застосуємо сказане вище до випадку лінійних однорідних систем, тобто рівнянь, всі вільні члени яких рівні нулю. Така система завжди сумісна, бо має тривіальний (нульовий) розв'язок $(0\ \ 0\ \ ...\ \ 0)$. Нехай в системі, яка розглядається, число рівнянь менше за число невідомих. Тоді наша система не може бути приведеною до трикутної форми, бо в результаті гауссових

перетворень число рівнянь може лише зменшитися; значить вона приводиться до трапецієподібного вигляду, тобто є невизначеною.

Іншими словами, якщо в системі лінійних однорідних рівнянь число рівнянь менше від числа невідомих, то ця система має, окрім нульового розв'язку, також і ненульовий розв'язок, в якому значення деяких (і навіть усіх) невідомих відмінні від нуля; таких розв'язків буде нескінченно багато.

При практичному розв'язуванні СЛАР методом Гаусса слід виписати матрицю з коефіцієнтів системи, приєднати до неї стовпчик вільних членів, для зручності відділений вертикальною рискою, і всі перетворення виконувати над рядками цієї „розширеної" матриці.

Приклад 3.1.

Розв'яжемо систему

$$\begin{cases} x_1 + 2x_2 + 5x_3 = -9 \\ x_1 - x_2 + 3x_3 = 2 \\ 3x_1 - 6x_2 - x_3 = 25 \end{cases}.$$

Здійснімо перетворення Гаусса над розширеною матрицею системи:

$$\left(\begin{array}{ccc|c} 1 & 2 & 5 & -9 \\ 1 & -1 & 3 & 2 \\ 3 & -6 & -1 & 25 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & 2 & 5 & -9 \\ 0 & -3 & -2 & 11 \\ 0 & -12 & -16 & 52 \end{array}\right) \to \left(\begin{array}{ccc|c} 1 & 2 & 5 & -9 \\ 0 & -3 & -2 & 11 \\ 0 & 0 & -8 & 8 \end{array}\right).$$

Значить ми прийшли до СЛАР

$$\begin{cases} x_1 + 2x_2 + 5x_3 = -9 \\ -3x_2 - 2x_3 = 11 \\ -8x_3 = 8 \end{cases},$$

яка має єдиний розв'язок $x_1 = 2,\ x_2 = -3,\ x_3 = -1.$

Приклад 3.2.

Розв'язати систему

$$\begin{cases} x_1 - 5x_2 - 8x_3 + x_4 = 3 \\ 3x_1 + x_2 - 3x_3 - 5x_4 = 1 \\ x_1 - 7x_3 + 2x_4 = -5 \\ 11x_2 + 20x_3 - 9x_4 = 2 \end{cases}.$$

Перетворимо розширену матрицю цієї системи:

$$\left(\begin{array}{cccc|c} 1 & -5 & -8 & 1 & 3 \\ 3 & 1 & -3 & -5 & 1 \\ 1 & 0 & -7 & 2 & -5 \\ 0 & 11 & 20 & -9 & 2 \end{array}\right) \to \left(\begin{array}{cccc|c} 1 & -5 & -8 & 1 & 3 \\ 0 & 16 & 21 & 8 & -8 \\ 0 & 5 & 1 & 1 & -8 \\ 0 & 11 & 20 & -9 & 2 \end{array}\right) \to \left(\begin{array}{cccc|c} 1 & -5 & -8 & 1 & 3 \\ 0 & -89 & 0 & -29 & 160 \\ 0 & 5 & 1 & 1 & -8 \\ 0 & -89 & 0 & -29 & 2 \end{array}\right) \to$$

$$\to \left(\begin{array}{cccc|c} 1 & -5 & -8 & 1 & 3 \\ 0 & -89 & 0 & -29 & 160 \\ 0 & 5 & 1 & 1 & -8 \\ 0 & 0 & 0 & 0 & 2 \end{array}\right).$$

Ми прийшли до системи, яка містить рівняння 0 = 2. Це означає, що вихідна система несумісна.

Приклад 3.3.

Розв'язати систему

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

Ця СЛАР однорідна, число рівнянь менше від числа невідомих, тому вона невизначена. Оскільки всі вільні члени рівні нулю, то ми будемо здійснювати перетворення Гаусса лише над матрицею системи:

$$\left(\begin{array}{cccc} 4 & 1 & -3 & -1 \\ 2 & 3 & 1 & -5 \\ 1 & -2 & -2 & 3 \end{array}\right) \to \left(\begin{array}{cccc} 0 & 9 & 5 & -13 \\ 0 & 7 & 5 & -11 \\ 1 & -2 & -2 & 3 \end{array}\right) \to \left(\begin{array}{cccc} 0 & 2 & 0 & -2 \\ 0 & 7 & 5 & -11 \\ 1 & -2 & -2 & 3 \end{array}\right).$$

Ми дійшли до СЛАР

$$\begin{cases} 2x_2 - x_4 = 0 \\ 7x_2 + 5x_3 - 11x_4 = 0 \\ x_1 - 2x_2 - 2x_3 + 3x_4 = 0 \end{cases}.$$

В якості вільної змінної можна прийняти будь-яке з невідомих x_2 або x_4. Нехай $x_4 = \gamma$, γ – довільне число, $x_2 = \frac{1}{2}\gamma$. Тоді з другого рівняння одержимо $x_3 = \frac{3}{2}\gamma$, а з третього рівняння $x_1 = \gamma$. Таким чином, загальний вигляд розв'язку має вигляд:

$$x_1 = \gamma,\ x_2 = \frac{1}{2}\gamma,\ x_3 = \frac{3}{2}\gamma,\ x_4 = \gamma.$$

На кінець зауважимо, що якщо перетворення Гаусса здійснювати, починаючи з найбільшого за абсолютною величиною

коефіцієнта (це можна здійснити за допомогою перестановки рівнянь і перенумерації невідомих), то сумарна похибка округлення буде мінімальною серед усіх інших гауссових процедур виключення. Кількість арифметичних операцій є $K\left(\frac{n^3}{3}\right)$, де K коефіцієнт, який від n не залежить.

§4. ІТЕРАЦІЙНІ МЕТОДИ РОЗВ'ЯЗУВАННЯ СЛАР

Метод Гаусса є скінченним методом розв'язування СЛАР: n-1 кроків прямого ходу Гаусса і n кроків оберненого ходу Гаусса. Розв'язок отримуємо з точністю до похибок округлень, якщо таких похибок немає (обчислення ведуться точно), то й роз'вязок отримуємо точним. Для СЛАР високої розмірності (числа n та m великі) обчислювальна трудоємність методу Гаусса може бути великою. При економічних дослідженнях виникають СЛАР з розмірностями рівними мільйонам і навіть десяткам мільйонів рівнянь та невідомих. При цьому випадку трудоємність методу Гаусса може бути настільки величезна, що навіть на найсучасніших обчислювальних комплексах за потрібний часовий проміжок розв'язок одержати не вдається. У такому разі зручно використати ітераційні методи розв'язування СЛАР. Крім того, точний розв'язок не завжди потрібний. Нерідко замовника розв'язку СЛАР задовольняє розв'язок, одержаний з точністю до ε (наперед задане число, для якого виконується, наприклад, умова $\max\limits_{i=\overline{1,n}} |x_i - x_i^*| \le \varepsilon$, де x_i – точна компонента розв'язку системи, а x_i^* – наближена компонента до точного розв'язку).

§5. МЕТОД ПРОСТОЇ ІТЕРАЦІЇ

Подамо в (1.1) $m = n$ і запишемо СЛАР у вигляді

$$x = b + \overline{A} \cdot x$$

або покомпонентно

$$\begin{aligned} x_1 &= b_1 - (a_{11} - 1)x_1 - a_{12}x_2 - a_{13}x_3 - \ldots - a_{1n}x_n \\ x_2 &= b_2 \quad - a_{21}x_1 \ - (a_{22} - 1)x_2 - a_{23}x_3 - \ldots - a_{2n}x_n \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n &= b_n \quad - a_{n1}x_1 \ - a_{n2}x_2 - a_{n3}x_3 - \ldots - (a_{nn} - 1)x_n. \end{aligned} \qquad (5.1)$$

Далі подамо $x_1 = x_1^0$, $x_2 = x_2^0, \ldots, x_n = x_n^0$, де x_i^0 – довільні наперед задані числа. Обчислимо праві частини (5.1) і одержимо

$$\begin{aligned} x_1' &= b_1 - (a_{11} - 1)x_1^0 - a_{12}x_2^0 - a_{13}x_3^0 - \ldots - a_{1n}x_n^0 \\ x_2' &= b_2 \quad - a_{21}x_1^0 \ - (a_{22} - 1)x_2^0 - a_{23}x_3^0 - \ldots - a_{2n}x_n^0 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n' &= b_n \quad - a_{n1}x_1^0 \ - a_{n2}x_2^0 - a_{n3}x_3^0 - \ldots - (a_{nn} - 1)x_n^0. \end{aligned}$$

Далі одержуємо числа x_1'', x_2'', ..., x_n'', обчисливши праві частини (5.1) при одержаних x_1', x_2', ..., x_n' і т. д. Процес простого ітерування в загальному випадку відображається у векторному вигляді

$$x^k = b + \overline{A} \cdot x^{k-1},$$

а в деталізованому вигляді як:

$$\begin{aligned} x_1^k &= b_1 - (a_{11} - 1)x_1^{k-1} - a_{12}x_2^{k-1} - a_{13}x_3^{k-1} - \ldots - a_{1n}x_n^{k-1} \\ x_2^k &= b_2 \quad - a_{21}x_1^{k-1} \ - (a_{22} - 1)x_2^{k-1} - a_{23}x_3^{k-1} - \ldots - a_{2n}x_n^{k-1} \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^k &= b_n \quad - a_{n1}x_1^{k-1} \ - a_{n2}x_2^{k-1} - a_{n3}x_3^{k-1} - \ldots - (a_{nn} - 1)x_n^{k-1}. \end{aligned}$$

Після кожного кроку простої ітерації (індекс вгорі означає номер кроку) обчислюємо різниці $\left|x_1^k - x_1^{k-1}\right|, \left|x_2^k - x_2^{k-1}\right|, \ldots, \left|x_n^k - x_n^{k-1}\right|$. Якщо найбільше з цих чисел $\le \varepsilon$, тоді $x_i^k \ \ \forall\, i = \overline{1,n}$ призначаємо розв'язком системи (5.1), а значить і (1.1).

Описана процедура одержання наближеного розв'язку можлива не завжди. Говорять, що метод простої ітерації збігається з розв'язком системи (1.1), якщо $\lim\limits_{k\to\infty} x_i^k \to x_i, \ \forall\, i = \overline{1,n}$. Зауважимо, що ця умова виконується, якщо найбільше за модулем власне число матриці $\overline{A}$ $\max\limits_{i=\overline{1,n}} \left|\lambda_{i\,\overline{A}}\right| < 1$. Оскільки знаходження власних чисел – процедура досить трудоємна, то використовуємо нерівність

$$\lambda_{\bar{A}} \le \max_{i=\overline{1,n}} \sum_{j=1}^{n} \left|\bar{a}_{ij}\right| < 1\,, \qquad (5.2)$$

де $\bar{a}_{ij}$ – елементи матриці $\bar{A}$, а (5.2) – достатня умова збіжності методу простої ітерації.

Умова (5.2) автоматично виконується для систем лінійних балансових рівнянь В.В. Леонтьєва, якщо балансові рівняння записані у вартісному вигляді.

Приклад 5.1.

Нехай маємо систему рівнянь

$$\begin{aligned} x_1 &= 0{,}2x_2 + 0{,}3x_3 + 0{,}1x_4 + 4 \\ x_2 &= 0{,}1x_1 + 0{,}2x_3 + 0{,}2x_4 + 5 \\ x_3 &= 0{,}3x_2 + 0{,}4x_4 + 3 \\ x_4 &= 0{,}3x_1 + 0{,}2x_2 + 0{,}3x_3 + 2 \end{aligned},$$

яку розв'язуємо з точністю до $\varepsilon = 0{,}1$.

Легко бачити, що

$$S_1 = \sum_{j=1}^{4} a_{1j} = 0{,}2 + 0{,}3 + 0{,}1 = 0{,}6 < 1,$$

$$S_2 = \sum_{j=1}^{4} a_{2j} = 0{,}1 + 0{,}2 + 0{,}2 = 0{,}5 < 1,$$

$$S_3 = \sum_{j=1}^{4} a_{3j} = 0{,}3 + 0{,}4 = 0{,}7 < 1,$$

$$S_4 = \sum_{j=1}^{4} a_{4j} = 0{,}3 + 0{,}2 + 0{,}3 = 0{,}8 < 1.$$

Тобто метод простої ітерації можна застосовувати.

Візьмемо $x_1^0 = 4,\ x_2^0 = 5,\ x_3^0 = 3,\ x_4^0 = 2.$

$$\begin{aligned} x_1^1 &= 0{,}2 \cdot 5 + 0{,}3 \cdot 3 + 0{,}1 \cdot 2 + 4 = 6{,}1 \\ x_2^1 &= 0{,}1 \cdot 4 + 0{,}2 \cdot 3 + 0{,}2 \cdot 2 + 5 = 6{,}3 \\ x_3^1 &= 0{,}3 \cdot 5 + 0{,}4 \cdot 2 + 3 = 5{,}3 \\ x_4^1 &= 0{,}3 \cdot 4 + 0{,}2 \cdot 5 + 0{,}3 \cdot 3 + 2 = 5{,}1 \end{aligned}.$$

Розраховуємо різницю $x_1^1 - x_1^0 = 6{,}1 - 4 = 2{,}1 > \varepsilon = 0{,}1$. Тобто продовжуємо ітерування.

$$x_1^2 = 0{,}2 \cdot 6{,}3 + 0{,}3 \cdot 5{,}3 + 0{,}1 \cdot 5{,}1 + 4 = 7{,}36$$
$$x_2^2 = 0{,}1 \cdot 6{,}1 + 0{,}2 \cdot 5{,}3 + 0{,}2 \cdot 5{,}1 + 5 = 7{,}69$$
$$x_3^2 = 0{,}3 \cdot 6{,}3 + 0{,}4 \cdot 5{,}1 + 3 = 6{,}93 \quad ,$$
$$x_4^2 = 0{,}3 \cdot 6{,}1 + 0{,}2 \cdot 6{,}3 + 0{,}3 \cdot 5{,}3 + 2 = 6{,}71$$

Розраховуємо різницю $x_1^2 - x_1^1 = 7{,}36 - 6{,}1 = 1{,}26 > \varepsilon = 0{,}1$. Тобто продовжуємо ітерування.

$$x_1^3 = 0{,}2 \cdot 7{,}69 + 0{,}3 \cdot 6{,}93 + 0{,}1 \cdot 6{,}71 + 4 = 8{,}288$$
$$x_2^3 = 0{,}1 \cdot 7{,}36 + 0{,}2 \cdot 6{,}93 + 0{,}2 \cdot 6{,}71 + 5 = 8{,}464$$
$$x_3^3 = 0{,}3 \cdot 7{,}69 + 0{,}4 \cdot 6{,}71 + 3 = 7{,}991 \quad ,$$
$$x_4^3 = 0{,}3 \cdot 7{,}36 + 0{,}2 \cdot 7{,}69 + 0{,}3 \cdot 6{,}93 + 2 = 7{,}825$$

розраховуємо різницю $x_1^3 - x_1^2 = 8{,}288 - 7{,}36 = 0{,}928 > \varepsilon = 0{,}1$. Тобто продовжуємо ітерування.

$$x_1^4 = 0{,}2 \cdot 8{,}464 + 0{,}3 \cdot 7{,}991 + 0{,}1 \cdot 7{,}825 + 4 = 8{,}8726$$
$$x_2^4 = 0{,}1 \cdot 8{,}288 + 0{,}2 \cdot 7{,}991 + 0{,}2 \cdot 7{,}825 + 5 = 9{,}792$$
$$x_3^4 = 0{,}3 \cdot 8{,}464 + 0{,}4 \cdot 7{,}825 + 3 = 8{,}6692 \quad ,$$
$$x_4^4 = 0{,}3 \cdot 8{,}288 + 0{,}2 \cdot 8{,}464 + 0{,}3 \cdot 7{,}991 + 2 = 8{,}5765$$

розраховуємо різницю $x_1^4 - x_1^3 = 8{,}8726 - 8{,}288 = 0{,}5846 > \varepsilon = 0{,}1$. Тобто продовжуємо ітерування.

$$x_1^5 = 0{,}2 \cdot 9{,}792 + 0{,}3 \cdot 8{,}6692 + 0{,}1 \cdot 8{,}5765 + 4 = 9{,}41681$$
$$x_2^5 = 0{,}1 \cdot 8{,}8726 + 0{,}2 \cdot 8{,}6692 + 0{,}2 \cdot 8{,}5765 + 5 = 9{,}2364$$
$$x_3^5 = 0{,}3 \cdot 9{,}792 + 0{,}4 \cdot 8{,}5765 + 3 = 9{,}3682 \quad ,$$
$$x_4^4 = 0{,}3 \cdot 8{,}8726 + 0{,}2 \cdot 9{,}792 + 0{,}3 \cdot 8{,}6692 + 2 = 9{,}22094$$

розраховуємо різницю $x_1^5 - x_1^4 = 9{,}41681 - 8{,}8726 = 0{,}544221 > \varepsilon = 0{,}1$. Тобто продовжуємо ітерування.

$$x_1^6 = 0{,}2 \cdot 9{,}2364 + 0{,}3 \cdot 9{,}3682 + 0{,}1 \cdot 9{,}22094 + 4 = 9{,}579834$$
$$x_2^6 = 0{,}1 \cdot 9{,}41681 + 0{,}2 \cdot 9{,}3682 + 0{,}2 \cdot 9{,}22094 + 5 = 9{,}66408$$
$$x_3^6 = 0{,}3 \cdot 9{,}2364 + 0{,}4 \cdot 9{,}22094 + 3 = 9{,}459296 \quad ,$$
$$x_4^6 = 0{,}3 \cdot 9{,}41681 + 0{,}2 \cdot 9{,}2364 + 0{,}3 \cdot 9{,}3682 + 2 = 9{,}742866$$

розраховуємо різницю $x_1^6 - x_1^5 = 9{,}579834 - 9{,}41681 = 0{,}163024 > \varepsilon = 0{,}1$. Тобто продовжуємо ітерування.

$$
\begin{aligned}
x_1^7 = {} & 0,2\cdot 9,66408 + 0,3\cdot 9,4592296 + 0,1\cdot 9,742866 + 4 = 9,75481354\\
x_2^7 = {} & 0,1\cdot 9,579834 + 0,2\cdot 9,4592296 + 0,2\cdot 9,742866 + 5 = 9,8251486\\
x_3^7 = {} & 0,3\cdot 9,66408 + 0,4\cdot 9,742866 + 3 = 9,7963704\\
x_4^7 = {} & 0,3\cdot 9,579834 + 0,2\cdot 9,66408 + 0,3\cdot 9,4592296 + 2 = 9,729626
\end{aligned}
,
$$

розраховуємо різницю $x_1^7 - x_1^6 = 9,75481354 - 9,579834 = 0,17497954 > \varepsilon = 0,1$. Тобто продовжуємо ітерування.

$$
\begin{aligned}
x_1^8 = {} & 0,2\cdot 9,8251486 + 0,3\cdot 9,7963704 + 0,1\cdot 9,729626 + 4 = 9,87691092\\
x_2^8 = {} & 0,1\cdot 9,75481354 + 0,2\cdot 9,7963704 + 0,2\cdot 9,729626 + 5 = 9,880677954\\
x_3^8 = {} & 0,3\cdot 9,8251486 + 0,4\cdot 9,729626 + 3 = 9,83939498\\
x_4^8 = {} & 0,3\cdot 9,75481354 + 0,2\cdot 9,8251486 + 0,3\cdot 9,7963704 + 2 = 9,8778441
\end{aligned}
,
$$

розраховуємо різницю $x_1^8 - x_1^7 = 9,87691092 - 9,75481354 = 0,12209738 > \varepsilon = 0,1$. Тобто продовжуємо ітерування.

$$
\begin{aligned}
x_1^9 = {} & 0,2\cdot 9,880677954 + 0,3\cdot 9,83939498 + 0,1\cdot 9,8778441 + 4 = 9,9157384948\\
x_2^9 = {} & 0,1\cdot 9,87691092 + 0,2\cdot 9,83939498 + 0,2\cdot 9,8778441 + 5 = 9,9311058736\\
x_3^9 = {} & 0,3\cdot 9,880677954 + 0,4\cdot 9,8778441 + 3 = 9,9153171502\\
x_4^9 = {} & 0,3\cdot 9,87691092 + 0,2\cdot 9,880677954 + 0,3\cdot 9,83939498 + 2 = 9,8910273608
\end{aligned}
.
$$

Перевіряємо виконання наступних умов:

$$
\begin{aligned}
x_1^9 - x_1^8 &= 9,9157384948 - 9,87691092 = 0,03882757 < \varepsilon = 0,1,\\
x_2^9 - x_2^8 &= 9,931058736 - 9,880677954 = 0,05380782 < \varepsilon = 0,1,\\
x_3^9 - x_3^8 &= 9,9153171502 - 9,83939498 = 0,0759321702 < \varepsilon = 0,1,\\
x_4^9 - x_4^8 &= 9,8910273608 - 9,8778441 = 0,0131832608 < \varepsilon = 0,1.
\end{aligned}
$$

Звідси випливає, що наближений розв'язок СЛАР має вигляд:

$x_1 = 9,9157384948,\ x_2 = 9,931058736,\ x_3 = 9,9153171502,\ x_4 = 9,8910273608.$

Слід зауважити, що похибки при обчисленні методом простої ітерації суттєво не впливають на розв'язок, тому що за умов (5.2) метод простої ітерації збігається при будь-яких початкових наближеннях x^0 (x^k).

§6. МЕТОД ГАУССА-ЗЕЙДЕЛЯ

Оскільки в методі простої ітерації при обчисленнях компоненти вектора $x^k = (x_1^k \quad x_2^k \quad ... \quad x_n^k)$ обчислюються послідовно (зразу x_1^k, потім x_2^k і т. д.), то одержані значення компонент векторів можна використати при обчисленні наступних компонент. Проілюструємо процедуру оперативного використання обчислених компонент, при обчисленні наступних, на прикладі систем В. В. Леонтьєва.

Розглянемо систему, векторний вигляд якої:

$$x = A \cdot x + b, \tag{6.1}$$

де $A = \{a_{ij}\}_{i,j=1}^{n}$ – матриця прямих затрат, $a_{ij} \geq 0 \ \forall i, j = \overline{1,n}$. В деталізованому вигляді маємо

$$\begin{aligned}
x_1 &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + ... + a_{1n}x_n + b_1 \\
x_2 &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + ... + a_{2n}x_n + b_2 \\
x_3 &= a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + ... + a_{3n}x_n + b_3 \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
x_n &= a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + ... + a_{nn}x_n + b_n.
\end{aligned} \tag{6.2}$$

Будемо вважати, що $x_1 = x_1^0,\ x_2 = x_2^0, ...,\ x_n = x_n^0,$ де $x_i^0 \ \ \forall i = \overline{1,n}$ наперед задані числа. Знайдемо

$$x_1' = a_{11}x_1^0 + a_{12}x_2^0 + a_{13}x_3^0 + ... + a_{1n}x_n^0 + b_1,$$

далі обчислимо x_2' використавши x_1'

$$x_2' = a_{21}x_1' + a_{22}x_2^0 + a_{23}x_3^0 + ... + a_{2n}x_n^0 + b_2,$$

потім

$$\begin{aligned}
x_3' &= a_{31}x_1' + a_{32}x_2' + a_{33}x_3^0 + ... + a_{3n}x_n^0 + b_3, \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
x_n' &= a_{n1}x_1' + a_{n2}x_2' + a_{n3}x_3' + ... + a_{nn}x_n^0 + b_n.
\end{aligned}$$

У загальному вигляді ітераційний процес можна записати:

$$\begin{aligned}
x_1^k &= a_{11}x_1^{k-1} + a_{12}x_2^{k-1} + a_{13}x_3^{k-1} + ... + a_{1n}x_n^{k-1} + b_1 \\
x_2^k &= a_{21}x_1^k + a_{22}x_2^{k-1} + a_{23}x_3^{k-1} + ... + a_{2n}x_n^{k-1} + b_2 \\
x_3^k &= a_{31}x_1^k + a_{32}x_2^k + a_{33}x_3^{k-1} + ... + a_{3n}x_n^{k-1} + b_3 \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
x_n^k &= a_{n1}x_1^k + a_{n2}x_2^k + a_{n3}x_3^k + ... + a_{nn}x_n^{k-1} + b_n.
\end{aligned} \tag{6.3}$$

У векторному вигляді процес виглядає так:

$$x^k = B \cdot x^k + C \cdot x^{k-1} + b, \tag{6.4}$$

де

$$B = \begin{pmatrix} 0 & 0 & 0 & ... & 0 & 0 \\ a_{21} & 0 & 0 & ... & 0 & 0 \\ a_{31} & a_{32} & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... \\ a_{n1} & a_{2n} & a_{3n} & ... & a_{nn-1} & 0 \end{pmatrix},$$

а

$$C = \begin{pmatrix} a_{11} & a_{12} & a_{13} & ... & a_{1n-1} & a_{1n} \\ 0 & a_{22} & a_{23} & ... & a_{2n-1} & a_{2n} \\ 0 & 0 & a33 & ... & a_{3n-1} & a_{3n} \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 0 & a_{nn} \end{pmatrix}.$$

Запишемо (6.4) у вигляді

$$(E - B) \cdot x^k = C \cdot x^{k-1} + b. \quad (6.5)$$

Оскільки матриця $E - B$ має обернену матрицю, тому що її визначник рівний 1 (діагональні елементи трикутної матриці $E - B$ дорівнюють одиниці). Запишемо (6.5) у вигляді

$$x^k = (E - B)^{-1} \cdot C \cdot x^{k-1} + (E - B)^{-1} \cdot b. \quad (6.6)$$

Формула (6.6) відображає просту ітерацію з матрицею $(E - B)^{-1} \cdot C$. В [9] доведено, що сума елементів рядків цієї матриці має вигляд

$$S_1 = \sum_{j=1}^{n} a_{1j},$$

$$S_2 = a_{21} \cdot S_1 + \sum_{j=2}^{n} a_{2j} = \sum_{j=1}^{1} a_{2j} \cdot S_j + \sum_{j=2}^{n} a_{2j},$$

$$S_3 = a_{31} \cdot S_1 + a_{32} \cdot S_2 + \sum_{j=3}^{n} a_{3j} = \sum_{j=1}^{2} a_{3j} \cdot S_j + \sum_{j=3}^{n} a_{3j},$$

.....

$$S_n = a_{n1} \cdot S_1 + a_{n2} \cdot S_2 + ... + a_{nn-1} \cdot S_{n-1} + a_{nn} = \sum_{j=1}^{n-1} a_{nj} \cdot S_j + a_{nn}.$$

Достатніми умовами збіжності процесу (6.3), (6.4) в такому разі є $S_i < 1, \ \forall i = \overline{1,n}$. В роботі [10] показано, що

$$\left\| (E - B)^{-1} \cdot C \right\|_1 \le \|A\|_1.$$

Крім того, в цій роботі показано, що перенумерація рівнянь та невідомих в (6.2) може прискорити ітераційний процес Гаусса-

Зейделя, якщо рівняння розташовувати в такому порядку: на першому місці поставимо рівняння (а, значить, і перенумеруємо відповідно невідомі), рядок, коефіцієнти якого задовольняють умову

$$\min_{i=\overline{1,n}} \sum_{j=1}^{n} a_{ij} = S_{l_1} = S_1;$$

на другому місці поставимо рівняння (відповідно перенумерувавши змінні), для якого досягається

$$\min_{i \neq l_1} \left(a_{i1} S_1 + \sum_{j=2}^{n} a_{ij} \right) = S_{l_2} = S_2;$$

на третьому –

$$\min_{i \neq l_1, l_2} \left(a_{i1} S_1 + a_{i2} S_2 + \sum_{j=3}^{n} a_{ij} \right) = S_{l_3} = S_3;$$

і т. д.

$$S_i = \max_{i \neq l_1, l_2, \ldots, l_{i-1}} \left(\sum_{j=1}^{i-1} a_{ij} S_j + \sum_{j=i}^{n} a_{ij} \right). \tag{6.7}$$

Розв'яжемо задачу з §5.

Приклад 6.1.

$$\begin{aligned} x_1 &= 0{,}2x_2 + 0{,}3x_3 + 0{,}1x_4 + 4 \\ x_2 &= 0{,}1x_1 + 0{,}2x_3 + 0{,}2x_4 + 5 \\ x_3 &= 0{,}3x_2 + 0{,}4x_4 + 3 \\ x_4 &= 0{,}3x_1 + 0{,}2x_2 + 0{,}3x_3 + 2 \end{aligned} \tag{6.8}$$

Розмістємо рівняння згідно з (6.7). Для цього треба переставити місцями перше та друге рівняння, змінивши при цьому нумерацію невідомих $x_2 \Rightarrow y_1,\ x_1 \Rightarrow y_2$, одержимо:

$$\begin{aligned} y_1 &= 0{,}1y_2 + 0{,}2x_3 + 0{,}2x_4 + 5 \\ y_2 &= 0{,}2y_1 + 0{,}3x_3 + 0{,}1x_4 + 4 \\ x_3 &= 0{,}3y_1 + 0{,}4x_4 + 3 \\ x_4 &= 0{,}2y_1 + 0{,}3y_2 + 0{,}3x_3 + 2 \end{aligned} \tag{6.9}$$

Крок 1.

Оскільки $y_1^0 = x_2^0,\ y_2^0 = x_1^0,\ y_1^0 = 5,\ y_2^0 = 4,\ x_3^0 = 3,\ x_4^0 = 2$.

$$y_1^1 = \quad 0{,}1 \cdot 4 \quad + 0{,}2 \cdot 3 + 0{,}2 \cdot 2 + 5 = 6{,}4$$
$$y_2^1 = 0{,}2 \cdot 6{,}4 \quad + 0{,}3 \cdot 3 + 0{,}1 \cdot 2 + 4 = 6{,}38$$
$$x_3^1 = 0{,}3 \cdot 6{,}4 \quad + 0{,}4 \cdot 2 + 3 = 5{,}72 \,,$$
$$x_4^1 = 0{,}2 \cdot 6{,}4 + 0{,}3 \cdot 6{,}38 + 0{,}3 \cdot 5{,}72 \quad + 2 = 6{,}91$$

$y_1^1 - y_1^0 = 6{,}4 - 5 = 1{,}4 > \varepsilon = 0{,}1$.

Крок 2.

$$y_1^2 = \quad 0{,}1 \cdot 6{,}38 \quad + 0{,}2 \cdot 5{,}72 + 0{,}2 \cdot 6{,}91 + 5 = 8{,}164$$
$$y_2^2 = 0{,}2 \cdot 8{,}164 \quad + 0{,}3 \cdot 5{,}72 + 0{,}1 \cdot 6{,}91 + 4 = 8{,}0398$$
$$x_3^2 = 0{,}3 \cdot 8{,}164 \quad + 0{,}4 \cdot 6{,}91 + 3 = 8{,}2132 \,,$$
$$x_4^2 = 0{,}2 \cdot 8{,}164 + 0{,}3 \cdot 8{,}0398 + 0{,}3 \cdot 8{,}2132 \quad + 2 = 8{,}521$$

$y_1^2 - y_1^1 = 8{,}164 - 6{,}4 = 1{,}764 > \varepsilon = 0{,}1$.

Крок 3.

$$y_1^3 = \quad 0{,}1 \cdot 8{,}0398 \quad + 0{,}2 \cdot 8{,}2132 \quad + 0{,}2 \cdot 8{,}521 + 5 = 9{,}15082$$
$$y_2^3 = 0{,}2 \cdot 9{,}15082 \quad + 0{,}3 \cdot 8{,}2132 \quad + 0{,}1 \cdot 8{,}521 + 4 = 9{,}146224$$
$$x_3^3 = 0{,}3 \cdot 9{,}15082 \quad + 0{,}4 \cdot 8{,}521 + 3 = 9{,}153646 \,,$$
$$x_4^2 = 0{,}2 \cdot 9{,}15082 + 0{,}3 \cdot 9{,}146224 + 0{,}3 \cdot 9{,}153646 \quad + 2 = 9{,}3205846$$

$y_1^3 - y_1^2 = 9{,}15082 - 8{,}164 = 0{,}98682 > \varepsilon = 0{,}1$.

Крок 4.

$$y_1^4 = \quad 0{,}1 \cdot 9{,}146224 + 0{,}2 \cdot 9{,}153646 + 0{,}2 \cdot 9{,}3205846 + 5 \approx 9{,}61$$
$$y_2^4 = 0{,}2 \cdot 9{,}61 \quad + 0{,}3 \cdot 9{,}153646 + 0{,}1 \cdot 9{,}3205846 + 4 \approx 9{,}6$$
$$x_3^4 = 0{,}3 \cdot 9{,}61 \quad + 0{,}4 \cdot 9{,}3205846 + 3 \approx 9{,}611 \,,$$
$$x_4^4 = 0{,}2 \cdot 9{,}61 + 0{,}3 \cdot 9{,}6 \quad + 0{,}3 \cdot 9{,}61 \quad + 2 \approx 9{,}986$$

$y_1^4 - y_1^3 = 9{,}61 - 9{,}15082 = 0{,}46 > \varepsilon = 0{,}1$.

Крок 5.

$$y_1^5 = \quad 0{,}1 \cdot 9{,}6 \quad + 0{,}2 \cdot 9{,}611 + 0{,}2 \cdot 9{,}686 + 5 \approx 9{,}82$$
$$y_2^5 = 0{,}2 \cdot 9{,}82 \quad + 0{,}3 \cdot 9{,}611 + 0{,}1 \cdot 9{,}686 + 4 \approx 9{,}82$$
$$x_3^5 = 0{,}3 \cdot 9{,}82 \quad + 0{,}4 \cdot 9{,}686 + 3 \approx 9{,}82 \,,$$
$$x_4^5 = 0{,}2 \cdot 9{,}82 + 0{,}3 \cdot 9{,}82 + 0{,}3 \cdot 9{,}82 \quad + 2 \approx 9{,}856$$

$y_1^5 - y_1^4 = 9{,}82 - 9{,}61 = 0{,}21 > \varepsilon = 0{,}1$.

Крок 6.

$$\begin{aligned} y_1^6 &= 0{,}1\cdot 9{,}82 + 0{,}2\cdot 9{,}82 + 0{,}2\cdot 9{,}856 + 5 \approx 9{,}9172 \\ y_2^5 &= 0{,}2\cdot 9{,}9172 + 0{,}3\cdot 9{,}82 + 0{,}1\cdot 9{,}856 + 4 \approx 9{,}91504 \\ x_3^5 &= 0{,}3\cdot 9{,}9172 + 0{,}4\cdot 9{,}856 + 3 \approx 9{,}91756 \\ x_4^5 &= 0{,}2\cdot 9{,}9172 + 0{,}3\cdot 9{,}91504 + 0{,}3\cdot 9{,}91756 + 2 \approx 9{,}93522 \end{aligned},$$

Перевіримо умови виходу з процедури:

$$\begin{aligned} y_1^6 - y_1^5 &= 9{,}9172 - 9{,}82 = 0{,}0972 < \varepsilon = 0{,}1, \\ y_2^6 - y_2^5 &= 9{,}91504 - 9{,}82 = 0{,}09504 < \varepsilon = 0{,}1, \\ x_3^6 - x_3^5 &= 9{,}91756 - 9{,}82 = 0{,}09756 < \varepsilon = 0{,}1, \\ x_4^6 - x_4^5 &= 9{,}93522 - 9{,}856 = 0{,}07922 < \varepsilon = 0{,}1. \end{aligned}.$$

Тобто розв'язок системи (6.9), а значить і (6.8):

$$x_1 = 9{,}91504,\ x_2 = 9{,}9172,\ x_3 = 9{,}91756,\ x_4 = 9{,}93522.$$

Використання наближених значень (шляхом заокруглень) не заважає одержанню розв'язку, бо ітераційні методи (в умовах збіжності процесу) стійкі відносно змін початкових чи проміжних значень ітерацій.

Порівнюючи процеси простої ітерації та методу Гаусса-Зейделя при розв'язуванні одного і того ж прикладу (в §5 та §6), переконуємося, що для розв'язку СЛАР методом простої ітерації необхідно 9 ітерацій, а методом Гаусса-Зейделя лише 6 (в 1,5 раза швидше).

Розв'язки в цілих числах систем лінійних балансових рівнянь можна одержати за допомогою методу послідовного аналізу варіантів викладеного в [11].

При розв'язуванні СЛАР нерідко зустрічаються випадки, коли малі похибки, при заданні коефіцієнтів або правих частин, приводять до великих похибок в розв'язках. Похибки можуть виникати при вимірюванні, обчисленні чи заокругленні елементів матриць систем або правих частин. Такі СЛАР будемо називати некоректно поставленими або погано обумовленими. Ця термінологія виникла з часів Ж. Адамара, який вважав вивчення некоректно поставлених задач недоцільним, тому що їх завжди можна, уточненням математичної моделі, поставити коректно (розумно). Але Ж. Адамар не вказав способу як „коректизувати" задачу, а в той же час ці задачі зустрічаються частіше ніж коректно поставлені задачі. Ця обставина випливає із труднощів при побудові адекватних діючим процесам математичних моделей.

В якості погано обумовленої задачі приведемо СЛАР з матрицею Гілберта

$$A = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \dots & \frac{1}{n+1} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \dots & \frac{1}{n+2} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \dots & \frac{1}{n+3} \\ \dots & \dots & \dots & \dots & \dots \\ \frac{1}{n+1} & \frac{1}{n+2} & \frac{1}{n+3} & \dots & \frac{1}{2(n+1)} \end{pmatrix}.$$

Довгий час СЛАР з цією матрицею, при $n > 7$, на сучасних обчислювальних комплексах одержувались такі, що не мали в компонентах розв'язку жодної правильної цифри.

У даний час така ж ситуація спостерігається для $n \geq 35$. Матриці Гілберта є одними з модельних задач для характеристики якості алгоритмів розв'язку СЛАР.

Тому виникає проблема розв'язку погано обумовлених задач (в тому числі СЛАР).

Зробимо деякі пояснення.

§7. МІРА КОРЕКТНОСТІ (ОБУМОВЛЕНОСТІ) МАТРИЦІ

Введемо важливу характеристику $\mu(A)$ матриці СЛАР – число обумовленості

$$\mu(A)=\max_{x\neq 0,\ \xi\neq 0}\left[\frac{\|Ax\|}{\|A\xi\|}\cdot\frac{\|\xi\|}{\|x\|}\right], \tag{7.1}$$

яку можна записати у вигляді

$$\mu(A)=\left[\max_{x\neq 0}\frac{\|Ax\|}{\|x\|}\Big/\min_{\xi\neq 0}\frac{\|A\xi\|}{\|\xi\|}\right]=\sqrt{\frac{\sigma_{\max}}{\sigma_{\min}}}, \tag{7.2}$$

якщо використовується векторна норма – евклідова; тут $\sigma_{\max},\sigma_{\min}$ – максимальне та мінімальне власне значення матриці A^*A, де A^* транспонована матриця до матриці A.

Основні властивості числа обумовленості очевидні і виводяться безпосередньо із визначення:

1) $\mu(A)\geq 1,\ \mu(E)=1,$ де E – одинична матриця;
2) $\mu(c\cdot A)=\mu(A),\ c-const$;
3) $\mu(A)=\dfrac{\max|d_{ii}|}{\min|d_{ii}|}$, якщо A –діагональна матриця;
4) $\mu(A)=\mu\left(A^{-1}\right)$.

З'ясуємо роль величини $\mu(A)$ в оцінці похибки при збуренні вихідних даних системи (1.1): вектора b і матриці A.

Розглянемо поряд з b, вектор $b+\Delta b$. Нехай x, $x+\Delta x$ – відповідно розвязок рівнянь $A\cdot x=b$, $A(x+\Delta x)=b+\Delta b$. Тоді $A\cdot\Delta x=\Delta b$. Згідно з визначенням $\mu(A)$ із (7.1)

$$\mu(A)=\max_{\xi,\Delta\xi}\left[\frac{\|A\cdot\xi\|}{\|A\cdot\Delta\xi\|}\cdot\frac{\|\Delta\xi\|}{\|\xi\|}\right]\geq\left[\frac{\|A\cdot x\|}{\|A\cdot\Delta x\|}\cdot\frac{\|\Delta x\|}{\|x\|}\right],$$

звідки

$$\frac{\|\Delta x\|}{\|x\|}\leq\mu(A)\cdot\frac{\|\Delta b\|}{\|b\|}, \tag{7.3}$$

тобто $\mu(A)$ – найменша константа, для якої при всіх $\Delta x, x, \Delta b, b$ виконується нерівність (7.3).

Таким чином, $\mu(A)$ виконує роль множника при зростанні відносної похибки розв'язку. Це значить, що зміна вектора в правій частині системи має як наслідок зміни в розв'язку, більші в $\mu(A)$ разів. Іншими словами, нерівність (7.3) означає, що $\mu(A)$ обмежує зверху відношення відносної невизначеності вектора x до відносної

невизначеності вектора b. Важливо підкреслити, що оцінка (7.3) точна. Це свідчить про те, що при відповідному підборі векторів b, Δb можна досягти рівності. Оцінка (7.3) досяжна. Значить, не можна дати більш точної оцінки, ніж (7.3) для довільних векторів b, Δb незалежно від їх величини.

Аналогічний зміст числа обумовленості в загальній ситуації, коли збурюється і вектор b і матриця A: $(A+\Delta A)\cdot(x+\Delta x)=b+\Delta b$. Не зупиняючись на висновку, приведемо заключну оцінку [25]:

$$\frac{\|\Delta x\|}{\|x\|} \le \frac{\frac{\|\Delta A\|}{\|A\|}+\frac{\|\Delta b\|}{\|b\|}}{1-\frac{\|\Delta A\|}{\|A\|}\cdot\mu(A)}\cdot\mu(A), \qquad (7.4)$$

котра одержана в допущенні невиродженості матриці A – та умови

$$\frac{\|\Delta A\|}{\|A\|}\cdot\mu(A)<1.$$

Доречно зауважити, що якщо A – вироджена матриця, то отримаємо крайній випадок $\mu(A)=\infty$ (оскільки $min\ \frac{\|A\cdot\xi\|}{\|\xi\|}=0$ і оцінки (7.3), (7.4) втрачають сенс), а якщо A – невироджена матриця, то

$$\mu(A)=\|A\|\cdot\left\|A^{-1}\right\|, \qquad (7.5)$$

де A^{-1} – обернена до A матриця, тобто матриця, яка задовольняє відношенню: $A\cdot A^{-1}=A^{-1}\cdot A=E$.

Отже, із оцінок (7.3), (7.4) випливає, що число обумовленості матриці А є характеристикою того, наскільки розв'язок системи $A\cdot x=b$ стійкий до збурень компонент вектора правих частин b та матриці коефіцієнтів A. При великих значеннях $\mu(A)$ розв'язок $\tilde{x}$ системи із збуреними даними

$$A\cdot\tilde{x}=\tilde{b} \qquad (7.6)$$

(навіть якщо всі обчислення при знаходженні $\tilde{x}$ проводяться абсолютно точно) може значно відрізнятися від розв'язку x системи (1.1), не зважаючи на те, що $\tilde{A}$ мало відрізняється від A, а $\tilde{b}$ – від b.

Тому, якщо $\{\tilde{A},\tilde{b}\}$ – збурені дані задачі (1.1), то приймаючи за наближений розвязок задачі (1.1) вектор $\tilde{x}$ – точний розв'язок задачі (7.6), ми не застраховані від великої відносної похибки в розв'язку ($\frac{\|x-\tilde{x}\|}{\|x\|}$, якщо $\mu(A)$ велике). Розуміється, ситуація ще

більше ускладниться (в сенсі накопичення похибок), якщо при знаходженні $\tilde{x}$ обчислення проводяться наближено (з заокругленнями), що практично завжди має місце при реалізації методу на обчислювальних комплексах. Слід зауважити, що сам процес вводу (запису) елементів матриці A і вектора b в ЕОМ і заокруглення, пов'язаного з обмеженою розрядністю ЕОМ, може привести до недопустимо великих спотворень розв'язку (див., наприклад, [13]).

Проілюструємо роль числа обумовленості $\mu(A)$ на таких прикладах.

Приклад 7.1.

$$A=\begin{bmatrix}4{,}1 & 2{,}8\\ 9{,}7 & 6{,}6\end{bmatrix},\quad b=\begin{bmatrix}4{,}1\\ 9{,}7\end{bmatrix},\quad x=\begin{bmatrix}1\\ 0\end{bmatrix}.$$

Очевидно, що $A\cdot x=b$, $\|b\|_{\text{II}}=13{,}8$, $\|x\|_{\text{II}}=1$.

Замінимо тепер праву частину на $b'=\begin{bmatrix}4{,}11\\ 9{,}7\end{bmatrix}$, тоді розв'язок буде

$$x'=\begin{bmatrix}0{,}34\\ 0{,}97\end{bmatrix}.$$

Позначимо $\Delta b=b-b'$, $\Delta x=x-x'$, тоді

$$\|\Delta b\|_{\text{II}}=0{,}01,\quad \|\Delta x\|_{\text{II}}=1{,}63.$$

Мале збурення, внесене в b, значно змінило розв'язок x.

Дійсно, відносні зміни правої частини і розв'язку рівні, відповідно:

$$\frac{\|\Delta b\|_{\text{II}}}{\|b\|_{\text{II}}}=0{,}0007246,\quad \frac{\|\Delta x\|_{\text{II}}}{\|x\|_{\text{II}}}=1{,}63.$$

Із визначення $\mu(A)$ і оцінки (7.3) випливає, що

$$\mu(A)\ge\frac{\|\Delta x\|}{\|x\|}:\frac{\|\Delta b\|}{\|b\|}=1{,}63:0{,}0007246=2249{,}5.$$

В дійсності вектори та Δb вибрані таким чином, що вони дають максимальне зростання похибки, так що для даного прикладу

$$\mu(A)=2249{,}5,$$

у чому легко можна переконатись, використовуючи для обчислень $\mu(A)$ формулу (7.5).

Дуже важливо зрозуміти, що в наведеному прикладі мова йде про точний розв'язок двох систем рівнянь, які мало відрізняються. Значить, метод, за допомогою якого були одержані ці розв'язки, зовсім не важливий. Незважаючи на штучність вибору прикладу таким, щоб ефект збурення b був яскраво виражений, подібну

ситуацію можна очікувати в будь-якій задачі з великим числом обумовленості.

Приклад 7.2.

Припустимо, що необхідно розв'язати систему, в якій $a_{11} = 0{,}1$, а всі інші елементи матриці A і b – цілі числа степеня 2. Число обумовленості матриці $\mu(A)=10^7$. Будемо вважати, що в нашому розпорядженні обчислювальний комплекс, який використовує 32 біти для дробової частини і ми якимось чином уміємо обчислити точний розв'язок системи, яка записана в пам'яті машини. Тоді єдина помилка буде пов'язана з двійчастим зображенням числа 0,1, і однак можна очікувати, що відносна похибка в розв'язку

$$\frac{\|\Delta x\|}{\|x\|} \approx 10^7 \cdot 2^{-32} \approx 2 \cdot 10^{-3},$$

це означає, що вже запис коефіцієнтів має в машині можуть привести до змін третьої (десяткової) значущої цифри компонент правильного розв'язку.

Число обумовленості має ґрунтовне значення при аналізі похибок заокруглень в процесі розв'язку СЛАР, наприклад, методом Гаусса. Нехай елементи матриці А і вектора правих частин b є точно зображеними числами з плаваючою комою, а x^* – вектор розв'язку, одержаний на виході підпрограм розв'язку СЛАР, які реалізують метод виключення Гаусса. Через *x*, як і вище, позначимо точний розвязок системи $A \cdot x = b$.

Тоді справедливі наступні нерівності ([13])

$$\begin{aligned} \frac{\left\|A \cdot x^* - b\right\|}{\|A\| \cdot \left\|x^*\right\|} &\le \rho \cdot \beta^{-t}, \\ \frac{\left\|x - x^*\right\|}{\left\|x^*\right\|} &\le \rho \cdot \mu(A) \cdot \beta^{-t}, \end{aligned} \tag{7.7}$$

де β – основа числення системи, яка використовується (звичайно $\beta = 2$) для зображення чисел з плаваючою комою,

t – число розрядів мантиси, так що β^{-t} відіграє роль машинного ε (тобто найменшого числа $\varepsilon \neq 0$, зображеного в ЕОМ),

ρ – деяка константа, не більша за β.

Величину $\left\|Ax^* - b\right\|$ будемо називати *нев'язкою*.

Із першої нерівності (7.7) випливає, що відносна величина нев'язки порівнювана з помилкою заокруглень, незалежно від

величини обумовленості. Друга нерівність, одержана при допущенні невиродженості матриці, говорить, що відносна похибка розв'язку може бути великою, якщо число обумовленості $\mu(A)$ велике. Це пов'язане з тим, що існують приклади, на яких досягається рівність в другій нерівності (7.7).

Вищевикладене дозволяє дати таку класифікацію матриць СЛАР. Матриця A (системи $A \cdot x = b$) називається добре обумовленою, якщо $\mu(A)$ відносно невелике. Матриця A (системи $A \cdot x = b$) називається погано обумовленою, якщо $\mu(A)$ відносно велике.

Що означає $\mu(A)$ „відносно мале" чи „відносно велике"? Однозначної відповіді на це запитання дати поки що не можна, але можна дещо уточнити ці поняття.

Припустимо, що вихідні дані СЛАР (тобто, матриця і праві частини задані з точністю δ: $\frac{\|A - \widetilde{A}\|}{\|A\|} \le \delta$, $\frac{\|b - \widetilde{b}\|}{\|b\|} \le \delta$). Треба знайти розв'язок $\widetilde{x}$, який апроксимує (наближає) точний розв'язок x з похибкою, не більшою ε, $\frac{\|x - \widetilde{x}\|}{\|x\|} \le \varepsilon$ (природно, що $\delta \le \varepsilon$). На основі оцінок (7.3), (7.4) розв'язуючи систему $\widetilde{A} \cdot \widetilde{x} = \widetilde{b}$ звичайним методом Гаусса, не можна, взагалі кажучи, одержати точність вищу, ніж $\mu(A) \cdot \delta$.

З цієї причини, якщо $\mu(A) \cdot \delta < \varepsilon$, то матрицю (систему) природно вважати добре обумовленою відносно заданих рівнів похибок δ вихідних даних і точності ε розв'язку, а якщо $\mu(A) \cdot \delta \ge \varepsilon$, то – погано обумовленою.

Практично, системи з $\mu(A)$, які не більші від декількох сотень, відносять до добре обумовлених, а системи з величиною $\mu(A) \ge 1000$ вважають вже погано обумовленими. Неважко бачити, що такий поділ є, звісно, умовним.

Прикладом дуже погано обумовленої системи може служити система з матрицею Гілберта (див. §6), яка виникає при апроксимації функції $f(x)$ на відрізку [0;1] поліномом степеня n [13]. Вже для n=6 $\mu(A) = 1{,}5 \cdot 10^7$, а при n=10 $\mu(A) = 1{,}6 \cdot 10^{13}$. Ефективний розв'язок настільки погано обумовлених систем складає нелегку обчислювальну проблему.

§8. АНАЛІЗ ПОХИБОК ОБЧИСЛЕНЬ

Розглянемо різні випадки, в яких відбувається велика втрата точності при знаходженні розв'язку або виникають явища чисельної нестійкості, коли розв'язок значно змінюється при незначних збуреннях правих частин та елементів матриць СЛАР.

8.1 Неправильна організація обчислень в методі виключення

Приклад 8.1.

Будемо розв'язувати наступну систему, виконуючи арифметичні операції з 4 десятковими знаками:

$$\begin{cases} 0{,}0001 \cdot x_1 + 0{,}5 \cdot x_2 = 0{,}5 \\ 0{,}4 \cdot x_1 \quad - 0{,}3 \cdot x_2 = 0{,}1 \end{cases}.$$

Справжній розв'язок, заокруглений до 4-х правильних знаків, є

$$x_1 = \frac{2}{2{,}0003} = 0{,}9999, \;\; x_2 = \frac{1999{,}9}{2000{,}3} = 0{,}9998.$$

Виключаючи невідоме x_1 з 2-го рівняння і заокруглюючи до 4-х знаків, одержимо методом Гаусса

$$\begin{cases} 0{,}0001 \cdot x_1 + \;\; 0{,}5 \cdot x_2 = 0{,}5 \\ \qquad\qquad - 2000 \cdot x_2 = 2000 \end{cases} \Rightarrow x_2 = 1, \; x_1 = 0.$$

Переставимо рівняння і знову проведемо виключення x_1 (заокруглюючи до 4-х знаків):

$$\begin{cases} 0{,}4 \cdot x_1 - \;\; 0{,}3 \cdot x_2 = 0{,}1 \\ \qquad\quad - 2000 \cdot x_2 = -2000 \end{cases} \Rightarrow x_2 = 1, \; x_1 = 1.$$

Цей приклад показує, що слід уникати малих за абсолютною величиною головних (ведучих) елементів a_{ii} (в першому випадку це $a_{11} = 0{,}0001$). Вихідна система, добре обумовлена, і правильна реалізація методу виключення (з вибором головного елемента, максимального в стовпчику) дає гарний результат. Якщо ж головний елемент вибирати по всій матриці,

$$\begin{cases} \;\; 0{,}5 \cdot x_2 + 0{,}0001 \cdot x_1 = 0{,}5 \\ - 0{,}3 \cdot x_2 + \quad 0{,}4 \cdot x_1 = 0{,}1 \end{cases} \Rightarrow 0{,}2003 \cdot x_1 = 0{,}2,$$

$$x_1 = 0{,}9999, \; x_2 = 0{,}9998.$$

Правильний розв'язок одержаний з точністю до 4-х знаків.

Таким чином, накопичення похибки в першому варіанті виключення обумовлено не „поганими" властивостями системи, а непродуманої схеми реалізації методу Гаусса.

8.2 Близькість до нуля визначника системи

Приклад 8.2.

Приведемо три системи рівнянь з двома невідомими з коефіцієнтами при них, які мало відрізняються один від одного. Праві частини також мало відрізняються

$$\begin{cases} 3\cdot x_1 - 7{,}0001\cdot x_2 = 0{,}9998 \\ 3\cdot x_1 - \quad 7\cdot x_2 = 1 \end{cases}, \quad x_1 = 5,\ x_2 = 2;$$

$$\begin{cases} 3\cdot x_1 - 7{,}0001\cdot x_2 = 1 \\ 3\cdot x_1 - \quad 7\cdot x_2 = 1 \end{cases}, \quad x_1 = \frac{1}{3},\ x_2 = 0;$$

$$\begin{cases} 3\cdot x_1 - 7\cdot x_2 = 0{,}9999 \\ 3\cdot x_1 - 7\cdot x_2 = 1 \end{cases}, \quad \text{система несумісна}.$$

Праві частини відрізняються на величину $2\cdot 10^{-4}$, а їх розв'язки відрізняються значно. Третя система, котра відрізняється від двох попередніх (в коефіцієнті матриці та в правій частині) на величину 10^{-4}, вже не має розв'язку. В наявності сильна чутливість розв'язку до збурення вхідних даних. Причиною цього є малість визначника системи (майже виродженість)

$$\Delta = \begin{vmatrix} 3 & 7{,}0001 \\ 3 & 7 \end{vmatrix} = -0{,}0003.$$

Таким чином, близькість до нуля визначника може бути причиною поганої обумовленості. Однак це не завжди так. Наприклад, діагональна матриця A порядку 100 і числом $\frac{1}{10}$ на діагоналі має визначник $\det(A) = 10^{-100}$. Число ж обумовленості для неї $\mu(A) = 1$ (див. властивість 3) в §3). Компоненти ж вектора $b = A\cdot x$ відрізняються тільки множником 0,1 від компонент вектора x. Значить, проблем зі знаходженням розв'язку не виникає.

8.3 Малі за модулем власні значення

Нехай A – квадратна матриця n x n. Нагадаємо, що власним значенням матриці A є число λ, для якого рівняння $A\cdot x = \lambda\cdot x$ має ненульовий розв'язок x_λ; цей вектор називають власним вектором СЛАР. Власними значеннями $\{\lambda_i\}_{i=1}^n$, є корені вікового рівняння

$$\det(A - \lambda\cdot E) = 0.$$

Якщо всі корені λ_i дійсні та різні, то відповідні власні вектори u_i утворюють базис в просторі R^n.

Розкладемо вектор b по цьому базису $b = \sum_{k=1}^{n} b_k u_k$ і будемо шукати розв'язок у вигляді $x = \sum_{k=1}^{n} x_k u_k$. Тоді

$$A \cdot x = \sum_{k=1}^{n} x_k \cdot A \cdot u_k = \sum_{k=1}^{n} x_k \cdot \lambda_k \cdot u_k = \sum_{k=1}^{n} b_k u_k ,$$

звідки через лінійну незалежність векторів

$$x_k = \frac{b_k}{\lambda_k}. \tag{8.1}$$

Очевидно, що якщо серед λ_i є близькі до нуля власні значення λ_{i_0} і вектор b заданий або обчислений з похибкою, то відповідна компонента вектора розв'язку може бути обчислена з великою похибкою (оскільки малі похибки b_{i_0} у формулі (8.1) діляться на дуже малі λ_{i_0}). В цьому випадку говорять, що розв'язок „розбовтується" в напрямку власних векторів, які відповідають малим власним значенням.

Справа ускладнюється ще й тим, що саме задача знаходження власних значень λ_i (які використовуються у формулі (8.1)), взагалі кажучи, нестійка відносно збурень елементів матриці.

Пояснимо це наступним прикладом. Розглянемо дві близькі матриці розмірності 25×25.

$$A = \begin{bmatrix} -1 & 10 & 0 & \dots & 0 \\ 0 & -1 & 10 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 10 \\ 0 & 0 & 0 & \dots & -1 \end{bmatrix}, \qquad A_\varepsilon = \begin{bmatrix} -1 & 10 & 0 & \dots & 0 \\ 0 & -1 & 10 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 10 \\ \varepsilon & 0 & 0 & \dots & -1 \end{bmatrix}.$$

Із рівнянь

$$\det(A - \lambda \cdot E) = (-1 - \lambda)^{25} = 0,$$

$$\det(A_\varepsilon - \lambda \cdot E) = (-1 - \lambda)^{25} + \varepsilon \cdot 10^{24} = (-1 - \lambda)^{25} + \left(\frac{10}{8}\right)^{25} = 0,\ \varepsilon = \frac{10}{8^{25}}.$$

Знаходимо відповідно $\lambda = -1,\ \lambda_\varepsilon = \frac{1}{4}$, тобто власні числа відрізняються на величину $\frac{5}{4}$, в той час як похибка в заданні коефіцієнтів матриці складає число, менше 10^{-22}. Зауважимо, що на ЕОМ, які використовують десяткові числа з мантисою, меншою 22 знаків, ці матриці при заданні тотожні. З цієї причини, при знаходженні розв'язку за формулами (8.1) помилка в розв'язку

може значно збільшитися за рахунок наближеного знаходження власного числа λ_i.

8.4 Наявність великих елементів в оберненій матриці

Розглянемо спочатку, як зв'язані зміни елементів оберненої матриці зі змінами елементів вхідної матриці. Нехай $A^{-1} = \{\alpha_{ij}\}_{i,j=1}^{n}$ – обернена до А матриця. Тоді

$$A \cdot A^{-1} = E, \quad \frac{\partial A}{\partial a_{ij}} \cdot A^{-1} + A \cdot \frac{\partial A^{-1}}{\partial a_{ij}} = 0, \quad \frac{\partial A^{-1}}{\partial a_{ij}} = -A^{-1} \cdot \frac{\partial A}{\partial a_{ij}} \cdot A^{-1},$$

$$\frac{\partial A}{\partial a_{ij}} = e_{ij} = i \begin{bmatrix} 0 & \dots & \overset{j}{0} & \dots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \dots & 1 & \dots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix} = i \begin{bmatrix} 0 & \dots & 0 \\ \vdots & & \vdots \\ 1 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & \dots & \overset{j}{1} & \dots & 0 \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix} = e_i \cdot e_j,$$

$$\frac{\partial A^{-1}}{\partial a_{ij}} = -\begin{bmatrix} \alpha_{1i} & 0 & \dots & 0 \\ \alpha_{2i} & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ \alpha_{ni} & 0 & \dots & 0 \end{bmatrix} \cdot \begin{bmatrix} \alpha_{j1} & \alpha_{j2} & \dots & \alpha_{jn} \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{bmatrix} = -\begin{bmatrix} \alpha_{1i} \cdot \alpha_{j1} & \dots & \alpha_{1i} \cdot \alpha_{jn} \\ \alpha_{2i} \cdot \alpha_{j1} & \dots & \alpha_{2i} \cdot \alpha_{jn} \\ \dots & \dots & \dots \\ \alpha_{ni} \cdot \alpha_{j1} & \dots & \alpha_{ni} \alpha_{jn} \end{bmatrix}$$

$$\frac{\partial \alpha_{kl}}{\partial a_{ij}} = -\alpha_{ki} \cdot \alpha_{jl}. \tag{8.2}$$

Звідси

$$d\,\alpha_{kl} = -\sum_{i,j=1}^{n} \alpha_{ki} \cdot \alpha_{jl} \cdot d\,a_{ij}.$$

Провівши аналогічні викладки, одержимо

$$x = A^{-1} \cdot b,$$

$$\frac{\partial x}{\partial a_{ij}} = \frac{\partial \left(A^{-1}\right)}{\partial a_{ij}} \cdot b = -A^{-1} \cdot e_{ij} \cdot A^{-1} \cdot b = A^{-1} \cdot e_{ij} \cdot x = -A^{-1} \cdot \begin{bmatrix} 0 \\ \vdots \\ x_j \\ \vdots \\ 0 \end{bmatrix} i = -\begin{bmatrix} \alpha_{1i} & x_j \\ \alpha_{2i} & x_j \\ \dots & \dots \\ \alpha_{ni} & x_j \end{bmatrix},$$

$$\frac{\partial x_k}{\partial a_{ij}} = -\alpha_{ki} \cdot x_j,$$

$$d\,x_k = -\sum_{i,j=1}^{n} \alpha_{kj} \cdot x_j \cdot d\,a_{ij}\,, \tag{8.3}$$

$$\frac{\partial x}{\partial b_i} = A^{-1} \cdot \frac{\partial b}{\partial b_i} = A^{-1} \cdot \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} i = \begin{bmatrix} \alpha_{1i} \\ \vdots \\ \alpha_{ii} \\ \vdots \\ \alpha_{ni} \end{bmatrix},$$

$$\frac{\partial x_k}{\partial b_i} = \alpha_{ki}, \tag{8.4}$$

$$d\,x_k = \sum_{i,j=1}^{n} \alpha_{ki} \cdot d\,b_i .$$

Обернену матрицю A^{-1} будемо називати нестійкою., якщо малим збуренням елементів матриці A відповідають великі зміни елементів оберненої матриці. Якщо A^{-1} містить великі за абсолютною величиною елементи α_{ij}, то на підставі формули (8.2) вона буде нестійкою. Як показують співвідношення (8.3), (8.4), в цьому випадку малі зміни коефіцієнтів системи та вільних членів можуть потягнути значні спотворення в розв'язках. Ця обставина також дає підстави характеризувати погано обумовлену систему рівнянь (поряд з визначенням в §7), як систему з нестійкою оберненою матрицею [13]. Такі системи можуть бути практично виродженими, якщо її елементи задані з похибками.

Приклад 8.3.

Розглянемо матрицю $n \times n$ (n=40)

$$A = \begin{bmatrix} 1 & 2 & 0 & \dots & 0 & 0 \\ 0 & 1 & 2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 2 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix},$$

для якої $\det(A) = 1$ і всі значення $\lambda_i = 1$. Розглянемо поряд з A збурену матрицю

$$A_\varepsilon = \begin{bmatrix} 1 & 2 & 0 & \dots & 0 & 0 \\ 0 & 1 & 2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 2 \\ \varepsilon & 0 & 0 & \dots & 0 & 1 \end{bmatrix},$$

для якої уже $\det(A_\varepsilon) = 1 - \varepsilon \cdot 2^{39}$; при $\varepsilon = 2^{-39} \approx 10^{-12}$, $\det(A_\varepsilon) = 1 - \varepsilon \cdot 2^{39}$, тобто система рівнянь з матрицею A_ε після запису її елементів у пам'ять ЕОМ з точністю $\varepsilon \approx 10^{-12}$ буде виродженою. Тоді при деяких правих частинах b ця система буде несумісна. Причиною нестійкості є наявність великих коефіцієнтів в оберненій матриці $A^{-1} = \{\alpha_{ij}\}_{i,j=1}^{n}$, наприклад, $\alpha_{11} = 2^{39} > 10^{11}$.

Приклад 8.4.

Розглянемо дві близькі матриці [26]

$$A = \begin{bmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{bmatrix}, \qquad A_\varepsilon = \begin{bmatrix} 5+\varepsilon & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{bmatrix},$$

при $\varepsilon = 10^{-2}$ обернені матриці до них є, відповідно,

$$A^{-1} = \begin{bmatrix} 68 & -41 & -17 & 10 \\ -41 & 25 & 10 & -6 \\ -17 & 10 & 5 & -3 \\ 10 & -6 & -3 & 2 \end{bmatrix}, \qquad A_\varepsilon = \begin{bmatrix} 212{,}5 & -128{,}12 & -53{,}12 & 31{,}25 \\ -128{,}12 & 77{,}53 & 31{,}78 & -18{,}81 \\ -53{,}12 & 31{,}78 & 14{,}03 & -8{,}31 \\ 31{,}25 & -18{,}81 & -8{,}31 & 5{,}12 \end{bmatrix}.$$

Великі розбіжності елементів A^{-1} і A_ε^{-1} показують, що матриця A^{-1} нестійка, система $A \cdot x = b$ буде погано обумовленою. Більше того, при $\varepsilon = -\frac{1}{68}$ неважко обчислити, що $\det(A_\varepsilon) = 1 + 68\varepsilon \approx 0$, тобто в границях точності до 0,02 система повинна розглядатися як вироджена, з усіма наслідками, що звідси випливають.

Приклади 8.3, 8.4 дають нам підстави ввести наступну міру виродженості матриці

$$d(A) = \min_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Якщо A вироджена, то, очевидно, $d(A) = 0$. Для невироджених матриць A – справедлива формула

$$d(A) = \frac{1}{\|A^{-1}\|},$$

з якої випливає, що чим більше $\|A^{-1}\|$ (тобто, чим більше число обумовленості $\mu(A)$), тим менше $d(A)$, і, значить, тим ближче матриця А до виродженості.

§9. УЗАГАЛЬНЕННЯ ПОНЯТТЯ РОЗВ'ЯЗКУ. ПСЕВДОРОЗВ'ЯЗОК

Перш ніж перейти до викладу методів наближеного розв'язку погано обумовлених систем, необхідно розв'язати питання про те, що треба розуміти під розв'язком системи рівнянь (1.1), котра в загальному випадку може бути перевизначеною, недовизначеною, тобто такою, коли апріорі невідомо існування і єдність розв'язку.

Позначимо через $\bar{x}$ вектор, який реалізує мінімум нев'язки

$$\|A \cdot \bar{x} - b\|^2 = \min_x \left\{ \|A \cdot x - b\|^2 : x \in Rn \right\}. \qquad (9.1)$$

Величина $\bar{x}$ називається розв'язком системи (1.1) в сенсі методу найменших квадратів. Необхідна умова мінімуму функціонала $I(x) = \|A \cdot x - b\|^2$ в (9.1) є $\delta I(\bar{x}) = 0$, де I – варіація функціонала. Оскільки для приросту функціонала має місце представлення

$$I(\bar{x} + h) - I(\bar{x}) = 2(Ax, Ah) - 2(Ah, b) + (Ah, Ah),$$

тоді $\delta I(\bar{x}) = 2(A^* A\bar{x} - A^* b) = 0$. Значить вектор $\bar{x}$ задовольняє системі рівнянь

$$A^* A\bar{x} = A^* b \qquad (9.2)$$

(нагадаємо, що A^* – транспонована до A матриця). Неважко показати, що правильне і обернене твердження: кожний розв'язок (9.2) мінімізує нев'язку в (9.1), так що задачі (5.1) і (5.2) еквівалентні. Із елементарних геометричних міркувань можна встановити, що задача (9.1) завжди має розв'язок, можливо не один. З цієї причини через встановлений факт еквівалентності система (9.2) також має розв'язок для будь-якої матриці A та вектора b.

Позначимо $\bar{X}$ – множину розв'язків системи (9.2) (а значить і задачі (9.1)). Задавши деякий фіксований елемент x^0, який відіграє роль пробного розв'язку (наприклад, можна покласти $x^0 = 0$), розглянемо задачу на мінімум:

$$min \left\{ \left\| x - x^0 \right\|^2 : x \in \bar{X} \right\}. \qquad (9.3)$$

Розв'язок $\tilde{x}$ задачі (9.3) існує і єдиний, що випливає з того факту, що суворо опуклий функціонал досягає найменшого значення на опуклій замкнутій множині в єдиній точці. *Вектор* $\tilde{x}$ будемо називати *псевдорозв'язком* здачі (1.1).

Із співвідношень

$$A^* Au = 0,$$
$$\left(A^* Au, u\right) = \|Au\|^2 = 0,$$

$$Au = 0,$$
$$Ax = 0,$$
$$A^* Ax = 0$$

випливає, що множина розв'язків однорідних систем $A^* Au = 0$, $Ax = 0$ збігається. Звідси негайно випливає наступний важливий факт. Якщо система (1.1) сумісна, то псевдорозв'язок $\tilde{x}$ збігається з нормальним розв'язком цієї системи, тобто є розв'язком, який найменше відхиляється по нормою від вектора x^0. І в частинному випадку, якщо (1.1) однозначно розв'язується, то псевдорозв'язок співпадає зі звичайним розв'язком.

Переконаймося, що псевдорозв'язок не стійкий по відношенню до збурень елементів матриці.

Приклад 9.1.

Для несумісної системи [16, стор. 299]

$$\begin{cases} x_1 + 0 \cdot x_2 = 1 \\ 0 \cdot x_1 + 0 \cdot x_2 = 1, \end{cases}$$
$$A^* A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A^* b = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

тому множина розв'язків системи (5.2) в даному прикладі $x_1 = 1,\ x_2$ – довільне число, тобто при $x^0 = 0$ псевдорозв'язком буде вектор $x_1 = 1,\ x_2 = 0$.

Збурену систему візьмемо у вигляді

$$\begin{cases} x_1 + 0 \cdot x_2 = 1 \\ 0 \cdot x_1 + \varepsilon \cdot x_2 = 1, \end{cases} \quad \varepsilon \text{ – мале число.}$$

Оскільки ця система має єдиний розв'язок $\tilde{x}_\varepsilon = \left(1, \frac{1}{\varepsilon}\right)^T$ (Т – значок транспонування вектора $\tilde{x}_\varepsilon$), то воно і буде псевдорозв'язком збуреної системи. При $\varepsilon \to 0$ $\tilde{x}_\varepsilon \to \infty$ і не апроксимує псевдорозв'язку $\tilde{x}_\varepsilon = (1, 0)^T$.

§10. МЕТОД РЕГУЛЯРИЗАЦІЇ АКАД. А. М. ТИХОНОВА

В основі побудови стійких методів розв'язку некоректних задач лежить поняття регуляризуючого алгоритму (РА) і зв'язаного з ним поняття регуляризованого сімейства розв'язків, введеного А. М. Тихоновим [13]. Погано обумовлені СЛАР слід розглядати як некоректно поставлені задачі і при їх наближеному розв'язку необхідно застосовувати ідеї регуляризації.

Умовимося розрізняти точні дані – пару $\{A, b\}$, котрі формують задачу (1.1) і нам невідомі, і наближені дані $\{A_h, b_\delta\}$, $\|A_h - A\| \le h$, $\|b_\delta - b\| \le \delta$ з рівнем похибок h, δ, якими ми володіємо. Суть регуляризованого методу наближеного розв'язку стосується побудови послідовності векторів $x_{h,\delta}$, яка збігається до розв'язку або псевдорозв'язку рівняння (1.1) при $\delta \vee h \to 0$. Як було показано в §§7-9, за наближений розв'язок $x_{h,\delta}$ не можна, взагалі кажучи, брати точний розв'язок або квазірозв'язок рівняння з індивідуальними даними

$$A_h \cdot x = b_\delta \tag{10.1}$$

(див. приклад 7.1).

Нехай в нашому розпорядженні є спосіб (правило), який по парі $\{A_h, b_\delta\}$ і додатньому параметру α однозначно будує вектор $x^\alpha(A_h; b_\delta)$. Якщо існує залежність параметра $\alpha(\delta, h)$ від похибок δ, h вихідних даних така, що

$$\lim_{\substack{\delta \to 0 \\ h \to 0}} \left\| x^{\alpha(\delta,h)}(A_h, b_\delta) - \tilde{x} \right\| = 0, \tag{10.2}$$

тоді множина $\left\{x^{\alpha(\delta,h)}(A_h; b_\delta)\right\}$ називається *регуляризованим сімейством* наближених розв'язків, а сам *спосіб побудови* $x^\alpha(A_h; b_\delta)$ – *регуляризуючим алгоритмом* для задачі (1.1). Тут вектор $\tilde{x}$ – розв'язок, нормальний розв'язок або псевдорозв'язок системи (1.1) залежно від того, чи розв'язується ця система однозначно, чи має множину розв'язків чи не має.

Співвідношення (10.2) засвідчує, що наближений розв'язок $x^\alpha(A_h; b_\delta)$ тим краще апроксимує точний розв'язок $\tilde{x}$, чим менша похибка вихідних даних δ, h. Таким чином, регуляризуючий алгоритм дає теоретичну базу для конструювання стійкого до збурень вихідних даних наближеного розв'язку системи (1.1) загального виду, включаючи погано обумовлені системи.

Важливо розуміти наступну обставину. Якщо A^{-1} існує, тобто A – невироджена матриця, то для достатньо малих h A_h^{-1} теж існує і розв'язок $x_{h,\delta}$ рівняння (10.1) теоретично буде збігатися до розв'язку рівняння (1.1) при $\delta, h \to 0$. Однак якщо A – погано обумовлена матриця ($\mu(A)$ велике), то величина похибки $\|\tilde{x} - x_{\delta,h}\|$, навіть при малих δ, h, може бути недопустимо великою і задачу слід вважати практично нестійкою (некоректною). Метод регуляризації саме і направлений на те, щоб зменшити вплив похибок (вхідних даних, обчислень) і одержати практично стійкий наближений розв'язок в цих несприятливих обставинах.

Тепер перейдемо до опису конкретних процедур побудови регуляризованих наближених розв'язків $x^{\alpha}(A_h; b_\delta)$. Далі для скорочення запису будемо опускати залежність $x^{\alpha(\delta,h)}(A_h; b_\delta)$ від $(A_h; b_\delta)$ і записувати просто $x^{\alpha(\delta,h)}$.

Розглянемо спочатку частинний випадок – схему М. М. Лаврент'єва [15], коли A –симетрична додатня напіввизначена матриця, для якої система (1.1) при заданому векторі b може бути розв'язана.

Перейдемо від (1.1) до регуляризованої системи

$$(A + \alpha E)x^{\alpha} = b + \alpha x^0,$$

де α – додатній параметр, E – одинична матриця, x^0 – пробний розв'язок, тобто деяке наближення до шуканого розв'язку (якщо інформація про розв'язок відсутня, то можна позначити $x^0 = 0$).

При зроблених застереженнях, СЛАР (10.3) має єдиний розв'язок x^{α}, який збігається при $\alpha \to 0$ до нормального розв'язку $\tilde{x}$ (див. §9).

Твердження (10.1) [15].

Нехай $\{A_h, b_\delta\}$, $\|A_h - A\| \le h$, $\|b_\delta - b\| \le \delta$ – наближені дані задачі і A_h – симетрична додатньо напіввизначена матриця. Тоді СЛАР

$$(A_h + \alpha E)x^{\alpha} = b_\delta + \alpha x^0 \qquad (10.4)$$

однозначно розв'язана і при зв'язку параметра α з похибками δ, h такими, що $\alpha(\delta, h) \to 0$, $\dfrac{h+\delta}{\alpha(\delta,h)} \to 0$, коли $\delta \to 0$, $h \to 0$, тобто $x^{\alpha(\delta,h)}$ збігається до нормального розв'язку $\tilde{x}$ рівняння (1.1). Це і є розв'язок, який найменше ухиляється від вектора x^0.

Таким чином, згідно з визначенням, даним вище, розв'язки СЛАР (10.4) $\left\{x^{\alpha(\delta,h)}\right\}$ утворюють регуляризоване сімейство наближених розв'язків для системи (1.1); причому, вибір параметра за

формулою $\alpha = \sqrt[p]{\delta + h}$ $(p > 1)$ задовольняють необхідним вимогам, оскільки $\alpha = \sqrt[p]{\delta + h} \to 0$, $\dfrac{\delta + h}{\sqrt[p]{\delta + h}} = (\delta + h)^{1 - \frac{1}{p}}$, коли $\delta,\ h \to 0$.

Зауваження (10.1).

Нехай A – додатньо напіввизначена вироджена матриця і $\|A\| = 1$ (цього можна завжди досягти підходящим масштабуванням рівнянь системи (1.1)). Коли $\mu(A) = \infty$, в той час $\mu(A + \alpha E) \le \dfrac{1 + \alpha}{\alpha}$. З цієї причини при розумному виборі параметра α можна досягти гарної обумовленості систем (10.3), (10.4) і задовільної апроксимації $x^\alpha \approx \tilde{x}$, не зважаючи на те, що ці вимоги мають протиріччя.

Роль параметра регуляризації α добре видно, якщо записати розв'язок системи (10.3) (при $x^0 = 0$) у вигляді

$$x^\alpha = \sum_{i=1}^{n} \frac{b_i}{\lambda_i + \alpha} \cdot u_i,$$

де λ_i власні значення ($\lambda_i \ge 0$), а u_i ортонормовані власні вектори матриці A (виведення формули аналогічне п. 3 §8). Це зображення показує, що при малих λ_i додавання додатного параметра суттєво збільшує знаменник і тим самим послаблює вплив можливих похибок у відповідних компонентах b_i ($\tilde{b}_i = b_i + \Delta b_i$). Одночасно для $\lambda_i \ge 0$ вплив малого параметра α незначне (і навіть таке, що ним можна знехтувати).

Тепер відмовимося від вимоги симетричності і додатності матриці A. Нехай матриця B така, що для деякого α_0 $A + \alpha_0 B$ є не виродженою матрицею і, значить, існує її обернена матриця. Тоді можлива регуляризація в наступній формі:

$$(A + \alpha B) x^\alpha = b,$$

де параметр α довільного знака і $|\alpha| \le |\alpha_0|$.

Справедливо наступне.

Зауваження 10.2 [16].

Нехай $\left\|(A + \alpha B)^{-1} A\right\| \le c < \infty$ (при $\alpha \to 0$), $\|A_h - A\| \le h$, $\|B_\mu - B\| \le \mu$, $\|b_\delta - b\| \le \delta \|b\|$.

Тоді при достатньо малих $h,\ \mu,\ \delta$ СЛАР

$$(A_h + \alpha B_\mu) x^\alpha = b_\delta \tag{10.6}$$

має єдиний розв'язок x^α і справедлива оцінка похибки

$$\left\|x^\alpha - \tilde{x}_B\right\| \le c \cdot \frac{|\alpha| + \mu + \delta + h}{|\alpha|}, \tag{10.7}$$

де $\widetilde{x}_B$ – розв'язок системи (1.1), що задовольняє умові ($\overline{X}$ – множина розв'язків системи (1.1))

$$\|B\widetilde{x}\| = min\{\|Bx\| : x \in \overline{X}\}.$$

Із оцінки (10.7) одразу випливає, що якщо $\alpha(\delta, h, \mu) \to 0$, $\dfrac{\delta + h}{|\alpha(\delta, h, \mu)|} \to 0$ при $\delta, h, \mu \to 0$, має місце збіжність

$$\lim_{\delta, h, \mu \to 0} \left\| x^{\alpha} - \widetilde{x}_B \right\| = 0.$$

Найбільш важливим моментом в описаній регуляризації є підбір матриці B, для якої $A + \alpha_0 B$ невироджена і $\left\|(A + \alpha B)^{-1} A\right\| < \infty$. В роботі [15] можна знайти деякі способи побудови матриць з такою властивістю.

Дослідимо, врешті, загальну ситуацію, коли система (1.1), взагалі кажучи, нерозв'язна. В цьому випадку шуканим є псевдорозв'язок, який був визначений в §9. Розв'язується задача стійкої апроксимації цього псевдорозв'язку в умовах задання вхідних даних з похибкою. Приклад 9.1 демонструє, що псевдорозв'язок $\widetilde{x}$ нестійкий до збурень елементів матриці, тому необхідно і тут використати принцип регуляризації. В якості регуляризованого наближення розв'язку приймемо вектор x^{α}, що задовольняє СЛАР

$$\left(A_h^* A_h + \alpha E\right) x^{\alpha} = A_h^* b_{\delta} + \alpha x^0. \qquad (10.8)$$

Твердження 10.3 [17].

Хай $\|A_h - A\| \le h$, $\|b_{\delta} - b\| \le \delta$, $\alpha > 0$. Тоді СЛАР (10.8) однозначно розв'язна і справедлива оцінка

$$\left\|\widetilde{x} - x^{\alpha}\right\| \le c_1 \alpha + \frac{h}{\alpha}\left(\|A\widetilde{x} - b\| + 2c_2^2 \alpha^2\right)^{\frac{1}{2}} + \frac{1}{\sqrt{\alpha}}(c_3 h + \delta), \qquad (10.9)$$

де c_i $(i = 1, 2, 3)$ константи, які залежать від норми $\|\widetilde{x}\|$ псевдорозв'язку.

Наслідок.

Нехай h, δ величини порядку ε, причому, ε достатньо мале число. Якщо точне рівняння (1.1) має розв'язок, (тобто $\|A\widetilde{x} - b\| = 0$), тоді права частина оцінки (10.9) за характером залежності від α та ε, є функція вигляду

$$\varphi(\alpha) = \alpha + \varepsilon + \frac{\varepsilon}{\sqrt{\alpha}}. \qquad (10.10)$$

При $\alpha = \varepsilon^{\frac{2}{3}}$ вона набуває значення порядку $\varepsilon^{\frac{2}{3}}$. Якщо СЛАР (1.1) нерозв'язна ($\|A\tilde{x} - b\| \neq 0$), то права частина нерівності (10.9) є функція вигляду

$$\psi(\alpha) = \alpha + \frac{\varepsilon}{\alpha} + \frac{\varepsilon}{\sqrt{\alpha}}. \qquad (10.11)$$

При $\alpha = \sqrt{\varepsilon}$ вона набуває значення порядку $\varepsilon^{\frac{1}{2}}$. Ці порядки одержуються за допомогою мінімізації функцій $\varphi(\alpha)$, $\psi(\alpha)$ (тобто є розв'язками рівнянь $\varphi'(\alpha) = 0$, $\psi'(\alpha) = 0$).

Таким чином, якщо вхідні дані рівняння (1.1) задані з точністю порядку ε, то псевдорозв'язок може бути визначений з точністю порядку $\varepsilon^{\frac{2}{3}}$, у випадку розв'язності точного рівняння, і з точністю порядку $\varepsilon^{\frac{1}{2}}$ – в оберненому випадку. Помітимо, однак, що більш складний спосіб вибору параметра α дозволяє апроксимувати псевдорозв'язок з точністю порядку апроксимації $h + \delta$ [64].

Задача (10.8) еквівалентна задачі на мінімум

$$min\left\{\|A_h x - b_\delta\|^2 + \alpha\left\|x - x^0\right\|^2 : x \in R^n\right\}, \qquad (10.12)$$

де L – невироджена матриця розмірності $n \times n$, вибором якої можна розумно розпорядитися, щоб підвищити точність регуляризованого розв'язку.

При $\alpha = 0$ (10.12) переходить в метод найменших квадратів (МНК), який, як показано в §9, не стійкий відносно збурень матриці. Перехід від МНК до його регуляризованого аналога (10.12) відтворює стійкість наближеного розв'язку.

§11. ЗАСТОСУВАННЯ МЕТОДУ РЕГУЛЯРИЗАЦІЇ

Приклад 11.1.

Звернемося до системи рівнянь, розглянутої в прикладі 8.2. Візьмемо, наприклад, систему

$$\begin{cases} 3 \cdot x_1 - 7 \cdot x_2 = 0{,}9999 \\ 3 \cdot x_1 - 7 \cdot x_2 = 1, \end{cases} \tag{11.1}$$

яка несумісна, оскільки ранг матриці $A = \begin{bmatrix} 3 & -7 \\ 3 & -7 \end{bmatrix}$ $(r(A) = 1)$ системи (11.1) не рівний рангу розширеної матриці $B = \begin{bmatrix} 3 & -7 & 0{,}9999 \\ 3 & -7 & 1 \end{bmatrix}$, $r(B) = 2$.

Нагадаємо, що ранг матриці визначається як максимальний порядок мінорів матриці, відмінних від нуля.

Розв'язку системи (11.1) в звичайному сенсі не існує, тому знайдемо спочатку псевдорозв'язок, який було визначено в §9. Для цього необхідно знайти вектор-розв'язок системи $A^* Ax = A^* b$ (A^* – транспонована до A матриця) з найменшою евклідовою нормою (модулем) (див. (9.1) – (9.3)).

Для системи (11.1) маємо

$$A = \begin{bmatrix} 3 & -7 \\ 3 & -7 \end{bmatrix},\ A^* = \begin{bmatrix} 3 & 3 \\ -7 & -7 \end{bmatrix},\ A^* A = \begin{bmatrix} 18 & -42 \\ -42 & 98 \end{bmatrix},$$

$$b = \begin{bmatrix} 0{,}9999 \\ 1 \end{bmatrix},\ A^* b = \begin{bmatrix} 5{,}9997 \\ -13{,}9993 \end{bmatrix}.$$

В одержаній системі

$$\begin{cases} 18 \cdot x_1 - 42 \cdot x_2 = 5{,}9997 \\ -42 \cdot x_1 + 98 \cdot x_2 = -13{,}9993 \end{cases}$$

друге рівняння є наслідком першого і одержане з нього множенням на $-\frac{7}{3}$. Тому, подаючи x_1 довільним, знаходимо з І-го рівняння $x_2 = \frac{(18x_1 - 5{,}9997)}{42}$. Так як норма розв'язку повинна бути мінімальною, то псевдорозв'язок знаходиться на основі розв'язку задачі

$$min\left\{ x_1^2 + \left(\frac{18x_1 - 5{,}9997}{42} \right)^2 \right\}.$$

Позначивши мінімізуючу функцію через $\varphi(x_1)$ і використовуючи необхідну умову екстремуму $\varphi'(x_1) = 0$, знаходимо

$$116x_1 = 5{,}9997, \quad \widetilde{x}_1 \approx 0{,}05172,$$

а $\widetilde{x}_2 = \dfrac{18x_1 - 5{,}9997}{42} = -0{,}12068$ після заокруглення до 4-х знаків.

Далі будемо вважати систему (11.1) точною (тобто система з точними даними), а систему

$$\begin{cases} 3 \cdot x_1 - 7{,}0001 \cdot x_2 = 1 \\ 3 \cdot x_1 \qquad - 7 \cdot x_2 = 1, \end{cases} \tag{11.2}$$

наближеною, так що

$$A_h = \begin{bmatrix} 3 & -7{,}0001 \\ 3 & -7 \end{bmatrix}, \; b_\delta = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \; \|b_\delta - b\| = 0{,}0001, \; \delta = 0{,}0001, \; h = 0{,}0001.$$

Для цих даних застосуємо метод регуляризації (10.8), вибираючи, згідно з наслідком із твердження (10.3), параметр регуляризації $\alpha = \sqrt{\delta} = \sqrt{h} = 0{,}01$.

Справа приводить до розв'язку системи $\left(A_h^* A_h + \alpha E\right)x = A_h^* b_\delta$

$$\begin{cases} 18{,}01 \cdot x_1 - 42{,}0003 \cdot x_2 = 6 \\ -42{,}0003 \cdot x_1 + 98{,}0114 \cdot x_2 = -14{,}0001, \end{cases}$$

$$x_1^\alpha = \frac{\begin{vmatrix} 6 & -42{,}0003 \\ -14{,}0001 & 98{,}0114 \end{vmatrix}}{\begin{Vmatrix} 18{,}01 & -42 \\ -42 & 98{,}01 \end{Vmatrix}} = \frac{0{,}06}{1{,}161} \approx 0{,}0511719,$$

$$x_2^\alpha = \frac{\begin{vmatrix} 18{,}01 & 6 \\ -42{,}0003 & -14{,}0001 \end{vmatrix}}{\begin{vmatrix} 18{,}01 & 42{,}0003 \\ -42{,}0003 & 98{,}0114 \end{vmatrix}} = \frac{0{,}140001}{1{,}161} \approx -0{,}11938.$$

Одержаний регуляризований розв'язок $\widetilde{x} = \left(x_1^\alpha, x_2^\alpha\right)^T$ апроксимує псевдорозв'язок $\widetilde{x} = (\widetilde{x}_1, \widetilde{x}_2)^T$ з точністю до 3-х правильних знаків, тобто з точністю навіть вищою, ніж гарантується оцінками (10.9) – (10.11).

Приклад 11.2.

Якщо взяти в якості наближеної системи

$$\begin{cases} 3 \cdot x_1 - 7{,}0001 \cdot x_2 = 0{,}9998 \\ 3 \cdot x_1 \qquad - 7 \cdot x_2 = 1, \end{cases} \tag{11.3}$$

$$A_h = \begin{bmatrix} 3 & -7{,}0001 \\ 3 & -7 \end{bmatrix}, \; b_\delta = \begin{bmatrix} 0{,}9998 \\ 1 \end{bmatrix}$$

і застосувати той же метод при $\alpha = 0{,}01$, тоді одержимо наближений розв'язок $(0{,}04546;\ -0{,}123347)^T$, який апроксимує шуканий псевдорозв'язок з точністю 0,006. Цей результат слід вважати досить задовільним, оскільки при $\delta = \varepsilon = 0{,}0002$, $h = 0{,}0001 < \varepsilon$, $\alpha = \sqrt{\varepsilon}$ згідно з оцінцкою (10.11) можна розраховувати лиш на точність $\Delta \approx \sqrt{\varepsilon} = 0{,}014$. Таким чином, для наближеної системи (11.3) результат встановлено дещо гірший, ніж для (11.2).

Приклад 11.3.

Дослідимо процедуру регуляризації для СЛАР, розглянутої в прикладі 9.1. Система

$$\begin{cases} x_1 + 0 \cdot x_2 = 1 \\ 0 \cdot x_1 + 0 \cdot x_2 = 1, \end{cases} \tag{11.4}$$

несумісна, тому не існує звичайного розв'язку. Псевдорозв'язок для (11.4) було знайдено в прикладі 9.1 і він рівний $\widetilde{x} = (1, 0)^T$. Наближена система вибиралася у вигляді

$$\begin{cases} x_1 + 0 \cdot x_2 = 1 \\ 0 \cdot x_1 + \varepsilon \cdot x_2 = 1, \end{cases} \tag{11.5}$$

ε – параметр похибки.

Позначимо матрицю цієї системи через A_ε, а вектор правих частин через b, маємо

$$A_\varepsilon = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \end{bmatrix},\ A_\varepsilon^* = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \end{bmatrix},\ A_\varepsilon^* A_\varepsilon = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon^2 \end{bmatrix},\ A_\varepsilon^* b = \begin{bmatrix} 1 \\ \varepsilon \end{bmatrix}.$$

Регуляризована система (10.8) набуває вигляду

$$\begin{cases} (1+\alpha) \cdot x_1 + 0 \cdot x_2 = 1 \\ 0 \cdot x_1 + (\varepsilon^2 + \alpha) \cdot x_2 = \varepsilon, \end{cases} \quad x_1^\alpha = \frac{1}{1+\alpha},\ x_2^\alpha = \frac{\varepsilon}{\varepsilon^2 + \alpha},\ x^\alpha = \left(x_1^\alpha, x_2^\alpha\right)^T.$$

Беручи $\alpha = \sqrt{\varepsilon}$, знаходимо $x^{\alpha=\sqrt{\varepsilon}} = \left(\frac{1}{1+\sqrt{\varepsilon}}, \frac{\sqrt{\varepsilon}}{\varepsilon^{\frac{3}{2}} + 1} \right)^T$.

Знаходимо оцінку похибки

$$\left\| \widetilde{x} - x^{\alpha=\sqrt{\varepsilon}} \right\| = \sqrt{\left(\widetilde{x}_1 - x_1^{\alpha=\sqrt{\varepsilon}}\right)^2 + \left(\widetilde{x}_2 - x_2^{\alpha=\sqrt{\varepsilon}}\right)^2} = \sqrt{\frac{\varepsilon}{\left(1+\sqrt{\varepsilon}\right)^2} + \frac{\varepsilon}{\left(\varepsilon^{\frac{3}{2}} + 1\right)^2}} < \sqrt{2\varepsilon},$$

яка показує, що прямування помилки вихідних даних до нуля призводить до того, що похибка розв'язку прямує також до нуля. Не буде зайвим нагадати, що якщо взяти за наближений розв'язок

псевдорозв'язок системи (11.5) (так як вона розв'язна, то псевдорозв'язок її збігається зі звичайним розв'язком $x_\varepsilon = \left(1, \frac{1}{\varepsilon}\right)^T$), то він взагалі не апроксимує псевдорозв'язок $\tilde{x} = (1,0)^T$ точної системи.

Приклад 11.4.

Розглянемо систему з прикладу 8.4 з точною матрицею A і збуреною матрицею A_ε. Оскільки А симетрична, то допустима схема регуляризації (10.4). Подамо $\alpha = (0{,}01)^{\frac{2}{3}} \approx 0{,}048$, а вектор правої частини $b = (23{,}038 \quad 32{,}048 \quad 33{,}048 \quad 31{,}048)^T$. Тоді розв'язок системи (10.4) при $A_h = A_\varepsilon$, $b_\delta = b$, $\alpha = 0{,}048$, $x^0 = 0 \in x^\alpha = (1,1,1,1)^T$. Так як в прикладі 8.4 були обчислені A^{-1}, A_ε^{-1}, то легко знаходимо $\tilde{x} = A^{-1}b = (1{,}28 \quad 0{,}84 \quad 0{,}93 \quad 1{,}04)^T$,

$x_\varepsilon = A^{-1}b = (4{,}00 \quad -0{,}81 \quad 0{,}25 \quad 1{,}44)^T$.

Переконуємося, що x^α апроксимує точний розв'язок $\tilde{x}$ з відносною похибкою $\Delta^\alpha = \frac{\|x^\alpha - \tilde{x}\|}{\|\tilde{x}\|} \approx 0{,}16$, а нерегуляризований розв'язок x_ε – апроксимує з точністю $\Delta^\varepsilon = \frac{\|x_\varepsilon - \tilde{x}\|}{\|x\|} \approx 1{,}8$, тобто на порядок гірше.

Приклад 11.5.

Застосуємо метод регуляризації (10.8) до погано обумовленої СЛАР із прикладу 7.1 при $\alpha = (0{,}01)^{\frac{2}{3}} \approx 0{,}046$, $b = (4{,}11 \quad 9{,}7)^T$, $A_h = A$, $\delta = 0{,}01$. Одержимо розв'язок $x^\alpha = (0{,}68 \quad 0{,}46)^T$ з відносною помилкою $\Delta^\alpha \approx 0{,}78$ в нормі $\|\cdot\|_{\text{II}}$, що в два рази менша від відносної помилки нерегуляризованого розв'язку x', для якого $\Delta' = 1{,}66$ (див. приклад 7.1).

Приклад 11.6.

Система рівнянь [18]

$$\begin{cases} 100 \cdot x_1 + 500 \cdot x_2 = 1700 \\ 15 \cdot x_1 + 75{,}01 \cdot x_2 = 255, \end{cases} \qquad \det(A) = 1$$

має розв'язок $\tilde{x} = (17{,}0 \quad 0{,}0)^T$, а близька до неї (наближені дані) системи

$$\begin{cases} 100 \cdot x_1 + 500 \cdot x_2 = 1700 \\ 15 \cdot x_1 + 75{,}01 \cdot x_2 = 255{,}03, \end{cases} \qquad \Delta b = b_\delta - b = (0{,}0 \quad 0{,}03)^T,\ \delta = 0{,}03$$

вже має розв'язок $x_\delta = (2{,}0 \quad 3{,}0)^T$, тобто $\dfrac{\|\tilde{x} - x_\delta\|}{\|\tilde{x}\|} \approx 1$. Регуляризації у формі (10.5) при $B = E$, $\alpha = 0{,}1$ дає розв'язок $x^\alpha = (10{,}1 \quad 1{,}5)^T$, для якого $\dfrac{\|x^\alpha - \tilde{x}\|}{\|\tilde{x}\|} \approx 0{,}4 < 1$.

Приклад 11.7.

Нарешті розглянемо систему рівнянь

$$\begin{cases} x_1 + 0{,}99 \cdot x_2 = 1{,}99 \\ 0{,}99 \cdot x_1 + 0{,}98 \cdot x_2 = 1{,}97, \end{cases}$$

з числом обумовленості $\mu(A) = 39600$ (див. [13], стор. 37), її розв'язок $\tilde{x} = (1{,}0 \quad 1{,}0)^T$. Однак $x_\delta = (3{,}0 \quad -1{,}0203)^T$ є розв'язком дуже близької до вихідної системи (наближені дані)

$$\begin{cases} x_1 + 0{,}99 \cdot x_2 = 1{,}989903 \\ 0{,}99 \cdot x_1 + 0{,}98 \cdot x_2 = 1{,}970106, \end{cases} \qquad \Delta b = b_\delta - b = (-0{,}000097 \quad 0{,}000106)^T,$$

тобто $\delta \approx 0{,}0001$. Регуляризований розв'язок за формулою (10.4) з цими наближеними даними, заокруглений до 5 знаків, є $x^\alpha = (0{,}98985 \quad 1{,}0002)^T$, який апроксимує розв'язок $\tilde{x}$ з точністю до 1% при $\alpha = 0{,}01$.

Зробимо висновки. У всіх розглянутих прикладах регуляризований розв'язок суттєво краще апроксимує точний розв'язок, ніж звичайний розв'язок або псевдорозв'язок з наближеними даними. Для подальшого підвищення точності необхідно уточнити вихідні дані.

§12. СПОСОБИ ВИБОРУ ПАРАМЕТРА РЕГУЛЯРИЗАЦІЇ

Згідно з висновками, одержаними в §10, параметр регуляризації α необхідно пов'язувати з похибками δ, h вихідних даних. Там же наведені асимптотичні залежності $\alpha(\delta, h)$, які гарантують збіжність наближених розв'язків при $\delta, h \to 0$. Однак тут допускається значна довільність. Наприклад, поряд з $\alpha = \varepsilon^{\frac{2}{3}}$ або $\varepsilon^{\frac{1}{2}}$ (див. наслідок з твердження 10.3) допустимі також залежності $\alpha = c \cdot e^{\frac{2}{3}}$, $\alpha = c \cdot e^{\frac{1}{2}}$, де c – довільна константа, а $\varepsilon \to 0$, звідси незрозуміло, яким вибрати c. Крім того, в прикладних задачах звичайно рівень похибок δ, h фіксований (δ, h не прямує до нуля) і треба вказати конкретне $\alpha(\delta, h)$, у визначеному сенсі, найкраще. Справа в тім, що при зменшенні α погіршується обумовленість матриць регуляризованих задач, а значить, можуть виникнуть обчислювальні похибки, а при збільшенні α наближений розв'язок погано апроксимує точний розв'язок, що випливає із оцінок (10.10), (10.11). Тому потрібний розумний компроміс.

Опишемо деякі практичні способи вибору параметру α.

1) ***Метод нев'язки.***

Припустимо, що похибка є лиш в правій частині СЛАР $\|b_\delta - b\| \le \delta$, $A_h = A$ $(h = 0)$. Через x^α, як і раніше, позначимо розв'язок системи (10.8). Легко бачити, що функція $\varphi(\alpha) = \left\|Ax^\alpha - b^\delta\right\|^2$ – строго зростаюча неперервна функція параметра α, а як функція від $\frac{1}{\lambda}$ вона буде опуклою вниз (див. [33]).

Значить, для знаходження кореня рівняння

$$\alpha_{нев.}: \quad \varphi(\alpha) = \left\|Ax^\alpha - b^\delta\right\|^2 = \delta^2$$

можна використати метод Ньютона. Тут і далі x^α – розв'язок рівняння (10.8) при $x^0 = 0$ і відповідних наближених даних.

2) ***Метод узагальненої неув'язки***.

Цей метод охоплює загальний випадок появи похибок: $\|A - A_h\| \le h$, $\|b_\delta - b\| \le \delta$. Параметр $\alpha_{0,нев.}$ знаходиться із рівняння

$$\alpha_{нев.}: \quad \left\|Ax^\alpha - b^\delta\right\|^2 = \left(h \cdot \left\|x^\alpha\right\| + \delta\right)^2,$$

для якого розроблені чисельні методи та матзабезпечення (див. [12]).

3) ***Вибір за квазіоптимальним розв'язком***

Даний метод ґрунтується на розв'язку оптимальної задачі (випадок $A_h = A$)

$$\alpha_{к.опт.}: \quad \min_{\alpha} \max_{b_\delta : \|b - b_\delta\| \le \delta} \left\| \alpha \frac{dx^\alpha}{d\alpha} \right\|,$$

де *max* віднаходиться серед всього сімейства наближених правих частин b_δ, але практично достатньо взяти досить багатий набір $\{b_\delta\}$, генеруючи його методом псевдовипадкових чисел. Що стосується обчислення вектора $y_\alpha = \alpha \dfrac{dx^\alpha}{d\alpha}$, то він знаходиться із СЛАР

$$\left(A^* A + \alpha E\right) y_\alpha = A^* A x^\alpha - A^* b_\delta .$$

4) ***Метод співвідношень*** [13].

Параметр $\alpha_{опт.}$ визначається, як розвязок задачі на максимум (випадок $A_h = A$)

$$\alpha_{відн.}: \quad \max_{\alpha} \frac{\left\| A_\alpha \dfrac{dx_\alpha}{d\alpha} - \left(Ax^\alpha - b_\delta\right) \right\|}{\left\| Ax^\alpha - b_\delta \right\|}.$$

5) ***Оптимальний критерій.***

Якщо б нам був відомий точний розв'язок $\tilde{x}$, то параметр $\alpha_{опт.}$, при якому реалізується

$$\alpha_{опт.}: \quad \min_{\alpha} \max_{b_\delta : \|b - b_\delta\| \le \delta} \left\| x^\alpha - \tilde{x} \right\|,$$

природно назвати оптимальним.

На великому класі СЛАР, що виникають при апроксимації інтегральних рівнянь I роду [14], були експериментально встановлені нерівності

$$\alpha_{відн.} < \alpha_{опт.} < \alpha_{к.опт.} < \alpha_{нев.},$$

котрі можуть бути корисними при практичному пошуку найкращого параметра регуляризації в методі А. М. Тихонова. Наприклад, віднайшовши $\alpha_{відн.}$ і $\alpha_{к.опт.}$, можна уточнити параметр α, взявши середнє арифметичне

$$\overline{\alpha} = \frac{\alpha_{відн.} + \alpha_{к.опт.}}{2}.$$

§13. ІТЕРАЦІЙНІ РЕГУЛЯРИЗУЮЧІ АЛГОРИТМИ

При розв'язуванні погано обумовлених систем можуть бути використані ітераційні методи при відповідній організації процесу ітерування. Принциповим тут є той факт, що необхідно сформулювати правило зупинки ітераційного процесу залежно від рівня спотворення вихідних даних, тоді як для добре обумовлених систем в цьому необхідності немає. Роль керуючого параметра (регуляризації) відіграє кількість кроків (ітерацій), виконаних за даною ітераційною схемою.

1) Дослідимо це питання на прикладі простої ітерації

$$x^k = \left(E - A^* A\right)x^{k-1} + A^* b, \quad k = 1, 2, \ldots \tag{13.1}$$

У випадку наближених даних $\{A_h, b_\delta\}$ замість (13.1) маємо ітераційну схему

$$x^k = \left(E - A_h^* A_h\right)x^{k-1} + A_h^* b_\delta, \quad k = 1, 2, \ldots, \tag{13.2}$$

де $\|A - A_h\| \le h, \ \|b - b_\delta\| \le \delta$.

Припустимо, що система (1.1) розв'язувана і $\|A\| \le 1, \ \|A_h\| \le 1$. Остання умова не обмежує класу розв'язуваних задач, поскільки його виконання можна завжди досягти, помноживши рівняння (1.1) на масштабуючу константу. Як і раніше, через $\tilde{x}$ позначимо нормальний розв'язок, фактично розв'язок системи (1.1), для якого норма $\left\|\tilde{x} - x^0\right\|$ найменша, де x^0 – початкове наближення в процесі (13.2).

Визначимо тепер правило зупинення ітерування.

Π_{30}: задамо довільні числа $a_1 > 0$ та $a_2 > 0$; ітерування зупинимо на такому номері $K(\delta, h)$, для якого перший раз виконано

$$\|x_k - x_{k-1}\| \le a_1\delta + a_2 h.$$

Π_{31}: задамо довільні числа $b_1 < 1, \ \tilde{b} \ge \|\tilde{x}\|$; ітерування зупинимо на такому номері $K(\delta, h)$, для якого перший раз виконується нерівність

$$\left\|A_h x^k - b_\delta\right\| \le b_1\delta + \tilde{b}h.$$

Π_{32}: задамо довільні числа $b_1 > 1, b_2 > 1$ і $a > 0$; ітерування призупинимо на такому номері $K(\delta, h)$, для якого вперше буде виконано хоч одну з умов:

$$\left\|A_h x^k - b_\delta\right\| \le b_1\delta + b_2\left\|x^k\right\|h,$$

$$K(\delta, h) \ge \frac{a}{\left(b_1\delta + b_2\left\|x^k\right\|h\right)^2}.$$

Твердження 13.1 [20].

Нехай послідовні наближення (13.2) призупиняються за будь-яким з правил **П30**, **П31**, **П32**. Тоді

$$\lim_{\delta,h\to 0}\left\|x^{K(\delta,h)}-\tilde{x}\right\|=0. \qquad (13.3)$$

При цьому для числа ітерацій $K(\delta,h)$ у випадку **П30** справедливе співвідношення

$$(\delta+h)\cdot K(\delta,h)\to 0 \text{ при } \delta,h\to 0,$$

у випадку правил зупинення **П31**, **П32** – виконується співвідношення

$$(\delta+h)^2\cdot K(\delta,h)\to 0 \text{ при } \delta,h\to 0.$$

Співвідношення (13.3) означають, що правила зупинення **П30**, **П31**, **П32** послідовних наближень (13.2) визначають регуляризуючі алгоритми розв'язку рівнянь (1.1).

Зауваження 13.1.

Позначимо через $x_{\delta,h}$ – нормальний відносно x^0 розв'язок СЛАР

$$A_h^*A_hx=A_h^*b_\delta. \qquad (13.4)$$

Тоді $\lim_{k\to\infty}\left\|x^k-x_{\delta,h}\right\|=0$.

Значить, із збільшенням номера k, x^k все точніше апроксимує розв'язок $x_{\delta,h}$. Але, як було показано раніше (див. приклади 8.2, 8.4), точні розв'язки рівняння (13.4) з наближеними даними можуть неймовірно відрізнятися від розв'язку рівняння (1.1) навіть у випадку однозначної розв'язності (помітимо, що якщо A_h^{-1} існує, то розв'язок $x_{\delta,h}$ СЛАР (13.4) збігається з розв'язком системи $A_hx=b_\delta$). Тому недоцільно виконувати дуже багато ітерацій згідно зі схемою (13.2). Процедури **П3i** (i=0, 1, 2) правильно регулюють при заданих δ, h число ітерацій і, у визначеному сенсі, цей вибір є оптимальним (див. [20]).

Зауваження 13.2.

Твердження 13.1 справедливе і для так званої неявної ітераційної схеми

$$x^k=x^{k-1}-\left(A_h^*A_h+B\right)^{-1}\left(A_h^*A_hx^{k-1}-A_h^*b_\delta\right), \qquad (13.5)$$

де B –додатньо визначена матриця і допускає перестановку з $A_h^*A_h$.

Для одержання необхідної точності, число ітерацій за схемою (13.5), як правило, вимагається менше, ніж за схемою (13.2). Але, зрозуміло, крок ітерування тут буде більш трудоємний, поскільки

треба обертати матрицю $A_h^* A_h + B$. Помітимо також, що в якості матриці B можна брати матрицю $B = \alpha E,\ \alpha > 0$.

Зауваження 13.3.

Якщо виникає потреба знаходження розв'язку системи (1.1), який повинен задовольняти додатковим обмеженням $x \in Q$ (Q – опукла множина), що відображають деякі властивості шуканого розв'язку, то необхідно, замість схем (13.2), (13.5) використати нелінійні ітераційні процеси [16]

$$z^k = P_{r_Q}\left\{\left[E - A_h^* A_h\right] z^{k-1} + A_h^* b_\delta\right\},$$

$$z^k = P_{r_Q}\left\{\left(A_h^* A_h + B\right)^{-1}\left(B z^{k-1} + A_h^* b_\delta\right)\right\},$$

де P_{r_Q} – матрична проекція, яка кожному вектору ставить у відповідність вектор із множини Q, найближчий до x.

Наведемо приклади апріорних множин Q, для яких P_{r_Q} виписується явно:

$$Q_1 = \left\{x \in R^n : x \geq 0\right\},\ P_{r_{Q_1}} x = x^+,\ \text{ де } x_i^+ = \begin{cases} x_i, & \text{якщо } x_i \geq 0 \\ 0, & \text{якщо } x_i < 0 \end{cases};$$

$$Q_2 = \left\{x \in R^n : \left\|x - x^0\right\| \geq r\right\},\ P_{r_{Q_2}} x = x_0 + \frac{x - x^0}{\left\|x - x^0\right\|} r;$$

$$Q_3 = \left\{x \in R^n : a_i \leq x_i \leq b_i\right\},\ P_{r_{Q_3}} x = z,$$

$$\text{де } z_i = \begin{cases} x_i, & \text{якщо } a_i \leq x_i \leq b_i \\ a_i, & \text{якщо } x_i < a_i \\ b_i, & \text{якщо } x_i > b_i \end{cases}.$$

Приклад 13.1.

Застосуємо метод простої ітерації для розв'язку системи $Ax = b$ на прикладі 8.4 з точною матрицею і збуреною

$$A = \begin{bmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{bmatrix},\quad A_h = \begin{bmatrix} 4{,}99 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{bmatrix}$$

для вектора правих частин $b = (23{,}038 \quad 32{,}048 \quad 33{,}048 \quad 31{,}048)^T$.

Як встановлено в кінці §11, точний розв'язок (а саме, розв'язок системи $Ax = b$) є $x = [1{,}28 \quad 0{,}84 \quad 0{,}93 \quad 1{,}04]^T$.

Оскільки матриця A_h симетрична і додатньо визначена, то метод простої ітерації можна використати у формі

$$x^k = (E - A_h)x^{k-1} + b, \quad (k = 1, 2, \ldots). \quad (13.6)$$

Попередньо нормувавши систему (поділивши в кожному рівнянні елементи матриці і праві частини на 1000 так, щоб $\|A_h\| \le 1$, і задавши початковий вектор $x^0 = (0 \quad 0 \quad 0 \quad 0)^T$), виконаємо ітерації за формулою (13.6). Після 42000 ітерацій одержимо наближений розв'язок $x^{42000} = (1{,}21 \quad 0{,}878 \quad 0{,}949 \quad 1{,}03)^T$, який апроксимує точний розв'язок $\tilde{x} = (1{,}28 \quad 0{,}84 \quad 0{,}93 \quad 1{,}04)^T$ краще, ніж розв'язок, одержаний за методом Тихонова $x^\alpha = (1 \quad 1 \quad 1 \quad 1)^T$ при $\alpha = 0{,}048$.

Враховуючи, що $h = 0{,}01$, $\left\|x^k\right\| \approx 2$, $\left\|\left\|Ax^k - b\right\|\right\| \approx 0{,}01$, можна вважати, що зупинка ітераційного процесу відбулася згідно з правилом $П_{32}$ при $b_2 = 1{,}1$.

§14. ІНШІ МЕТОДИ РЕГУЛЯРИЗАЦІЇ

Викладені в §10-13 регуляризуючі алгоритми, зрозуміло, не вичерпують всіх стійких методів розв'язування погано обумовлених СЛАР. У цьому параграфі дамо поверховий огляд деяких інших поглядів на розв'язування погано обумовлених СЛАР.

14.1. Метод квазірозв'язків

Будемо вважати, що апріорі відомо про належність розв'язку системи рівнянь (1.1) опуклій замкненій обмеженій множині $Q \subset R^n$, $\tilde{x} \in Q$. Опуклість множини означає, що якщо дві точки $x_1, x_2 \in Q$, то і відрізок їх з'єднуючий їх, належить Q. Замкненість множини означає, що якщо $x_n \in Q$ і $x_n \to x$, то $x \in Q$, а саме множина містить всі свої граничні точки (вектори).

Будемо вважати, що система (1.1) має на множині Q єдиний розв'язок $\tilde{x}$.

Позначимо через $x^{\delta,h}$ розв'язок задачі

$$min\left\{\left\|A_h x - b_\delta\right\| : x \in Q\right\}, \tag{14.1}$$

де $\|b - b_\delta\| \le \delta$, $\|A - A_h\| \le h$. При наших допущеннях задача на мінімум розв'язувана для будь-яких A_h, b_δ, можливо, не єдиним чином (матриця A_h оберненої може і не мати) і справедливе співвідношення

$$\lim_{\delta,\, h \to 0} \left\|x^{\delta,h} - \tilde{x}\right\| = 0.$$

Спосіб побудови наближених розв'язків за допомогою (14.1) відомий, як метод квазірозв'язків В. К. Іванова [14].

Якщо система (1.1) має на Q множину розв'язків $\tilde{X}$, то можна стверджувати, що сукупність граничних точок послідовностей $\left\{x^{\delta,h}\right\}$ належить $\tilde{x}$, а саме, якщо для послідовності $x^{\delta,h} \to \bar{x}$, то обов'язково $\bar{x} \in \tilde{X}$.

14.2 Метод неув'язки

Вважатимемо, що збурена тільки права частина b, $\|b - b_\delta\| \le \delta$, а матриця A системи (1.1) відома точно $A_h \equiv A$. За наближений розв'язок x^δ приймемо розв'язок задачі на умовний екстремум

$$min\left\{\left\|x - x^0\right\| : \left\|A_h x - b_\delta\right\|^2 \le \delta^2\right\} \tag{14.2}$$

Задача 14.2 завжди має єдиний розв'язок і справедливе співвідношення

$$\lim_{\delta\to 0}\left\|x^{\delta}-\tilde{x}\right\|=0, \qquad (14.3)$$

де $\tilde{x}$ – нормальний відносно x^0 розв'язок системи (1.1), котру вважаємо розв'язуваною.

Якщо відома оцінка для норми розв'язку СЛАР (1.1), $\|\tilde{x}\|\le c$, то метод неув'язки можна узагальнити на випадок, коли матриця A_h теж відома з помилкою $\|A-A_h\|\le h$. Тоді замість (14.2) необхідно розглянути задачу на мінімум

$$min\left\{\left\|x-x^0\right\|:\left\|A_h x-b_\delta\right\|^2\le(ch+\delta)^2\right\}.$$

Для її розв'язку $x^{\delta,h}$ справедливий аналог співвідношення (14.3)

$$\lim_{\delta,h\to 0}\left\|x^{\delta,h}-\tilde{x}\right\|=0.$$

Деталі відносно методів квазірозв'язків і нев'язки можна знайти в монографії [14].

Зауваження 14.1.

Методи неув'язки і квазірозв'язків приводяться до методу Тихонова при деякому виборі параметра регуляризації α (див. [14]).

Зрозуміло, що знаходження векторів $x^{\delta,h}$ в (14.1), (14.2) можна здійснити методами математичного програмування (див., наприклад, [13]).

Зауваження 14.2.

При розв'язуванні вироджених СЛАР (1.1) в деяких випадках необхідно знайти не розв'язок з мінімальною нормою, а розв'язок, що задовольняє визначеним додатковим обмеженням у вигляді лінійних (або опуклих) нерівностей, наприклад,

$$x\in K=\left\{x:\left(h_j,x\right)-l_j\le 0,\ j=1,2,...,m\right\}\ (*)$$

ця задача в рамках методу найменших квадратів досліджувалася в [18]. Для загального випадку опуклих обмежень

$$x\in K=\left\{x:\ g_j\le 0,\ j=1,2,...,m\right\}\ (**)$$

в [16] запропоновано ітераційні процеси для сумісного розв'язку СЛАР (1.1) і нерівність (**) (в частинному випадку, нерівностей (*)).

14.3 Узагальнений метод неув'язки

Якщо інформація про вихідну СЛАР (1.1) відома у вигляді індивідуальної наближеної системи $\left\{\tilde{A};\tilde{b}\right\}$, тоді, як видно з прикладів, при скільки завгодно малих збуреннях $\tilde{A}$, $\tilde{b}$ розв'язок $\tilde{x}$ може терпіти які завгодно великі зміни. Таким чином, наближена

система $\{\widetilde{A};\widetilde{b}\}$ не містить достатньої інформації для побудови стійкого наближеного розв'язку.

А. М. Тихонов запропонував в [24] під наближеними даними системи (1.1) розуміти: індивідуальну наближену систему $\{\widetilde{A};\widetilde{b}\}$, рівень похибки δ, h і сукупність СЛАР, еквівалентних за точністю парі $\{\widetilde{A};\widetilde{b}\}$:

$$\widetilde{\Sigma}(\delta,h)=\{(A,b):\|A-\widetilde{A}\|\le h,\ \|b-\widetilde{b}\|\le\delta\} \qquad (14.4)$$

Вектор x називають *допустимим розв'язком* з $\widetilde{\Sigma}(\delta,h)$., якщо існує пара $(A,b)\in\widetilde{\Sigma}(\delta,h)$ така, що $Ax=b$. Сукупність усіх допустимих розв'язків позначимо

$$\widetilde{Z}(\delta,h)=\{x:\ Ax=b,(A,b)\in\widetilde{\Sigma}(\delta,h)\},$$

або в еквівалентній формі

$$\widetilde{Z}(\delta,h)=\{x:\ \|b-\widetilde{b}\|\le\delta,\ \|A-\widetilde{A}\|\le h\}.$$

Наближені дані $\widetilde{\Sigma}(\delta,h)$. системи (1.1) назвемо *зіставними*, якщо $\widetilde{Z}(\delta,h)\not\equiv\varnothing$. *Вектор* $\widetilde{x}_{\delta,h}$, який реалізує мінімум в задачі

$$min\ \{\|x\|:\ x\in\widetilde{Z}(\delta,h)\}, \qquad (14.5)$$

назвемо *нормальним наближеним розв'язком* системи (1.1).

Прямий розв'язок (14.5) є досить важкою проблемою, оскільки структура множини $\widetilde{Z}(\delta,h)$ може бути складною і наперед нам невідомою.

У роботі [24] було проведено дослідження цієї задачі і запропоновано обчислювальну процедуру для знаходження $\widetilde{x}_{\delta,h}$. Її суть полягає в наступному.

Візьмемо до розгляду функції:

$$\gamma(\delta)=min\{\|x\|:\ \|\widetilde{A}x-\widetilde{b}\|\le\delta\},$$
$$\beta(r)=min\{\|\widetilde{A}x-\widetilde{b}\|:\ \|x\|\le r\}.$$

1 етап: розв'язуємо рівняння

$$\beta(r)-hr=\delta \qquad (14.6)$$

і знаходимо корінь $r=\widetilde{r}$.

2 етап: розв'язуємо задачу на екстремум

$$min\{\|\widetilde{A}x-\widetilde{b}\|:\ \|x\|\le\widetilde{r}\},$$

або, що еквівалентно,

$$min\{\|x\|:\ \|\widetilde{A}x-\widetilde{b}\|\le\delta+h\widetilde{r}\}.$$

Одержаний екстремальний вектор і буде нормальним наближеним розв'язком $\|\widetilde{x}_{\delta,h}\|$.

Наведемо умови розв'язуваності рівняння (14.6).

Нехай $x^* = min\left\{ \left\| \widetilde{A}x - \widetilde{b} \right\| : x \in R^n \right\} = min\left\{ \left\| b - \widetilde{b} \right\| : b = \widetilde{A}x, x \in R^n \right\}$.

Нехай x^* – розв'язок задачі $min\left\{ \left\| \widetilde{A}x - \widetilde{b} \right\| : x \in R^n \right\}$, а $\left\| x^* \right\| = r^*$, тим самим $x^* = \left\| \widetilde{A}x^* - \widetilde{b} \right\|$. Величину x^* природно називати порогом розв'язуваності (або мірою несумісності). Умови розв'язуваності рівняння (14.6) відносно h виглядають так:

a) Якщо $x^* \le hr^*$, то рівняння розв'язуване для всіх δ: $0 < \delta \le \left\| \widetilde{b} \right\|$,

b) Якщо $x^* > hr^*$, тоді рівняння розв'язуване для всіх δ: $x^* - hr^* \le \delta \le \left\| \widetilde{b} \right\|$.

Твердження 14.1.[17]

Нехай $(\overline{A}, \overline{b}) \in \widetilde{\sum}(\delta, h)$, тобто система $\overline{A}x = \overline{b}$ розв'язувана і нехай $\overline{x}$ – її нормальний розв'язок. Тоді

$$\lim_{\delta, h \to 0} \left\| \widetilde{x}_{\delta,h} - \overline{x} \right\| = 0.$$

Крім того, якщо $\widetilde{\sum}(\delta, h)$ – інші наближені дані, породжені індивідуальною наближеною системою $\{\widetilde{A}; \widetilde{b}\}$, а $\widetilde{x}'_{\delta,h}$, побудований за цією системою, – нормальний наближений розв'язок, тоді справедлива оцінка

$$\left\| \widetilde{x}_{\delta,h} - \widetilde{x}'_{\delta,h} \right\| \le \varphi(\delta, h), \ \varphi(\delta, h) \to 0 \text{ при } \delta, h \to 0$$

§15. МЕТОД СИНГУЛЯРНОГО РОЗКЛАДУ

15.1. Основи методу

Зупинимось ще на одному алгоритмі розв'язку довільної системи, який дозволяє знаходити множину всіх розв'язків. Обмежимось описом цього методу на ідейному рівні, звертаючись за деталями до відповідної літератури [31, 33].

Сингулярним розкладом матриці A з дійсними елементами розмірності $m \times n$, називається всякий розклад вигляду

$$A = USV^* \qquad \left(S = U^* AV\right), \tag{15.1}$$

де U – ортогональна (тобто, $UU^* = E$, $U^{-1} = U^*$) $m \times m$ матриця,

V – ортогональна $n \times n$ матриця,

S – діагональна $m \times n$ матриця, у якої $\sigma_{ij} = 0$ при $i \neq j$, а $\sigma_{ii} = \sigma_i \geq 0$.

Величини σ_i називаються *сингулярними числами* матриці A. Відомо, що $\sigma_i = \sqrt{\lambda_i}$, де λ_i – власні числа матриці A^*A або AA^*. Мовою лінійної алгебри матриця A є зображенням деякого лінійного оператора в конкретній координатній системі. Виконуючи одне ортогональне перетворення координат в діяльності визначення оператора, а друге ортогональне перетворення в діяльності значень, перетворюємо зображення в діагональне. Опис алгоритму одержання сингулярного розкладу, на якому тут ми не зупиняємося, можна знайти в [31, 33].

Використовуючи розклад (15.1), перепишемо систему $Ax = b$ у вигляді

$$USV^* x = b, \tag{15.2}$$

або в еквівалентній формі

$$Sz = d, \;\; z = V^* x, \;\; d = U^* b. \tag{15.3}$$

Оскільки матриця S діагональна, ця система легко розв'язується. Перепишемо її в наступному вигляді (вважаючи для визначеності $m \geq n$)

$$\begin{aligned} &\sigma_j z_j = d_j, \text{ якщо } j \leq n, \; \sigma_j \neq 0; \\ &0 \cdot z_j = d_j, \text{ якщо } j \leq n, \; \sigma_j = 0; \\ &0 \cdot z_j = d_j, \text{ якщо } j > n, \end{aligned} \tag{15.3}$$

де 2-га підсистема не має сенсу, якщо n=2 (система повного роду)

3-я підсистема не має сенсу, якщо n=m.

Зауважимо, що вихідне рівняння розв'язуване тоді і тільки тоді, коли $d_j = 0$ щоразу, коли $\sigma_j = 0$ або $j \geq n$. Якщо $\sigma_j = 0$, то

невідомому z_j, яке відповідає нульовому $\sigma_j = 0$, можна надати довільне значення (яке відіграє роль параметра). Розв'язок вихідної системи обчислюється за формулою $x = Vz$.

Нагадаємо, що ядро матриці A є множина K векторів x, для яких Ax=0, а ділянка значень – множина R векторів b, для яких система $Ax = b$ має розв'язок. Сингулярний розклад дозволяє описати множини K і R.

Дійсно, позначимо u_j, v_j стовпчики матриці U, V, тоді (15.1) можна переписати наступним чином

$$Av_j = \sigma_j u_j, \quad j = 1, 2, ..., n \tag{15.5}$$

(для цього достатньо помножити рівність (15.1) справа на матрицю V і врахувати, що $V^* V = E$).

Якщо $\sigma_j = 0$, то із (15.5) маємо $Av_j = 0,$ тобто $v_j \in K$. Якщо $\sigma_j \neq 0$, то $u_j \in R$.

Нехай U_1 – система стовпчиків u_j, для яких $\sigma_j \neq 0$, а U_0 – система інших стовпців. Аналогічно визначаємо V_1, V_0. Тоді

a) V_0 – ортонормований базис ядра K,

b) U_1 – ортонормований базис для ділянки значень R.

Вкажемо на одну суттєву перешкоду, що може виникнути при реалізації методу сингулярного розкладу, коли в результаті обчислень одержані малі сингулярні числа. В цьому випадку при визначенні z_j, за формулою $z_j = \dfrac{d_j}{\sigma_j}$ можемо отримати велику похибку. Тому ключем до вірного використання сингулярного розкладу є введення границі τ, яка відображає точність вихідних даних і машинних обчислень, а саме: всяке $\sigma_i > \tau$ придатне і відповідне $z_j = \dfrac{d_j}{\sigma_j}$, а довільне $\sigma_i < \tau$ слід вважати нульовим і відповідному z_j може бути надано довільне значення (або можна позначити $z_j = 0$, щоб одержати розв'язок з мінімальною нормою). Величина τ відіграє тут роль параметра регуляризації.

15.2 Застосування до задачі найменших квадратів

На підставі сингулярного розкладу можна одержати розв'язок в сенсі найменших квадратів (див. (9.1))

$$min \left\{ \|Ax - b\|^2 : \ x \in R^n \right\}.$$

Оскільки ортогональні матриці зберігають норму і $V^*V = E$, то

$$\|Ax - b\| = \left\|U^*\left(AV^*x\right) - b\right\| = \|Sz - d\|, \quad d = U^*b.$$

Вектор z, який забезпечує мінімум неув'язки, виражається формулою $z_j = \dfrac{d_j}{\sigma_j}$, якщо $\sigma_j \neq 0 \ \ (j = 1, 2, ..., k)$, z_j – довільне, якщо $\sigma_j = 0$, причому, $\|Ax - b\|^2 = \sum\limits_{j=k+1}^{m} d_j^2$. Множина всіх розв'язків в сенсі найменших квадратів обчислюється за формулою $x = Vz$. Якщо $z_j = 0$, коли $\sigma_j = 0$, то x буде збігатися з псевдорозв'язком.

При виникненні малих сингулярних чисел для уникнення накопичення похибки, як і в попередньому пункті, необхідно передбачити операцію „обнулювання" малих сингулярних чисел, яку можна інтерпретувати як спеціальну процедуру регуляризації.

Зауважимо, що чисельному дослідженню методу найменших квадратів присвячено монографію [18].

§16. ПРЯМІ МЕТОДИ РОЗВ'ЯЗКУ РЕГУЛЯРИЗОВАНИХ СИСТЕМ

16.1 Метод квадратного кореня

При використанні схем регуляризації Лаврентьєва-Тихонова наближений розв'язок знаходиться із систем (10.4), (10.8) з симетричною додатньо визначеною, а при розумному виборі α, і гарно обумовленою матрицею A розмірності $n \times n$. Для їх використання, поряд з ітераційними процесами, можна використати традиційні прямі методи типу Гаусса, Жордана і інші (див. [2, 65]). Тут ми опишемо лиш метод квадратного кореня (МКК), який пристосований для рівнянь з симетричними матрицями і відзначаються економічністю та стійкістю.

МКК ґрунтується на розкладі Холеського симетричної матриці

$$A = U^* U, \tag{16.1}$$

де U – верхня трикутна матриця.

Позначимо через u_{ij} коефіцієнти матриці U і, розписуючи рівність (16.1) поелементно, маємо

$$\sum_{k=1}^{n} u_{ki} u_{kj} = a_{ij} \qquad (i = 1, 2, ..., n;\ j = 1, 2, ..., n) \tag{16.2}$$

Розв’язуючи (16.2), одержуємо розрахункові формули в МКК

$$u_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} u_{ki}^2},$$

$$u_{ii} = \frac{a_{ij} - \sum_{k=1}^{i-1} u_{ki} u_{kj}}{u_{ii}},$$

$$(j = i + 1, ..., n;\ i = 1, 2, ..., n).$$

Після одержання розкладу для обчислення розв'язку системи достатньо розв’язати дві трикутні системи:

$$U^* z = b,\ Ux = z \tag{16.3}$$

Особливо зручний МКК для рядкових матриць (див. [6]), бо схема методу дозволяє легко врахувати структуру матриці і виключити операції з нулями.

Проста модифікація МКК, $u_{ij} = 0$ для будь-якого значення і, при якому $u_{ii} = 0$, зберігає обчислювальну процедуру і для випадку тільки додатньо напіввизначеної матриці. Крім того, клітинні аналоги МКК [24, 32] дозволяють ефективно розв’язувати системи високого порядку, притягуючи пам’ять ЕОМ другого рівня.

16.2 Ітераційне уточнення

У результаті застосування якого-небудь методу до розв'язку (вихідної або регуляризованої) системи може виникнути проблема, коли через вплив обчислювальної похибки, замість точного розв'язку $\tilde{x}$, $A\tilde{x} = b$, буде отримано деяке наближення $\breve{x}$. Припустимо, що матриця А – невироджена і спосіб розв'язку такий, що

$$\frac{\|\breve{x} - \tilde{x}\|}{\|x\|} \le q < 1 \qquad (16.4)$$

незалежно від вектора правих частин b. Пере позначимо $\breve{x} = x^k$ і визначимо величини $\tilde{x} - x^k = \Delta^k$, $b - Ax^k = r^k$. Тоді $A\Delta^k = b - Ax^k = r^k$. У результаті розв'язку прийнятним методом системи $A\Delta^k = r^k$, одержимо вектор $\tilde{\Delta}^k$ і, значить, $x^{k+1} = x^k + \tilde{\Delta}^k$. На підставі допущень

$$\left\|x^{k+1} - \tilde{x}\right\| = \left\|x^k + \tilde{\Delta}^k - \tilde{x}\right\| = \frac{\left\|\tilde{\Delta}^k - \Delta^k\right\|}{\left\|\Delta^k\right\|}\left\|\Delta^k\right\| \le q \cdot \left\|x^k - \tilde{x}\right\| \qquad (16.5)$$

Таким чином, процес (16.4)-(16.5)

$$x^{k+1} = x^k + \Delta^k, \quad A\Delta^k = r^k, \quad r^k = b - Ax^k \qquad (16.6)$$

при обумовлених застереженнях збігається із швидкістю геометричної прогресії. Важливо відмітити, що поправки Δ^k слід знаходити з максимальною точністю, для чого треба обчислити r^k з подвійною точністю.

§17. РЕКОМЕНДАЦІЇ ДО ВИБОРУ АЛГОРИТМУ РОЗВ'ЯЗКІВ СЛАР

Для вибору придатного методу розв'язку СЛАР необхідна якісна та кількісна інформація про вхідні дані: обумовленість матриці, точність задання коефіцієнтів, довжина мантиси машинного слова і т. д. Для СЛАР, що виникають в прикладних дослідженнях при описі конкретних фізичних та економічних процесів, найбільш типовою є ситуація, коли апріорі властивості матриці нам невідомі.

Найбільш важливою характеристикою матриці (системи), від якої залежить точність розв'язку, є число обумовленості $\mu(A)$. Оскільки формула $\mu(A)=\|A\|\cdot\left\|A^{-1}\right\|$ – неконструктивна, вкажемо практичні способи одержання наближених апостеріорних оцінок числа обумовленості системи.

Помітимо тільки спочатку, що при використанні різноманітних матричних норм одержимо і різні числа обумовленості. Але це – не принципово, так як при невеликих розмірностях матриці при різних нормах величини $\mu(A)$ мають, як правило, один і той же порядок, так що можна обмежитись обчисленням $\mu(A)$ для якої–небудь матричної норми.

У роботі [32] (див. також [58]) для оцінки $\mu_{\text{II}}(A)=\|A\|_{\text{Л}}\cdot\left\|A^{-1}\right\|_{\text{II}}$ використовується наближена формула

$$\mu_{\text{II}}(A)\approx \max_j \left\|a_j\right\|_{\text{II}} \frac{\|z\|_{\text{II}}}{\|y\|_{\text{II}}},$$

де a_j – стовпці матриці A,

z, y – розв'язки систем

$$R^* y = e, \quad Rz = y$$

при спеціально підібраному векторі e з компонентами ± 1,

R – верхня трикутна матриця, одержана при розкладі $A = QR$,

Q – ортогональна матриця.

У монографії [32] пропонується така наближена оцінка числа обумовленості матриці

$$\mu(A)\approx \frac{1}{\varepsilon}\frac{\left\|\Delta^1\right\|}{\left\|x^1\right\|}\approx \frac{1}{\varepsilon}\cdot 10^{-2},$$

де ε – найбільше число, для якого рівність $1\oplus\varepsilon = 1$ справедливо в обчисленнях на заданій ЕОМ з плаваючою комою, а Δ^1, x^1 знайдені із співвідношень (16.6) та

$$\eta = \left[-\ln \frac{\left\| \Delta^1 \right\|_{\mathrm{I}}}{\left\| x^1 \right\|_{\mathrm{I}}} \right] \qquad \left(\|x\|_{\mathrm{I}} = max\,|x_i| \right).$$

Значок $\oplus$ означає машинне додавання з заокругленням.

Практично ж систему слід розглядати як погано обумовлену, якщо спостерігається не зменшення норм двох послідовних поправок Δ^k або надмірно повільне зменшення в ітераційному процесі (16.6): наприклад, за один крок ітерації уточнюється менше половини десяткового розряду.

В книзі [31] детально викладено алгоритм обчислення числа обумовленості довільної системи з гарантованої точності на основі алгоритму сингулярного розкладу (див. §15), який, ясна річ, є більш працеємним, ніж згадувані вище способи.

Якщо на основі тієї чи іншої апостеріорної оцінки встановлений факт поганої обумовленості системи, то необхідно застосовувати методи регуляризації (див. §§10-15). Якщо для цієї мети використовується схема Лаврентьєва-Тихонова, то регуляризовані системи (10.6), (10.8), що тут виникають, при належному виборі параметра α (§12) можна розв’язати звичайними прямими методами (див. §16) або ітераційними, типу простої ітерації, Гаусса-Зейделя [9,10].

Природно, що допустимі і інші регуляризуючі алгоритми, описані в §§14, 15. Вибір того чи іншого методу звичайно диктується міркуваннями зручності, типом апріорної інформації, наявними технічними і програмними засобами і цілями, котрі при цьому ставляться (точність, економічність і т.д.).

Якщо ж у результаті чисельного аналізу встановлено, що система гарно обумовлена, то її розв'язок здійснюється традиційними методами на базі стандартного програмного забезпечення для розв'язку СЛАР. Для добре обумовлених систем порівняно невисокого порядку практично всі прямі методи (інколи в поєднанні з процедурою ітераційного уточнення) дозволяють знаходити розв’язки з необхідною точністю, яка визначається точністю вихідних даних. Тут не виникає проблеми катастрофічного накопичення обчислювальної похибки, якщо операції на ЕВМ з належною довжиною машинного слова.

Опис різноманітних підходів до побудови алгоритмів і вибору тактики розв’язування СЛАР загального виду можна знайти в [31, 32, 28], а інформацію про якість математичного забезпечення для розв’язування задач лінійної алгебри – в [28, 7, 57].

§18. СТІЙКІ МЕТОДИ РОЗВ'ЯЗУВАННЯ ЗАДАЧ ЛІНІЙНОГО ПРОГРАМУВАННЯ

На жаль, проблема стійкості розв'язку характерна і задачам лінійного програмування (ЗЛП). Часто математичні моделі будуються у вигляді лінійних агрегатів (лінійних залежностей) за допомогою методу найменших квадратів (МНК) на основі спостережень на деякому інтервалі часу. Дані спостережень на різних проміжках інтервалу можуть бути ідентичними – близькі одні до одного, але не тотожно рівні. З цієї причини в обмеженнях ЗЛП можуть бути обмеження, які мало відрізняються одне від одного, але малим відхиленням коефіцієнтів матриці обмежень можуть відповідати великі розбіжності в розв'язках ЗЛП. Фактично в таких задачах ми маємо справу з недовизначеністю розв'язку. Це свідчить про неповноту постановки задачі. Уточнення постановки задачі має суттєве прикладне значення, так як розв'язок ЗЛП лежить в основі застосування математичного апарата до багатьох задач планування.

Наведемо класичну постановку типової задачі оптимального планування.

Нехай x_j – кількість виробів j-го виду, $j = 1, 2, ..., n$; α_j – максимально можлива кількість виробів j-го виду; c_j – сумарна трудоємність для виробів j-го виду по всіх основних видах устаткування (група станків). Тоді скалярний добуток векторів $c = (c_1, c_2, ..., c_n)$ та $x = (x_1, x_2, ..., x_n)$ $L = (c, x)$ характеризує завантаження станків за плану x випуску виробів.

Нехай, далі, b_i – фонди часу i-ої групи станків, $i = 1, 2, ..., m$; a_{ij} – трудоємність для j-го виробу на i-ій групі станків. Тоді $Ax \le b$, $0 \le x \le \alpha$, де α і b вектори $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$, $b = (b_1, b_2, ..., b_m)$; $A = \{a_{ij}\}$ – матриця з елементами a_{ij}.

Задача визначення оптимального плану $\bar{x}$ може складатися зі знаходження такого вектора із множини векторів x (планів) $G = \{x : Ax \le b\}$, для якого завантаження устаткування буде максимальним, то є

$$\max_{x \in G} (L(x) = (c, x)) = L(\bar{x})$$

Функція $L(x)$ називається цільовою функцією задачі.

За даними одного із заводів добре відомим симплекс-методом були розраховані з різною точністю оптимальні квартальні плани, які відповідали деякому набору вхідних даних [1]. Результати

розрахунків наведені в таблиці 1, де римськими цифрами пронумеровані варіанти розв'язків, які відповідають різним вхідним даним, арабськими – компоненти розв'язків (векторів $\bar{x}$). Із таблиці видно, що для порівняно близьких оптимальних значень цільової функції $L(\bar{x})$ (при відхиленнях близько 1%) кількість виробів, які будуть випущені згідно з цими оптимальними планами, по окремих видах виробів коливається в межах декількох сотень.

Таблиця 1.

$\bar{x}$	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**
1	614	590	596	765	638	507	469	446	373	642	383
2	638	634	634	684	644	611	604	548	583	444	581
3	418	424	423	376	412	446	555	479	479	479	479
4						49	49				
5	66			36		105		66		232	238
6		36					60		106		
7								105			
8											
9											
10			56								
11											
12	296	297	294	314	298	293	295	291	255	286	237
13	50	49	50		47	61	59	69	68	68	78
14									54		72
$L(\bar{x})$, **в тис. крб**	2391	2390	2390	2388	2387	2382	2380	2376	2373	2371	2370

Таким чином, сформована задача (ЗЛП) нестійка.

Нестійкі ЗЛП природно також називати некоректно поставленими. Очевидно, що точні розв'язки $\bar{x}$ некоректно поставленої задачі з наближеними вхідними даними не несуть в собі достатньої інформації про розв'язок задачі.

Таким чином, точні розв'язки задач такого виду не служать надійним способом розв'язку задач оптимального планування з наближеними початковими даними.

Така ситуація часто виникає в обчислювальній процедурі при використанні багатьох методів розв'язку задач математичного програмування.

Зауважимо, що при цьому, не зважаючи на великі розбіжності розв'язків $\bar{x}'$ та $\bar{x}''$, значення цільової функції $L(\bar{x}')$ і $L(\bar{x}'')$ можуть

відрізнятись одне від іншого мало. Сказане знаходить своє відображення в результатах, наведених в таблиці 1.

Однак, в ЗЛП з наближеними вхідними даними, як завгодно близькими до точних, мінімальні значення цільової функції можуть суттєво відрізнятися. Різниця може бути як завгодно велика.

Приклад 18.1.

Розглянемо наступну ЗЛП. Нехай треба знайти мінімум функції $L(x)=y$ на частині прямої $y=\lambda_0 x+y_0$, розміщеної в діяльності $\{y\ge 0, x\ge 0\}$. Тут λ_0 і y_0 наперед задані числа і $y_0>0$.

Нехай $\lambda_0=0$, але замість λ_0 ми маємо деяке число λ_δ таке, що $|\lambda_\delta-\lambda_0|\le\delta$, δ – наперед задане мале число. Розглянемо два випадки.

1) $\lambda_\delta>0$. У цьому випадку замість прямої $y=y_0$ (бо $\lambda_0=0$) ми маємо пряму l_1: $y=\lambda_\delta x+y_0$ (див. Рис.1).

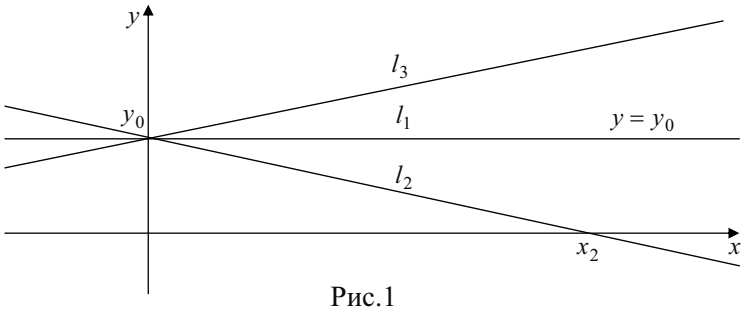

Рис.1

Мінімум функції $L(y)$ на частині прямої l_1 розміщений в діяльності $\{y\ge 0, x\ge 0\}$, досягається в точці $(0, y_0)$, при $x=0$, і рівний y_0.

2) $\lambda_\delta<0$. У цьому випадку замість прямої $y=y_0$ ми маємо пряму l_2: $y=\lambda_\delta x+y_0$ (див. Рис.1). мінімум функції $L(y)$ на частині прямої l_2, розміщеної в діяльності $\{y\ge 0, x\ge 0\}$, досягається в точці $(x_2, 0)$, то є при $x=x_2$, і дорівнює нулю.

При $\delta\to 0$, $\lambda_\delta\to 0$ $x_2\to\infty$. Таким чином, ця задача нестійка. При цьому нестійка як задача мінімізації функції $L(y)$ по аргументу, так і задача мінімізації по значенню функції $L(y)$.

Сформулюємо ЗЛП в наступному вигляді.

Необхідно знайти

$$\min_{x\in G}\left(L(x)=(c,x)\right) \tag{18.1}$$

Якщо область G утворена системою рівностей

$$Ax = y \tag{18.2}$$

$$x \geq 0 \tag{18.3}$$

де $y=(y_1,y_2,...,y_n)$, $c=(c_1,c_2,...,c_n)$, $A=\left\{a_{ij}\right\}_{i,j=1}^{m,n}$.

Якщо умова (18.2) містить в собі лінійно залежні стрічки, то задача (18.1) – (18.3), як правило, некоректно поставлена. Зазвичай робиться припущення, що стрічки наведених умов (18.2) лінійно незалежні. Однак це допущення при наближеному заданні початкових даних практично перевірити неможливо. З цієї причини в майбутньому ми не будемо робити припущень про лінійну незалежність наведених умов і будемо за замовчанням вважати, що задача (18.1) – (18.3) некоректно поставлена. Слід зауважити, що ця задача може мати не єдиний розв'язок. Приклад наведений нижче.

Приклад 18.2.

Нехай $L(x)=x_3$, а умова (18.2) має вигляд

$$x_1 - x_2 = 0$$

Очевидно, що мінімум функції $L(x)=x_3$ на множині $G:\left\{x; x_j \geq 0; j=1,2,3\right\}$ рівний нулю і він досягається в точках напівпрямої $x_1 \geq 0$, що визначається рівняннями

$$x_2 = x_1,\ x_2 = 0.$$

За наявності множини розв'язків для визначеності задачі можна накласти додаткові умови на шуканий розв'язок. Нехай мова йде про задачу оптимального планування. Припустимо, що робота виконується у відповідності до плану $x^{(0)}$ і його треба змінити у зв'язку з тим, що змінилися вхідні (початкові) дані. Новим вхідним даним відповідають інші оптимальні плани. Природно вибрати той із них, що найменше ухиляється від $x^{(0)}$. Подібний критерій вибору пов'язаний з мінімумом затрат на організаційні перебудови, які не були враховані в самій постановці задачі. Мірою відхилення нового плану від старого $\bar{x}^{(0)}$ і $x^{(0)}$ вибирається зважене квадратичне відхилення

$$\left\|\bar{x}^{(0)} - x^{(0)}\right\| = \left\{\sum_{j=1}^{n} P_j\left(\bar{x}_j^{(0)} - x_j^{(0)}\right)^2\right\}^{\frac{1}{2}}$$

або взагалі додатньо визначена квадратична форма.

Маючи це на увазі, введемо наступне **означення**. Нехай дано деякий вектор $x^{(0)} \in R^n$. Вектор $\bar{x}^{(0)}$ будемо називати нормальним розв'язком (18.1) – (18.3) (по відношенню до $x^{(0)}$), якщо

$$\left\|\bar{x}^{(0)} - x^{(0)}\right\| \le \left\|x - x^{(0)}\right\|,$$

де x – довільний розв'язок задачі. Якщо розв'язок задачі (18.1) – (18.3) єдиний, то він, очевидно, збігається з нормальним. Якщо задача має не один розв'язок, то існування нормального розв'язку очевидне, так як множина H, на елементах (векторах) котрої досягається мінімум функції $L(x)$, замкнута, оскільки вона є спільною частиною трьох замкнених множин:

$$G := \{x, Ax = \bar{y}\},\ R_1 \equiv \{x; x_j \ge 0, j = 1, 2, ..., n\},\ S \equiv \{x; L(x) = L_0\}.$$

Легко довести, що нормальний розв'язок завжди єдиний.

Розгляд некоректно поставлених екстремальних задач, а значить і ЗЛП, детальніше буде розглядатися в наступних розділах. Тут ми наведемо лише метод регуляризації ЗЛП [1].

Вхідні дані в задачах лінійного програмування задаються, як правило, наближено. Будемо в подальшому термінологічно розрізняти розв'язувану (точну) і задану (задану наближено до точної) задачі.

Задана задача не дозволяє робити висновки ні про стійкість розв'язуваної задачі, ні про єдиність її розв'язку, навіть якщо задана задача такими властивостями володіє. Точний розв'язок заданої задачі, як видно, неефективний для дослідження розв'язуваної задачі. З точки зору наявної інформації в якості вхідних даних розв'язуваної (точної) задачі може служити будь-який набір вхідних даних з ділянки, що задається нерівностями:

$$\|c_\delta - c\| \le \delta_1,\ \|A_\delta - A\| \le \delta_2,\ \|y_\delta - y\| \le \delta_3 \qquad (18.4)$$

Слід зауважити, що поповнення умов $Ax = y$ лінійно залежними рівняннями робить задачу нестійкою (навіть якщо вона була нестійкою), хоч вона і залишається еквівалентною в класичному розумінні. Для ЗЛП з досить великим числом умов $Ax = y$ ми, як правило, не можемо практично встановити факту їх лінійної залежності. Значить необхідний такий підхід до розв'язку ЗЛП, який не вимагає допущення про лінійну незалежність умов $Ax = y$. Виявилося, що це можливо. ЗЛП є екстремальна задача: знайти мінімум функції $L(x)$, або (за умови $L(x) \ge 0$), що одне і те ж, функції $L^2(x) = (c, x)^2$ на множині:

$$G \equiv \{x; x \in R_1, Ax = y\}.$$

Якщо $L^2(x)$ є для множини G стабілізуючою функцією, то є множина G_α елементів $x \in G$, для яких $L^2(x) \le d$, компактна, тоді існування елемента x_0, реалізуючого мінімум $L(x)$, очевидна. Однак, $L^2(x)$ не завжди є стабілізуючою функцією.

Зупинимося детальніше на визначенні міри похибки задання $L(x)$. Нехай на R^n задані стабілізуючий функціонал $\Omega[x]$ (наприклад, $\Omega[x]=\sum_{i=1}^{n} p_i\left(x_i - x_i^0\right)^2$) і цільові функції $L(x)$ та $\tilde{L}(x)$. Міру відхилення $L^2(x)$ і $\tilde{L}^2(x)$ визначимо, як найменше число, для якого виконується

$$\left|L^2(x) - \tilde{L}^2(x)\right| \le \delta(1+\Omega[x]).$$

Наявність у правій частині доданка, рівного одиниці, необхідна, так як, якщо $\Omega[x_0]=0$, то $L^2(x)$, взагалі кажучи, не рівне нулю.

Візьмемо допоміжну функцію

$$\Phi^2(x) = \tilde{L}(x) + \lambda(1+\Omega[x]),\ \lambda > 0 \qquad (18.5)$$

і будемо наближено розв'язувати, тепер уже задачу квадратичного програмування з цільовою функцією (18.5). Така заміна допустима з точки зору точності задання цільової функції, якщо $0 < \lambda \le \delta$, так як

$$\left|\Phi^2(x) - \tilde{L}^2(x)\right| = \lambda(1+\Omega[x]) \le \delta(1+\Omega[x]).$$

Таким чином, серед векторів x, які належать множині R_1 і таких, що $\|Ax - \tilde{y}\| \le \delta$, треба знайти вектор x_δ, що мінімізує функцію $\Phi^2(x)$. Допоміжна цільова функція $\Phi^2(x)$ є, очевидно, квазімонотонним стабілізуючим функціоналом на множині R_1 (де $\Phi^2(x)$ – квадратична функція).

Одним із часткових випадків ЗЛП є задача оптимального планування. Як правило, в математичній постановці задач оптимального планування при виборі цільової функції $L(x)=(c,x)$ враховуються не всі фактори. До них, наприклад, може відноситися вимога найменшого відхилення шуканого оптимального плану, що відповідає новим (мало відмінними від попередніх) вхідним даним, від попереднього плану (вимога найменших організаційних переобладнань). Введення в новій цільовій функції $\Phi^2(x)$ доданка $\lambda(1+\Omega[x])$ можна розглядати як поправку на вплив не врахованих факторів у цільовій функції $\tilde{L}^2(x)$, а λ – як величину експертної оцінки їх впливу.

Аналогічним чином мотивується побудова наближеного розв'язку задачі оптимального планування з наближеними вхідними даними $(\tilde{A},\tilde{c},\tilde{y})$ як розв'язок задачі на мінімум згладжуючої функції

$$M_\lambda^\alpha\left[x,\tilde{y},\tilde{c},\tilde{A}\right] = \left\|\tilde{A}x - \tilde{y}\right\|^2 + \alpha\left\{L^2(x) + \lambda(1+\Omega[x])\right\},$$

де α визначається за узагальненою неув'язкою з умови

$$\left\|\tilde{A}x - \tilde{y}\right\|^2 = \left\{\delta + h\Phi(x_\alpha)\right\}^2 - \mu,$$

де

$$\mu = \inf_{x \in R_1} \left\| \widetilde{A}x - \widetilde{y} \right\|^2 .$$

Тут h характеризує похибку в заданні оператора $\widetilde{A}$:

$$\left\| A - \widetilde{A} \right\| \le h .$$

Повернемося знову до ЗЛП: знайти елемент $\bar{x}^0$, що мінімізує функцію $L(x) = (c, x)$ на множині

$$R_2 \equiv \{x; Ax = \bar{y}, z \in R_1\},$$

де A – матриця.

Нехай x_0 – елемент, щодо якого шукається нормальний розв'язок. Розглянемо допоміжну задачу I_λ: знайти елемент x_λ, що мінімізує функцію

$$\Phi^2(x) = L^2(x) + \lambda \Omega[x]$$

на множині R_2. Тут $\Omega[x]$ – додатньо визначена квадратична форма (наприклад, $\Omega[x] = \sum_{i=1}^{n} p_i \left(x_i - x_i^0\right)^2$, $p_i \ge 0, \forall i = \overline{1,n}$ або $\Omega[x] = \left\| x - x^0 \right\|^2$).

Існування елементів x_1 очевидне, якщо умови, які визначають R_2, сумісні. Легко переконатися, що елемент x_λ єдиний. І справді, для ЗЛП множина R_2 опукла. Нехай існує два елементи – x_λ^1 та x_λ^2, які мінімізують квадратичну функцію $\Phi^2(x)$. На відрізку прямої

$$x = x_\lambda^1 + \beta\left(x_\lambda^2 - x_\lambda^1\right), \; -\infty < \beta < \infty,$$

який належить R_2, значення функції $\Phi^2(x)$ є квадратичною функцією від β, яка не може мати двох мінімальних значень (парабола). Можна показати [1], що розв'язок x_λ задачі з функцією $\Phi^2(x)$ прямує при $\lambda \to 0$ до нормального розв'язку задачі.

Зауваження. Ми інколи подавали, що $\Omega[x] = \left\| x - x^0 \right\|^2$ або $\Omega[x] = \sum_{i=1}^{n} p_i \left(x_i - x_i^0\right)^2$ і фактично використовували тільки те, що множина елементів з X, для яких $\Omega[x] \le d$, компактна для будь-якого $d > 0$. Усі висновки зберігаються, якщо в якості $\Omega[x]$ брати довільну додатньо визначену квадратичну форму

$$\Omega[x] = \sum_{i,j=1}^{n} P_{ij} \left(x_i - x_i^0\right)\left(x_j - x_j^0\right),$$

видозмінивши, відповідно, визначення нормального розв'язку.

Розв'язок задачі квадратичного програмування, що виникає в результаті регуляризації ЗЛП, можна здійснювати методом Пауелла.

§19. АЛГОРИТМИ ВІДТВОРЕННЯ ФУНКЦІЇ ТА ЧИСЕЛЬНОГО ДИФЕРЕНЦІЮВАННЯ

19.1. Умови коректності задачі обчислення значень необмеженого оператора

19.1.1. Постановка задачі. Задача наближеного зображення функції та її похідних по вхідних даних, які містять помилки (неточності), добре відома будь-якому досліднику, що стикається з прикладними проблемами. При цьому зашумлені дані можуть бути відомі як аналітично, так і дискретно, наприклад у вигляді наближених значень на деякій фіксованій сітці. Будемо вважати, що оператор диференціювання діє на парі лінійних нормованих просторів (ЛНП) $T = \dfrac{d^m}{dt^m}$: $Y \to X$. Далі будемо обмежуватись, як правило, випадком $X = Y = C[a,b]$ або $X = Y = L_2[a,b]$, $D(T) \subset Y$ або їх дискретними аналогами, де $C[a,b]$ – простір неперервних на відрізку $[a,b]$ функцій, $L_2[a,b]$ – простір інтегрованих з квадратом функцій на відрізку $[a,b]$, $D(T)$ – підмножина операторів Y.

Задача 19.1.

Нехай $\|y - y_\delta\|_r \le \delta$, де значок $\|\cdot\|$ означає деяку із норм елемента, який міститься в ньому, $y \in D(T)$, $y_\delta \in Y$. Треба побудувати послідовність x_δ, $x_\delta \in X$ таку, що

$$\lim_{\delta \to 0} \|x_\delta - Ty\|_X = 0, \qquad (19.1.1.)$$

де $Ty = \dfrac{d^m y}{dt^m}$, $x_\delta = Ry_\delta$, $R : Y \to X$.

Задача 19.2.

Нехай задана деяка підмножина $M \subseteq Y$ (як правило, обмежена компактна [14, с.39] і спосіб побудови (оператор R) $x_\delta = R_{y_\delta}$. Треба дати оцінку величини

$$\sup_{y \in M,\, y_\delta \in Y,\, \|y - y_\delta\| \le \delta} \|Ty - Ry_\delta\| = r_\delta(T; R; M) \qquad (19.1.2)$$

найбільшої похибки методу R на класі M, тут $T_y = \dfrac{d^m y}{dt^m}$.

Помітимо, що задачі 1, 2 мають сенс і при $m = 0$ (задача відтворення функції), якщо норма $\|\cdot\|_X$ сильніша від норми $\|\cdot\|_Y$.

Основне затруднення при побудові наближеного зображення похідної виражається в тому, що операція диференціювання $T = \frac{d^m}{dt^m}$ розривна при вибраній парі ЛНП, у чому легко переконатися, розглянувши послідовність

$$y_n(t) = \frac{\sin nt}{\sqrt{n}} \quad \lim_{n\to\infty} \|y_n(t)\|_{C[a,b]} = 0,$$

однак

$$\left\| \frac{d^m y_n(t)}{dt^m} \right\|_{C[a,b]} = n^{\frac{m-1}{2}} \left\| \begin{matrix} \sin nt \\ \cos nt \end{matrix} \right\|_{C[a,b]} \to \infty, \text{ якщо } n \to \infty$$

Звідси, подавши $y_{\delta_n} = y(t) + y_n(t)$, доходимо висновку, що функції y_{δ_n} і $y(t)$ як завгодно близькі за нормою $C[a,b]$, а норма різниці похідної будь-якого порядку як завгодно велика.

З цієї причини за наближений розв'язок не можна брати $x_\delta(t) = \frac{d^m y_\delta(t)}{dt^m}$, навіть якщо $y_\delta(t)$ диференційована потрібне число разів. Крім того, слід пам'ятати, що $y_\delta(t)$ – елемент простору $C[a,b]$ або $L_2[a,b]$, і не зобов'язана мати похідної.

Узагальнюючи задачу чисельного диференціювання, можна говорити про обчислення значень деякого лінійного абстрактного оператора

$$Ty = x, \tag{19.1.3}$$

де елемент y заданий своїм δ-наближенням y_δ, $\|y_\delta - y\| \le \delta$, і для цієї спільної ситуації мають сенс задачі 1,2.

19.1.2. Коректність за Адамаром

Означення 1.1. Задача (19.1.3) називається коректною за Адамаром [14, 35], якщо виконані умови:

1. зона визначення оператора T $D(T) = Y$;
2. T – однозначне відображення (оператор);
3. T – неперервний оператор.

Якщо порушена хоча б одна з умов 1-3, то задача (19.1.3) вважається некоректно поставленою (некоректною).

Вище було показано, що оператор $T = \frac{d^m}{dt^m}$ розривний (необмежений), то у відповідності до означення 1.1 задача диференціювання є типовою некоректно поставленою задачею.

Означення 1.2. Множина відображень $\{R_\delta\}$, $R_\delta : Y \to X$ називається регуляризуючою множиною операторів або регуляризуючим алгоритмом (РА) задачі (19.1.3) в точці y, якщо виконані умови:

1) $\forall \delta > 0 \quad D(R_\delta) = Y$;

2) $\lim\limits_{\delta \to 0} \sup\limits_{y_\delta : \|y - y_\delta\| \le \delta} \|R_\delta(y_\delta) - Ty\| = 0$

Якщо співвідношення 2 виконано для будь-якого $y \in D(t)$, то говорять про РА для задачі (19.1.3). Якщо $\{R_\delta\}$ – РА для (19.1.3), то сукупність $\{x_\delta : x_\delta = R_\delta(y_\delta), 0 < \delta \le \delta_0\}$ називається регуляризованим сімейством наближених розв'язків, а кожний оператор R_δ – регуляризатором задачі.

Таким чином, в задачах 1, 2 мова йде про побудову регуляризованого сімейства розв'язків і оцінку максимальної похибки для цих розв'яків.

19.2. Задача про найкраще наближення необмеженого оператора

19.2.1. Постановка задачі. При дослідженні конкретних методів чисельного диференціювання нам знадобляться деякі факти, які відносяться до загальної задачі обчислення значень необмеженого оператора (19.1.3). Вважатимемо, якщо не обумовлено окремо, що множину m (клас допустимих елементів в задачі 2) можна зобразити у вигляді

$$M = M_z = \{y : \|Ly\| \le z\}, \qquad (19.2.1)$$

де L – деякий лінійний (необмежений) оператор, діючий із $D(L) \subseteq Y \to X$.

Сформулюємо задачу про знаходження найкращого (оптимального) регуляризатора на множині M при заданому рівні похибки δ.

Задача А.

Знайти величину

$$\inf_{R \in [Y \to X]} \sup_{y \in M,\, y_\delta \in Y,\, \|y - y_\delta\| \le \delta} \|Ty - Ry_\delta\| = \inf_{R \in [Y \to X]} (T; R; M) = \Omega_\delta(T; x) \quad (19.2.2)$$

і екстремальний оператор R_δ^*, на якому реалізується нижня грань в (19.2.2); тут $[Y \to X]$ – множина лінійних обмежених операторів, діючих з Y в X .

Таким чином, це – задача пошуку оптимального лінійного РА і обчислення максимальної похибки, яку допускає РА на множині M при заданій можливій похибці δ елемента y.

Задача А тісно пов'язана з наступною задачею про найкраще наближення необмеженого оператора обмеженими [20,36].

Задача В.

Знайти величину

$$\inf_{\|R\le N\|}\sup_{y\in M}\left\|Ty-R_y\right\|\le E_N(T,z) \qquad (19.2.3)$$

і визначити екстремальний оператор R_N^*, для якого досягається нижня грань в (19.2.3).

19.2.2. Оцінка похибки знизу. Введемо ще одну функцію

$$\Phi_\tau(T,z)=\sup\left\{\left\|Ty\right\|: y\in M, \left\|y\right\|\le\tau\right\}, \qquad (19.2.4)$$

яку будемо називати **модулем неперервності оператора *T*** в нулі на множині *M*, і оцінимо величини $\Omega_\delta(T,z)$, $E_N(T,z)$ знизу через цю функцію [2, 36].

Лема 2.1. Справедлива нерівність

$$\Omega_\delta(T,z)\ge\Phi_\delta(T,z) \qquad (19.2.5)$$

Доведення. Враховуючи, що нульовий елемент θ належить множині *M*, заданій у формі (19.2.1), а оператори лінійні (однорідні та адитивні), то при $y_\delta=\theta$ одержимо

$$\Omega_\delta(T,z)=\inf_{s\in[Y\to X]}\sup_{y\in M,\,y_\delta\in Y}\left\|Ty-R_{y_\delta}\right\|\ge\sup_{y\in M,\,\|y\|\le\delta}\left\|Ty\right\|=\Phi_{\delta(T,z)} \qquad (19.2.6)$$

Можна одержати оцінку типу (19.2.6) і в суттєво більш загальній ситуації [34, стор.159]

Лема 2.1. Нехай T – довільний (нелінійний оператор), M – довільна непуста множина із $D(T)$, $\{Y\to X\}$ – множина віх відображень із Y в X, тоді

$$\Omega'_\delta(T;M)=\inf_{R\in\{Y\to X\}}\sup_{y\in M,\,y_\delta\in Y,\,\|y-y_\delta\|\le\delta}\left\|Ty-R_{y_\delta}\right\|\ge\frac{1}{2}\Phi'_\delta(T,M), \qquad (19.2.7)$$

де $\Phi'_\delta(T;M)=\sup\left\{\left\|Ty_1-Ty_2\right\|: y_1,y_2\in M, \left\|y_1-y_2\right\|\le\tau\right\}$.

Зауваження. Для лінійного оператора T і множини M, зображеного в формі (19.2.1), оцінки знизу для $\Omega_\delta(T,z)$ і $\Omega'_\delta(T,M_\tau)$, в (19.2.6), (19.2.7) співпадають.

Лема 2.2. Справедливі співвідношення

$$E_N(T;z)\ge\sup_{\tau\ge0}\left\{\Phi_\tau(T;z)-N_\tau\right\}$$

$$\Phi_\tau(T;z) \le \inf_{N\ge 0}\left\{E_M(T;z)+N_\tau\right\}$$

Доведення випливає із наступних нерівностей

$$E_N(T;z)=\inf_{\|R\|\le N}\sup_{y\in M}\left\|Ty-R_y\right\| \ge \inf_{\|R\|\le N}\sup_{y\in M}\left\{\|Ty\|-\left\|R_y\right\|\right\} \ge \sup_{y\in M}\left\{\|Ty\|-N\|y\|\right\} \ge$$
$$\ge \Phi_\tau(T;z)-N_\tau$$

19.2.3. Зв'язок задач A і B. Тепер встановимо взаємозв'язок між екстремальними операторами в задачах (19.2.2), (19.2.3) при деякому узгодженні параметрів δ і N.

<u>Теорема 2.1.</u> Якщо R_δ^* – екстремальний оператор в задачі A, задача B розв'язна при $N=N_\delta=\left\|R_\delta^*\right\|$ і виконано співвідношення

$$\Omega_\delta(T;z)=\sup_{y\in M}\left\|Ty-R_\delta^* y\right\|+\left\|R_\delta^*\right\|\delta,$$

тоді R_δ^* – екстремальний оператор в задачі B при $N=N_\delta$.

Навпаки, якщо R_N^* – розв'язок задачі B, а параметр $\delta=\delta(N)$ задовольняє співвідношенню

$$\Phi_\delta(T;z)=E_N(T;z)+N_\delta,$$

тоді R_N^* буде екстремальним оператором в задачі A при $\delta=\delta(N)$ і $\Omega_{\delta(n)}(T,z)=E_N(T;\delta)+N\delta$.

<u>Доведення.</u> Позначимо через R_N^* екстремальний оператор задачі B при $N=\left\|R_\delta^*\right\|$. Тоді маємо наступні нерівності:

$$\Omega_\delta(T;z)=\sup_{y\in M,\,y_\delta\in Y,\,\|y-y_\delta\|\le\delta}\left\|Ty-R_\delta^* y_\delta\right\|=\inf_{R\in[Y\to X]}\sup_{y\in M,\,y_\delta\in Y,\,\|y-y_\delta\|\le\delta}\left\|Ty-R_{y_\delta}\right\|\le$$
$$\le \sup_{y\in M,\,y_\delta\in Y,\,\|y-y_\delta\|\le\delta}\left\|Ty-R_N^* y_\delta\right\|\le \sup_{y\in M}\left\|Ty-R_N^* y\right\|+\left\|R_N^*\right\|\delta\le$$

$$\le \sup_{y\in M}\left\|Ty-R_\delta^* y\right\|+\left\|R_N^*\right\|\delta\le \sup_{y\in M}\left\|Ty-R_\delta^* y\right\|+\left\|R_\delta^*\right\|\delta.$$

Оскільки $\left\|R_\delta^*\right\|\ge\left\|R_N^*\right\|$, то на підставі умов теореми

$$\sup_{y\in M}\left\|Ty-R_\delta^* y\right\|\le \sup_{y\in M}\left\|Ty-R_N^* y\right\|\le \inf_{\|R\|\le\left\|R_\delta^*\right\|}\left\|Ty-R_y\right\|,$$

то R_δ^* – екстремальний оператор в задачі B при $N=\left\|R_\delta^*\right\|$.

Обернене твердження з урахуванням (19.2.5) одержується з наступних очевидних нерівностей

$$\Phi_\delta(T;z)\le\Omega_\delta(T;z)=\inf_{R\in[T\to X]}\sup_{y\in M,\,y_\delta\in Y,\,\|y-y_\delta\|\le\delta}\left\|Ty-R_{y_\delta}\right\|\le$$

$$\sup_{y\in M,\, y_\delta\in Y,\, \|y-y_\delta\|\le\delta}\left\|Ty - R_N^* y_\delta\right\| \le \sup_{y\in M}\left\|Ty - R_N^* y\right\| + \sup_{\|y-y_\delta\|\le\delta}\left\|R_N^*(y - y_\delta\right\| \le$$
$$\le E_N(T;z) + N\delta$$

19.3. Оптимальна кінцево-різницева регуляризація в просторі $C(-\infty,\infty)$

Нехай $C(-\infty,\infty)$ – простір неперервних на дійсній прямій функцій $y(t)$ з нормою $\|y\| = \sup\limits_{-\infty<t<\infty}|y(t)|$. Будемо вважати, що функція $y(t)$, похідну якої треба знайти, має обмежену другу похідну, тобто належить множині

$$M = \left\{ y : \frac{d^2y}{dt^2} \in C(-\infty,\infty), \left\|\frac{d^2y}{dt^2}\right\|_C \le r \right\} \qquad (19.3.1)$$

Будемо розв'язувати задачі 1, 2 з п.1 для $X = Y = C(-\infty,\infty)$, $m = 1$ для множини *M*, визначеної в (19.3.1).

Визначимо кінцево-різницевий оператор:

$$\Delta_h y = \frac{y(t+h) - y(t-h)}{2h}.$$

Лема 3.1. Правильна оцінка

$$r_\delta\left(\frac{d}{dt}; \Delta h; M\right) = \sup_{y\in M,\, \|y-y_\delta\|\le\delta}\left\|\frac{dy}{dt} - \Delta_h y_\delta\right\|_C \le \frac{rh}{2} + \frac{\delta}{h} \qquad (19.3.2)$$

Доведення. Розглянемо розклад Тейлора

$$y(t+h) = y(t) + y'(t)h + y''(\widetilde{t})\frac{h^2}{2},$$
$$y(t-h) = y(t) - y'(t)h + y''(\widetilde{\widetilde{t}})\frac{h^2}{2}$$

Віднімаючи із першого співвідношення друге і враховуючи (19.3.2) множини *M*, одержимо

$$\left| y'(t) - \frac{y(t+h) - y(t-h)}{2h} \right| \le \frac{h}{4}\left(\left|y''(\widetilde{t})\right| + \left|y''(\widetilde{\widetilde{t}})\right| \right),$$

звідки

$$\left\|\frac{dy}{dt} - \Delta_h y(t)\right\|_C \le \frac{rh}{2} \qquad (19.3.3)$$

Очевидно, що

$$\left\|\Delta_h y - \Delta_h y_\delta\right\|_C = \sup_{-\infty<t<\infty}\left|\frac{\left[y(t+h) - y_\delta(t+h)\right] - \left[y(t-h) - y_\delta(t-h)\right]}{2h}\right| \qquad (19.3.4)$$

Із нерівності трикутника для норми оцінок (19.3.3), (19.3.4) одержуємо (19.3.2).

Лема 3.2. Для модуля неперервності оператора $\frac{d}{dt}$ на множині M (19.3.1) справедлива рівність

$$\Phi_\delta\left(\frac{d}{dt};r\right)=\sqrt{2r\delta}\,. \tag{19.3.5}$$

Доведення. Згідно з нерівністю Ландау-Адамара-Колмогорова [34]

$$\left\|\frac{dy}{dt}\right\|_C \le \sqrt{2}\|y\|_C^{\frac{1}{2}}\left\|\frac{d^2y}{dt^2}\right\|_C^{\frac{1}{2}} \tag{19.3.6}$$

і для будь-якої трійки чисел m_0, m_1, m_2, яка задовольняє умові $m_1=\sqrt{2m_0m_2}$, знайдеться така функція $\varphi(t)\in C(-\infty,\infty)$, що

$$\left\|\frac{d^i m}{dt}\right\|_C = m_i,\quad (i=0,1,2)$$

і, значить, в (19.3.5) реалізується рівність. Звідси випливає (19.3.5).

Теорема 3.1. Оператор $\Delta_{h(\delta)}: y(t)\to\frac{[y(t+h)-y(t-h)]}{2h}$ при співвідношенні $h=\sqrt{2\frac{\delta}{r}}$ є оптимальним регуляризатором в задачі диференціювання на множині (19.3.1), тобто реалізує нижню грань в задачі (19.2.2) при $T=\frac{d}{dt}$, $\Omega_\delta\left(\frac{d}{dt};r\right)=\sqrt{2\delta r}$.

Доведення. Застосовуючи оцінки (19.3.3), (19.3.4), маємо

$$\left\|\frac{dy}{dt}-\Delta_h y_\delta(t)\right\|_C \le \left\|\frac{dy}{dt}-\Delta_h y(t)\right\|_C + \|\Delta_h y(t)-\Delta_h y_\delta(t)\|_C \le \frac{rh}{2}+\frac{\delta}{h}=\beta(h,\delta)$$

Із умови мінімуму правої частини, $\beta_h'(h,\delta)=0$, одержуємо залежність $h=\sqrt{\frac{2\delta}{h}}$, а, значить, $\beta(h,\delta)=\sqrt{2\delta h}$. З цієї причини

$$\Omega_\delta\left(\frac{d}{dt};r\right)\le \sup_{y\in M,\,y_\delta\in C,\,\|y-y_\delta\|\le\delta}\left\|\frac{dy}{dt}-\Delta_{h(\delta)}y_\delta(t)\right\|_C \le \sqrt{2\delta r}\,.$$

З іншого боку, згідно з (19.2.5), (19.3.5)

$$\Omega_\delta\left(\frac{d}{dt};r\right)\ge\Phi_\delta\left(\frac{d}{dt};r\right)=\sqrt{2\delta r}$$

Звідки й випливає доведення теореми.

За такою ж схемою досліджується задача А для $T=\frac{d^2}{dt^2}$ і множини

$$M = \left\{ y : \frac{d^3 y}{dt^3} \in C(-\infty, \infty), \left\| \frac{d^3 y}{dt^3} \right\|_C \le r \right\} \tag{19.3.7}$$

Для апроксимізації $T = \frac{d^2}{dt^2}$ використовується оператор другої різниці

$$\overline{\Delta}_h y = \frac{y(t+h) - 2y(t) + y(t-h)}{h^2}.$$

При співвідношенні $h = 2\sqrt[3]{\frac{3\delta}{r}}$ є оптимальним регуляризатором в задачі диференціювання $T = \frac{d^2}{dt^2}$ на множині (19.3.7), тобто реалізує нижню грань в задачі (19.2.2), причому $\Omega_\delta\left(\frac{d^2}{dt^2}; r\right) = \sqrt[3]{3}\sqrt[3]{r^2}\sqrt[3]{r}$.

Доведення. Додаючи два розклади

$$y(t+h) = y(t) + y'(t)h + y''(t)\frac{h^2}{2} + y'''(\bar{t})\frac{h^3}{3!}$$

$$y(t-h) = y(t) - y'(t)h + y''(t)\frac{h^2}{2} - y'''(\bar{\bar{t}})\frac{h^3}{3!}$$

знаходимо оцінки

$$\left\| \frac{d^2 y}{dt^2} - \overline{\Delta}_h(t) \right\|_C \le \frac{rh}{3} \tag{19.3.9}$$

Норма оператора $\overline{\Delta}_h : C \to C$ легко оцінюється

$$\left\| \overline{\Delta}_h \right\| \le \frac{4}{h^2}, \left\| \overline{\Delta}_h y - \overline{\Delta}_h y_\delta \right\|_C \le \frac{4\delta}{h^2}. \tag{19.3.10}$$

Із співвідношення (19.3.9), (19.3.10) випливає:

$$\Omega_\delta\left(\frac{d^2}{dt^2}; r\right) \le \min_h \left\{ \frac{2h}{3} + \frac{48}{h^2} \right\} = \sqrt[3]{3r^2\delta}.$$

Разом з нерівністю Колмогорова

$$\left\| \frac{d^3 y}{dt^3} \right\|_C \le \sqrt[3]{3\|y\|} \left\| \frac{d^3 y}{dt^3} \right\|$$

і нерівністю (19.2.5) завершуємо доведення теореми.

§20. ІНТЕРПОЛЯЦІЙНІ СПЛАЙНИ

Обмежимося розглядом кубічних сплайнів, як найбільш поширених в застосуванні. На відрізку $[a,b]$, $a<b$ задамо сітку:

$$\Delta: a=t_0<t_1<\ldots<t_n=b \tag{20.1}$$

Нехай P_m – множина поліномів ступеня не вище $m\geq 0$ і $C^k=C^k[a,b]$ – множина неперервних на $[a,b]$ функцій, які мають неперервну k-ту похідну.

Визначення 20.1. Функцію $S_3(t)=S_3(t,y)=S_3^{\Delta}(t,y)$ називають інтерполяційним сплайном відносно сітки (20.1) для функції $y(t)$, якщо:

1) $S_3(t)\in P_3$, $t\in(t_{j-1},t_j)$, $j=1,2,\ldots,N$;

2) $S_3(t)\in C^2[a,b]$;

3) $S_3(t)y(t_j)$, $j=1,2,\ldots,N$.

Введемо позначення $M_j=S_3''(t_j)$, $y(t_j)=y_j$, $h_j=t_j-t_{j-1}$ і знайдемо зображення сплайна $S_3(t)$ на відрізку $[t_{j-1},t_j]$.

На підставі лінійності другої похідної на $[t_{j-1},t_j]$

$$S_3''(t)=M_{j-1}\frac{t_j-t}{h_j}+M_j\frac{t-t_{j-1}}{h_j}. \tag{20.2}$$

Інтегруючи обидві частини рівності (20.2)

$$S_3'(t)=-M_{j-1}\frac{(t_j-t)^2}{2h_j}+M_j\frac{(t-t_{j-1})^2}{2h_j}+c_1$$

$$S_3(t)=M_{j-1}\frac{(t_j-t)^3}{6h_j}+M_j\frac{(t-t_{j-1})^3}{6h_j}+c_1t+c_2$$

і визначаючи константи c_1, c_2 із умови інтерполяції 3)

$$y_{j-1}=S_3(t_{j-1})=M_{j-1}\frac{h_j^2}{6}+c_1t_{j-1}+c_2,$$

$$y_j=S_3(t_j)=M_j\frac{h_j}{6}+c_1t_j+c_2,$$

одержуємо

$$S_3(t)=M_{j-1}\frac{(t_j-t)^3}{6h_j}+M_j\frac{(t-t_{j-1})^3}{6h_j}+$$

$$+\left(y_{j-1}-\frac{M_{j-1}h_j^2}{6}\right)\frac{t_j-t}{h_j}+\left(y_j-\frac{M_jh_j^2}{6}\right)\frac{t-t_{j-1}}{h_j},$$

$$S_3'(t) = -M_{j-1}\frac{(t_j - t)^2}{2h_j} + M_j\frac{(t - t_{j-1})^2}{2h_j} + \frac{y_j - y_{j-1}}{h_j} - \frac{M_j - M_{j-1}}{6}h_j, \quad (20.3)$$

із (20.3) знаходимо односторонні границі

$$S_3'(t_j - 0) = M_{j-1}\frac{h_j}{6} + M_j\frac{h_j}{3} + \frac{y_j - y_{j-1}}{h_j}, \quad (20.4)$$

$$S_3'(t_j + 0) = -M_j\frac{h_{j+1}}{3} - M_{j+1}\frac{h_{j+1}}{6} + \frac{y_{j+1} - y_j}{h_{j+1}}$$

звідки з урахування неперервності $S_3'(t)$ в точках t_j $S_3'(t_j - 0) = S_3'(t_j + 0)$ одержуємо систему

$$\frac{h_j}{6}M_{j-1} + \frac{h_j + h_{j+1}}{3}M_j + \frac{h_{j+1}}{6}M_{j+1} = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j} \quad (20.5)$$

$$\forall j = \overline{1, N-1}$$

$N-1$ рівнянь з $N+1$ невідомими $\{m_j\}$, $j = 1, 2, ..., N-1$.

Щоб замкнути систему, необхідно поповнити її крайовими умовами. Розглянемо деякі з них:

$$S_3''(a) = M_0 = a_0,\ S_3''(b) = M_N = a_N, \quad (20.6)$$

де a_0, a_N – невідомі числа, зокрема, $a_0 = a_N = 0$;

$$S_3'(a) = a_0'\ \ S_3'(b) = a_N' \quad (20.7)$$

Тоді із (20.4) одержуємо два додаткових рівняння:

$$2M_0 + M_1 = \frac{6}{h_1}\left(\frac{y_1 - y_0}{h_1} - a_0'\right),\ M_{N-1} + 2M_N = \frac{6}{h_N}\left(a_N' - \frac{y_N - y_{N-1}}{h_N}\right)$$

Можна також використати й інші співвідношення, наприклад, ті, що випливають із умови перпендикулярності [37, 38].

Таким чином, побудова кубічного сплайна звелася до розв'язку системи (20.5) з трьох діагональною матрицею, яка успішно може бути розв'язана методом прогонки за $8N$ арифметичних операцій.

Інтерполяційні сплайни широко використовуються при здійсненні чисельного диференціювання.

У відповідності до постановки задачі чисельного диференціювання в задачах I, II замість функції $y(t)$ нам відомо тільки її δ-наближення $y_\delta(t)$: $\|y_\delta(t) - y(t)\|_c \le \delta$. З цієї причини інтерполяційний сплайн будується по значеннях $\tilde{y}_j = y_\delta(t_j)$, то є $S_3(t, \tilde{y})$. Як і у випадку кінцево-різницевих процедур, для одержання РА на ґрунті сплайнів необхідно зв'язувати крок сітки з рівнем похибки.

Це питання розв'язує наступна теорема.

Теорема 20.1. Нехай сітка (20.1) рівномірна і $t_j - t_{j-1} = h$ ($j = \overline{1, N}$). Нехай $y \in C^4[a,b]$, $\|y - \widetilde{y}\|_c \le \delta$. Тоді справедлива оцінка

$$\left\| y^{(i)}(t) - S_3^{(i)}(t, \widetilde{y}) \right\|_c \le c\delta^{\frac{4-i}{4}} \tag{20.8}$$

при $h = c\sqrt[4]{\delta}$, де $i = 0,1,2,3, \varepsilon$ – порядок обчислювальної похідної.

Доведення. У відповідності до результатів [38] (§3, Розділ III)

$$\left\| y^i(t) - S_3^{(i)}(t, \widetilde{y}) \right\|_c \le \left\| y^{(i)} - S_3^{(i)}(t, y) \right\|_c +$$
$$+ \left\| S_3^{(i)}(t, y) - S_3^{(i)}(t, \widetilde{y}) \right\|_c \le c_2 h^{4-i} + c_3 \delta h^{-i}$$

Мінімізувавши по h праву частину нерівності, знаходимо найкращий за порядком крок сітки $h = c\sqrt[4]{\delta}$ і найкращу за порядком оцінку похибки (20.8) для даної мажорантної оцінки.

Нарешті, наведемо формулювання теореми про властивість мінімальності кубічного сплайна [37, 38].

Теорема 20.2. Єдиним розв'язком задачі на мінімум

$$\min\left\{ \int_a^b [y''(t)]^2 \, dt : y \in W_2^2[a;b], y(t_i) = y_i \right\} \tag{20.9}$$

є інтерполяційний кубічний сплайн $S_3(t, y)$ з граничними умовами $S_3''(a, y) = S''(b, y) = 0$.

Зауваження 20.2. Якщо замість $W_2^2[a;b]$ в задачі (20.9) розглянути класи

$$\widetilde{W}_2^2[a;b] = \{ \text{періодичні (з періодом } b-a) \text{ функції із } W_2^2 \}$$
$$\widetilde{W}_2^2[a;b] = \{ y \in W_2^2[a;b] : y'(a) = a_0', y'(b) = a_N' \},$$

то мінімум реалізується на сплайні із того ж класу, тобто з тими ж граничними умовами.

§21. ЗГЛАДЖУЮЧІ І АПРОКСИМУЮЧІ СПЛАЙНИ

Будемо вважати, що наближені вихідні дані про функцію $\{\widetilde{y}_i\}$ задовольняють умовам апроксимації

$$|y(t_i) - \widetilde{y}_i| < \delta_i \quad (i = \overline{1, N}) \tag{21.1}$$

Якщо похибки δ_i відносно великі, то ця обставина погано впливає на поведінку інтерполяційного сплайна і особливо його похідних. Спостерігається сильна осциляція сплайна, зумовлена великими розбіжностями вихідних даних $\widetilde{y}_i$, по яких будується інтерполяційний сплайн.

Тому природно побудувати сплайн, графік якого проходить поблизу заданих значень $\{\widetilde{y}_i\}$, але більш „гладкий", ніж інтерполяційний. Такий сплайн називається згладжуючим (див. Рис. 21.1)

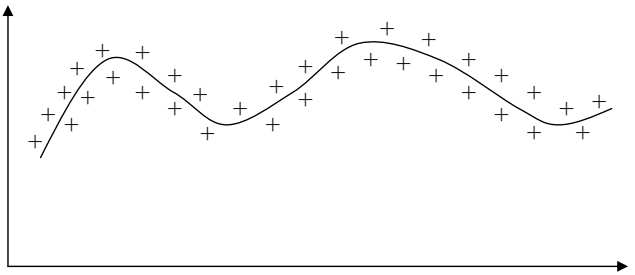

Рис. 21.1

Математично згладжуючий сплайн знаходиться із розв'язку задачі на мінімум [39, 40]:

$$\min\left\{\int_a^b |y''(t)|^2\,dt + \sum_{i=0}^{N}\left(\frac{y_i - \widetilde{y}_i}{p_i}\right)^2 : y \in W_2^2\right\}, \tag{21.2}$$

де $p_i, \widetilde{y}_i$ – задані наперед величини.

Зрозуміло, що чим менше p_i, тим ближче підходить функція, яка мінімізує функціонал в (21.2), до заданих значень $\widetilde{y}_i$. При збільшенні p_i буде спостерігатись у розв'язку в більшій мірі ефект згладжування.

Теорема 21.1. Єдиним розв'язком задачі (21.2) є кубічний сплайн $S(t)$, який задовольняє умові:

$$S''(a) = S''(b) = 0 \tag{21.3}$$

Доведення. Позначимо цільовий функціонал в задачі (21.2) через $I(y)$. Нехай $\varphi(t) \in W_2^2[a,b]$ мінімізує $J(y)$. Покажемо, що $\varphi(t)$ – кубічний сплайн з крайовими умовами (21.3). Дійсно, нехай ця обставина не має місця. Візьмемо інтерполяційний кубічний сплайн $S_3(t;\varphi(t))$ з граничними умовами (21.3). Другий доданок у виразі $J(y)$ однаковий для $\varphi(t)$ і $S_3(t;\varphi)$. Згідно теоремою (20.1):

$$\int_a^b [S_3''(t);\varphi]^2\,dt < \int_a^b [\varphi''(t)]^2\,dt$$

Значить, $J[S_3(t,\varphi)] < J[\varphi]$, що суперечить тому, що $\varphi(t)$ мінімізує $J(y)$.

Вияснимо тепер необхідну умову мінімуму. Нехай $S(t)$ – кубічний сплайн, що мінімізує $J(y)$. Подамо $\widetilde{S}(t) = S(t) + \alpha F_k(t)$, де $F_k(t)$ $(0 \le k \le N)$ – фундаментальний кубічний сплайн, що задовольняє умовам:

$$F_k(t_i) = \delta_{ki} = \begin{cases} 1, & k = i \\ 0, & k \ne i \end{cases}, \quad i = \overline{0,N}; F_k''(a) = F_k''(b) = 0$$

Тоді маємо

$$J[\widetilde{S}] - J[S] = \alpha^2 a_k + 2\alpha b_k,$$

де $a_k = \int_a^b [S''(t)]^2\,dt + \frac{1}{\rho_k^2}$, $b_k = \int_a^b F_k''(t)S''(t)dt + \frac{S_k - \widetilde{y}_k}{\rho_k^2}$.

Тут $a_k > 0$. Покажемо, що $b_k = 0$. І справді, якщо припустити, що $b_k \ne 0$, то вибираючи α так, щоб $|\alpha| < 2|b_k|a_k^{-1}$, $sign\,\alpha = -sign\,b_k$, одержуємо $J[\widetilde{S}] - J[S] < 0$. Це суперечить тому, що сплайн $S(t)$ мінімізує $J(y)$. Тоді

$$b_k = 0,\ k = \overline{0,N} \tag{21.4}$$

є необхідні умови мінімуму.

Перетворимо інтеграл, що входить в b_k, наступним чином:

$$\int_a^b F_k''(t)S''(t)dt = \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} S''(t)F_k''(t)dt =$$

$$= \sum_{i=1}^{N-1} \left[F_k'(t)S''(t)\Big|_{t_i}^{t_{i+1}} - \int_{t_i}^{t_{i+1}} F_k'(t)S'''(t)dt \right] =$$

$$= \sum_{i=0}^{N-1} \left(S'''(t_{i+0})[F_k(t_{i+1}) - F_k(t_i)] \right) =$$

$$= S'''(t_0 + 0)F_k(t_0) + \sum_{i=1}^{N-1} F_k(t_i)[S'''(t_i + 0) - S'''(t_{i-1} + 0)] - S'''(t_{N-1} + 0)F_k(t_N) = D_k$$

Із властивості функції $F_k(t)$ випливають формули:

$$D_k = \begin{cases} S'''(t_0+0), & k=0 \\ S'''(t_k+0) - S'''(t_k-0), & k=\overline{1,N-1} \\ -S'''(t_N-0), & k=N \end{cases}$$

Таким чином, необхідні умови (21.4) набувають вигляду:

$$\begin{cases} S'''(t_0+0) = \dfrac{\widetilde{y}_0 - S(t_0)}{\rho_0^2} \\ S'''(t_k+0) - S'''(t_k-0) = \dfrac{\widetilde{y}_k - S(t_k)}{\rho_k^2}, \quad k=\overline{1,N-1} \\ -S'''(t_N-0) = \dfrac{\widetilde{y}_N - S(t_N)}{\rho_N^2} \end{cases} \tag{21.5}$$

Вище було встановлено, що $S(t)$ – кубічний сплайн з граничними умовами (21.3), тоді $S \in C^2$ і тому виконані співвідношення (теж необхідні умови):

$$S^{(j)}(t_i+0) - S^{(j)}(t_i-0) = 0, \ j=0,1,2, \ i=\overline{0,N} \tag{21.6}$$

Покажемо далі, що співвідношення (21.5), (21.3) є достатніми умовами мінімуму.

Нехай існує задовольняючий їм сплайн $S(t)$. Для будь-якої функції $y(t) \in W_2^2[a,b]$ справедлива тотожність

$$\widetilde{J}[y-S] = J[y] - J[S] - 2\left[I + \sum_{i=0}^{N} \frac{(y_i - s_i)(s_i - \widetilde{y}_i)}{\rho_i^2} \right] \tag{21.7}$$

де $\widetilde{J}[y-S] = \int_a^b [y''(t) - S''(t)]^2 \, dt + \sum_{i=0}^{M} \dfrac{(y_i - s_i)^2}{\rho_i^2}$,

$$I = (y_0 - S_0)S'''(t_0+0) + \sum_{i=1}^{N-1} (y_i - S_i)\left[S'''(t_i+0) - S'''(t_{i-1}+0)\right] -$$

$$-(y_N - S_N)S'''(t_{N-1}+0)$$

Отже, видно, що при виконанні умов (21.5) вирази в квадратних дужках перетворюються в нуль. Значить

$$J[y] = J[S] + \widetilde{J}[y-S]$$

Оскільки $\widetilde{J}[y-S] \geq 0$, то сплайн S забезпечує мінімум функціоналу $J[y]$.

Для завершення доведення необхідно показати, що сплайн, який задовольняє умовам (21.3), (21.5), існує і єдиний. З цією метою, задавшись зображення сплайна $S(t)$ на відрізку $[t_i, t_{i+1}]$ у формі

$$S(t) = a_i + b_i(t-t_i) + c_i(t-t_i)^2 + d_i(t-t_i)^3 \tag{21.8}$$

встановимо, що знаходження коефіцієнтів a_i, b_i, c_i, d_i зводиться до роз'язку СЛАР з невиродженою матрицею. При $k=2$ і (21.6), (21.8) знаходимо

$$S''(a)=S''(t_0)=2c_0=0,\ S''(b)=S''(t_N-0)=2c_N=0,$$

$$S''(t_{i+1}-0)=2c_i+6d_i(t-t_i)\big|_{t=t_{i+1}}=S''(t_{i+1}+0)=2c_{i+1}+6d_{i+1}(t-t_{i+1})\big|_{t=t_{i+1}},$$

звідки

$$d_i=\frac{c_{i+1}-c_i}{3h_i},\ h_i=t_{i+1}-t_i. \qquad (21.9)$$

При $k=0$ аналогічно одержуємо

$$b_i=\frac{a_{i+1}-a_i}{h_i}-c_ih_i-d_ih_i^2. \qquad (21.10)$$

При $k=1$ маємо співвідношення

$$b_i+2c_ih_i+3d_ih_i^2=b_{i+1}. \qquad (21.11)$$

При $k=3$ з врахуванням $S(t_i)=a_i$

$$6d_{i+1}-6d_i=\frac{\widetilde{y}_i-a_i}{\rho_i^2}. \qquad (21.12)$$

Введемо наступні векторно-матричні позначення:

$$c=(c_1,c_2,...,c_{N-1})^T\quad \widetilde{y}=(\widetilde{y}_0,\widetilde{y}_1,...,\widetilde{y}_N)^T\quad a=(a_0,a_1,...,a_N)^T$$

$$D=diag\left(\frac{\rho_1}{\sqrt{2}},\frac{\rho_2}{\sqrt{2}},...,\frac{\rho_N}{\sqrt{2}}\right)$$

$$T=\begin{pmatrix}
\frac{2(h_0+h_1)}{3} & \frac{h_1}{3} & 0 & .....0 & 0\\
\frac{h_1}{3} & \frac{2(h_1+h_2)}{3} & \frac{h_2}{3} & .....0 & 0\\
..... & & & & \\
0 & 0 & 0 & .....0 & 0\\
0 & 0 & 0 & .....\frac{h_{N-2}}{3} & \frac{2(h_{N-2}+h_{N-2})}{3}
\end{pmatrix},$$

де матриця T має розмірність $(N-1)\times(N-1)$;

$$Q=\begin{pmatrix} \frac{1}{h_0} & -\left(\frac{1}{h_0}+\frac{1}{h_1}\right) & \frac{1}{h_1} & 0\ .....0 & 0 & 0 \\ 0 & \frac{1}{h_1} & -\left(\frac{1}{h_1}+\frac{1}{h_2}\right) & \frac{1}{h_2}.....0 & 0 & 0 \\ ..... & & & & & \\ 0 & 0 & 0 & 0..... \ 0 & 0 & 0 \\ 0 & 0 & 0 & .....\frac{1}{h_{N-2}} & -\left(\frac{1}{h_{N-2}}+\frac{1}{h_{N-1}}\right) & \frac{1}{h_{N-1}} \end{pmatrix},$$

де матриця Q має розмірність $(N-1)\times(N+1)$.

Об'єднуючи співвідношення (21.9) – (21.11), приходимо до рівності:

$$Tc = Q^T a, \tag{21.13}$$

а з (21.9), (21.12) випливає

$$Qc = D^{-2}(\tilde{y} - a) \tag{21.14}$$

Із (21.13), (21.14) одержуємо систему:

$$\left(Q^T D^2 Q + T\right)c = Q^T \tilde{y} \tag{21.15}$$

з симетричною додатньо визначеною 5-и діагональною матрицею для визначення вектора c. За знайденим c із (21.14) знаходимо

$$a = \tilde{y} - D^2 Qc,$$

а із рівностей (21.9), (21.10) послідовно виразяться вектори d і b.

Що і треба було довести.

Зауваження 21.1. Поряд з інтерполяційними і згладжуючими сплайнами розроблений підхід [41] до задачі відтворення функції і чисельного диференціювання, заснований на використання явної апроксимації сіткової (періодичної) функції комбінацією B-сплайнів [39]:

$$S_{2n-1,\rho}(t) = \sum_{i=1}^{N}\left(\sum_{j=-p}^{P} c_{ij}\, y_{i+j}\right) S_{i-n,2n-1}(t), \tag{21.16}$$

де $S_{i,m}(t)$ – базові сплакни [39] ступеня $m \le 2n-1$, зображувані формулою

$$S_{i,m}(t) = (t_{i+m+1} - t_i)\sum_{j=i}^{i+m+1} \frac{(t_i - t)^+}{\omega_{i,m}(t_j)},$$

де

$$(u)^+ = \begin{cases} u, \text{ якщо } u > 0 \\ 0, \text{ якщо } u \le 0 \end{cases} \quad \omega_{i,m}(t) = \prod_{j=1}^{i+m+1}(t - t_j).$$

Коефіцієнти c_{ij} знаходяться із умови апроксимації для всіх функцій $y(t)$ одночасно.

Наприклад, для $n=2$ і $\rho=1,2,3$ маємо відповідно:

$$S_{3,1}(t)=\sum_{i=1}^{N}\left(-\frac{y_{i-1}}{6}+\frac{4y_i}{3}-\frac{y_{i+1}}{6}\right)S_{i-2,3}(t),$$

$$S_{3,2}(t)=\sum_{i=1}^{N}\left(\frac{y_{i-2}}{36}-\frac{5y_{i-1}}{18}+\frac{3y_i}{2}-\frac{5y_{i+1}}{18}+\frac{y_{i+2}}{36}\right)S_{i-2,3}(t),$$

$$S_{3,3}(t)=\frac{1}{216}\sum_{i=1}^{N}\left(-y_{i-3}+12y_{i-2}-75y_{i-1}+344y_i-75y_{i+1}+12y_{i+2}-y_{i+3}\right)S_{i-2,3}(t).$$

Завдяки використанню явної апроксимації (21.16) немає необхідності розв'язувати СЛАР (на відміну від інтерполяційних та згладжуючих сплайнів). З цієї причини досягається висока економічність, особливо в багатовимірному випадку, для якого побудовані аналоги формул (21.16) (див. [41]).

Математичне забезпечення, що реалізує процедури згладжування і диференціювання наближено заданих функцій, засновані на сплайн-апроксимації, описані і опубліковані в роботах [41, 39].

§22. МЕТОД СЕРЕДНІХ ФУНЦІЙ

Кінцево-різницеві РА, розглянуті в попередньому параграфі, породжують наближені розв'язки $\Delta_{h(\delta)} y_\delta$, які мають ту ж гладкість, що і y_δ. Більш гладкі (навіть нескінченно диференційовані) апроксимації похідних можна конструювати за допомогою апарата середніх функцій. Побудовані на їх основі РА, на відміну від кінцево-різницевих, є оптимальними за порядком.

Визначення 22.1. Регуляризуючий алгоритм $\{R_\delta\}$ називається оптимальним за порядком на множині M, якщо

$$\sup_{\substack{y\in M,\, y_\delta\in Y \\ \|y-y_\delta\|\le\delta}} \left\| R_\delta y_\delta - \frac{d^m y}{dt^m} \Big/ \Omega_\delta\left(\frac{d^m}{dt^m}; Z\right) \right\| \le K < \infty .$$

При $K = 1$ одержуємо оптимальність в сенсі задачі A.

Нижче побудуємо РА, для якого $K < 1{,}9$, а $x_\delta(t) = R_\delta y_\delta(t)$ – нескінченно диференційовані функції.

Визначимо функцію двох змінних:

$$\omega_\alpha(t,s) = \begin{cases} c_\alpha \exp\left(\dfrac{(t-s)^2}{(t-s)^2-\alpha^2}\right) & при\ |t-s| < \alpha \\ 0 \ при\ |t-s| \ge \alpha \end{cases},$$

де

$$c_\alpha = \frac{1}{\int\limits_{-\alpha}^{\alpha} \exp\left[\dfrac{r^2}{r^2-\alpha^2}\right] dr}.$$

Для неї справедливі наступні властивості (див. [42], стор.18):

1. функція $\omega_\alpha(t,s)$ неперервна разом зі своїми похідними на площині R^2;
2. при $|t-s| \ge \alpha$ функція $\omega_\alpha(t,s)$ і всі її похідні рівні нулю;
3. $\int\limits_{|t-s|\le\alpha} \omega_\alpha(t,s)ds = \int\limits_{|t-s|\le\alpha} \omega_\alpha(t,s)dt = 1$;
4. середня функція

$$y^\alpha(t) = \int\limits_{|t-s|\le\alpha} \omega_\alpha(t,s) y(s) ds$$

від функції $y(t) \in C(-\infty,\infty)$ нескінченно диференційована, крім того,

$$\frac{d^k}{dt^k}\left[y^\alpha(t)\right] = \int\limits_{|t-s|\le\alpha} \frac{d^k}{dt^k}\left[\omega_\alpha(t,s)\right] y(s) ds .$$

Для задачі обчислення першої похідної визначимо регуляризуюче сімейство $\{R_\alpha\}$ операторів

$$R_\alpha y(t) = \int\limits_{|t-s|} \frac{d}{dt}\omega_\alpha(t,s)y(s)ds. \qquad (22.1)$$

Лемма 22.1. Нехай $y(t)$ належить множині M, визначеній в (19.3.1), тоді справедлива оцінка

$$\sup_{y\in M}\left\|R_\alpha y(t) - \frac{d}{dt}y(t)\right\|_c \le \alpha r.$$

Доведення. Беручи до уваги властивості 1, 2 і інтегруючи по частинах, одержуємо:

$$R_\alpha y(t) = \int\limits_{|t-s|\le\alpha} \frac{d}{dt}\omega_\alpha(t,s)y(s)ds = -\int\limits_{|t-s|\le\alpha} \frac{d}{ds}\omega_\alpha(t,s)y(s)ds =$$

$$= \int\limits_{|t-s|\le\alpha} \omega(t,s)\frac{d}{ds}y(s)ds. \qquad (22.2)$$

Далі, із умов леми, властивості 3 і формули (22.2) знаходимо

$$Sup_{y\in M}\left\|R_\alpha y(t) - \frac{d}{dt}y(t)\right\|_c = Sup_{y\in M}\ Sup_{-\infty<t<\infty}\left|\int\limits_{|t-s|\le\alpha}\omega_\alpha(t,s)\frac{d}{ds}y(s)ds - \int\limits_{|t-s|\le\alpha}\omega_\alpha(t,s)\frac{d}{dt}y(s)ds\right| \le$$

$$\le Sup_{y\in M}\ Sup_{|t-s|\le\alpha}|y'(s)-y'(t)| \int\limits_{|t-s|\le\alpha}\omega_\alpha(t,s) \le rd.$$

Лема 22.2. Нехай $y_\delta(t)\in C(-\infty,\infty)$, $y(t)\in M$ і $\|y_\delta(t)-y(t)\|_c \le \delta$, тоді справедлива оцінка

$$Sup_{y\in M,\|y-y_\delta\|_c\le\delta}\|R_\alpha y_\delta(t) - R_\alpha y(t)\|_c \le \frac{\delta h}{\alpha} \qquad (22.3)$$

де

$$h = \left\{\int_0^1 \exp\left[\frac{r^2}{r^2-1}\right]dr\right\}^{-1} \approx 1{,}65$$

Доведення. Маємо очевидні нерівності.

$$\|R_\alpha y_\delta(t) - R_\alpha y(t)\|_c \le Sup_{-\infty<t<\infty}\int\limits_{|t-s|\le\alpha}\left|\frac{d\omega_\alpha(t,s)}{dt}\left[y_\delta(s)-y(s)\right]\right|ds \le$$

$$\le \delta\ Sup_{-\infty<t<\infty}\int\limits_{|t-s|\le\alpha}\left|\frac{d\omega_\alpha(t,s)}{dt}\right|ds = \delta c_\alpha\int_{-\alpha}^{\alpha}\left|\left[\frac{e^{\xi^2}}{\xi^2-\alpha^2}\right]\cdot\frac{2\alpha^2\xi}{\xi^2-\alpha^2}\right|d\xi =$$

$$= 2\delta c_\alpha\int_{-\infty}^{0}e^u du = 2\delta c_\alpha = \delta\left\{\int_0^1\frac{e^{\eta^2}}{\eta^2-1}d\eta\right\}^{-1} = \frac{\delta h}{\alpha}$$

Теорема 22.1. Якщо виконані умови Лем 22.1, 22.2 і $\alpha(\delta)=\sqrt{\frac{n\delta}{2}}$, справедлива оцінка

$$\Omega_\delta\left(\frac{d}{dt};r\right)\le \underset{y\in M,\|y-y_\delta\|_c\le\delta}{Sup}\left\|R_{\alpha(\delta)}y_\delta-\frac{dy}{dt}\right\|_c<1{,}83\Omega\left(\frac{d}{dt};r\right) \tag{22.4}$$

Доведення. Із нерівності

$$\left\|\frac{dy(t)}{dt}-R_\alpha y_\delta(t)\right\|_c\le\left\|\frac{dy(t)}{dt}-R_\alpha y(t)\right\|_c+\left\|R_\alpha y(t)-R_\alpha y_\delta(t)\right\|_c$$

і тверджень лем 22.1, 22.2 випливає оцінка

$$\sup\left\|\frac{dy(t)}{dt}-R_\alpha y_\delta(t)\right\|_c\le r\alpha+\frac{n\delta}{2} \tag{22.5}$$

Залежність $\alpha=\sqrt{\frac{n\delta}{r}}$ надає мінімум правій частині (22.5) і забезпечує оцінку зверху в (22.4).

Таким чином, справедливість теореми встановлена і розв'язані задачі I, II.

Зауваження 22.1. Регуляризатор R, визначений формулою (22.1), можна зобразити у вигляді добутку $R_\alpha=TS_\alpha$, де

$$S_\alpha y(t)=\int\limits_{|t-s|\le\alpha}\omega_\alpha(t,s)y(s)ds\quad T_r=\frac{dz(t)}{dt}.$$

Користуючись аргументацією, наведеною при доведенні Лем 22.1, 22.2, легко довести, що для оператора S_α: $c(-\infty,\infty)\to c(-\infty,\infty)$ справедливі наступні властивості:

а) $\|S_\alpha\|\le 1$;

б) $\|S_\alpha-Ey(t)\|_c\to 0$ при $\alpha\to 0$ для любої рівномірно неперервної функції.

Сімейство S_α з властивостями а),б) називається фільтруючим в $D(T)$, регуляризуюче сімейство $\{R_\alpha\}$, зображуване у вигляді $R_\alpha=TS_\alpha$, називається нормальним регуляризуючим сімейством операторів [43].

Задача відтворення похідної n-ого порядку розв'язується заміною оператора (22.1) оператором

$$R_\alpha y(t)=\int\limits_{|t-s|\le\alpha}\frac{\alpha^n\omega_\alpha(t,s)}{dt^n}y(s)ds$$

Для чисельного знаходження наближеного значення похідної $y'(t)$ в точці $\bar{t}$ достатньо застосувати до інтегралу в (22.1) яку-небудь квадратурну формулу:

$$\sum_{k=1}^{n} A_k \frac{d\omega_\alpha(\bar{t},S_k)}{dt} y_\delta(S_k) \overset{n\to\infty}{\to} R_{\alpha(\delta)} y_\delta(t) \to \frac{dy(\bar{t})}{dt}$$

Зауваження 22.2. Властивості а), б) підказують, що замість $\omega_\alpha(t,s)$ можна використати інші усереднюючі ядра, які утворюють δ-подібні послідовності (див. [44]).

Зауваження 22.3. Задача I із §19 при $m=0$, то є задача відтворення функції в просторі X із більш сильною нормою, ніж в Y, розв'язується за допомогою послідовності $x_\delta = S_{\alpha(\delta)} y_\delta$, де сімейство $\{S_\alpha\}$, $S_\alpha : Y \to X$ задовольняє наступним умовам:

а') $\forall \alpha > 0 \; \|S_\alpha\| \le c(\alpha) < \infty$;

б') $\forall y \in X \; \lim\limits_{\alpha\to 0} \|(S_\alpha - E)y\|_X = 0$.

Дійсно, це випливає із нерівності

$$\|S_\alpha y_\delta - y\|_X \le \|S_\alpha\| \cdot \|y_\delta - y\|_Y + \|S_\alpha y - y\|_X$$

за допомогою вибору залежності $\alpha(\delta)$ такою, що $c(\alpha(\delta))\cdot\delta \to 0$ при $\delta \to 0$.

Зауваження 22.4. Можна довести, що послідовність $x_\delta = S_{\alpha(\delta)} y_\delta$ розв'язує задачу I при $m=0$, де

$$S_\alpha y = \frac{1}{2\alpha} \int_{t-\alpha}^{t+\alpha} y(\tau)\, d\tau,$$

$\alpha(\delta) = \delta^\gamma \; (0 < \gamma < 2)$, $Y = L_2(-\infty,\infty)$, $X = \widetilde{C} \in C(-\infty,\infty)$ – простір рівномірно неперервних функцій.

Аналогічні результати можна одержати і для $Y = L_p[a,b]$ ($p>1$), $X = C[a,b]$.

§23. ЧИСЕЛЬНІ ЕКСПЕРИМЕНТИ

Наведемо результати чисельних реалізацій розглянутих методів на деяких прикладах.

Приклад 23.1.

$y(t)=\sin t$, $0\le t\le\pi$: із мережевими значеннями, обчисленими в точках $t_k=\dfrac{k\pi}{180}$ $(k=\overline{0,180})$ з точністю до 4-х правильних знаків, трьома методами відтворювалися функції і її похідні до третього порядку. В таблиці 1 наведена середньоквадратична похибка (детальніше див. у [40]).

Таблиця 23.1

Порядок похідної	**Кінцево-різницевий метод**	**Інтерполяційний кубічний сплайн**	**Згладжуючий кубічний сплайн**
0	$3\cdot10^{-5}$	$3\cdot10^{-5}$	$1.3\cdot10^{-5}$
1	$2.6\cdot10^{-3}$	$3.4\cdot10^{-3}$	$0.21\cdot10^{-3}$
2	0.26	0.67	$0.42\cdot10^{-2}$
3	28	74	0.16

Із наведених числових даних видно, що $y(t)$, $y'(t)$ всіма методами відтворюються приблизно однаково, тоді як для $y''(t)$, $y'''(t)$ згладжуючий сплайн дає набагато кращі результати.

Приклад 23.2.

$y(t)=\sin t$, $0\le t\le\dfrac{\pi}{2}$, $y'(t)=\cos t$.

Інтерполяційний кубічний сплайн, побудований по 100 точках мережі з заокругленням значень функції до 5-ти знаків ($\delta\le5\cdot10^{-5}$), дає при відтворенні похідної абсолютну похибку $\Delta\approx2\cdot10^{-4}$.

Метод середніх функцій для тієї ж функції при $\alpha=10^{-4}$ з використанням формули Сімпсона з числом точок $m=100$ дає абсолютну похибку $\Delta\approx1.5\cdot10^{-4}$. Це означає, що обидва методи працюють однаково добре для гладкої функції при малих похибках.

Приклад 23.3.

$y(t)=1-|t|$, $-1\le t\le1$, $y'(t)=\begin{cases}-1, & t<0\\ 1, & t>0\end{cases}$.

На відміну від попереднього прикладу ця функція негладка (Кусково-диференційована). Для інтерполяційного сплайна – це сприятлива ситуація; похибка чисельного диференціювання зростає до $\Delta \approx 0.26$. Метод середніх функцій не реагує на зміну ситуації і відтворює при тих же параметрах ($m = 100$, $\alpha = 10^{-4}$) похідну з точністю $\Delta \approx 10^{-4}$.

Приклад 23.4.

$y(t) = e^t$, $0 \le t \le 1$, $y'(t) = e^t$.

Значення y_i заокруглювалися з точністю $\delta = 0.05$ і обчислювалися на мережі $t_k = \dfrac{k}{20}$ ($k = \overline{0,20}$). Інтерполяційний сплайн з граничними даними дає похибку чисельного диференціювання $\Delta \approx 0.9$, а згладжуючий сплайн при деякому виборі параметрів ρ_i покращує ситуацію, похибка тут $\Delta \approx 0.1$.

Модельні розрахунки підтверджують міркування §21, що для сильно „зашумлених" вихідних даних для чисельного диференціювання переважає застосування згладжуючи сплайнів.

Приклад 23.5.

$y(t) = e^{-t^2}$, $t \in [-1,1]$, $h = t_{k+1} - t_k = 0.25; 0.1$.

Використовувалась апроксимуюча формула (21.16) при $n = 1,2,3,4$ і $\rho = 0$. Обчислювалися похідні до порядку $2n-2$ на сітці з кроком h. У таблиці 23.2 наведені похибки

$$\Delta_e = \max_{1 \le i \le N} \left| y^{(l)}(t_i) - S^{(l)}_{2n-1,0} \right|, \; l = 0,1,2,3,4$$

Таблиця 23.2

	n	Δ_0	Δ_1	Δ_2	Δ_3	Δ_4
$h = 0.25$	$n = 1$	$2 \cdot 10^{-2}$				
	$n = 2$	$2.02 \cdot 10^{-2}$	$4.0 \cdot 10^{-2}$	$1.3 \cdot 10^{-1}$		
	$n = 3$	$3 \cdot 10^{-2}$	$5.8 \cdot 10^{-2}$	0.14	0.47	1.78
	$n = 4$	$3.95 \cdot 10^{-2}$	$7.53 \cdot 10^{-2}$	$2.18 \cdot 10^{-1}$	0.61	2.18
$h = 0.1$	$n = 1$	$2.47 \cdot 10^{-3}$				
	$n = 2$	$3.32 \cdot 10^{-3}$	$6.46 \cdot 10^{-3}$	$2.45 \cdot 10^{-2}$		
	$n = 3$	$5.0 \cdot 10^{-3}$	$9.67 \cdot 10^{-3}$	$2.97 \cdot 10^{-2}$	$8.06 \cdot 10^{-2}$	$3.38 \cdot 10^{-1}$
	$n = 4$	$6.6 \cdot 10^{-3}$	$1.28 \cdot 10^{-2}$	$3.94 \cdot 10^{-2}$	$1.07 \cdot 10^{-1}$	$3.91 \cdot 10^{-1}$

§24. ЗАДАЧІ НА ЕКСТРЕМУМ ФУНКЦІОНАЛА. ОСНОВНІ ВИЗНАЧЕННЯ І УМОВИ КОРЕКТНОСТІ

24.1. Постановка задачі

Задачі оптимізації (мінімізації функціоналів) є математичними моделями широкого кола прикладних проблем техніки, економіки, керування (див. наприклад [22,45]). Серед численних постановок оптимізаційних задачі особливе місце посідають нестійкі (некоректно поставлені) задачі. Характерною рисою цих задач є відсутність неперервної залежності розв'язку від цільового функціонала і допустимої множини і, як наслідок, неможливість ефективного їх розв'язку традиційними методами.

Тут сформулюємо основну задачу оптимізації, визначимо два типи коректності і дамо достатні умови для їх виконання.

Нехай Ω – підмножина (скалярний добуток позначимо $<,>$) гільбертового простору X, а f – дійсно значний функціонал із зоною визначення $D(f) \supseteq \Omega$, обмежений знизу на Ω. Розглядається задача мінімізації

$$\min\{f(x) : x \in \Omega\} = F > -\infty \qquad (24.1)$$

Число F називається оптимальним значенням $f(x)$, Ω – допустимою множиною, а сукупність елементів $M = \{y \in \Omega : f(y) = F\}$ – оптимальною множиною.

Визначення 24.1. Послідовність елементів $\{x_n\}$, $x_n \in \Omega$ називається множиною наближених розв'язків задачі (24.1) по функціоналу, якщо

$$\lim_{n\to\infty} f(x_n) = F$$

або по функціоналу, якщо

$$\lim_{n\to\infty} \rho(x_n, M) = \lim_{n\to\infty} \inf_{y\in M} \rho(x_n, y) = 0$$

Спосіб побудови послідовності $\{x_n\}$ назвемо алгоритмом розв'язку задачі (24.1) по функціоналу (по аргументу).

24.2. Коректність за Адамаром та Тихоновим

У відповідності до [46] дамо наступне:

Визначення 24.2. Задача (24.1) називається коректно поставленою (або, коротко – коректною) за Адамарам, якщо:

1. розв'язок існує, то є $M \neq \varnothing$;
2. єдиний, то є M – одноточкова множина;

3. розв'язок неперервно залежить від f і Ω.

Попередньо в роботах [13, 18] було дано інше визначення коректності.

Визначення 24.3. Задача (24.1) називається коректною за Тихоновим, якщо:

1. розв'язок існує;
2. розв'язок єдиний;
3. кожна мінімізуюча послідовність x_n (то є $x_n \in \Omega$ і $f(x_n) \to F$) збігається до єдиного розв'язку $\bar{x}$ задачі (24.1).

Визначення 24.4. Якщо порушена одна з умов 1-3 у визначенні 24.2, 24.3, то говорять, що задача некоректно поставлена (або коротко – некоректна) за Адамаром (Тихоновим). Як правило, тут мова йде про порушення умови 3.

Наведені поняття коректності тісно пов'язані між собою. Перш ніж перейти до точних формулювань, введемо два види збіжності множин.

Визначення 24.5. Послідовність множин $\Omega_n \subset X$ збігається до множини $\Omega \subset X$ в сенсі Хаусдорфа (позначимо $\Omega_n \overset{H}{\to} \Omega$), якщо

а) $\forall \varepsilon > 0\ \exists N : n \geq N\ \Omega_n \subset \Omega^{\varepsilon} = \{x \in X : \rho(x,\Omega) < \varepsilon\}$;

б) $\forall \varepsilon > 0\ \exists N : n \geq N\ \Omega \subset \Omega_n^{\varepsilon} = \{x \in X : \rho(x,\Omega_n) < \varepsilon\}$

або еквівалентно

$$\lim_{n\to\infty} \alpha(\Omega_n,\Omega) = \lim_{n\to\infty} \max\{\beta(\Omega_n,\Omega), \beta(\Omega,\Omega_n)\} = 0$$

де $\beta(\Omega_n,\Omega) = \sup\limits_{y\in\Omega_n} \inf\limits_{x\in\Omega} \rho(x,y)$ (див. Рис. 24.1)

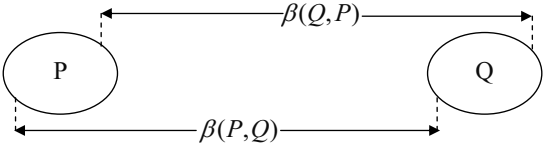

Рис.24.1

Величину $\beta(Q,P)$ називають напіввідхиленням множини Q від P (в сенсі Хаусдорфа).

Визначення 24.5. Послідовність множин $\Omega_n \subseteq X$ збігається до множини $\Omega \subset X$ в сенсі Моско (позначимо $\Omega_n \overset{M}{\to} \Omega$), якщо

а) $\forall x \in \Omega\ \exists \{x\} : x_n \in \Omega, x_n \to x$;

б) $\forall \{x_{nj}\}, x_{nj} \subset \Omega_{nj}, x_{nj} \to x \Rightarrow x \in \Omega$,

де значок „$\to$" означає слабку збіжність в X.

Твердження 24.1. Нехай f – опуклий рівномірно неперервний функціонал. Якщо задача мінімізації (19.1.1) коректно поставлена за Адамаром відносно збіжності $\Omega_n \overset{H}{\to} \Omega$ на всякій замкненій опуклій множині Ω, то ця задача коректно поставлена за Тихоновим на кожні опуклій замкненій множині Ω.

Наслідок 24.1. Якщо задача (19.1.1) некоректна за Тихоновим на множині Ω, то ця задача, взагалі кажучи, некоректна за Адамаром.

Твердження 24.2. Нехай f – опуклий рівномірно неперервний функціонал на будь-який опуклій обмеженій множині. Тоді коректність за Тихоновий на всякій замкненій опуклій множині тягне за собою коректність за Адамаром відносно Моско-збіжності, то є

$$\Omega_n \overset{M}{\to} \Omega \Rightarrow \arg\min\{f(x): x \in \Omega_n\} \to \arg\min\{f(x): x \in \Omega\}$$

Доведення цих тверджень наведено в [66].

24.3. Достатні умови коректності за Тихоновим

Мета цього пункту – пояснити умови на вихідні дані (24.1), які гарантують властивості коректності за Тихоновим і які, згідно з твердженням 24.2., тягнуть за собою неперервну залежність розв'язку при варіації допустимої множини.

Визначення 24.7. Функціонал f називається сильно опуклим на замкнутій множині Ω, якщо $f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1) + (1-\lambda) f(x_2)$ при будь-яких $x_1 \neq x_2$, $x_1, x_2 \in \Omega$, $0 < \lambda < 1$.

Визначення 24.8. Функціонал f називається сильно опуклим на замкнутій множині Ω, якщо існує константа $r > 0$ так, що

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2) - r\lambda(1-\lambda)\|x_1 - x_2\|^2 \qquad (24.2)$$
$$x_1 \neq x_2 \quad x_1, x_2 \in \Omega \quad 0 < \lambda < 1$$

Приклади: 1. $f(x) = \|x\|$ – опуклий, але не суворо опуклий функціонал; 2. $f(x) = \|x\|^2$, $f(x) = < Bx, x > + < p, x >$ (де $< Bx, x > \ge c\|x\|^2$, $c > 0$) – сильно опуклі і, значить, суворо опуклі функціонали.

Сформулюємо послідовність лем, в яких виясняються деякі властивості опуклих функціоналів. Перші дві з них повністю очевидні.

Лема24.1. Якщо задача мінімізації (24.1) розв'язувана і функціонал f – строго опуклий, то розв'язок єдиний.

Лема24.2. Якщо f – опуклий неперервний функціонал, а Ω – опукла замкнута множина, тоді множина розв'язків M (оптимальна множина) задачі (24.1) опукла, замкнута або пуста.

Лема 24.3. Якщо функціонал f – сильно опуклий на Ω, то справедлива нерівність

Лемма 24.3. Якщо функціонал f – сильно опуклий на Ω, то справедлива нерівність

$$\|x-\tilde{x}\| \le \frac{2}{r}\left(f(x)-f(\tilde{x})\right), \tag{24.3}$$

де $\tilde{x}$ – розв'язок задачі (24.1), x – довільний елемент із Ω.

Доведення. Оскільки f сильно опуклий, із нерівності (24.2) при $\lambda=\frac{1}{2}$ випливає:

$$f\left(\frac{1}{2}x+\frac{1}{2}\tilde{x}\right) \le \frac{1}{2}f(x)+\frac{1}{2}f(\tilde{x})-\frac{r}{4}\|x-\tilde{x}\|^2.$$

звідки

$$\frac{r}{4}\|x-\tilde{x}\|^2 \le \frac{1}{2}f(x)+\frac{1}{2}f(\tilde{x})-f\left(\frac{1}{2}x+\frac{1}{2}\tilde{x}\right) \le$$

$$\le \frac{1}{2}f(x)+\frac{1}{2}f(\tilde{x})-f(\tilde{x}) \le \frac{1}{2}f(x)-\frac{1}{2}f(\tilde{x}) \tag{24.4}$$

Лема 24.4. Якщо f – дужеопуклий функціонал, тоді

$$\lim_{\|x\|\to\infty} f(x)=\infty.$$

Доведення. Визначимо ε – оптимальний елемент x_ε співвідношенням $F \le f(x_\varepsilon) \le F+\varepsilon$, $x_\varepsilon \in \Omega$. Замінимо в першій нерівності із ланцюга (24.4) елемент $\tilde{x}$ (розв'язність не гарантується) на x_ε і замінимо $f\left(\frac{1}{2}x+\frac{1}{2}\tilde{x}\right)$ на $f(x_\varepsilon)-\varepsilon \le F$ із одержаної нерівності

$$\|x-\tilde{x}\|^2 \le \frac{4}{r}\left\{\frac{1}{2}f(x)-\frac{1}{2}f(x_\varepsilon)+\varepsilon\right\},$$

із якого і випливає потрібна властивість.

Лема 24.5. Якщо f – сильно опуклий неперервний функціонал, а Ω – опукла замкнена множина, то задача (24.1) розв'язувана єдиним способом.

Доведення. Переконаємось, що множина $\Omega_0=\{x\in\Omega: f(x)\le f(x_0)\}$, де x_0 – довільна точка із Ω, обмежена. І справді, допустимо протилежне, що існує $x_n\in\Omega$, таке, що $\lim\limits_{n\to\infty}\|x_n\|=\infty$, згідно з лемою (24.4) будемо мати $\lim\limits_{n\to\infty} f(x_n)=\infty$. Тоді при $n\ge n_0$, то є $x_n\notin\Omega$, що

суперечить вибору x_n. Оскільки $F \le f(x_0)$, то задача (24.1) еквівалентна:

$$\min\{f(x): x \in \Omega_0\} = F_0 = F$$

Нехай x^k – мінімізуюча послідовність, то є $x^k \in \Omega_0$, $\lim\limits_{k\to\infty} f(x^k) = F$.

Внаслідок обмеженості Ω_0 $\{x^k\}$ слабо компактна ([47], стор. 254), то є $x^k \to \tilde{x}$. Множина Ω_0, як і Ω, опукла і замкнена, значить слабо замкнена ([35], стор. 196), це означає, що $\tilde{x} \in \Omega_0$. Опуклий неперервний функціонал слабо неперервний знизу, тому

$$F \le f(\tilde{x}) \le \lim_{i\to\infty} \inf f(x^{ki}) = F,$$

це означає, що $\tilde{x}$ – розв'язок. Єдиність випливає із леми (24.1).

Визначення 24.9. Нехай x^0 – деякий елемент із X, M – оптимальна множина задачі (24.1). Розв'язок задачі на мінімум

$$\min\left\{\left\|x - x^0\right\|^2 : x \in M\right\}$$

називається нормальним розв'язком (в подальшому, як правило, позначаємо $x^0 = 0$).

Лема 24.6. Якщо f – опуклий неперервний функціонал, то існує єдиний нормальний розв'язок.

Доведення. Безпосередньою перевіркою можна легко переконатися, що $f(x) = \left\|x - x^0\right\|^2$ – сильно опуклий функціонал. Згідно з лемою 24.2 M – опукла і замкнена. Тоді згідно з лемою 24.2 справедливе твердження леми 24.6.

Теорема 24.1. Нехай f – сильно опуклий неперервний функціонал, Ω – замкнена опукла множина. Тоді задача (24.1) коректно поставлена за Тихоновим (визначення 24.3).

Доведення. Розв'язуваність і єдиність розв'язку $\tilde{x}$ є наслідок Леми 24.5. Нехай $\{x^k\}$ – мінімізуюча послідовність: $\lim\limits_{k\to\infty} f(x^k) = F$, $x^k \in \Omega$. Підставляючи в нерівність (24.3) $x = x^k$, робимо висновок, що

$$\lim_{k\to\infty} \left\|x^k - \tilde{x}\right\| \le \lim_{k\to\infty} \frac{2}{r}\left(f(x^k) - f(\tilde{x})\right) = 0.$$

§25. РЕГУЛЯРИЗАЦІЯ ЕКСТРЕМАЛЬНИХ ЗАДАЧ

25.1. Регуляризація задачі з точними вхідними даними

Якщо функціонал f – тільки опуклий, але не сильно опуклий, то, взагалі кажучи, малі похибки цільового функціоналу і допустимої множини можуть суттєво спотворити розв'язок, що приводить до значних труднощів при побудові стійкого обчислювального алгоритму. Суть методу регуляризації є в апроксимації некоректної задачі параметричним сімейством коректних задач і розумною узгодженістю параметра регуляризації з похибкою вхідних даних.

Теорема 25.1. Нехай f – випуклий неперервний функціонал, Ω – опукла замкнена множина і $M \neq \varnothing$. Тоді для будь-якого $\alpha > 0$ існує єдиний розв'язок x_α задачі

$$\min\left\{f(x) + \alpha\|x\|^2 : x \in \Omega\right\} = F_\alpha \tag{25.1}$$

і $\lim\limits_{\alpha \to 0}\|x_\alpha - \tilde{x}\| = 0$, де $\tilde{x}$ – нормальний розв'язок задачі (24.1) (визначення 25.9 при $x^0 = 0$). Крім того, $\lim\limits_{\alpha \to 0} f(x_\alpha) = F$.

Доведення. Оскільки функціонал f – опуклий, то $f_\alpha = f(x) + \alpha\|x\|^2$ – сильно опуклий функціонал. Тому існує і єдиний розв'язок $x_\alpha \in \Omega$, що випливає із леми 24.5. Із сукупності нерівностей

$$f(x_\alpha) + \alpha\|x_\alpha\|^2 \le f(\tilde{x}) + \alpha\|\tilde{x}\|^2 \le f(x_\alpha) + \alpha\|\tilde{x}\|^2$$

випливає оцінка $\|x_\alpha\| \le \|\tilde{x}\|$, що означає існування слабо збіжної послідовності $x_{\alpha_k} \to \bar{x} \in \Omega$ (що випливає із слабо замкнутості Ω) і нерівності

$$\|\tilde{x}\| \le \|\bar{x}\| \le \lim_{k \to \infty}\inf\|x_{\alpha_k}\| \le \|\tilde{x}\| \tag{25.2}$$

$$F \le f(\bar{x}) \le \lim_{k \to \infty}\inf f(x_{\alpha_k}) \le \lim_{k \to \infty}\inf\left\{f(x_{\alpha_k}) + \alpha\|x_{\alpha_k}\|^2\right\} \le$$

$$\le \lim_{k \to \infty}\left\{f(\tilde{x}) + \alpha_k\|\tilde{x}\|^2\right\} \le F,$$

де використана слаба напівнеперервність знизу функціоналів $\varphi(x) = \|x\|^2$, $f(x)$. Отож, $f(\bar{x}) = F$, $\bar{x} \in \Omega$. Із (25.2) і леми 24.6 випливає, що $\bar{x} = \tilde{x}$, а із слабкої збіжності $\{x_{\alpha_k}\}$ і збіжності норм у гільбертовому просторі випливає $\lim\limits_{k \to \infty}\|x_{\alpha_k} - \tilde{x}\| = 0$. Міркуючи від протилежного, переконуємося в сильній збіжності всієї послідовності.

Наслідок 25.1. Якщо регуляризуючий функціонал взяти у формі $f_\alpha(x) = f(x) + \alpha\left\|x - x^0\right\|^2$, то послідовність $\{x_\alpha\}$ буде збігатись до розв'язку $\tilde{x} \in M$, який найменше ухиляється від x^0.

25.2. Регуляризація з наближеними даними

Будемо вважати, що замість точного функціонала f відомо лише його ε-наближення $f^\varepsilon(x)$, яке задовольняє умові апроксимації:

$$\left|f^\varepsilon(x) - f(x)\right| \le \varepsilon,\ \forall x \in \Omega. \tag{25.3}$$

Тоді регуляризована задача набуває вигляду:

$$\min\left\{f^\varepsilon(x) + \alpha\|x\|^2 : x \in \Omega\right\} = F_\alpha^\varepsilon. \tag{25.4}$$

Зауважимо, що допустима множина тут задана точно, ця обставина принципово важлива.

Теорема 25.2. Нехай f, f^ε – опуклі неперервні функціонали і виконана умова апроксимації (25.3). Тоді для будь-яких $\alpha > 0$ і f^ε задача (25.4) має єдиний розв'язок x_α^ε і $\lim\limits_{\varepsilon\to 0}\left\|x^\varepsilon_{\alpha(\varepsilon)} - \tilde{x}\right\| = 0$, $\lim\limits_{\varepsilon\to 0} f(x^\varepsilon_{\alpha(\varepsilon)}) = F$, якщо залежність $\alpha(\varepsilon)$ вибрана так, що $\lim\limits_{\varepsilon\to 0}\alpha(\varepsilon) = 0$, $\lim\limits_{\varepsilon\to 0}\dfrac{\varepsilon}{\alpha(\varepsilon)} = 0$.

Доведення. Оскільки f^ε – опуклий, то $f^\varepsilon(x) + \alpha\|x\|^2$ – суворо опуклий функціонал, тому згідно з лемою 24.5 розв'язок x_α^ε задачі 25.4 існує і єдиний. Маємо

$$f^\varepsilon(x_\alpha^\varepsilon) + \alpha\left\|x_\alpha^\varepsilon\right\|^2 \le f^\varepsilon(\tilde{x}) + \alpha\|\tilde{x}\|^2 \le \left|f^\varepsilon(\tilde{x}) - f(\tilde{x})\right| + f^\varepsilon(\tilde{x}) + \alpha\|\tilde{x}\|^2 \le \varepsilon + F + \alpha\|\tilde{x}\|$$

де $\tilde{x}$ – нормальний розв'язок задачі 24.1.

З урахуванням одержаної оцінки отримаємо:

$$F - \varepsilon + \alpha\left\|x_\alpha^\varepsilon\right\|^2 \le F - \left|f_\alpha^\varepsilon(x_\alpha^\varepsilon) - f^\varepsilon(x_\alpha^\varepsilon)\right| + \alpha\left\|x_\alpha^\varepsilon\right\| \le$$

$$\le F + f^\varepsilon(x_\alpha^\varepsilon) - f(x_\alpha^\varepsilon) + \alpha\left\|x_\alpha^\varepsilon\right\|^2 \le f^\varepsilon(x_\alpha^\varepsilon) + \alpha\left\|x_\alpha^\varepsilon\right\|^2 \le \varepsilon + F + \alpha\|\tilde{x}\|^2,$$

звідки

$$\left\|x_\alpha^\varepsilon\right\|^2 \le \|\tilde{x}\|^2 + \frac{2\varepsilon}{\alpha}. \tag{25.5}$$

Із обмеженості $\left\{x_\alpha^\varepsilon\right\}$ випливає існування слабо збіжної підпослідовності

$$x^\varepsilon_{\alpha(\varepsilon_i)} \equiv x^i \to \bar{x} \in \Omega \tag{25.6}$$

(множина Ω опукла і замкнена, значить, слабо замкнена). Із співвідношень (25.5), (25.6) і умов теореми одержуємо

$$\|\tilde{x}\| \le \|\bar{x}\|^2 \le \lim_{i\to\infty}\left\|x^i\right\| \le \|\tilde{x}\| \qquad (25.7)$$

Беручи до уваги слабу напів неперервність знизу f, маємо

$$F \le f(\bar{x}) \le \lim_{i\to\infty}\inf f(x^i) \le \lim_{i\to\infty}\inf\left\{f(x^i) + \alpha\left\|x^i\right\|^2\right\} \le$$

$$\le \lim_{i\to\infty}\inf\left\{f^{\varepsilon_i}(x^i) + \alpha(\varepsilon_i)\left\|x^i\right\|^2 + \varepsilon_i\right\} \le \lim_{i\to\infty}\inf\left\{f(\tilde{x}) + \alpha(\varepsilon_i)\|\tilde{x}\|^2 + 2\varepsilon_i\right\} \le F$$

Це означає, що $\bar{x} \in M$. Разом з нерівністю (25.7) і лемою 24.6 одержуємо $\bar{x} = \tilde{x}$. Об'єднуючи тепер співвідношення (25.6), (25.7), одержуємо $\lim_{i\to\infty}\left\|x^i - \tilde{x}\right\| = 0$. Оскільки $\left\{x_\alpha^\varepsilon\right\}$ має єдину граничну точку, то збігається вся послідовність $\lim_{\varepsilon\to 0}\left\|x_{\alpha(\varepsilon)}^\varepsilon - \tilde{x}\right\| = 0$, значить, враховуючи неперервність f, маємо $\lim_{\varepsilon\to 0} f(x_{\alpha(\varepsilon)}^\varepsilon) = F$.

Зауваження 25.1. Нехай $x_\alpha^{\varepsilon,r}$ – елемент, що задовольняє нерівності $F_\alpha^\varepsilon \le f(x_\alpha^{\varepsilon,r}) + \alpha\left\|x_\alpha^{\varepsilon,r}\right\|^2 \le F_\alpha^\varepsilon + r$, то є елемент, реалізуючий мінімум з точністю r. Тоді при допущеннях теореми 25.2. $\lim_{\varepsilon\to 0}\left\|x_{\alpha(\varepsilon)}^{\varepsilon,r} - \tilde{x}\right\| = 0$, якщо $\alpha(\varepsilon)\to 0$, $r(\varepsilon)\to 0$, $\dfrac{\varepsilon + r(\varepsilon)}{\alpha(\varepsilon)} \to 0$ при $\varepsilon \to 0$.

25.3. Канонічна задача опуклого програмування (ОП)

Коректність сильна та слаба. Стандартна постановка задачі ОП записується в наступному вигляді:

$$\min\{f(x) : f_i(x) \le 0, i = 1,2,...,m\} = F, \qquad (25.8)$$

де f, f_i – опуклі функції, а $x \in R^n$. Більш загальне формулювання запису передбачає також поряд з обмеженнями типу нерівностей також обмеження вигляду $x \in S$, де S – опукла множина. Для простоти викладу обмежимося частинним випадком (25.8) – $S \equiv R^n$. Відповідна задача з наближеними даними набуває вигляду:

$$\min\{f^\varepsilon(x) : f_i^\varepsilon(x) \le 0, i = 1,2,...,m\} = F^\varepsilon, \qquad (25.9)$$

де

$$\left|f^\varepsilon(x) - f(x)\right| \le \varepsilon, \ \left|f_i^\varepsilon(x) - f_i(x)\right| \le \varepsilon, \ \forall x \in R^n. \qquad (25.10)$$

Через Ω^ε, M^ε позначимо допустиму та оптимальну множину в задачі (25.9).

Оскільки на відміну від попередніх пунктів дозволяється задання функціоналів з похибкою, які визначають допустиму множину, то

доцільно ввести декілька інших визначень коректності [45], які враховують специфіку задачі.

Визначення 25.1. Задача (25.8) називається слабо коректною, якщо $\lim\limits_{\varepsilon\to 0}\sup\limits_{y^\varepsilon\in M^\varepsilon}\inf\left\|y^\varepsilon - y\right\| = 0$ або рівнозначно: для будь-якого $\delta > 0$ існує $\varepsilon_0 > 0$ таке, що для будь-якого $\varepsilon \in (0,\varepsilon_0)$ і будь-якого $y^\varepsilon \in M^\varepsilon$ знайдеться $y \in M$, яке задовольняє нерівності $\left\|y^\varepsilon - y\right\| < \delta$.

Звідси випливає близькість оптимальних значень $\left|f^\varepsilon(y^\varepsilon) - f(y)\right| < \delta_1(\varepsilon)$. З цієї причини слаба коректність відповідає стійкості (неперервній залежності) по функціоналу.

Визначення 25.2. Задача (25.8) називається сильно коректною, якщо

$$\lim_{\varepsilon\to 0}\sup_{y',y''\in M\cup M^\varepsilon}\left\|y' - y''\right\| = 0$$

або еквівалентно: для будь-якого $\varepsilon > 0$ існує ε_0 таке, що для будь-якого $\varepsilon \in (0,\varepsilon_0)$ та будь-яких $y \in M$, $y^\varepsilon \in M^\varepsilon$ виконується $\left\|y^\varepsilon - y\right\| < \varepsilon$.

Із цього визначення випливає, що M – множина одноточкова і M_ε стягується в одну точку при $\varepsilon \to 0$.

Приклад 25.1.

$$\min\{-(x_1 + x_2) :\ x_1 + x_2 \le 1, x_1 \ge 0, x_2 \ge 0\} = F,$$

$$\min\{-(x_1 + x_2) : (1 + |\varepsilon| + \varepsilon)x_1 + (1 + |\varepsilon| - \varepsilon)x_2 \le 1, x_1 \ge 0, x_2 \ge 0\} = F^\varepsilon$$

При $\varepsilon < 0$, $M^\varepsilon = \{(1,0)\}$, $F^\varepsilon = -1$.

При $\varepsilon > 0$, $M^\varepsilon = \{(0,1)\}$, $F^\varepsilon = -1$, зрозуміло також, що $F = -1$.

Тому задача слабо коректна, але не сильно коректна (множина M – не одноточкова).

У наступному прикладі розглянута задача лінійного програмування (ЗЛП), яка не є слабо коректною, тим більше сильно коректною.

Приклад 25.2. Розглянемо задачу

$$\min\{-x_2 : 0 \le x_1 \le 1, x_1 - x_2 \le 0, -x_1 + x_2 \le 0\} = F,$$

Відповідна їй задача з наближеними даними має вигляд:

$$\min\{-x_2 : 0 \le x_1 \le 1, (1 + \varepsilon)x_1 - x_2 \le 0, (1 - \varepsilon)x_1 + x_2 \le 0\} = F^\varepsilon$$

Тут $F = -1$, $M = \{(1,1)\}$, $F^\varepsilon = 0$, $M^\varepsilon = \{(0,0)\}$.

Оскільки $F^\varepsilon \Rightarrow F$, $\lim\limits_{\varepsilon\to 0}\left\{\sup\limits_{y^\varepsilon\in M^\varepsilon}\inf\limits_{y\in M}\left\|y^\varepsilon - y\right\|\right\} \Rightarrow 0$, то задача не слабо коректна.

Обґрунтуємо для задачі (25.8) аналог теореми 24.1. при більш загальних допущеннях на f^ε, але при $\Omega \equiv \Omega^\varepsilon$, то є $f_i(x) \equiv f_i^\varepsilon(x)$.

Теорема 25.3. Нехай f – неперервна і сильно опукла функція на Ω, f^ε – неперервна (не обов'язково опукла), задовольняюча умовам (25.10), і $\lim\limits_{\varepsilon\to 0}\eta(\varepsilon)=0$. Тоді задача (25.8) має єдиний розв'язок y і $\limsup\left\|\widetilde{y}^\varepsilon - y\right\| = 0$, де $\widetilde{M}_r = \left\{\widetilde{y}_\varepsilon : f^\varepsilon(\widetilde{y}_\varepsilon) \le F^\varepsilon + \eta(\varepsilon)\right\}$.

Доведення. Так як із (25.10) видно, що $f^\varepsilon(x) \ge f(x) - \varepsilon$, то $\lim\limits_{\|x\|\to\infty} f^\varepsilon(x) = \infty$ і, значить, множина $Q = \left\{x : f^\varepsilon(x) \le \varepsilon\right\}$ обмежена і замкнута. Нехай $y \in M^\varepsilon$, $y \in M$, тоді з урахуванням $f^\varepsilon(y^\varepsilon) \le f^\varepsilon(y)$:

$$f(\widetilde{y}_\varepsilon) - f(y) \le \left[f(\widetilde{y}^\varepsilon) - f^\varepsilon(\widetilde{y}^\varepsilon)\right] + \left[f^\varepsilon(\widetilde{y}^\varepsilon) - f^\varepsilon(y^\varepsilon)\right] + \left[f^\varepsilon(y^\varepsilon) - f(y)\right] \le$$
$$\le \varepsilon + \eta(\varepsilon) + \left[f^\varepsilon(y) - f(y)\right] \le 2\varepsilon + \eta(\varepsilon)$$

Враховуючи (24.3), знаходимо

$$\sup_{\widetilde{y}_\varepsilon \in \widetilde{M}_\varepsilon} \left\|\widetilde{y}^\varepsilon - y\right\|^2 \le \frac{2}{r}\left[f(\widetilde{y}^\varepsilon) - f(y)\right] \le \frac{4\varepsilon + 2\eta(\varepsilon)}{r} \to 0$$

Зауваження 25.2. Із тверджень Теорем 24.1, 25.1, 25.2, 25.3 випливає, що задачі (24.1), (25.8) коректні по Тихонову і стійкі до збурень цільового функціоналу при фіксованій допустимій множині. З другого боку, згідно з Твердження 24.2, коректність по Тихонову тягне за собою коректність за Адамаром, то є стійкість розв'язку до збурень допустимої множини відносно збіжності за Моско. Але для задачі ОП (25.8) малі збурення функції f можуть сильно змінити допустиму множину і , значить, оптимальне значення і оптимальну множину. Тому задача буде некоректною в сенсі Визначень 25.1, 25.2, незважаючи на те що функціонал сильно опуклий, що гарантує коректність за Тихоновим. Наприклад:

Приклад 25.3.

$$\min\left\{\|x\|^2 : 0 \le x_1 \le 1, x_1 - x_2 \le 0, -x_1 + x_2 \le 0\right\} = F,$$

де $M = \{(0,0)\}$, $F = 0$. Збурена задача має вигляд:

$$\min\left\{\|x\|^2 : 0 \le x_1 \le 1, x_1 - (1-\varepsilon)x_2 - \varepsilon \le 0, -x_1 + (1+\varepsilon)x_2 - \varepsilon \le 0\right\} = F^\varepsilon,$$

для якої $M^\varepsilon = \{(1,1)\}$, $F^\varepsilon = 2$, для $\varepsilon > 0$.

Таким чином, задача не є коректною ні в слабому, ні в сильному сенсі (Визначення 25.1, 25.2), хоч $f(x) = \|x\|^2$ суворо опуклий функціонал.

Цей приклад засвідчує, між іншим, що для задачі із прикладу 25.2. стандартна схема регуляризації у формі $\min\left\{f^{\varepsilon}(x)+\alpha\|x\|^2 : x\in\Omega^{\varepsilon}\right\}$ не приводить до мети і не породжує стійкого регуляризуючого алгоритму (див. Рис. 25.1).

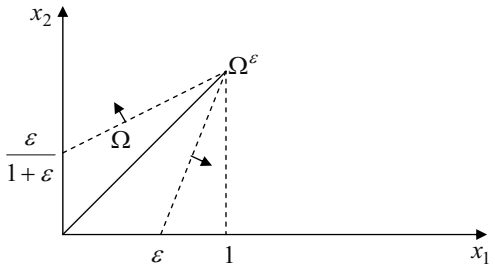

Рис. 25.1

Для побудови стійких алгоритмів розв'язку таких задач знадобляться більш складні схеми регуляризації, які включають в себе поряд з тихоновською ідеєю методи штрафних функцій (МШФ). Тому в наступному пункті 25.4 дамо основи цього методу, а в пункті 25.5 опишемо загальну схему регуляризації.

25.4. Основи методу штрафних функцій

Перехід від оптимізаційної задачі з обмеженнями до задачі без обмежень можна здійснити різними методами. Найпростіші обмеження можна зняти заміною змінних:

$|x|\le 1 \Rightarrow x=\sin t \ \ (-\infty<t<\infty)$

$0\le x<\infty \Rightarrow x=t^2$ або $x=2^t \ \ (-\infty<t<\infty)$

$a\le x\le b \Rightarrow x=a+(b-a)\cos^2 t \ \ (-\infty<t<\infty)$, a і b – дійсні числа.

Очевидно, що такий підхід не є універсальним і не годиться в загальному випадку опуклих обмежень задачі (25.8).

Один із варіантів МШФ – метод зовнішньої точки – зводиться до такого перетворення цільової функції $f(x)$, при якому значення перетвореної цільової функції в допустимій зоні Ω точно або наближено рівне $f(x)$, в той час як значення поза зоною Ω дуже велике в порівнянні із значеннями $f(x)$ (дуже великий штраф за межею зони).

Загальна схема МШФ для задачі (25.8) виражається в наступному: вибирається функція $\psi(x)$, яка володіє властивістю:

$$\psi(x)=\begin{cases}>0,\ x\notin\Omega\\ 0,\quad x\in\Omega\end{cases} \tag{25.11}$$

і утворюється функція $f_\beta(x)=f(x)+\dfrac{\psi(x)}{\beta}$, де $\beta\to 0$ $(\beta>0)$, і називається параметром штрафу. Вихідній задачі (25.8) зіставляється задача на безумовний мінімум

$$\min\left\{f(x)+\frac{\psi(x)}{\beta}: x\in R^n\right\}=F_\beta \tag{25.12}$$

Розв'язок x_β і оптимальне значення F_β цієї задачі розглядається в якості апроксимації розв'язків вихідної задачі (25.8) і значення F.

Термін „метод штрафних функцій" можна обґрунтувати наступним чином: коли x порушує одне або декілька обмежень, на цільову функцію накладається штраф величиною $\dfrac{\psi(x)}{\beta}$, збільшуючи її значення.

Розглянемо більш загальну задачу, ніж (25.8), додавши обмеження типу рівностей:

$$\min\{f(x): f_i(x)\le 0, i=1,2,...,S, f_i(x)=0, i=S+1,S+2,...,m\} \tag{25.13}$$

вважаючи f_i $(i=S+1,S+2,...,m)$ такими, що Ω – опукла множина. Задамо функцію $\psi(x)$ в наступній формі:

$$\psi(x)=\sum_{i=1}^{S}\left[\max\left(f_i(x),0\right)\right]^P+\sum_{i=S+1}^{m}\left|f_i(x)\right|^P\quad P>0$$

і сформулюємо задачу (25.12), в якій позначимо

$$f(x)+\frac{\psi(x)}{\beta}=\Phi(x,\beta).$$

Теорема 25.4. Нехай f, f_i $(i=1,2,...,m)$ – неперервні функції, $x\in R^n$ і $\lim\limits_{\|x\|\to\infty} f(x)=\infty$. Тоді задачі (25.12), (25.13) мають розв'язок і всі граничні точки послідовності розв'язків задачі (25.12) є розв'язками задачі (25.13).

Доведення. Оскільки $\lim\limits_{\|x\|\to\infty} f(x)=\infty$, а $\psi(x)\ge 0$, то також $\lim\limits_{\|x\|\to\infty}\Phi(x,\beta)=\infty$. Тому існування розв'язку доводиться стандартно. Позначимо через $\tilde{x}$, x_β ці розв'язки. Так як $\tilde{x}$ задовольняє обмеженням, то $\psi(\tilde{x})=0$ і

$$\Phi\left(x_\beta,\beta\right)=f\left(x_\beta\right)+\frac{\psi(x)}{\beta}\le f(\tilde{x}), \tag{25.14}$$

звідки $\left\|x_\beta\right\|\le const$, а значить,

$$x_{\beta_k}\to\bar{x},\ k\to\infty. \tag{25.15}$$

Внаслідок неперервності f, f_i

$$f(\bar{x}) = \lim_{k\to\infty} f(x_{\beta_k}) \le \lim_{k\to\infty} \sup\{\Phi(x_{\beta_k}, \beta_k)\} \le f(\tilde{x}),$$

звідки $f(x_{\beta_k}) \ge f(\bar{x}) - c$ ($c = const$), що разом з (2.15) тягне оцінку

$$\frac{\psi(x_{\beta_k})}{\beta} \le f(\tilde{x}) - f(\bar{x}) + c < \infty.$$

Для кожного f_i, $i = 1,2,...,S$ справедливо

$$f_i(\bar{x}) = \lim_{k\to\infty} f_i(x_{\beta_k}) \le \lim_{k\to\infty} \sup\left[\psi(x_{\beta_k})\right]^{\frac{1}{P}} \le$$

$$\le \lim_{k\to\infty} \sup \beta_k^{\frac{1}{P}} \left[f_i(\tilde{x}) - f_i(\bar{x}) + c\right]^{\frac{1}{P}} = 0, \qquad (25.16)$$

а при $i = S+1, S+2, ..., m$

$$0 < |f_i(\bar{x})| = \lim_{k\to\infty} |f_i(x_{\beta_k})| \le \lim_{k\to\infty} \left[\psi(x_{\beta_k})\right]^{\frac{1}{P}} = 0. \qquad (25.17)$$

Об'єднуючи (25.16), (25.17), робимо висновок, що $\bar{x} \in \Omega$ – допустимій множині задачі (25.13). Враховуючи встановлену вище нерівність $f(\bar{x}) \le f(\tilde{x})$, одержуємо, що $\bar{x}$ – розв'язок задачі (25.13). Крім того,

$$\lim_{k\to\infty} f(x_{\beta_k}) \le \lim_{k\to\infty} F_{\beta_k} = F.$$

Разом з (25.15) доведення завершується.

25.5. Регуляризація в загальному випадку

Введемо параметричну функцію.

$$\Phi_{\alpha\beta}^{\varepsilon}(x) = \beta\left[f^{\varepsilon}(x) + \alpha\|x\|^2 + \psi^{\varepsilon}(x)\right],$$

де

$$\psi^{\varepsilon}(x) = \sum_{i=1}^{S}\left[\max\left(f_i^{\varepsilon}(x), 0\right)\right] + \sum_{i=S+1}^{m}\left[f_i^{\varepsilon}(x)\right],$$

$$\left|f^{\varepsilon}(x) - f(x)\right| \le \varepsilon, \ \left|f_i^{\varepsilon}(x) - f_i(x)\right| \le \varepsilon, \ \forall x \in R^n \qquad (25.18)$$

Позначимо

$$\psi(x) = \sum_{i=1}^{S}\left[\max\left(f_i(x), 0\right)\right] + \sum_{i=S+1}^{m}\left[f_i(x)\right]$$

Лема 25.1. Для будь-якого $x \in R^n$ $\left|\psi^{\varepsilon}(x) - \psi(x)\right| \le \varepsilon m$.

Доведення випливає з очевидних нерівностей:

$$\left|\,|f_i(x)| - |f_i^{\varepsilon}(x)|\,\right| \le |f_i(x) - f_i^{\varepsilon}(x)| \le \varepsilon$$

$$\left|\max\{f_i(x), 0\} - \max\{f_i^{\varepsilon}(x), 0\}\right| - |f_i(x) - f_i^{\varepsilon}(x)| \le \varepsilon$$

і способу задання ψ, ψ^{ε}.

Лема 25.2. Нехай f – опуклий функціонал. Тоді для будь-яких $0 < \varepsilon < \varepsilon_0 < \infty$, $0 \le \beta \le \beta_0 < \infty$ і фіксованого $\alpha > 0$

$$\Phi^* = \inf_{x \in R^n} \Phi^{\varepsilon}_{\alpha\beta}(x) > -\infty$$

Доведення. Функція $N_\alpha(x) = f(x) + \alpha\|x\|^2$ при $\alpha > 0$ сильно опукла, тому з урахуванням нерівності (24.3)

$$N^{**} = \min_{x \in R^n} N_\alpha(x) > -\infty .$$

Так як $f_\varepsilon(x) \ge f(x) - \varepsilon$ (див. 25.18), одержимо співвідношення

$$\Phi^{\varepsilon}_{\alpha\beta}(x) \ge \beta\left[f(x) - \varepsilon + \alpha\|x\|^2\right] \ge \beta N_\alpha(x) - \beta\varepsilon \ge \beta N^{**} - \beta\varepsilon > -\infty ,$$

що й забезпечує доведення.

Будемо мінімізувати функцію наближено з точністю $\eta(\varepsilon)$ і позначимо через $\tilde{y}^{\varepsilon}_{\alpha\beta}$ елемент, який задовольняє умові

$$\Phi^{\varepsilon}_{\alpha\beta}(\tilde{y}^{\varepsilon}_{\alpha\beta}) \le \Phi^* + \eta(\varepsilon),\ \eta(\varepsilon) > 0$$

Введемо позначення:

$$\hat{y} = \arg\min\left\{\|x\|^2 : y \in M\right\}$$

$$y_\alpha = \arg\min\{N_\alpha(x) : x \in \Omega\},\ N_\alpha(y_\alpha) = N^*$$

Теорема 25.5. Нехай виконані наступні умови:

а) f, f_i – опуклі неперервні функції, $x \in R^n$;

б) виконуються умови апроксимації (25.18);

в) задача (25.8) має розв'язок, то є $M \ne \varnothing$.

Тоді, якщо параметри β, η узгоджені з ε так, що виконуються умови:

$$\lim_{\varepsilon \to 0} \beta(\varepsilon) = 0,\ \lim_{\varepsilon \to 0} \frac{\varepsilon}{\beta(\varepsilon)} = 0,\ \lim_{\varepsilon \to 0} \frac{\eta(\varepsilon)}{\beta(\varepsilon)} = 0,$$

тоді

$$\lim_{\alpha \to 0} \lim_{\varepsilon \to 0} \left\| \tilde{y}^{\varepsilon}_{\alpha\beta} - \hat{y} \right\| = 0$$

Доведення. Беручи до уваги співвідношення, що

$$\psi^{\varepsilon}(x) \ge 0,\ f(x) \le f^{\varepsilon}(x) + \varepsilon,\ f^{\varepsilon}(x) \le f(x) + \varepsilon,$$

$$\Phi^{\varepsilon}_{\alpha\beta}(\tilde{y}^{\varepsilon}_{\alpha\beta}) \le \Phi^* + \eta(\varepsilon)$$

виконуються для будь-яких $x \in R^n$, одержуємо сукупність нерівностей:

$$N^{**} \le N_\alpha(\tilde{y}^{\varepsilon}_{\alpha\beta}) \le N_\alpha(\tilde{y}^{\varepsilon}_{\alpha\beta}) + \frac{\psi_\varepsilon(\tilde{y}^{\varepsilon}_{\alpha\beta})}{\beta} = f(\tilde{y}^{\varepsilon}_{\alpha\beta}) + \alpha\left\|\tilde{y}^{\varepsilon}_{\alpha\beta}\right\|^2 + \frac{\psi_\varepsilon(\tilde{y}^{\varepsilon}_{\alpha\beta})}{\beta} \le$$

$$\le f(\tilde{y}^{\varepsilon}_{\alpha\beta}) + \alpha\left\|\tilde{y}^{\varepsilon}_{\alpha\beta}\right\|^2 + \frac{\psi_\varepsilon(\tilde{y}^{\varepsilon}_{\alpha\beta})}{\beta} + \varepsilon = \frac{\Phi^{\varepsilon}_{\alpha\beta}(\tilde{y}^{\varepsilon}_{\alpha\beta})}{\beta} + \varepsilon \le \frac{\Phi^* + \eta(\varepsilon)}{\beta} + \varepsilon \le$$

$$\le \frac{\Phi^{\varepsilon}_{\alpha\beta}(y_\alpha)}{\beta} + \frac{\eta(\varepsilon)}{\beta} + \varepsilon = f^{\varepsilon}(y_\alpha) + \alpha\|y_\alpha\|^2 + \frac{\psi^{\varepsilon}(y_\alpha)}{\beta} + \frac{\eta(\varepsilon)}{\beta} + \varepsilon \quad (25.19)$$

Так як $y_\alpha \in \Omega$, то $\psi(y_\alpha) = 0$ і, значить, з урахуванням леми 25.1 $\psi^{\varepsilon}(y_\alpha) \le \varepsilon m$, крім того, $f^{\varepsilon}(y_\alpha) \le f(y_\alpha) + \varepsilon$, то

$$N^{**} \le N_\alpha\left(\tilde{y}^{\varepsilon}_{\alpha\beta}\right) \le f(y_\alpha) + \alpha\|y_\alpha\|^2 + \frac{\varepsilon m}{\beta} + \frac{\eta(\varepsilon)}{\beta} + 2\varepsilon = N^* + \frac{\varepsilon m}{\beta} + \frac{\eta(\varepsilon)}{\beta} + 2\varepsilon \quad (25.20)$$

Звідси, а також з умов теореми випливає, що $f\left(\tilde{y}^{\varepsilon}_{\alpha\beta}\right) + \alpha\left\|\tilde{y}^{\varepsilon}_{\alpha\beta}\right\|^2 \le c < \infty$, а оскільки N_α суворо опукла функція, то $\left\|\tilde{y}^{\varepsilon}_{\alpha\beta}\right\|^2 \le c_1$ ($c_1 = const$, α фіксовано). З нерівності

$$N^{**} \le N_\alpha\left(\tilde{y}^{\varepsilon}_{\alpha\beta}\right) \le N_\alpha\left(\tilde{y}^{\varepsilon}_{\alpha\beta}\right) + \frac{\psi^{\varepsilon}\left(\tilde{y}^{\varepsilon}_{\alpha\beta}\right)}{\beta} = N^* + \frac{\varepsilon m}{\beta} + \frac{\eta(\varepsilon)}{\beta} + 2\varepsilon \quad (25.21)$$

яка є наслідком нерівностей (25.19), (25.20), випливає, що $\lim\limits_{k\to\infty} \psi^{\varepsilon_k}\left(\tilde{y}^{\varepsilon_k}_{\alpha\beta_k}\right) = 0$.

Із леми 25.1 одержуємо

$$\psi(\tilde{y}) = \lim_{k\to\infty} \psi\left(y^{\varepsilon_k}_{\alpha\beta_k}\right) \le \lim_{k\to\infty} \left[\psi^{\varepsilon_k}\left(y^{\varepsilon_k}_{\alpha\beta_k}\right) + \varepsilon_k m\right] = 0,$$

то є $\tilde{y} \in \Omega$. Переходячи в (25.21) до границі при $\varepsilon_k \to 0$, маємо

$$N_\alpha(\tilde{y}) \le N^* = \min\{N_\alpha(x) : x \in \Omega\} = N_\alpha(y_\alpha),$$

це означає, що $N_\alpha(\tilde{y}) = N_\alpha(y_\alpha)$. Із єдиної точки мінімуму для сильно опуклого функціоналу можемо зробити висновок, що $\tilde{y} = y_\alpha$. Оскільки встановлено, що будь-яка гранична точка $\left\{\tilde{y}^{\varepsilon}_{\alpha\beta}\right\}$ збігається з y_α, то $\lim\limits_{\varepsilon\to 0} \tilde{y}^{\varepsilon}_{\alpha\beta} = y_\alpha$. Згідно з теоремою 25.1 $\lim\limits_{\alpha\to 0} y_\alpha = \hat{y}$, тому

$$\lim_{\alpha\to 0}\lim_{\varepsilon\to 0}\left\|\tilde{y}^{\varepsilon}_{\alpha\beta} - \hat{y}\right\| = 0.$$

Зауваження 25.3. Із нерівності

$$\left\|\hat{y} - \tilde{y}^{\varepsilon}_{\alpha\beta}\right\| \le \|\hat{y} - y_\alpha\| + \left\|y_\alpha - \tilde{y}^{\varepsilon}_{\alpha\beta}\right\|$$

випливає, що для будь-якого $\delta > 0$ знайдуться такі $\alpha(\delta)$, $\varepsilon(\delta)$, що

$$\left\|\hat{y} - y^{\varepsilon(\delta)}_{\alpha(\delta)}\beta(\varepsilon(\delta))\right\| < \delta$$

Таким чином, має місце апроксимація нормального розв'язку $\hat{y}$, з елементами $y^{\varepsilon}_{\alpha\beta}$, стійка по відношенню до збурень вихідних даних.

§26. ДИСКРЕТИЗАЦІЯ ОПТИМАЛЬНИХ ЗАДАЧ. ДИСКРЕТНА АПРОКСИМІЗАЦІЯ І ДИСКРЕТНА ЗБІЖНІСТЬ

26.1. Постановка задачі

Основна задача на оптимум, сформульована в §24, є, взагалі кажучи, нескінченномірною в тому сенсі, що аргумент x – елемент нескінченномірного простору, а обчислення цільового функціонала на цьому аргументі не можна реалізувати за кінцеве число арифметичних операцій, наприклад

$$\min \to f(x) = \int_0^1 \left[\left(\frac{dx(t)}{dt} \right)^2 + g(t)x^2(t) - 2p(t)x(t) \right] dt \tag{26.1}$$

при $x \in W_2^1[0,1]$, $x(0) = x(1) = 0$, де $W_2^1[0,1]$ – простір функцій з нормою

$$\|x\|^2_{W_2^1} = \int_0^1 \left(\frac{dx(t)}{dt} \right)^2 dt < \infty$$

Замінимо похідну кінцево-різницевими співвідношеннями, а до інтервалу застосуємо квадратурну формулу прямокутників, тоді, попадаючи $h_i = t_i - t_{i-1}$, маємо

$$\min \to f_n(x_n) = \sum_{i=1}^n h_i \left[\left(\frac{x(t_i) - x(t_{i-1})}{h_i} \right)^2 + g(t_i)x^2(t_i) - 2p(t_i)x(t_i) \right] \tag{26.2}$$

при $x_n \in W_2^{1,n}$, $x(0) = x(1) = 0$, $\|x_n\| = \sum_{i=1}^n h_i \left(\frac{x_{n,i} - x_{n,i-1}}{2} \right)^2$.

Для загальної задачі мінімізації

$$\min\{f(x) : x \in \Omega\} = F > -\infty, \tag{26.3}$$

під дискретною апроксимацією будемо розуміти послідовність кінцево-вимірних задач

$$\min\{f_n(x_n) : x_n \in \Omega_n\} = F_n > -\infty, \tag{26.4}$$

де f_n – звичайні функції n змінних, Ω_n – підмножина евклідового простору. Необхідно встановити умови, за яких має місце збіжність $F_n \to F$ і оптимальних (або ε-оптимальних) множин $M_n \to M$ ($M_n^\varepsilon \to M$).

26.2. Основні визначення і поняття

Перш ніж викласти загальну схему дискретної оптимізації, введемо деякі поняття, на слідуючи [48], [49].

Розглянемо поряд з лінійно-нормованим простором (ЛНП) X послідовність кінцево-вимірних ЛНП $\{X_n\}$ (які не обов'язково будуть підпросторами X). Перехід із простору X в простір X_n

здійснюється за допомогою зв'язуючих операторів (операторів звуження) $P_n\colon X - X_n$, $P_n X = X_n$, які задовольняють умовам

$$\forall x \in X \quad \lim_{n\to\infty} \|p_n x\|_{X_n} = \|x\|_X \tag{26.5}$$

$$\forall x,\ x',\ \forall a, a' = const \quad \lim_{n\to\infty} \|P_n(ax - a'x') - ap_n x - a' p_n x'\|_{X_n} = 0. \tag{26.6}$$

Визначення 26.1. Послідовність з ЛНП $\{X_n\}$ утворює дискретну апроксимацію ЛНП X, якщо існує сімейство зв'язуючих операторів $\{P_n\}$, які задовольняють властивостям (26.5), (26.6).

В подальшому для спрощення запису будемо опускати індекси при нормах, якщо це не викликає непорозумінь.

Визначення 26.2. Послідовність $\{X_n\}$, $x_n \in X_n$, називається дискретно збіжною до X (будемо позначати $X_n \to X$), якщо $\lim_{n\to\infty} \|x_n - p_n x\| = 0$.

Нехай X, X_n – гільбертові простори, в яких задані скалярні добутки $<,>$, $<,>_n$ відповідно.

Визначення 26.3. Послідовність $\{X_n\}$ дискретно слабо збігається до X (позначимо $X_n \to X$), якщо $\lim_{n\to\infty} \langle x_n, v_n \rangle_n = \langle x, v \rangle$, для будь-якої довільної дискретно збіжної послідовності $V_n \to V$.

Для оцінювання похибки наближеного розв'язку за нормою простору X визначають оператори поповнення $r_n : X_n \to X$. В необхідних випадках на оператори r_n накладаються деякі умови.

Визначення 26.4. Послідовність $\{r_n\}$ утворює сімейство сильних (слабких) операторів поповнення, якщо $r_n : X_n \to X$ лінійно обмежені і задовольняють співвідношенню

$$X_n \to X\,(X_n - \to X) \Rightarrow \lim_{n\to\infty} \|r_n x_n - x\|_{X_n} = 0 \ \ (r_n x_n \to x).$$

Пояснимо на прикладах сенс введених операторів.

Приклад 26.1.

Нехай $X = C[0,1]$ – простір неперервних функцій на відрізку $[0,1]$, $\|x\|_C = \max_{0\le t\le 1} |x(t)|$. На відрізку $[0,1]$ задамо розбиття

$$T_n : 0 = t_0^n < t_1^n < \ldots < t_n^n = 1 \ \ h_i^n = t_i^n - t_{i-1}^n \ \ \max_{n\to\infty} h_i^n \to 0.$$

Визначимо $X_n = C_n = \{x_n : x_n = (x_{n1}, x_{n2}, \ldots, x_{nn})\}$

$$\|x_n\|_{X_n} = \max_{1\le i\le n} |x_{ni}| \quad p_n x = \left(x\left(t_1^n\right), x\left(t_2^n\right), \ldots, x\left(t_n^n\right)\right).$$

Приклад 26.2.

$$X = L_2[0,1], \qquad \|x_n\|^2 = \sum_{i=1}^{n} h_i^n |x_{ni}|^2, \qquad h_i^n = t_i^n - t_{i-1}^n, \qquad \lim_{n\to\infty} \max_{1\le i\le n} h_i^n = 0,$$

$$P_n X = \left\{ \frac{1}{h_1^n} \int_{t_0=0}^{t_1} x(s)ds, ..., \frac{1}{h_n^n} \int_{t_{n-1}}^{t_n=1} x(s)ds \right\}.$$

Приклад 26.3.

$$X = W_2^1[0,1], \quad \|x\|^2 = \int_0^1 \left|\frac{dx}{dt}\right|^2 dt + \int_0^1 |x(t)|^2 dt, \quad X_n = \{x_n : x_n = (x_{n0}, x_{n1}, ..., x_{nn})\},$$

$$\|x_n\|^2 = \sum_{i=1}^{n} h_i^n |x_{n,i}|^2 + \sum_{i=1}^{n} h_n^n \left| \frac{x_{n,i} - x_{n,i-1}}{h_i^n} \right|^2.$$

Побудована $\{X_n\}$ утворює дискретну апроксимізацію відповідних просторів X. Для прикладу 26.1 це очевидно слідує із неперервності функції.

Щоб перевірити властивості (26.5), (26.6), для сімейства $\{p_n^i\}$ введемо ще одне визначення і встановимо декілька допоміжних тверджень.

Визначення 26.5. Дві системи операторів $\mathrm{P} = \{p_n\}$ і $\mathrm{P}' = \{p_n'\}$, які задовольняють властивостям (26.1), (26.2), називаються еквівалентними, якщо

$$\lim_{n\to\infty} \|p_n x - p_n' x\| = 0 \quad \forall x \in X.$$

Очевидно, що системи P і P′ еквівалентні тоді і тільки тоді, коли виконано співвідношення

$$\lim_{n\to\infty} \|p_n x - x_n\| = 0 \Leftrightarrow \lim_{n\to\infty} \|p_n' x - x_n\| = 0.$$

Лема 26.1. Нехай $\mathrm{P} = \{p_n\}$, $\mathrm{P}' = \{p_n'\}$ – дві системи операторів, які задовольняють властивостям (26.5), (26.6), і $\Phi \subset X$ щільна множина в X, то є $\overline{\Phi} = X$. Якщо

$$\lim_{n\to\infty} \|p_n x - p_n' x\| = 0 \quad \forall x \in \Phi,$$

то системи P і P′ еквівалентні.

Доведення. Для довільного $x \in X$ вибираємо $x_\varepsilon \in \Phi$ так, щоб $\|x - x_\varepsilon\| < \varepsilon$. Тоді для достатньо великого n має місце оцінка

$$\|p_n' x - p_n x\| \le \|p_n' x - p_n' x_\varepsilon\| + \|p_n' x_\varepsilon - p_n x_\varepsilon\| + \|p_n x_\varepsilon - p_n x\| \le 3\varepsilon,$$

оскільки на підставі (26.5), (26.6)

$$\|p_n' x - p_n' x_\varepsilon\| - \|p_n'(x - x_\varepsilon)\| < \|(p_n x - p_n' x_\varepsilon) - p_n'(x - x_\varepsilon)\| \to 0$$

$$\lim_{n\to\infty} \|p_n'(x - x_\varepsilon)\| = \|x - x_\varepsilon\| < \varepsilon.$$

Теорема 26.1[49]. Нехай $\mathrm{P}=\{p_n\}$ система операторів $p_n\colon X'\to X_n$, де X' щільне в X і p_n задовольняють умовам (26.5), (26.6) на X'. Тоді оператори p_n можна поширити на весь простір X зі збереженням властивостей (26.5), (26.6), при тому, якщо $\mathrm{P}'=\{p_n'\}$ і $\mathrm{P}''=\{p_n''\}$ – довільні так продовжені оператори, то ці системи еквівалентні.

Лема 26.2. Сімейства зв'язуючих операторів $\mathrm{P}=\{p_n\}$, $\mathrm{P}'=\{p_n'\}$ із прикладів 26.1, 26.2 еквівалентні і, значить, визначають еквівалентну дискретну збіжність.

Доведення. Оператори p_n лінійні і обмежені

$$\|p_n'x\|_{l_2^n}=\left(\sum_{i=1}^{n}\frac{h_i^n}{\left|h_i^n\right|^2}\left|\int_{t_{i-1}^n}^{t_i^n}x(s)ds\right|^2\right)^{\frac{1}{2}}\le\left(\sum_{i=1}^{n}\frac{h_i^n}{\left|h_i^n\right|^2}\int_{t_{i-1}^n}^{t_i^n}|x(s)|^2h_i^nds\right)^{\frac{1}{2}}=\|x\|_{L_2},\ \|p_n\|\le 1$$

Для кожного $x(t)\in C[a,b]$ справедливо

$$\|p_nx-p_n'x\|_{l_2^n}\le\|p_nx-p_n'x\|_{C_n}\to 0$$

$$\|p_nx\|_{l_2^n}\to\|x\|_{L_2}$$

(останнє співвідношення випливає із збіжності квадратурної формули прямокутників для неперервної функції $|x(t)|^2$).

Значить, $\|p_nx\|_{l_2^n}\to\|x\|_{L_2}\quad\forall x\in C[0,1]$, а так як $\|p_n'\|\le 1$, то збіжність зберігається для будь-якого $x\in X=L_2$. На підставі справедливості тверджень Леми 26.1 і Теореми 26.1 системи P і P′ еквівалентні.

При дискретизації задачі на мінімум (26.3) від множини $\Omega\subset X$ переходимо до множини Ω_n і кінцево-вимірного простору X_n. У наступному прикладі випишемо деякі множини Ω і апроксимуючі їх Ω_n.

Приклад 26.4.

$$\Omega=\{x\in L_2[a,b]\colon x(t)\ge 0\},\ \Omega_n=\{x_n\in l_2^n\colon x_{ni}\ge 0\}\qquad(26.7)$$

$$\Omega=\left\{x\in W_2^1[a,b]\colon\frac{dx}{dt}\ge 0\right\},\ \Omega_n=\{x_n\in W_2^{1,n},x_{n,i}-x_{n,i-1}\ge 0\}\qquad(26.8)$$

$$\Omega=\{x\in L_2[a,b]\colon\|x\|_{L_2}\le r\},\ \Omega_n=\{x_n\in l_2^n\colon\|x_n\|_{l_2^n}\le r\}\qquad(26.9)$$

В якості прикладу задання r_n можна розглянути оператор кусково-постійного відтворення

$$r_n\colon x_n=(x_{n1},x_{n2},\ldots,x_{nn})\to\hat{x}_n(t)=(x_{ni}\ \ при\ t_{i-1}\le t\le t_i)\qquad(26.10)$$

або оператори кусково-лінійного відтворення

$$r_n\colon x_n=(x_{n0},x_{n1},\ldots,x_{nn})\to$$

$$\to\hat{x}_n(t)=\left(\frac{x_{ni}-x_{ni-1}}{t_i-t_{i-1}}t+\frac{x_{ni}t_{i-1}-x_{ni-1}t_i}{t_{i-1}-t_i},t_{i-1}\le t\le t_i\right)\qquad(26.11)$$

§27. ДОСТАТНІ УМОВИ ЗБІЖНОСТІ

Будемо розв'язувати задачу (26.4) наближено з точністю ε_n і позначимо через $x_n^{\varepsilon_n}$ елемент із Ω, який задовольняє співвідношення

$$f_n\left(x_n^{\varepsilon_n}\right) \le F_n + \varepsilon_n, \varepsilon_n \to 0, n \to \infty. \tag{27.1}$$

Теорема 27.1. Нехай X – гільбертів простір, $\{X_n\}$ – послідовність кінцево-вимірних просторів, f – слабо напівнеперервний знизу функціонал, Ω – замкнута випукла обмежена множина із X. Нехай виконані наступні умови:

1. $\forall n\ p_n\Omega \subseteq \Omega$, де $p_n : X \to X_n$ – зв'язуючі оператори;
2. $\forall n\ r_n\Omega \subseteq \Omega$, де $r_n : X_n \to X$ – оператори поповнення;
3. для деякого $\hat{x} \in M$ $\limsup\limits_{n\to\infty} f_n(p_n\hat{x}) \le f(\hat{x})$;
4. $\forall\{X_n\},\ X_n \in \Omega_n$ $\limsup\limits_{n\to\infty}\left[f(r_n x_n) - f_n(x_n)\right] \le 0$.

Тоді задача (26.3) розв'язувана, то є $M \neq \varnothing$ і всі слабо граничні точки послідовності $\left\{r_n x_n^{\varepsilon_n}\right\}$, які визначаються формулою (27.1), належать множині M задачі (26.3). Крім того,

$$\lim_{n\to\infty} f\left(r_n x_n^{\varepsilon_n}\right) = \lim_{n\to\infty} f_n\left(x_n^{\varepsilon_n}\right) = F. \tag{27.2}$$

Доведення. Розв'язуваність задачі встановлюється за стандартною схемою. Через слабку компактність обмеженої множини в гільбертовім просторі існує $\left\{x_{n_k}^{\varepsilon n_k}\right\}$, $r_{n_k}^{\varepsilon n_k} \to \bar{x}$, при цьому $\bar{x} \in \Omega$, бо множина Ω – випукла і замкнута, а значить, слабо замкнута і згідно з умовою 2) $r_n x_n^{\varepsilon} \in \Omega$. Враховуючи умови 3), 4) визначення $x_n^{\varepsilon_n}$ і властивості слабої неперервності f, маємо

$$F \le f(\bar{x}) \le \limsup_{k\to\infty} f\left(r_{n_k} x_{n_k}^{\varepsilon_{n_k}}\right) \le \limsup_{k\to\infty}\left|f\left(r_{n_k} x_{n_k}^{\varepsilon_{n_k}}\right) - f_{n_k}\left(x_{n_k}^{\varepsilon_{n_k}}\right)\right| + \limsup_{k\to\infty}\left|f_{n_k}\left(x_{n_k}^{\varepsilon_{n_k}}\right)\right| \le$$

$$\le \limsup_{k\to\infty}\left(F_{n_k} + \varepsilon_{n_k}\right) \le \lim_{k\to\infty} f_{n_k}\left(P_{n_k}\hat{x}\right) \le f(\hat{x}) = F,\ \hat{x} \in M,$$

звідки $\hat{x} \in M$, що й завершує доведення.

Зауваження 27.1. В умові 3) вимогу „для деякого $\hat{x} \in M$" можна замінити на „для будь-якого $x \in \Omega$", а якщо функціонал f неперервний, то достатньо вимагати „для будь-якого $x \in \Omega'$, де $\Omega' = \Omega$".

Зауваження 27.2. При відповідному виборі операторів P_n і r_n (див. приклади 26.1, 26.2) для пар множин Ω, Ω_n на (26.7)–(26.9) виконуються умови теореми 27.1.

Для задачі безумовної мінімізації, то є для $\Omega = X$, $\Omega_n = X_n$, Теорема 27.1 незастосовна, оскільки не виконуються допущення про обмеженість множини Ω. Для цього випадку необхідні додаткові вимоги на цільові функціонали. Тому подано в задачі (26.3) $\Omega = X$, а в задачі (26.4) $\Omega_n = X_n$.

Теорема 27.2. Нехай X – гільбертовий простір, X_n – послідовність гільбертових просторів. Нехай f – слабо напівнеперервний знизу функціонал і виконані наступні умови:

1) $\lim\limits_{m\to\infty} f\left(x^m\right) = 0$, коли $\lim\limits_{m\to\infty} \sup \left\|x^m\right\| = \infty$;

2) $\lim\limits_{n\to\infty} \sup f_n(x_n) = \infty$, коли $\lim\limits_{n\to\infty} \sup \|x_n\| = \infty$;

3) існує $a > 0$ таке, що $\|r_n x_n\| \le a\|x_n\|$ для будь-яких $x_n \in X_n$;

4) $\lim\limits_{n\to\infty} \sup\left[f(r_n x_n) - f_n(x_n)\right] \le 0$ для будь-яких обмежених послідовностей $\{r_n x_n\}$;

5) $\lim\limits_{n\to\infty} f_n(P_n\hat{x}) \le f(\hat{x})$ для деякого $\hat{x} \in M$.

Тоді справедливе твердження Теореми 27.1.

Доведення. Існування розв'язку задачі (26.3) доводиться подібно лемі 24.5. Нехай елемент $x_n^{\varepsilon_n}$ визначений співвідношенням (27.1) і $\hat{x} \in M$, тоді у відповідності до умови 5)

$$\lim_{n\to\infty} \sup f_n\left(x_n^{\varepsilon_n}\right) \le \lim_{n\to\infty} \sup\left(F_n + \varepsilon_n\right) \le \lim_{n\to\infty} \sup\left(f_n\left(P_n\hat{x}\right) + \varepsilon_n\right),$$

звідки на підставі умов 1), 3)

$$\left\|x_n^{\varepsilon_n}\right\| \le c < \infty, \ \left\|r_n x_n^{\varepsilon_n}\right\| \le a\left\|x_n^{\varepsilon_n}\right\|.$$

Слаба компактність обмеженої множини в X тягне за собою $r_{n_k} x_{n_k}^{\varepsilon_{n_k}} \to \bar{x} \in X$. Притягуючи умови 4), 5) теореми і слабку напівнеперервність знизу f, одержуємо

$$F \le f(\bar{x}) \le \lim_{k\to\infty} \sup f\left(r_{n_k} x_{n_k}^{\varepsilon_{n_k}}\right) \le \lim_{k\to\infty} \sup\left[f\left(r_{n_k} x_{n_k}^{\varepsilon_{n_k}}\right) - f_{n_k}\left(x_{n_k}^{\varepsilon_{n_k}}\right)\right] + \lim_{k\to\infty} \sup\left[f_{n_k}\left(x_{n_k}^{\varepsilon_{n_k}}\right)\right] \le$$

$$\le \lim_{k\to\infty} \sup\left(F_{n_k} + \varepsilon_{n_k}\right) \le \lim_{k\to\infty} \sup\left(f_{n_k}\left(P_{n_k}\hat{x}\right) + \varepsilon_{n_k}\right) \le f(\hat{x}) = F,$$

то є $\hat{x} \in M$.

На відміну від установлених вище теорем [50] сформулюємо твердження [51] про апроксимацію задач оптимізації в термінах дискретної збіжності, в якому явно використовуються властивості зв'язуючих операторів відтворення, а також умова дискретної апроксимації простору.

Теорема 27.3. Нехай сімейство $\{x_n\}_X$ утворює дискретну апроксимацію гільбертового простору X, f – опуклий неперервний функціонал, Ω – замкнута опукла множина і виконані наступні умови:

1) для будь-якого $x \in \Omega'$, де $\Omega' = \Omega$, існує $\{x_n\}$, $x_n \in \Omega_n$, таке, що $x_n \to x$;

2) для пари x, $\{x_n\}$ із умови 1) виконано співвідношення

$$\limsup_{n\to\infty} f_n(x_n) \le f(x);$$

3) справедливе співвідношення $x_n \in \Omega_n$, $\lim_{n\to\infty} f_n(x_n) < \infty \Rightarrow$ $\Rightarrow \limsup_{n\to\infty} \|x_n\| < \infty$;

4) $\{r_n\}$ – сімейство слабих операторів відтворення, для котрого $r_n\Omega_n \subseteq \Omega$ для будь-якого n;

5) $x_n \to x \Rightarrow f(x) \le \liminf_{n\to\infty} f_n(x_n)$.

Тоді задача 26.3 розв'язувана, $\lim_{n\to\infty} F_n = F$, всі дискретно слабо граничні точки $\left\{x_n^{\varepsilon_n}\right\}$, що визначається формулою (27.1), належать M, притому, якщо $x_{n_k}^{\varepsilon_n} \to \hat{x}$, то $\hat{x} \in M$ і $r_{n_k} x_{n_k}^{\varepsilon_{n_k}} \to \hat{x}$, то є слаба в X.

Доведення з деякими змінами проводиться подібно до доведень теорем 27.1, 27.2. При цьому суттєво використовується факт дискретної слабкої компактності обмеженої послідовності $\{x_n\}$, $x_n \in X$, $\|x_n\| \le c$ ([48], [49]).

Наслідок 27.1. Якщо цільові функції мають вигляд

$$f(x) = \varphi(x) + \alpha\|x\|^2,\ f_n(x_n) = \varphi_n(x_n) + \alpha\|x_n\|^2,$$

де φ, φ_n задовольняють умовам 2), 5) теореми, тоді $\lim_{n\to\infty} F_n = F$ і $x_n^{\varepsilon} \to \hat{x}$ при $n \to \infty$, $\varepsilon_n \to 0$, а якщо додатково $\{z_n\}$ – оператори сильного відтворення, то $\lim_{n\to\infty} \sup_{x_n^{\varepsilon} \subset M_n^{\varepsilon}} \left\| r_n x_n^{\varepsilon} - \hat{x} \right\| = 0$.

Зауваження 27.3. Якщо в умовах 1), 2) теореми 27.3 змінити „$\forall x \in \Omega'$" на „для деякого $x \in M$", то вимогу неперервності f можна пропустити. Крім того, якщо $\Omega = X$, $\Omega_n = X_n$, то вимога „r_n – оператори слабого відтворення" в умові 4) зайва.

§28. ЗАСТОСУВАННЯ ПРИВЕДЕНИХ ДОСТАТНІХ УМОВ ЗБІЖНОСТІ ДО ЗАДАЧІ ВАРІАЦІЙНОГО ЧИСЛЕННЯ

28.1. Формулювання задачі

Розглянемо дещо більш загальну задачу варіаційного числення, ніж (26.1):

$$\min\left\{\int_0^1 g\left(t,x(t),\frac{dx(t)}{dt}\right): x(t)\in W_2^1[0,1], x(0)=x(1)=0\right\}, \qquad (28.1)$$

де функції $g(t,x(t),y(t))$ задовольняють наступним властивостям [52]:

а) $g(t,x(t),y(t))$ – неперервна функція в зоні

$$G=\{0\le t\le 1,-\infty<x(t),y(t)<\infty\};$$

б) існують константи $a,b>0$ такі, що $g(t,x(t),y(t))\ge a+b|y|^2$ для будь-яких кінцевих значень t, $x(t)$;

в) для будь-яких t, $x(t)$ із зони значень існує неперервна похідна $g_y(t,x(t),y(t))$, не спадаюча по y;

г) існують додатні константи c,d і неперервна функція $S(t,x(t))$, такі що $|g(t_1,x(t_1),y)-g(t_2,x(t_2),y)|\le\left(c+d|y|^2\right)|S(t_1,x(t_1))-S(t_2,x(t_2))|$

для задовільних t_1, t_2, $x(t_1)$, $x(t_2)$ із зони визначення, $S\in C$.

Будемо також вважати, що існує розв'язок задачі (28.1) $x(t)\in[0,1]$, то є функція $x(t)$ – неперервно-диференційована.

28.2. Квадратурний (кінцево-різницевий) метод

Послідовність дискретних задач, зіставлених задачі (28.1) має вигляд

$$\min\left\{\sum_{i=1}^n hg\left(t_i x_n(t_i),\frac{x_n(t_i)-x_n(t_{i-1})}{h}\right): x_n(t_i)\in W_2^{1,n}, x_n(0)=x_n(t_n)=0\right\}, \qquad (28.2)$$

де $h=t_i-t_{i-1}=\frac{1}{n}$. В якості простору X приймається $\overset{0}{W}{}_2^1[0,1]=\left\{x\in W_2^1: x(0)=x(1)=0\right\}$ з нормою $\|x\|^2=\int_0^1\left|\frac{dx(t)}{dt}\right|^2 dt$, $X_n=W_2^{1,n}$ – $n-1$-мірний простір з нормою елемента $\|x\|^2=\sum_{i=1}^n h\left|\frac{x_n(t_i)-x_n(t_{i-1})}{h}\right|^2$, p_n – оператор перенесення на сітку $\{t_i\}$, то є $p_n x(t)\to(x(t_1),x(t_2),...,x(t_{n-1}))=(x_1,x_2,...,x_{n-1})$, r_n – оператор кусково-лінійного відтворення (див. (26.10)).

Тепер займемось перевіркою умов Теореми 27.2.

Умови 1), 2) випливають із нерівностей

$$f(x)=\int_0^1 g\left(t,x(t),\frac{dx(t)}{dt}\right)dt \ge a+b\int_0^1\left|\frac{dx(t)}{dt}\right|^2 dt$$

$$f_n(x_n)=\sum_{i=1}^n hg\left(t_i,x_{ni},\frac{x_{ni}-x_{ni-1}}{n}\right)\ge\sum_{i=1}^n h(a+b)\left|\frac{x_{ni}-x_{ni-1}}{n}\right|^2,$$

котрі є наслідками властивості б)

Умова 3) випливає із нерівності

$$\|r_n x_n\|^2=\int_0^1\left|\frac{d(r_n x_n)}{dt}\right|^2 dt=\sum_{i=1}^n\int_{t_{i-1}}^{t_i}\left|\frac{d(r_n x_n)}{dt}\right|^2 dt=$$

$$=\sum_{i=1}^n h\left|\frac{x_{ni}-x_{ni-1}}{n}\right|^2=\|x_n\|^2\ \ (a=1)$$

Звернемося до властивості 5). Оскільки розв'язок $x(t)\in C^1[0,1]$, то

$$\lim_{n\to\infty}\max_{t_{i-1}\le t\le t_i}\left|\frac{dx(t)}{dt}-\frac{x(t_i)-x(t_{i-1})}{h}\right|=0,\ \lim_{n\to\infty}\max_{t_{i-1}\le t\le t_i}x(t_i)-x(t_{i-1})=0,$$

значить, із рівномірної неперервності g на обмеженій замкнутій підмножині із G права частина оцінки

$$|f(x)-f_n(p_n x)|\le\sum_{i=1}^n\int_{t_{i-1}}^{t_i}\left|g\left(t,x(t),\frac{dx(t)}{dt}\right)-g\left(t_{i-1},x(t_{i-1}),\frac{x(t_i)-x(t_{i-1})}{h}\right)\right|dt$$

буде прямувати до нуля при $n\to\infty$.

На кінець, умову 4) одержуємо із властивості 2) і ланцюга нерівностей:

$$|f_n(r_n x_n)-f_n(x_n)|\le h\sum_{i=1}^n\int_{t_{i-1}}^{t_i}\left|g\left(t,r_n x_n,\frac{d(r_n x_n)}{dt}\right)-g\left(t_i,x_{n_i},\left(\frac{x_{n_i}-x_{n_{i-1}}}{n}\right)\right)\right|dt\le$$

$$\le h\sum_{i=1}^n\int_0^1\left|g\left(t_{i-1}+\xi h_1(1-\xi)x_{n_{i-1}}+\xi x_{n_i},\left(\frac{x_{n_i}-x_{n_{i-1}}}{n}\right)\right)-g\left(t_{i-1},x_{n_{i-1}},\left(\frac{x_{n_i}-x_{n_{i-1}}}{n}\right)\right)\right|d\xi\le$$

$$\le h\sum_{i=1}^n\int_0^1\left|c+d\left(\frac{\left|x_{n_i}-x_{n_{i-1}}\right|^2}{n}\right)\right|S\left(t_{i-1}+\xi h_1(1-\xi)x_{n_{i-1}}+\xi x_{n_i}\right)-S\left(t_{i-1},x_{n_{i-1}}\right)\Big|d\xi,$$

де $\frac{t_i-t_{i-1}}{n}=\xi,\ (1-\xi)x_{n_{i-1}}+\xi x_{n_i}-x_{n_{i-1}}=\xi\left(x_{n_i}-x_{n_{i-1}}\right)\left|x_{n_i}-x_{n_{i-1}}\right|\le\sqrt{n}\|x_n\|.$

Значить, права частина останньої нерівності прямує до нуля.

Зауваження 28.1. Якщо $p(t)$, $g(t)$ - функції неперервні і $g(t)\ge\overline{g}(t)>0$, тоді підінтегральна функція в задачі (26.1) задовольняє вимогам а) – г) для $g(t,x(t),y(t))$. Тому для пари задач

(26.1), (26.2) справедлива теорема 27.2 про слабку збіжність наближень (апроксимацій).

Зауваження 28.2. Слаба збіжність в просторі $W_2^1[0,1]$ тягне за собою сильну збіжність в просторі $C[0,1]$.

28.3. Проекційні методи Рітца та Ейлера

Нехай задана система функцій в просторі $W_2^1[0,1]$, для визначеності будемо вважати, що це тригонометрична система функцій. Розглянемо ту ж задачу (28.1) з тією лише різницею, що вважаємо, на підставі вибору системи, відрізок інтегрування $[-\pi,\pi]$, а $X \subset W_2^1$ – підпростір періодичних функцій, опускаючи для простоти граничні умови.

За X_n приймемо підпростір, утворений першими $2n+1$ елементами тригонометричними системами, а за P_n - оператор, який ставить у відповідність функції $x(t) \in W_2^1[-\pi,\pi]$ частинну суму Фейєра, то є:

$$P_n : x(t) \to \frac{S_0(t) + S_1(t) + \ldots + S_{n-1}(t)}{n},$$

де $S_k(t) = \frac{a_0}{2} + \sum_{j=1}^{k}\left(a_j \cos t_j + b_j \sin t_j\right)$, $\{a_j\}$, $\{b_j\}$, $\forall j = \overline{1,k}$ – коефіцієнти Фур'є функції $x(t)$. Визначимо f_n як звуження функціоналу на підпростір X_n

$$f_n(x_n) = \int_{-\pi}^{\pi} g\left(t, x_n, \frac{dx_n}{dt}\right)dt,\ x_n \in X_n, \qquad (28.3)$$

що відповідає методу Рітца.

Переконаємося, що виконані допущення Теореми 27.3, вважаючи $\Omega = X$, $\Omega_n = X_n$, $r_n = I$. Беручи до уваги, що згідно з теоремою Фейєра ([35], стор. 415) для будь-якої неперервно-диференційованої періодичної функції $x(t)$

$$x_n(t) \to \frac{S_0(t) + S_1(t) + \ldots + S_{n-1}(t)}{n} \in X_n$$

рівномірно збігається до $x(t)$, $\frac{dx_n(t)}{dt}$ прямує до $\frac{dx(t)}{dt}$ і той факт, що $g(t, x(t), y(t))$ рівномірно неперервна на обмеженій множині, одержуємо

$$\limsup_{n\to\infty}\left(f_n(x_n) - f(x)\right) \le \limsup_{n\to\infty} \int_{-\pi}^{\pi}\left|g\left(t, x(t), \frac{dx(t)}{dt}\right) - g\left(t, \alpha_n, \frac{dx_n}{dt}\right)\right|dt = 0.$$

Таким чином, умова 2) згаданої теореми має місце.

Для проекційних методів дискретна збіжність збігається зі звичайною збіжністю, тому умова 5) є наслідком слабої

неперервності функціоналу f. Що стосується умови 3), то вона безпосередньо одержується із умови 5) функції $g(t,x(t),y(t))$. Всі інші умови очевидні.

У варіаційному численні (див., наприклад, [53]) досить поширений ще один спосіб дискретизації - метод Ейлера, який легко вкладається в нашу схему. Цей метод є різновидністю проекційного методу Рітца з вибором в якості базисних кусково-лінійних функцій.

Нехай для функції g виконані властивості а) – г) і також задане ріномірне розбиття $\{t_i\}_0^n$ відрізка $[0,1]$. За X_n приймемо підпростір кусково-лінійних функцій з вершинами у вузлах сітки $\{t_i\}$. Оператор p_n функції $x(t)\subset W_2^1[0,1]$ ставить у відповідність кусково-лінійну функцію $\hat{x}_n(t)$, яка приймає значення $x(t_i)$ в точках t_i. Функціонал f_n визначається як звуження f на підпростір X_n

$$f_n(x_n)=\int_0^1 g\left(t,x_n(t),\frac{dx(t)}{dt}\right)dt,\ x_n\in X_n$$

Властивість 2) Теореми 27.3 з урахуванням неперервності $g(t,x(t),y(t))$, випливає із того, що якщо функції $x(t)\in C^1[0,1]$, то для будь-якого $\varepsilon>0$ знайдеться достатньо великий номер n(крок сітки h), при якому для $t\in[t_{i-1},t_i]$

$$|x(t)-\bar{x}_n(t)|<\varepsilon,\ \left|\frac{dx(t)}{dt}-\frac{d\bar{x}_n(t)}{dt}\right|=\left|\frac{dx(t)}{dt}-\frac{x(t_i)-x(t_{i-1})}{h}\right|<\varepsilon,$$

де $\bar{x}_n(t)$ - кусково-лінійна функція, яка приймає значення $x(t_i)$ в точках t_i і лінійна на $[t_{i-1},t_i]$. Звідси, до речі, випливає, що $\bigcup_{n=1}^{\infty}X_n=W_2^1$ згідно з нормою простору W_2^1, і для p_n виконані властивості (26.5), (26.6) на всюди щільній в $W_2^1[0,1]$ множині $C^1[0,1]$. З цієї причини існує єдине з точністю до еквівалентності продовження p_n на весь простір W_2^1 (див. теорему 26.1). Не важко переконатись у виконанні інших умов теореми.

28.4. Дискретний метод Рітца

Розглянуті в попередньому пункті методи Рітца і Ейлера не дають повної дискретизації задачі, бо залишається “неперервний об’єкт” – інтеграл. Якщо додатково провести аппроксимацію інтеграла по квадратурній формулі (прямокутників)

$$f_{mn}(x_n)=\sum_{i=1}^{m}hg(t_i,x_n(t_i),\dot{x}_n(t_i)),\qquad(28.4)$$

то приходимо до дискретного методу Рітца. Для його збіжності необхідні більш жорсткі вимоги до вхідних даних задачі 28.1. Для його збіжності, зокрема, достатньо допустити, щоб точний розв'язок задачі 26.1 належав $C^{(2)}[0,1]$. При цьому від функції $g(t,x(t),y(t))$ вимагається лише виконання властивості а) із пункту 28.1.

Зауваження 28.3. Дискретний метод Рітца, в окремому випадку, при $m=n$ і виборі в якості X підпростір кусково-лінійних функцій для сітки $\{t_i\}$, переходить у кінцево-різницевий метод, розглянутий у п.28.2.

§29. ОПЕРАТОРНІ ТА ІНТЕГРАЛЬНІ РІВНЯННЯ ПЕРШОГО РОДУ. ПОСТАНОВКА ЗАДАЧІ ТА УМОВА КОРЕКТНОСТІ

29.1. Формулювання проблеми

Нехай на парі ЛНП X, Y задане лінійне операторне рівняння першого роду

$$Ax = y,\ y \in R(A) \subset Y, \tag{29.1}$$

включення $y \in R(A)$ означає, що задача (29.1) розв'язувана.

Будемо вважати, що замість точних $\{A, y\}$ відомі наближені дані $\{A_h, y_\delta\} \subset m$ задачі (29.1) з умовою апроксимації

$$\|A - A_h\| \le h,\ \|A - A_h\| \le h,\ \{A_h, y_\delta\} \subseteq m, \tag{29.1'}$$

де $m \subset [X \to Y] \times Y$, $[X \to Y]$ - множина лінійних обмежених операторів на X в Y.

Треба за наближеними даними $\{A_h, y_\delta\} \subset m$ побудувати послідовність $\{x_\Delta : \Delta = (\delta, h)\}$, яка збігається до розв'язку рівняння (29.1).

Формалізуємо цю вимогу в двох наступних визначеннях [1], [2].

Визначення 29.1. Сімейство операторів $\{R_{\delta,h}\}$ (не обов'язково лінійних), заданих на множині m з зоною значень в X, називається *регуляризуючим алгоритмом* для задачі (29.1), якщо виконані умови:

1) Оператор $R_{\delta,h}$ визначений для будь-якої пари $\{A_h; y_0\}$ і який задовольняє співвідношенню (29.1');

2) $\sup\left(R_{\delta,h}[A_h, y_\delta], x_0\right) \to 0$ при $\Delta = \Delta(\delta, h) \to 0$, $y_\delta : \|y - y_\delta\| \le \delta$, $A_h : \|A - A_h\| \le h$, де x_0 – розв'язок (або множина розв'язків) задачі (29.1) (визначення напіввідхилення див. в [п.1.2 (коректність по Адамару і Тихонову, гл.II)]

Визначення 29.2. Якщо $R_{\delta,h}$ – регуляризуючий алгоритм (РА), то сукупність

$$\left\{x_\Delta : x_\Delta = R_{\delta,h}[A_h, y_\delta], \{A_h, y_\delta\} \subset m, 0 < \delta \le \delta_0, 0 < h \le h_0\right\}$$

називається *регуляризованим сімейством наближених розв'язків.*

У допущенні, що обернений оператор A^{-1} до A існує, проаналізуємо два випадки:

1) оператор A^{-1} обмежений, значить, неперервний;

2) оператор A^{-1} необмежений, значить, розривний.

У першій ситуації на підставі відомого факту (з [47], стор. 157) при достатньо малих h існує обмежений оператор A_h^{-1}, для якого виконується умова $\left\|A_h^{-1} - A^{-1}\right\| \to 0$, якщо $h \to 0$. Ця обставина

дозволяє прийняти в якості регуляризованого (наближеного) розв'язку $x_\Delta = A_h^{-1} y_\delta$. Дійсно, із цілком очевидної оцінки

$$\|x_\delta - x_0\| \le \left\|A_h^{-1} - A^{-1}\right\| \|y_\delta\| + \left\|A^{-1}\right\| \|y_\delta - y_0\|$$

одержуємо збіжність $x_\Delta \to x_0$ при $\Delta \to 0$.

У випадку 2) цей факт, взагалі кажучи, вже місця не має, тому за наближений розв'язок не можна брати розв'язок рівняння $A_h x = y_\delta$. Природно, що становище ускладнюється, якщо оператори A, A_h не мають обернених. Сказане вище дає підстави ввести наступні визначення коректності за Адамаром.

Визначення 29.3. Якщо виконані умови:

1) для будь-якого $y \in Y$ знайдеться елемент $x \in X$ такий, що $Ax = y$, то є $R(A) = Y$ (факт існування розв'язку);

2) елементом y розв'язок x визначається однозначно, тобто існує обернений оператор A^{-1} (факт єдиності розв'язку).

Тоді задача (29.1) називається *коректно поставленою* (або коректною) на парі ЛНП X, Y.

Якщо порушено хоча б одну з умов 1)-3), то задача називається *некоректно поставленою* (або некоректною), притому якщо не виконано умови 3), то говорять про *суттєво некоректну задачу.*

29.2. Рівняння, породжені інтегральними операторами

Приклад 29.1.

Інтегральне рівняння Фредгольма першого роду

$$Ax = \int_a^b K(t,s)x(s)ds = y(t),\ c \le t \le d, \qquad (29.2)$$

розглядуване на парі ЛНП $X = L_2[a,b]$, $Y = L_2[c,d]$ або $X = C[a,b]$, $Y = C[c,d]$. Нехай $K(t,s)$ – неперервна функція від t, s і така, що $Кег(A) = \{0\}$, і, значить, A^{-1} існує, тому умова 2) коректності виконана. Однак умови 1), 3) завідомо місця не мають (не виконуються).

І справді, оператор A в (29.2) в наших допущеннях цілком неперервний (тобто обмежену множину переводить в компактну), в чому легко переконатися, перевіривши критерії компактності в просторах L_2, C ([47], стор 236, 245). Якщо допустити, що A^{-1} неперервний (обмежений), то добуток $A^{-1} \cdot A = I$ зобов'язаний бути цілком неперервним оператором. Це суперечить тому, що сфера (яка переводить відображення $I = A \cdot A^{-1}$ в себе) некомпакта в просторах

L_2, C ([47], стор. 107, 110). Відомо також, що зона визначення цілком неперервного оператора не замкнута ([47], стор. 225).

У факті стійкості можна також переконатись із наступних простих міркувань. Нехай $x_1(S)$ і $x_2(S) = x_1(S) + N\sin\omega x$ $(N > 0, \omega > 0)$ розв'язок рівняння (29.2) для правих частин, $y_1 = Ax_1$; $y_2 = Ax_2$. Тоді в допущенні неперервності $K_S(t,s)$

$$\|x_2(S) - x_1(S)\|_{C[a,b]} = \max|N\sin\omega x| = N \quad \forall\omega,$$

$$\|y_2(t) - y_1(t)\|_{C[a,b]} = \max_t \left| N\int_a^b K(t,s)\sin\omega x ds \right| =$$

$$= N\max_t \frac{1}{\omega}\left| K(t,s)\cos\omega x\Big|_a^b + \int_a^b K'(t,s)\cos\omega x ds \right| \le \frac{N}{\omega}c,\ c = const,$$

звідки випливає, що розв'язки можуть бути як завгодно далекі (при великих N), в той час, як відповідні праві частини як завгодно близькі (за рахунок достатньо великих ω).

Установлений факт суттєвої некоректності задачі (29.2) приводить до сильної чисельної нестійкості, якщо розв'язати систему лінійних алгебраїчних рівнянь(СЛАР), одержану після дискретизації задачі квадратурним методом, наприклад, методом прямокутників:

$$\sum_{j=1}^{n} \Delta s_i K(t_i, s_j) x(s_j) = y(t_i) \ \ (i = \overline{1,n}). \tag{29.3}$$

Наведемо результати чисельного експерименту [54], в якому розв'язувалася СЛАР (29.3) для інтегрального рівняння

$$Ax = \int_0^1 \left\{ \frac{1}{2}(t+s) + ts + \frac{1}{3} \right\} x(s) ds = y(t) \equiv t + \frac{7}{12} \tag{29.4}$$

на сітці $\Delta t_j = \Delta s_j = 0.05 + 0.1(j-1)$, $j = \overline{1,n}$, $n = 10$.

Точний розв'язок рівняння (29.2) заданими (29.4) є $x(s) \equiv 1$. Машинний розв'язок $\{\tilde{x}(s_j)\}$, зображений в таблиці 29.1, показує, що воно не має нічого спільного з точним розв'язком.

Таблиця 29.1

s_j	0.05	0.15	0.25	0.35	0.45
$\tilde{x}(s_j)$	-8.00	-23.9	-7.75	7.62	0.00
s_j	0.55	0.65	0.75	0.85	0.95
$\tilde{x}(s_j)$	-12.5	4.00	-2.00	-14.0	-1.25

Та ж схема основана на розв'язку СЛАР (29.3) для інтегрального рівняння (див. [55]).

$$\frac{1}{\pi}\int_{-1}^{1}\frac{h}{(x-s)^2+h^2}x(s)ds=\bar{y}(t) \qquad (29.5)$$

$\bar{x}(s)=\left(1-s^2\right)^2$ – точний розв'язок на сітці із 41 точок ($s_j=-1+(j-1)h$, $h=\frac{2}{40}$, $j=\overline{1,41}$) дає абсолютно нерегулярний розв'язок $\{\tilde{x}(s_j)\}$, котрий змінюється в границях від $-9\cdot10^6$ до $2\cdot10^5$.

§30. РЕГУЛЯРИЗУЮЧІ МЕТОДИ

30.1. Варіаційний метод

Регуляризуючий алгоритм для операторного (інтегрального) рівняння Фредгольма першого роду будуємо на засадах тихонівської процедури у варіаційній формі (див. задачі на екстремум – регуляризація екстремальних задач)

$$\min\left\{\left\|A_h x - y_\delta\right\|^2 + \alpha\left\|x - x^0\right\|^2 : x \in X\right\} = F_\alpha, \qquad (30.1)$$

де x^0 – пробний розв'язок, роль якого складається з урахування якісної та кількісної інформації про розв'язок, одержаної *a priori*. Введемо також позначення $\hat{x}$ – для нормального розв'язку задачі (29.1) за аналогією до визначення 24.8.

Теорема 30.1. Нехай X, Y – гільбертові простори, A, A_h $(0 < h \le h_0)$ – лінійні обмежені оператори із X в Y і виконані умови апроксимації (29.1'). Тоді для будь-яких $\{A_h; y_\delta\} \subset m$, $x^0 \in X$, $\alpha > 0$ існує єдиний розв'язок x_Δ^α і

$$\lim_{\Delta \to 0}\left\|x_\Delta^\alpha - \hat{x}\right\| = 0,\ \Delta = (\delta, h), \qquad (30.2)$$

якщо $\dfrac{(\delta + h)^2}{\alpha(\Delta)} \to 0$, $\alpha(\Delta) \to 0$ при $\Delta \to 0$.

Доведення. Оскільки $f(x) = \left\|A_h x - y_\delta\right\|^2$ – опуклий, а $\varphi(x) = \left\|x - x^0\right\|^2$ – сильно опуклий функціонал, тоді цільовий функціонал в задачі (30.1) – сильно опуклий. Згідно з Лемою (15, гл.II) задача розв'язувана і її розв'язок x_Δ^α – єдиний. На підставі визначення x_Δ^α і умов апроксимації маємо

$$\left\|A_h x_\Delta^\alpha - y_\delta\right\|^2 + \alpha\left\|x_\Delta^\alpha - x^0\right\|^2 \le \left\|A_h \hat{x} - y_\delta\right\|^2 + \alpha\left\|\hat{x} - x^0\right\|^2 \le$$

$$\le \left(\left\|A_h \hat{x} - A\hat{x}\right\| + \left\|A\hat{x} - y_\delta\right\|\right)^2 + \alpha\left\|\hat{x} - x^0\right\|^2 \le \left(h\left\|\hat{x}\right\| + \delta\right)^2 + \alpha\left\|\hat{x} - x^0\right\|^2, \qquad (30.3)$$

звідки одержуємо оцінку

$$\left\|x_\Delta^\alpha - x^0\right\|^2 \le \frac{\left(h\left\|\hat{x}\right\| + \delta\right)^2}{\alpha} + \left\|\hat{x} - x^0\right\|^2. \qquad (30.4)$$

У відповідності до вибору $\alpha(\Delta)$ і слабої компактності обмеженої множини в X виділяємо послідовність

$$x_{\Delta_k}^{\alpha(\Delta_k)} - x^0 \to \bar{x} - x^0. \qquad (30.5)$$

Враховуючи слабу неперервність A, умов апроксимації і (30.3)-(30.5), знаходимо

$$\|A\bar{x}-y\| \le \lim_{k\to\infty}\inf\left\|Ax_{\Delta_k}-y\right\| \le \lim_{k\to\infty}\sup\left\{\left\|Ax_{\Delta_k}^{\alpha_k}-A_{h_k}x_{\Delta_k}^{\alpha_k}\right\|+\left\|A_{h_k}x_{\Delta_k}^{\alpha_k}-y_{\delta_k}\right\|+\left\|y_{\delta_k}-y\right\|\right\} \le$$

$$\le \lim_{k\to\infty}\sup\left\{h_k\left\|x_{\Delta_k}^{\alpha_k}\right\|+\delta_k\right\}+\lim_{k\to\infty}\sup\left\{\left\|A_{h_k}x_{\Delta_k}^{\alpha_k}-y_{\delta_k}\right\|^2+\alpha(\Delta_k)\left\|x_{\Delta_k}^{\alpha_k}-x^0\right\|^2\right\}^{1/2} \le$$

$$\le \lim_{k\to\infty}\sup\left\{\left(h_k\|\hat{x}\|+\delta_k\right)^2+\alpha(\Delta_k)\left\|\hat{x}-x^0\right\|^2\right\}^{1/2}=0$$

тобто $\bar{x}$ – розв'язок задачі (29.1). Тепер, об'єднуючи (30.4), (30.5), маємо

$$\left\|\bar{x}-x^0\right\| \le \lim_{k\to\infty}\inf\left\|x_{\Delta_k}^{\alpha_k}-x^0\right\| \le \left\|\hat{x}-x^0\right\|. \qquad (30.6)$$

А оскільки нормальний розв'язок єдиний, то $\bar{x}=\hat{x}$. Цей факт замість (30.5), (30.6) тягне сильну збіжність

$$\lim_{k\to\infty}\left\|x_{\Delta_k}^{\alpha_k}-\hat{x}\right\|=0, \qquad (30.7)$$

а так як $\hat{x}$ – єдина гранична точка, то (30.7) еквівалентне (30.2).

Наступна лема встановлює зв'язок між розв'язком задачі (30.1) і розв'язком операторного рівняння другого роду

$$\left(A_h^*A_h+\alpha E\right)x=A_h^*y_\delta+\alpha x^0, \qquad (30.8)$$

де $A^*: Y\to X$ – спряжений до A оператор.

Лема 30.1. Задачі (30.1) і (30.8) еквівалентні, тобто розв'язок задачі (30.1) є розв'язком задачі (30.8), і навпаки.

Доведення. Позначимо

$$J(x)=\|A_hx-y_\delta\|^2+\alpha\left\|x-x^0\right\|^2$$

і обчислимо першу варіацію δJ цього функціоналу. Так як

$$J(x+v)-J(x)=2\langle A_hx-y_\delta,A_hv\rangle+2\alpha\left\langle x-x^0,v\right\rangle+\|A_hv\|^2+\|v\|^2,$$

то $\delta Jv=2\left\langle A_h^*A_hx-A_hy_\delta+\alpha\left(x-x^0\right),v\right\rangle$. Використовуючи необхідну умову екстремуму $\delta J=0$, одержуємо (30.8). І так розв'язок задачі (30.1) задовольняє рівняння (30.8). Навпаки, нехай $\bar{x}$ – розв'язок рівняння (30.8). Оскільки

$$\left\|A_h^*A_h+\alpha E\right\|^2=\left\|A_h^*A_hx\right\|^2+\alpha\left\langle A_h^*A_hx,x\right\rangle+\alpha^2\|x\|^2\ge\alpha^2\|x\|^2,$$

то оператор $A_h^*A_h+\alpha E$ має обернений, значить, $\bar{x}$ – єдиний розв'язок ($Ker\left(A_h^*A_h+\alpha E=\theta\right)$). Тоді на підставі доведеного вище $\bar{x}$ буде єдиним розв'язком екстремальної задачі (30.1).

30.2. Зведення до рівняння другого роду

Як було показано в попередньому пункті, варіаційна постановка у формі (30.1) редукується в рівняння другого роду (30.8). При

деяких додаткових допущеннях регуляризація, заснована на зведенні до рівняння другого роду, може бути виконана в більш простій формі

$$(A_h + \alpha E)x = y_\delta + \alpha x^0. \qquad (30.9)$$

Теорема 30.2. Нехай X, Y – гільбертові простори, A, A_h – лінійні самоспряжені вектори (тобто $A^* = A$, $A_h^* = A_h$) і додатньо визначені (тобто $\langle Ax, x\rangle \geq 0 \quad \forall x \in X$) оператори, для яких виконані умови апроксимації (29.1'). Тоді рівняння (30.9) має єдиний розв'язок x_Δ^α для будь-яких $\{A_h; y_\delta\} \subset m$, $x^0 \in X$ і $\alpha > 0$

$$\lim_{\Delta \to 0} \left\| x_\Delta^\alpha - \hat{x} \right\| = 0, \ \Delta = (\delta, h),$$

коли $\dfrac{(\delta + h)^2}{\alpha(\Delta)} \to 0$, якщо $\alpha(\Delta) \to 0$ при $\Delta \to 0$.

Доведення можна знайти в книжці [2] (Теорема 2, стор.63).

30.3. Ітеративна регуляризація

Не обмежуючи загальності міркувань, можна вважати, що $\|A\| \leq 1$. Розглянемо дві ітераційні схеми:

$$x^k = \left(E - A_h^* A_h\right)x^{k-1} + A_h^* y_\delta \equiv U x^{k-1}, \qquad (30.10)$$

$$x^k = \left(A_h^* A_h + E\right)^{-1}\left(x^{k-1} + A_h^* y_\delta\right) \equiv U x^{k-1}. \qquad (30.11)$$

Відомо, що ітераційні послідовності (30.10), (30.11) збігаються до розв'язку x_δ^h рівняння $A_h x = y_\delta$, якщо воно розв'язуване. Але, як було показано в §29, якщо A^{-1} необмежений, то елемент x_δ^h не апроксимує розв'язку рівняння (29.1) при $\delta, h \to 0$. Тому не доцільно виконувати велику кількість кроків в процесах (30.10), (30.11), а необхідно формулювати правила зупинки Π залежно від рівня похибки [20].

Теорема 30.3. Нехай виконані умови Теореми 30.1. Якщо здійснювати зупинку по одному з правил:

$\Pi_0 : K(\delta, h)$, для якого вперше виконується $\left\| x^k - x^{k-1} \right\| \leq a_1\delta + a_2 h$, де $a_1 > 0$, $a_2 > 0$;

$\Pi_1 : K(\delta, h)$, для якого вперше виконується $\left\| A_h x^k - y_\delta \right\| \leq b_1\delta + b^*\eta$, де $b^* > \|\hat{x}\|$;

$\Pi_2 : K(\delta, h)$, для якого вперше виконується одна з нерівностей

$$\left\|A_h x^k - y_\delta\right\| \le b_1\delta + b_2\left\|x^k\right\|h,\ k \ge \frac{a}{\left(b_1\delta + b_2\left\|x^k\right\|h\right)^2},$$

де $a > 0$, $b_1 > 0$, $b_2 > 0$, тоді для ітерацій (30.10)

$$\lim_{\delta,h\to 0}\left\|x^{k(\delta,h)} - \hat{x}\right\| = 0,$$

де $\hat{x}$ – нормальний розв'язок (розв'язок, який найменше ухиляється від початкового наближення x^0). При цьому

$(\delta + h)K(\delta,h) \to 0$ для правила Π_0,

$(\delta + h)^2 K(\delta,h) \to 0$ для правил Π_1, Π_2.

Зауваження 30.1. Аналог Теореми 30.3 має місце і для неявної схеми (30.11).

30.4. Нелінійні ітераційні методи розв'язку задач з апріорною інформацією

У багатьох прикладних задачах, що описуються рівнянням (29.1), часто відома деяка додаткова інформація про властивості розв'язку, яку математично можна зобразити у вигляді належності шуканого розв'язку опуклій замкнутій множині Q.

Конкретний вигляд множини Q, як правило, визначається фізичною суттю розв'язку (29.1). Часто характерними прикладами задання множини Q є

$$Q_\varepsilon = \left\{x(t): \frac{d^l x(t)}{dt^l}\begin{pmatrix}\ge\\ \le\end{pmatrix}0\right\}, \qquad (30.12)$$

де $l = 0, 1, 2$, $x(t) \in W_2^l = X$; у деяких випадках в апріорні обмеження входять також значення похідних у фіксованих точках.

Введемо на розгляд більш загальний спосіб задання апріорної множини у формі

$$Q = \left\{x \in X : g_j(x) \le 0, j = 1, 2, ..., m\right\}, \qquad (30.13)$$

де g_j – опуклі диференційовані функціонали. Зауважимо, що після кінцево-різницевої апроксимації похідних множини (30.12) зображувані у формі (30.13) допускаються обмеження у вигляді рівнянь $g_j = \left\langle h_j, x\right\rangle - b_j = 0$.

Позначимо через M множину розв'язків задачі (29.1) при точних даних $\{A; y\}$.

Вимагається побудувати регуляризоване сімейство наближених розв'язків, регуляризуюче відносно слабої чи сильної збіжності елемент $\hat{x} \in Q \cap M$, тобто розв'язок, що задовольняє фізичним вимогам задачі.

Побудована задача розв'язується на основі однокрокових ітераційних процесів:

$$x^k = FUx^{k-1},\ x^0 \in X, k = 1, 2, ... \qquad (30.14)$$

$$x^k = [\lambda P + (1-\lambda)U]x^{k-1},\ x^0 \in X, k = 1, 2, ..., \qquad (30.15)$$

де U – оператор переходу розв'язуючої ітераційної процедури для задачі (29.1) без урахування умови $x \in Q$ (див., наприклад, (30.10), (30.11) при $y_\delta \equiv y$, $A_h \equiv A$), $P - Q$ – псевдозвужуючі відображення, конструктивно визначається по множині Q.

Визначення 30.1. Відображення $V : X \to X$ називається Q-*псевдозвужуючим*, якщо $Q = \{z : Vz = z\} \neq \varnothing$ і існує $v > 0$ таке, що $\|vx - z\|^2 \le \|x - z\|^2 - v\|vx - x\|$ для будь-яких $z \in Q$, $x \in X$; позначимо клас таких відображень через P_Q.

Визначення 30.2. Відображення $V : X \to X$ називається Q-*квазірозтягуючим*, якщо множина $Q = \{z : Vz = z\} \neq \varnothing$ і $\|vz - z\| \le \|x - z\|$ для будь-яких $z \in Q$, $x \in X$; позначимо клас таких відображень через K_Q.

Очевидно, що має місце суворе входження $P_Q \subset K_Q$, крім того, справедливе співвідношення

$$\{\lambda I + (1-\lambda)K_Q; 0 < \lambda < 1\} = P_Q.$$

Необхідно зауважити, що основна суть використання псевдозвуження в методах (30.14), (30.15) є в тому, що наближений розв'язок, одержаний після кожного кроку базової ітераційної схеми, зсувається відображенням P в напрямку множини Q, чим в результаті і досягається збіжність $\{x^k\}$ до елемента із $Q \cap M$.

Теорема 30.3. Нехай відображення U, P задовольняє умовам:

a) $U \in P_M$, $P \in P_Q$ $(0 < v < 1)$;

b) із того, що $z_k \to x$ ("$\to$" – знак слабої збіжності), $z_k - U_{z_k} \to 0$, $z_k - P_{z_k} \to 0$, випливає $x \in M \cap Q$.

Тоді для будь-яких $x^0 \in X$ для ітераційних послідовностей (30.14), (30.15) справедливі такі властивості:

1) $x^k \to \hat{x} \in M \cap Q$;

2) $\inf\limits_z \{\lim\limits_{k\to\infty} \|x^k - z\| : z \in M \cap Q\} = \lim\limits_{k\to\infty} \|x^k - \hat{x}\|$;

3) Або $\|x^{k+1} - \hat{x}\| < \|x^k - \hat{x}\|$, або $\{x^k\}$ стаціонарна, починаючи з деякого $k \ge k_0$, то є $x^{k_0} = x^{k_0+1} = ... = \hat{x}$;

4) $\sum\limits_{k=0}^{\infty}\left\|x^k - x^{k+1}\right\| \le \dfrac{2\left\|x^0 - y\right\|^2}{\nu} \quad \forall y \in M \cap Q$.

Доведення теореми можна знайти в [7].

Якщо права частина рівняння задана з похибкою, то є $\|y - y_\delta\| \le \delta$, то при деяких допущеннях і зв'язку параметрів ($\delta \cdot n(\delta) \to 0$, $\delta \to 0$) також має місце слаба збіжність ітерацій до елемента $x \in M \cap Q$ (див. §6 із [7]). З цієї причини можна стверджувати, що послідовність $\left\{x^{k(\delta)}\right\}$, одержана методами (30.14), (30.15), утворює регуляризоване сімейство наближених розв'язків відносно слабої топології.

Зауваження 30.2. Матрична проекція Pr_Q на множину Q (відображення, яке кожному $x \in X$ ставить у відповідність елемент із Q, найближчий до X) належить класу P_Q і, значить, може бути використана в якості відображення P у процесах (30.14), (30.15). Крім того, якщо $Q = Q_1 \cap Q_2 \cap ... \cap Q_m$, то

$$P = \mathrm{Pr}_{Q_1}\,\mathrm{Pr}_{Q_2}...\mathrm{Pr}_{Q_m},\ P = \sum_{i=1}^{m}\lambda_i\,\mathrm{Pr}_{Q_i},\ \sum_{i=1}^{m}\lambda_i = 1,\ \lambda_i > 0,$$

тобто операцію P можна конструювати в такій формі, знаючи матричну проекцію на кожну з множин Q_i (яке вважається обмеженням типу рівності або нерівності).

Так, наприклад, при розшифровуванні атомної структури сплавів на основі EXAFS-методики виникає інтегральне рівняння [52]:

$$Ax \equiv \int_a^b \exp\left[-\frac{2s}{\lambda(t)}\right]\sin\left[2ts + \varphi(t)\right]x(s)ds = X(t), \qquad (30.16)$$

де $x(s)$ – шукана функція радіального розподілу атомів (ФРРА), $x(t)$ – експериментально визначена, а $\lambda(t)$, $\varphi(t)$ – задані функції. Із фізичної суті задачі функція $x(s)$ задовольняє додатковим обмеженням:

$$x(s) \ge 0,\ \langle v, x\rangle \equiv \int_a^b s^2 x(s)ds = \frac{b^3 - a^3}{3} \equiv q.$$

Ці співвідношення визначають дві множини

$$Q_1 = \{x(s) : x(s) \ge 0\},\ Q_2 = \{x : \langle v, x\rangle = q\}.$$

Матрична проекція на ці множини обчислюється за явними формулами

$$\mathrm{Pr}_{Q_1}\,x = x^+(t) = \begin{cases} x(t), \text{ якщо } x(t) > 0 \\ 0, \quad \text{якщо } x(t) \le 0 \end{cases},\ \mathrm{Pr}_{Q_2}\,x = x - \frac{\left(\langle x, v\rangle - q\right)v}{\|v\|^2}.$$

Після дискретної апроксимації (30.16) методом колокацій (базис-ступеневої функції) з числом $m = 300$ вузлів по t, а по s $n = 100$,

застосовувалась ітераційна схема (30.10) (при $A_h \equiv A_{mn}$, $y_\delta = x_m$, де A_{mn} – матриця, апроксимуюча оператор A, x_m – вектор, апроксимуючий $x(t)$, то є

$$x^{k+1} = \left(E - A^*_{mn} A_{mn}\right) x^k + A^*_{mn} x_m \tag{30.17}$$

і її нелінійний аналог (30.14) з псевдозвуженням, то є

$$x^{k+1} = \Pr_{Q_1} \Pr_{Q_2} \left[\left(E - A^*_{mn} A_{mn}\right) x^k + A^*_{mn} x_m\right] \tag{30.18}$$

За модельним розв'язком $\bar{x}(s)$, яке вибиралось якісно подібним ФРРА для деякого сплаву, обчислювалася права частина $\bar{x}(t)$. З одержаною $\bar{x}(t)$ розв'язувалося рівняння (30.16) ітераційними методами (30.17), (30.18) з початковим наближенням $x^0 = 0$.

Схема (30.18) показала себе цілком ефективною (похибка наближеного розв'язку складала 3% після 400 ітерацій), в той час, як явна схема (30.17) дає значну похибку (33% після 400 ітерацій). Аналогічна картина спостерігається при використанні, замість базової, наявної схеми (30.11), а також при розв'язуванні рівнянь [51, 56, 57].

§31. КІНЦЕВО-ВИМІРНА АПРОКСИМАЦІЯ РА. КРИТЕРІЙ ЗБІЖНОСТІ

У попередньому параграфі встановлено, що побудова регуляризованого сімейства розв'язків редукується до пошуку оптимальних елементів в задачі мінімізації (30.1), яка є нескінченно вимірною (наприклад, по простору X і оператору A_h). Тому дискретизація (кінцево-вимірна апроксимація) задачі (30.1) – необхідний етап при розробці обчислювальних процедур.

Нижче наводяться необхідні і достатні умови збіжності кінцево-вимірних наближень в термінах дискретної збіжності. Для елементів це поняття введено в §26. Далі нам знадобиться також поняття дискретної слабої і сильної збіжності для операторів [48, 49].

Нехай послідовності $\{X_n\}$, $\{Y_m\}$ утворюють дискретну апроксимацію просторів X, Y з оператором звуження $\{p_n\}$ і $\{q_m\}$ відповідно (див. визначення 26.1).

Визначення 31.1. Послідовність операторів $A_{mn}: X_n \to Y_m$ дискретно слабо збігається до оператора $A: X \to Y$, якщо виконане співвідношення

$$x_n \to x \Rightarrow A_{mn}x_n \to Ax;$$

позначимо $A_{mn} \to A$.

У відповідності до загальної схеми дискретизації проблеми (30.1) запишемо послідовності апроксимуючих задач

$$\min\left\{\left\|A_{mn}x_n - y_m\right\|^2 + \alpha\left\|x_n - x_n^0\right\| : x_n \in X_n\right\} = F_{mn}^{\alpha} \qquad (31.1)$$

Введемо позначення x^{α}, x_{mn}^{α} для розв'язку задач (30.1), (31.1) відповідно. Зауважимо, що внаслідок сильної опуклості цільових функціоналів існує єдиний розв'язок згаданих задач.

Теорема 31.1. Нехай $A: X \to Y$, $A_{mn}: X_n \to Y_m$ – лінійні обмежені оператори, причому

$$\left\|A_{mn}\right\| \le C, \qquad (31.2)$$

де X, $\{X_n\}$, Y, $\{Y_m\}$ – гільбертові простори, які володіють властивостями дискретної апроксимації (визначення 26.1).

Для того щоб

$$x_{mn}^{\alpha} \to x^{\alpha}, \ \lim_{m,n\to\infty} F_{mn}^{\alpha} = F^{\alpha} \qquad (31.3)$$

для будь-яких $x^0 \in X$, $x_n^0 \in X_n$, $y \in Y$, $y_m \in Y_m$ таких, що $x_n^0 \to x^0$, $y_m \to y$, необхідно і достатньо, щоб

$$A_{mn} \to A \qquad (31.4)$$

Доведення. Достатність легко виводиться із наслідку 30.1 до Теореми 30.3, оскільки легко перевірити, що співвідношення (31.4) і властивості слабої збіжності в гільбертовім просторі

$$x_n - \to x \Rightarrow \|x\| \le \lim_{n\to\infty} \inf \|x_n\|;$$

$$x_n - \to x \Rightarrow \|x_n\| \le c - const;$$

$$\|x_n\| \le c \Rightarrow x_{n_k} - \to x;$$

$$x_n - \to x \Rightarrow x_n - \to x$$

тягнуть за собою виконання умов Наслідку 30.1.

Встановимо необхідність умов. Нехай виконано (31.3) для будь-яких

$$x_n^0 - \to x^0,\ x^0 \in X,\ x_n^0 \in X_n,\ y_m - \to y,\ y \in Y,\ y_m \in Y_m \tag{31.5}$$

Для розв'язку задач (30.1), (31.1) маємо

$$\hat{x} = \left(A^* A + \alpha E\right)^{-1} \left(\alpha x^0 + A^* y\right) \tag{31.6}$$

$$\hat{x}_{mn} = \left(A_{mn}^* A_{mn} + \alpha E\right)^{-1} \left(\alpha x_n^0 + A_{mn}^* y_m\right). \tag{31.7}$$

При нульових $y = \theta$, $y_m = \theta_m$ із (31.3), (31.5)-(31.7) одержуємо дискретну збіжність операторів

$$\left(A_{mn}^* A_{mn} + \alpha E\right)^{-1} - \to \left(A^* A + \alpha E\right)^{-1},\ m,n \to \infty,$$

а при нульових $x^0 = \theta$, $x_n^0 = \theta_n$ за тими ж міркуваннями маємо

$$\left(A_{mn}^* A_{mn} + \alpha E\right)^{-1} A_{mn}^* - \to \left(A^* A + \alpha E\right)^{-1} A^*. \tag{31.8}$$

Так як $\left\|\left(A_{mn}^* A_{mn} + \alpha E\right) x_n\right\|^2 \ge \alpha^2 \|x_n\|^2$, то з урахуванням (31.2) приходимо до нерівностей

$$\left\|\left(A_{mn}^* A_{mn} + \alpha E\right)^{-1}\right\| \le \frac{1}{\alpha},\ \left\|\left(A_{mn}^* A_{mn} + \alpha E\right)\right\| \le c - const.$$

Згідно з теоремою Ф. Штумеля ([49], стор. 55) можна зробити висновок, що

$$A_{mn}^* A_{mn} + \alpha E - \to A^* A + \alpha E \Rightarrow A_{mn}^* A_{mn} - \to A^* A. \tag{31.9}$$

Нехай $x_n - \to x$. Тоді внаслідок дискретної збіжності (31.9), визначення і властивостей слабої дискретної збіжності

$$\lim_{m,n\to\infty} \|A_{mn} x_n\|^2 = \lim_{m,n\to\infty} \langle A_{mn} x_n, A_{mn} x_n \rangle = \lim_{m,n\to\infty} \left\langle x_n, A_{mn}^* A_{mn} x_n \right\rangle = \left\langle x, A^* A x \right\rangle = \|Ax\|^2 \tag{31.10}$$

Далі переконаємося, що

$$A_{mn} x_n - \to Ax. \tag{31.11}$$

Попередньо зауважимо, що із (31.8), (31.9) випливає властивість

$$A_{mn}^* - \to A^*, \tag{31.12}$$

оскільки для будь-якої $y_m - \to y$ справедливі співвідношення

$$y_m - \to y \Rightarrow \left(A_{mn}^* A_{mn} + \alpha E\right)^{-1} A_{mn}^* y_m - \to \left(A^* A + \alpha E\right)^{-1} A^* y \Rightarrow A_{mn}^* y_m - \to A^* y.$$

Скористуємося критерієм дискретної слабої збіжності ([48], стор. 21) і (31.12)

$$\|A_{mn}x_n\| \le c - const,$$

$$\lim_{m,n\to\infty} \langle A_{mn}x_n, q_m z\rangle = \lim_{m,n\to\infty} \langle x_n, A_{mn}^* q_m z\rangle = \langle x, A^* z\rangle = \langle Ax, z\rangle \Rightarrow A_{mn}x_n - \to Ax.$$

Таким чином, властивість (31.11) встановлена. Так як в гільбертових просторах із (31.10), (31.11) випливає $A_{mn} - \to A$, а із (31.12) $- A_{mn} - \to A$ ([48], стор. 27), то доведення теореми завершено.

§32. РЕАЛІЗАЦІЯ ЗАГАЛЬНОЇ СХЕМИ ДИСКРЕТИЗАЦІЇ

Дослідимо три конкретні схеми кінцево-вимірної апроксимації інтегральних рівнянь Фредгольма І-го роду (29.2) на парі гільбертових просторів $X = L_2[a,b]$, $Y = L_2[c,d]$.

32.1. Метод механічних квадратур

Задамо деякий збіжний квадратурний процес, наприклад, за формулою прямокутників

$$\int_a^b u(s)ds = \sum_{j=1}^{n} h_j^n U\left(s_j^n\right) + R_n(u),\ h_j^n = s_j^n - s_{j-1}^n,$$

у якого залишок $\lim\limits_{n\to\infty} R_n(u) = 0$ для будь-якої неперервної функції. Сімейство зв'язуючих операторів $\{p_n\}$ і аппроксимуючих просторів $\{X_n\}$ визначимо, як у прикладі 26.2. Дискретну аппроксимацію $Y = L_2[c,d]$ визначимо також за допомогою формули прямокутників на сітці $\left\{t_i^m\right\}$ $\overline{h}_j^m = t_j^m - t_j^{m-1}$ і зв'язуючих операторів $\{q_m\}$, $q_m : Y \to Y_m$, які визначаються аналогічно $\{p_n\}$.

Аппроксимуючі оператори $A_{mn} : X_n \to Y_m$ задамо формулою

$$\left(A_{mn}x_n\right)_i = \sum_{j=1}^{n} h_j^n K\left(t_i^m, s_j^n\right)x_{nj} \qquad \left(i = \overline{1,m}\right). \tag{32.1}$$

Теорема 32.1. Для операторів A, A_{mn}, що визначаються формулами (29.2), (32.1).

Доведення. Для перевірки першого співвідношення в (31.4) застосуємо критерій дискретної збіжності операторів [48].

Перш за все із ланцюга нерівностей

$$\left\|A_{mn}x_n\right\|^2 = \sum_{i=1}^{m} \overline{h}_i^m \left[\sum_{j=1}^{n} h_j^n K\left(t_i^m, s_j^n\right)x_{nj}\right]^2 \le$$

$$\le \sum_{i=1}^{m} h_i^m \left[\left(\sum_{j=1}^{n} h_j^n \left|K\left(t_i^m, s_j^n\right)\right|^2\right)^{\frac{1}{2}} \left(\sum_{j=1}^{n} h_j^n \left|x_{nj}\right|^2\right)^{\frac{1}{2}}\right]^2 \le$$

$$\le \max_{\substack{a\le s\le b \\ c\le t\le d}} \left|K(t,s)\right|^2 \left(\sum_{j=1}^{n} h_j^n\right)\left(\sum_{i=1}^{m} \overline{h}_i^m\right)\left\|x_n\right\|^2$$

і обмеженості сум випливає висновок про рівномірну обмеженість норм $\left\|A_{mn}\right\| \le c - const$.

Введемо зв'язуючі оператори $\{\widetilde{p}_n\}$, $\{\widetilde{q}_m\}$ (див. Приклад 26.1), котрі еквівалентні $\{p_n\}$, $\{q_m\}$ відповідно. Тоді для неперервної функції маємо

$$\left\|A_{mn}\widetilde{p}_n x-\widetilde{q}_m Ax\right\|^2=\sum_{i=1}^{m}\overline{h}_i^m\left[\sum_{j=1}^{n}h_j^n K\left(t_i^m,s_j^n\right)x\left(s_j^n\right)-\int_a^b K\left(t_i^m,s\right)x(s)ds\right]^2\le$$

$$\le\max_{1\le i\le m}\left|\sum_{j=1}^{n}h_j^n K\left(t_i^m,s_j^n\right)x\left(s_j^n\right)-\int_0^1 K(t_i,s)x(s)ds\right|^2\cdot\sum_{i=1}^{m}\overline{h}_i^m\le(d-s)\sup_{w\in\Phi}R_n^2(w),$$

де $R_n(w)$ – залишок квадратурної формули, а сімейство $\Phi=\{w_t(s):w_t(s)=K(t,s)x(s),0\le t\le 1\}$ внаслідок неперервності $K(t,s)x(s)$ відносно компактне в $C[0,1]$. Значить ([48], стор. 97),

$$\lim_{n\to\infty}\sup_{w\in\Phi}R_n(w)=0 \qquad (32.2)$$

І перше співвідношення в (31.4) доведено.

Для перевірки другого співвідношення скористуємося критерієм дискретної слабої збіжності ([48], стор. 21)

$$x_n-\to x\Rightarrow\lim_{m,n\to\infty}\left\langle\widetilde{q}_m z,A_{mn}x_n\right\rangle=\langle z,Ax\rangle$$

для будь-якої функції $z(t)\in C[a,b]$, де $C[c,d]=L_2[a,b]$.

З урахуванням визначення 26.3 достатньо переконатися в дискретній збіжності

$$w_n^m-\to w,\ w_{nj}^m=\sum_{i=1}^{m}\overline{h}_i^m K\left(t_i^m,s_j^n\right)z\left(t_i^m\right),\ w=\int_c^d K(t,s)z(t)dt$$

Згідно з визначенням дискретної збіжності маємо

$$\left\|w_n^m-\widetilde{p}_n w\right\|^2=\sum_{j=1}^{n}h_i^n\left|\sum_{i=1}^{m}\overline{h}_i^m K\left(t_i^m,s_j^n\right)z\left(t_i^m\right)-\int_c^d K\left(t,s_j^n\right)z(t)dt\right|^2\le\sum_{j=1}^{n}h_j^n\sup_{y\in\Psi}R_m^2(y),$$

де сімейство $\Psi=\{y_s(t):y_s(t)=K(t,s)z(t),a\le s\le b\}$ відносно компактне в $C[c,d]$. З цієї причини, згідно з (32.2), права частина прямує до нуля при $m,n\to\infty$. Теорема доведена.

У відповідності до Теореми 32.1 це гарантує дискретну збіжність кінцево-вимірних розв'язків задач (32.1), одержаних на підставі квадратурного методу, до регуляризованого розв'язку x^α задачі (30.1).

У §29 було показано, що для інтегральних рівнянь (29.4), (29.5) пряме застосування квадратурних схем у формі (29.3) веде до абсолютно нестійких результатів. Однак використання тих же схем для регуляризованої задачі у вигляді (31.1), де A_{mn} задаються формулою (32.1), $y_m=q_m y_\delta$, дає регулярну процедуру. При цьому регуляризація допустима як у формі (29.1), так і в (29.9).

Так, протягуючи квадратурну формулу прямокутників до регуляризованого рівняння (29.9) з даними (29.4), тобто до рівняння

$$\alpha x(s)+\int_0^1\left\{\frac{1}{2}(s+t)+st+\frac{1}{3}\right\}x(s)ds=t+\frac{7}{12}$$

при $n=0$, одержуємо чисельні розв'язки (див. [54]), котрі зображені в наступній таблиці (точний розв'язок $x(s)\equiv 1$).

Таблиця 32.1

α \ s	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
$\frac{1}{40}$	0.62	0.688	0.756	0.824	0.892	0.96	1.028	1.096	1.164	1.235
$\frac{1}{70}$	0.671	0.731	0.791	0.851	0.911	0.971	1.031	1.091	1.151	1.211
$\frac{1}{1024}$	0.937	0.947	0.959	0.971	0.983	0.995	1.007	1.02	1.032	1.044

Отже, при $\alpha=\frac{1}{1024}$ розв'язок відтворюється з точністю $\Delta\approx 4\%-6\%$.

Формула прямокутників, застосована до регуляризованої задачі (30.1) для даних (30.5) при $n=41$ (то є крок сітки $h=\frac{2}{41}$) і $\alpha=5\cdot 10^{-3}$, відтворює розв'язок з похибкою, меншою 1% (див. [55]).

32.2. Метод колокацій

Цей метод, як правило, використовується для наближеного розв'язку операторних (диференційних, інтегральних та інших) рівнянь, які задовольняють умовам коректності Адамара. Нижче ми застосуємо його для дискретної апроксимації варіаційної задачі (30.1);його використання безпосередньо для рівнянь (29.1), взагалі кажучи, неможливе через нестійкість вихідної задачі.

Будемо вважати виконаними допущення попереднього пункту про вибір простору і властивостей ядра $K(t,s)$. Нехай $\{x_n\}$ – послідовність кінцево-вимірних підпросторів в $X=L_2[a,b]$ і сімейство $\{p_n\}$ ортогональних проекторів p_n на X із звичайними властивостями

$$\left.\begin{array}{l} p_nX=X_n,\ \ p_n^2=p_n\ \ (n=1,2,...) \\ \lim\limits_{n\to\infty}\|x-p_nx\|=0\ \ \forall x\in X \end{array}\right\}. \qquad (32.3)$$

Послідовність $\{p_n\}$ приймаємо за зв'язуючі оператори між $X = L_2[a,b]$ і X_n, а сіткові оператори q_m: $y(t) \to \left(y\left(t_1^m\right),...,y\left(t_m^m\right)\right)$ – між $Y = L_2[a,b]$ і простором $Y_m = l_2^m$ з нормою

$$\|y_m\|^2 = \sum_{i=1}^{m} h_i^m |y_{mi}|^2 ,\ h_i^m = t_i^m - t_{i-1}^m .$$

Апроксимуючі оператори $A_{mn} : X_n \to Y_m$ визначаються для $x_n \in X_n$ формулою

$$\left(A_{mn}x_n\right)_i = \int_a^b K\left(t_i^m, s\right)x_n(s)ds . \qquad (32.4)$$

Лема 32.1. Дискретна та дискретна слабка збіжність визначається сімейством проекційних операторів (32.3), еквівалентні збіжності за нормою і слабою збіжністю, відповідно, в довільному гільбертовому просторі.

Доведення. Нехай $x_n \to x$, тобто $\lim\limits_{n\to\infty}\|x_n - p_n x\| = 0$, тоді з урахуванням (32.3)

$$\|x_n - x\| \le \|x_n - p_n x\| + \|p_n x - x\|$$

і права частина прямує до нуля при $n \to \infty$. Навпаки, нехай $\lim\limits_{n\to\infty}\|x_n - x\| = 0$, тоді

$$\|x_n - p_n x\| \le \|x_n - x\| + \|x - p_n x\|$$

і перша частина твердження доведена.

Припустимо тепер, що $x_n \to x$. Це означає, згідно з визначенням 29.3, що справедливе співвідношення

$$\lim_{n\to\infty}\langle v_n, x_n\rangle = \langle v, x\rangle \qquad (32.5)$$

для будь-якої $v_n \to v$. Покажемо, що $\lim\limits_{n\to\infty}\langle z, x_n\rangle = \langle z, x\rangle$ для всіх $x \in X$ (слабка збіжність).

В очевидній нерівності

$$\left|\langle z, x_n\rangle - \langle z, x\rangle\right| \le \left|\langle z, x_n\rangle - \langle p_n z, x_n\rangle\right| + \left|\langle p_n z, x_n\rangle - \langle z, x\rangle\right|$$

другий доданок прямує до нуля, так як $p_n z \to z$ і виконано (32.5). Для першого ж доданка справедлива оцінка

$$\left|\langle z, x_n\rangle - \langle p_n z, x_n\rangle\right| \le \|x_n\| \cdot \|p_n z - z\|,$$

де $\|x_n\| \le const$, $\lim\limits_{n\to\infty}\|p_n z - z\| = 0$.

Обернена імплікація $x_n \to x \Rightarrow x_n \to x$ випливає із наступних співвідношень:

$$\left|\langle v_n, x_n\rangle - \langle v, x\rangle\right| \le \left|\langle v_n, x_n\rangle - \langle v, x_n\rangle\right| + \left|\langle v, x_n\rangle - \langle v, x\rangle\right|,$$

де $v_n \to v$, а, значить, $\lim\limits_{n\to\infty}\|v_n - v\| = 0$.

Звернемося тепер до співвідношень (31.4), які гарантують збіжність аппроксимацій. Дискретна збіжність $A_{mn} - \to A$ для операторів, що визначаються формулами (29.2) і (32.4), випливає із Леми 32.1 і оцінки

$$\left\|A_{mn}x_n - q_m A_*\right\|^2 = \sum_{i=1}^{m} h_i^m \left[\int_a^b K(t_i,s)\left(x_n(s) - x(s)\right)ds\right]^2 \le$$

$$\le \sum_{i=1}^{m} h_i^m \left(\int_a^b \left|K(t_i,s)\right|^2 ds\right)\cdot\left(\int_a^b \left|x_n(s) - x(s)\right|^2 ds\right) \le$$

$$\le (d-c)\cdot \max_{c\le t\le d} \int_a^b \left|K(t,s)\right|^2 ds \cdot \left\|x_n - x\right\|^2_{L_2[a,b]}.$$

Будемо вважати, що $x_n - \to x$. Для перевірки співвідношень $A_{mn} - \to A$,яке еквівалентне

$$x_n - \to x \Rightarrow \lim_{m,n\to\infty} \left\langle q_m z, A_{mn}x_n\right\rangle = \left\langle z, Ax\right\rangle,$$

де $z(t)$ – неперервна функція, достатньо переконатися, що

$$A_{mn}^* q_m z = \sum_{i=1}^{m} h_i^m K\left(t_i^m, s\right) z\left(t_i^m\right) \Rightarrow \int_a^b K(t,s)z(t)dt = A^* z.$$

Враховуючи передкомпактність сімейства

$$\Phi = \left\{\varphi_s(t) : \varphi_s(t) = z(t)K(t,s),\ a \le s \le b\right\},$$

одержуємо

$$\lim_{m\to\infty}\left\{\sup_{0\le s\le b}\left|\sum_{i=1}^{m} h_i^m K\left(t_i^m, s\right) z\left(t_i^m\right) - \int_a^b K(t,s)z(t)dt\right|\right\} = 0,$$

із якого і слідує бажане співвідношення.

32.3. Проекційні методи

Нехай задана послідовність кінцево-вимірних підпросторів $\{X_n\}$ ($\{Y_m\}$) простору X (Y); сімейство $\{p_n\}$ ($\{q_m\}$) проекторів, які володіють властивостями (32.3), приймаємо за зв'язуючі оператори між X (Y) і X_n (Y_m).

Вважаючи $A_{mn} = q_m A p_n$, прийдемо до проекційних методів. Зокрема, при $q_m \equiv E$ одержуємо метод Рітца. Так як p_n, q_m – оператори ортогонального проектування, то $\|p_n\| \le 1$, $\|q_m\| \le 1$.

Тоді для лінійного обмеженого оператора A (зокрема, оператора (29.2)) потрібні властивості (31.2), (31.4) випливають із Леми 32.1 і співвідношень

$$\left\|A_{mn}\right\| = \left\|q_m A p_n\right\| \le \left\|q_m\right\|\cdot\left\|A\right\|\cdot\left\|p_n\right\| \le \left\|A\right\|,$$

$$\left\|A_{mn}p_n x - q_m Ax\right\| \le \left\|q_m\right\|\cdot\left\|Ap_n x - Ax\right\|,$$

$$\left|\left\langle A_{mn}, q_m y\right\rangle - \left\langle Ax, y\right\rangle\right| = \left|\left\langle q_m A p_n x_n, q_m x\right\rangle - \left\langle Ax, y\right\rangle\right| = \left|\left\langle Ax_n, q_m x\right\rangle - \left\langle Ax, y\right\rangle\right|,$$

де $x_n \rightarrow x$, тому $x_n \to x$; внаслідок слабої неперервності A $Ax_n \to Ax$, значить,

$$\lim_{m,n\to\infty} \left|\langle Ax_n, q_m y\rangle - \langle Ax, y\rangle\right| = 0$$

Крім того, позначаючи $y_m = q_m y$, $x_n^0 = p_n x^0$, будемо мати $y_m \rightarrow y$, $x_n^0 \rightarrow x^0$.

Нехай X_n (Y_m) утворені першими n (m) елементами ортонормованого базису $\{l_j\}$ ($\{g_j\}$). Тоді для викладеного вище проекційного методу задачу (31.1) можна переписати у вигляді

$$\min\left\{\left\|q_m A p_n x - q_m y_\delta\right\|^2 + \alpha\left\|p_n\left(x - x^0\right)\right\|^2 : x \in X\right\} =$$

$$= \min\left\{\left\|q_m A x_n - y_m\right\|^2 + \alpha\left\|\left(x_n - x_n^0\right)\right\|^2 : x_n \in X_n\right\}. \qquad (32.6)$$

Застосовуючи Лему 30.1 робимо висновок, що задача (32.6) еквівалентна розв'язку СЛАР

$$\sum_{j=1}^{n}\left[\sum_{i=1}^{m}\left(Al_j, q_i\right)\left(Al_k, q_i\right)\right]x_n^j + \alpha x_n^k = \sum_{i=1}^{m}\left(Al_k, q_i\right)y_m^i + \alpha x_n^{0,k} \quad \left(k = \overline{1,n}\right), \qquad (32.7)$$

котра дозволяє обчислити коефіцієнти Фур'є $\{x_n^j\}$ наближеного розв'язку $x_n^\alpha = \sum_{j=1}^{n} x_n^j l_j$ за коефіцієнтами $\{y_m^i\}$ Фур'є правої частини і матриці

$$\left\langle Al_j, q_i\right\rangle = \int_c^d\int_a^b\left[K(t,s)l_j(s)ds \cdot q_i(t)\right]dt\,.$$

Для інтегрального рівняння (див. (29.4)):

$$Ax = \int_0^1\left\{\frac{1}{2}(t+s) + ts\right\}x(s)ds = \frac{7}{12} + t \equiv f(t),\ 0 \le t \le 1$$

з точним розв'язком $\tilde{x}(s) \equiv 1$ система функцій $\left\{1, \sqrt{2}\cos\pi t, \sqrt{2}\sin\pi t, \ldots\right\}$ приймалася за ортонормований базис $\{l_j = q_j\}_{j=1}^{m=n}$, $n = m = 10$ і розв'язувалася система (32.7) при $x_n^0 = 0$ для різноманітних значень параметра α.

Результати чисельних розрахунків зведені в таблиці 32.2, де прийняті позначення для неув'язки $\varepsilon = \left\|A_{mn}x_n^\alpha - y_m\right\|$ і відносної похибки розв'язку $\Delta = \max\limits_{1\le j\le 10}\left|x_n^\alpha(s_j) - \tilde{x}(s_j)\right|$ ($s_j = 0.05 + 0.1j$).

Таблиця 32.2

α	Δ	ε
10^{-4}	$3.4 \cdot 10^{-1}$	$0.114 \cdot 10^{-2}$
10^{-7}	$2 \cdot 10^{-3}$	$0.431 \cdot 10^{-5}$
10^{-9}	$8 \cdot 10^{-5}$	$0.426 \cdot 10^{-7}$
10^{-10}	$4 \cdot 10^{-3}$	$0.355 \cdot 10^{-8}$

Аналіз одержаних результатів показує, що найменшу похибку розв'язку одержуємо при $\alpha = 10^{-9}$, і при подальшому зменшенні α похибка починає зростати, хоч неув'язка (при $\alpha = 10^{-10}$) продовжує зменшуватись. Цей факт підтверджує, що неув'язка не є надійною характеристикою для оцінки якості розв'язку погано обумовлених СЛАР [58].

§33. АНАЛІЗ МЕТОДІВ ОБЧИСЛЕННЯ СУМ, ДОБУТКІВ ТА ЦІЛИХ СТЕПЕНІВ

З аналізу достовірності обчислень при розв'язку некоректно поставлених задач, здійсненого в попередніх параграфах, випливає, що в багатьох ситуаціях під час обчислень при розв'язку нестійких задач важливо мати якомога точнішу інформацію про вхідні дані сформульованої задачі. Нерідко вхідна інформація для сформульованої задачі є результатом розв'язку інших обчислювальних задач. Це означає, що неточність вхідних даних (спадкова помилка) пов'язана не тільки з точністю вимірювальних приладів (носіїв вихідної інформації), а й з неточністю попередніх обчислень, тому підвищення їх точності може суттєво збільшити якість результату (в сенсі точності розв'язку розв'язуваної в даний момент задачі). Проілюструємо це явище на прикладі розв'язування канонічних задач обчислювальної математики: обчисленні сум, добутків та цілих степенів. Аналіз будемо проводити (виходячи з простоти та наглядності міркувань) у форматі двійкової системи обчислень.

33.1. Високоточний алгоритм обчислення сум

На слідуючи [60, 63], будемо вважати, що треба обчислити

$$S = \sum_{i=1}^{n} x_i$$

Будемо також вважати, що обчислювальний пристрій (ЕОМ) працює в режимі плаваючої коми, використовує двійкову систему зображення чисел та стандартні правила заокруглення чисел при підсумовуванні. Похибками зображення чисел в ЕОМ будемо нехтувати, тобто будемо вважати, що числа в обчислювальній машині зображуються точно, а довжина машинного зображення чисел не змінюється в процесі обчислень.

Розглянемо наступний алгоритм обчислень:

а) розташуємо доданки x_i, $i = 1, 2, ..., n$ за зростанням порядків. У випадку рівності порядків розсортуємо числа на додатні та від'ємні. Ділянки однаковості порядків і знаків запам'ятаємо;

б) перевіряємо, чи утворюють останні доданки ділянку однокроковості порядків. Якщо так, то обчислюємо парні суми, доданки яких мають різні знаки. Потім одержану послідовність доданків (уже вкорочену) знову впорядковуємо згідно з пунктом а) і повторюємо процедуру, поки не зникнуть всі ділянки однаковості порядків;

в) якщо умова, сформульована в пункті б) не виконується, тоді обчислення починаємо з найменших за порядком доданків;

г) якщо всі одержані доданки та частинні суми мають різні порядки, тоді для обчислення S здійснюємо просте накопичення суми, тобто

$$S_1 = \widetilde{x}_1,\ S_{l+1} = S_l + \widetilde{x}_l. \tag{33.1}$$

Тут $\widetilde{x}_l$ є числа, що утворилися в результаті парного підсумовування, а також доданки x_i, які не перетерпіли зміни (до яких не було застосоване парне підсумовування);

д) якщо спочатку вищеописаної процедури на ділянках однаковості порядків не було чисел з різними знаками (додатних і від'ємних), тоді попарне підсумовування на цих ділянках здійснюється за принципом: парні числа сумуються (попарно) з непарними, а непарні з непарними з подальшою перестановкою доданків і частинних сум, утворених попарним підсумовуванням.

Теорема 33.1. Для описаного алгоритма обчислення S справедлива оцінка

$$\left|S - \overline{S}\right| \le \begin{cases} 2^{-\tau}\left(2^{s} - 2^{k_1}\right), & \text{якщо } s = k_{n+1} \\ 2^{-\tau}\left(2^{\lambda} - 2^{k_1}\right), & \text{якщо } s < k_{n+1} \\ 2^{-\tau}\left(2^{s+1} - 2^{k_1}\right), & \text{якщо } s > k_{n+1} \end{cases} \tag{33.2}$$

де $\overline{S}$ – наближене (обчислене) значення S, s – порядок S, k_i – порядок доданка x_i, $k_1 \le k_2 \le \ldots \le k_n$, τ – кількість двійкових розрядів (розмір мантиси) комірок ЕОМ, λ – найбільший порядок доданків та частинних сум, які зустрічалися під час обчислень S.

Доведення. Якщо всі доданки різних порядків, то S обчислюється шляхом поступового накопичення (33.1) і порядок суми може лише на одиницю бути більшим за порядок найбільшого доданка. Якщо доданки всі одного знака і одного порядку, то, враховуючи той факт, що при попарному додаванні парних і непарних чисел вирівнювання мантис не призводить до втрати останнього розряду, помилка обчислень оцінюється сумою

$$2^{-\tau+k_1}\left(1 + 2 + 2^2 + \ldots + 2^r + \ldots + 2^{[\log_2 n]^* + 1}\right), \tag{33.3}$$

де $r = [\log_2 n]$ доданок в сумі (33.3) дає оцінку помилки обчислень при $r-1$-ому турі обчислення парних сум. Але сума (33.3) рівна $2 \cdot \left(2^{-\tau+k_1+r} - 2^{-\tau+k_1}\right)$, де $[\log_2 n]$ – ціла частина $\log_2 n$, $k_{n+1} = [\log_2 n] + 1$, що й відповідає формулі три в (33.2).

Другий вираз в (33.2) одержується у випадку виникнення ситуації передбаченої пунктом б).

Зауваження 33.1. Для виконання пункту а) необхідно, згідно з алгоритмом Шелла, використати $n\lceil \log_2 n \rceil$ порівняльних операцій.

Зауваження 33.2. Наведений алгоритм забезпечує точність обчислення S, в якій можуть бути неправильними лише два останні (двійкові) розряди мантиси. Помітимо, що алгоритм, запропонований в [64], забезпечує лиш $\tau - \lceil \log_2 n \rceil - 1$ правильних розряди, тому твердження автора про його оптимальність помилкова. Запропонований в даному параграфі алгоритм є оптимальним за точністю серед алгоритмів, які відрізняються перестановкою доданків і частинних сум.

Зауваження 33.3. Обчислення сум є основною операцією при чисельному знаходженні визначених інтегралів, сум збіжних рядів, обробці варіаційних рядів в економічній статистиці і т.д. Алгоритм, як плату за підвищену точність, вимагає додаткової роботи за попереднім упорядкуванням доданків. Неважко побудувати приклад, в якому стандартний спосіб організації обчислення суми (спосіб поступового її накопичення) не забезпечує жодного правильного знака. Якщо перший доданок у сумі достатньо великий, а величезна кількість інших доданків має достатньо малий порядок, то при вирівнюванні порядків (що відбувається при будь-якому додаванні) мантиса обнуляється і всі доданки ніяким чином не впливають на результат підсумування. В той же час сума великої кількості доданків може бути величезною. Це означає, що помилка може бути як завгодно великою.

33.2. Алгоритм обчислення добутків

Нехай треба обчислити

$$P = \prod_{i=1}^{n} \alpha_i = (...(((\alpha_1 x \alpha_2) x \alpha_3)...) x \alpha_n) \qquad (33.4)$$

Будемо вважати, що співмножники α_i, $\forall i = \overline{1,n}$, зображені в ЕОМ точно.

Теорема 33.2. Якщо ніяка частинна комбінація співмножників не утворює добутку, рівного машинному нулю або машинній нескінченності, то алгоритм, при якому $\overline{\alpha}_1 \geq \overline{\alpha}_2 \geq ... \geq \overline{\alpha}_n$, де $\overline{\alpha}_i$ – двійкова мантиса α_i, буде оптимальним за точністю серії всіх алгоритмів, які відрізняються розташуванням співмножників.

Доведення. Можливе збурення алгоритму (33.4) має вигляд [65]

$$P - \overline{P} = \sum_{j=1}^{n-2} \eta_j \prod_{i=j+2}^{n} \alpha_i + \eta_{n-1}, \tag{33.5}$$

де $\overline{P}$ – наближене значення P, η_i – помилка при i-ому такті обчислення P (P_i-ий частинний добуток).

У (33.5) доданки одного порядку з $2^{-\tau+p}$, p – порядок добутку. Тому

$$\left|P - \overline{P}\right| \le 2^{-\tau+p}\left(\sum_{i=1}^{n-2}\left|\prod_{j=l+2}^{n}\overline{\alpha}_i\right| + 1\right). \tag{33.6}$$

Очевидно, що вираз у дужках досягає свого найменшого значення, якщо мантиси співмножників розташовані в порядку спадання.

Чітко видно, що, якщо $\overline{\alpha}_i = \overline{\alpha}_{i+1} = 1$ (чого не може бути в ЕОМ з плаваючою комою), оцінка (33.5) буде близька до аналогічної оцінки [65].

Якщо Δ – оцінка похибки обчислення P, то вона задовольняє умові

$$\Delta \le (n-1)\cdot 2^{-\tau+p} - (n-1)(n-2)\cdot 2^{-2\tau+p+1} - o\left(2^{-3\tau}\right). \tag{33.7}$$

Якщо хоч один послідовно одержуваний частинний добуток (при вказаному розташуванні співмножників) перетворюється в нуль або ∞, тоді слід перестановкою співмножників виключити виниклу ситуацію (якщо це можливо), але номінально порушити умови Теореми 33.2.

Зауваження 33.3. Обчислення добутків зі співмножниками малих і великих порядків виникає часто в теорії ймовірностей, коли виникає необхідність застосувати біноміальні розподіли випадкових величин. Біноміальні розподіли характеризуються необхідністю виконувати множення великих чисел (біноміальних коефіцієнтів) на малі числа (добутки степенів ймовірностей випадкових величин). Порядок виконання помножень при великих розмірностях оброблюваних чисел часто вирішує долю точності обчислень.

33.3. Алгоритм обчислення високих степенів

Нехай треба обчислити

$$R = L^n \tag{33.8}$$

де L може бути числом або будь-яким алгебраїчним виразом, n – ціле число.

Розглянемо наступний алгоритм обчислення (33.8) [62]:

1) Запишемо число n в двійковій системі (нулевий вигляд). Для цього слід виділити машинне слово такої довжини, скільки вимагає вказане зображення.

2) Нехай число (вираз) закодовано в слові a, а результат міститься в слові за номером r (спочатку це слово містить одиницю).

3) Перевіряємо значення молодшого розряду двійкового зображення n (значення найменшого розряду слова комірки a).

Якщо останній розряд в a містить одиницю, тоді множимо все, що міститься в слові r на L $(1\times L)$ і обчислюємо $L\cdot L = L^2$. Якщо ж останній розряд дорівнює 0, тоді обчислюємо L^2 і перевіряємо значення передостаннього розряду в r. Якщо передостанній розряд містить 1, то r домножуємо на L^2, переходимо до обчислення $L^2\cdot L^2 = L^4$ і повторюємо процес знову, якщо зустрічаються нульові розряди, домноження r на одержаний розряд не здійснюється, а обчислюємо наступний квадрат. Якщо вказана послідовність дій проведена над старшим розрядом з a, то обчислення R завершено.

Теорема 33.3. Описаний вище алгоритм вимагає для своєї реалізації не більше $2[\log_2 n]+1$ множень, причому $[\log_2 n]$ із них будуть піднесенням до квадрата.

Доведення. Зобразимо n у вигляді

$$n = \delta_k 2^k + \delta_{k-1}2^{k-1} + ... + \delta_0 2^0, \qquad (33.9)$$

δ_k може бути нулем або одиницею.

$$R = L^n = L^{\delta_k\cdot 2^k}\cdot L^{\delta_{k-1}\cdot 2^{k-1}}\cdot ....\cdot L^{\delta_0\cdot 2^0}. \qquad (33.10)$$

Неважко побачити, що кожний співмножник вищого степеня (якщо $\delta_i = 1$, $\forall i = \overline{1,k}$) одержується із множників наступного за старшинством степеня шляхом піднесення його в квадрат, то є $L^{2^{i-1}}\cdot L^{2^{i-1}} = L^{2^i}$. Звідси одержуємо, що описаний рекурентний спосіб обчислення R закінчується після $k+1$ послідовних помножень справа наліво (33.10) і k піднесень до квадрату. Якщо в розкладів (33.9) при якій-небудь степені двійки стоїть нуль, то в (33.10) множити на відповідний множник не треба, бо він дорівнює одиниці. З цієї причини кількість множень в (33.10) може лише зменшитися. Теорема доведена.

Наступним міркуванням роботи [63] щодо обчислення добутків можна одержати оцінку похибки заокруглень для розглядуваного способу обчислення R, якщо L дійсне число:

$$\left|R - \overline{R}\right| \le \{2[\log_2 n]+1\}\cdot 2^{-\tau+n\rho}, \qquad (33.11)$$

де $\overline{R}$ – наближене (обчислене) значення R, ρ – машинний порядок L.

Оцінка (33.11) досягається лише при R, розклад (33.9) яких має $\delta_i = 1 \ \forall i = \overline{1,k}$, то є $n = 2^{k+1} - 1$.

Особливу ефективність можна отримати в реалізації цього алгоритму в програмах і мікропрограмах, які здійснюють аналітичні перетворення, так як економія в кількості помножень тягне за собою значну економію в трудомісткій операції приведення подібних. При реалізації процедур, пов'язаних із застосуванням біноміальних розподілів в економічній статистиці цей алгоритм не замінний. Чисельні експерименти підтверджують зазначений висновок.

СПИСОК ЛІТЕРАТУРИ

1. *А. Н. Тихонов, В. Я. Арсенин.* Методы решения некорректных задач. – Москва: Наука, Главная редакция Физматгиз, 1979 г.,– 285 с.
2. *Иванов В. К., Васин В. В., Танана В. П.* Теория линейных некорректных задач и ее приложение. – М.: Наука, 1978 г. – 206 с.
3. *Васин В. В.* Методы решения неустойчивых задач. – Свердловск, Уральский государственный университет, 1989 г. – 94 с.
4. *Васин В. В.* Методы решения плохо обусловленных систем линейных алгебраических уравнений. – Свердловск: Изд-во СГИ, 1988 г. – 54 с.
5. *Курош А. И.* Курс высшей алгебры. – М.: Наука, 1965 г. – 431 с.
6. *Заборовець М. О.,* Левченко Ф. А., Охріменко М. Г. Сучасні методи розв'язування систем лінійних алгебраїчних рівнянь: Навч.-метод. посіб. – К.:КНЕУ, 2006. – 76 с.
7. *А. Г. Курош.* Курс высшей алгебры, М., Физматгиз, 1963 г.,– 432 с.
8. *М. Г. Охрименко.* К вопросу о решении систем линейных балансовых уравнений, ДАН СССР, 1977, Том 234, №1, 4с.
9. *М. Г. Охріменко, І. В. Сергієнко, О. С. Стукало.* Ефективна організація обчислень при розв'язку систем лінійних рівнянь, ДАН УРСР, серія „А" №1, 1978 р., 67-70 с.
10. *О. В. Волошин, С. О. Мащенко, М. Г. Охріменко.* Алгоритм послідовного аналізу варіантів для розв'язування балансових моделей, Доп. АН УРСР, Сер. „А" Фіз. – мат. та техн. науки, 1988, №9, 67-70 с.
11. *Воеводин В. В.* Численные методы алгебры. Теория и алгоритмы, М.: Наука, 1966. – 248 с.
12. *Тихонов А. Н.* О решении некорректно поставленых задач и методе регуляризации // Докл. АН СССР, - 1963. – Т. 151, №3.- С. 501-504.
13. *Іванов В. К., Васин В. В., Танана В. П.* Теория линейных некорректных задач и ее приложения. – М.: Наука, 1978. – 206 с.
14. *Назимов А. Б.* Исследования методов регуляризации сдвигом и его приложения. Дис. канд. физ. – мат. наук. М. 1986 г. – 95 с.
15. *Воеводин В. В.* Линейная алгебра. – М.: 1974.- 326 с.
16. *Тихонов А. Н., Гончаровский А. В., Степанов В. В., Яголе А. Г.* Регуляризующие алгоритмы и априорная информация. – М. : Наука, 1983. – 198 с.
17. *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. – М.: Наука, 1974. – 224 с.

18. *Заикин П. Н., Мечёнов А. С.* Некоторые вопросы численного решения интегральных уравнений первого порядка методом регуляризации // Отчет ВУМГУ, №144-Т3, Изд. МГУ, 1971. – 29с.//.
19. *Войникко Г. М.* Оценки погрешности метода последовательных приближений для некорректных задач // Автоматика и телемеханика. – 1980., №3. – С.84-92//.
20. *Васин В. В.* Дискретизация, итерационно-аппроксимационные алгоритмы решения неустойчивых задач и их приложения: Дис. докт. физ. – мат. наук. – Свердловск, 1980. – 299 с.
21. *Реклейтис Г., Рейвиндрен А., Рэгсдел К.* Оптимизация в технике: В 2 томах. – М.: Мир, 1986. – Т.1.-349 с; Т.2 – 32 с.
22. *Лоусон Ч., Херсон Р.* Численное решение задач метода наименьших квадратов. – М.: Наука, 1986. – 230 с.
23. *Тихонов А. Н.* О приближенных системах линейных алгебраических уравнений // Журн. вычисл. матем. и мат. физика. – 1980. -№6 – 1. 1373-1383//.
24. *Годунов С. К.* Решение систем линейных уравнений. – М.: Наука, 1980. – 177 с.
25. *Фадеев Д. К., Фадеева В. Н.* Вычислительные методы линейной алгебры. – М.: Физматгиз, 1963 г. – 734 с.
26. *Волович В. М.* О решении систем линейных уравнений клеточными методами // Вычисл. методы и программир. – М.: Изд. МГУ, 1965. – Вып. 3. – с. 106 – 133
27. *Cline A. K., Moler C. B., Stewart C. W., Wilkinson J. H.* An estimate for the condition number of matrix // SIAM J. Numer. Analysis. – 1979. – Vol. 16, no. 2.- P. 368 – 375.
28. *Молчанов И. Н.* Машинные методы решения прикладных задач. Алгебра, приближение функций. – Киев: Наукова думка, 1987. – 288 с.
29. *Молчанов И. Н.* Проблемы создания пакетов программ лмнейной алгебры // Вычислительные методы линейной алгебры: Тр. Всесоюзн. конф. (М. 23 – 25 августа 1982 г.) – М.: 1983. – С. 187 – 202.
30. *Уилкинсон Дж. Х. Райнш К.* Справочник алгоритмов на языке АЛГОЛ. Линейная алгебра. – М.: Машиностроение, 1976. – 390 с.
31. *Форсайт Дж. Маулер К.* Численное решение систем линейных алгебраических уравнений. – М.: Мир, 1969 г. – 167 с.
32. *Гордонова В. И., Морозов В. А.* Численные методы выбора параметра в методе регуляризации // Журн. вычисл. математики и мат. физики. – 1973. – Т. В. №3.– С. 539-545.

33. *Колмогоров А. Н.* О неравенствах между верхними гранями последовательных производных произвольной функции на бесконечном интервале // Учен. зап. МГУ, Математика, 1930. Т.30, кн.3. С. 3–16.
34. *Колмогоров А. Н., Фомин С. В.* Элементы теории функций и функционального анализа. – М.: Наука, 1976 г., 542. с.
35. *Стечкин С. Б.* Наилучшее приближение линейных операторов // Мат. заметки, 1967. Т.1. №2.– С. 137–148.
36. *Альберг Дж., Нельсон Э., Уолт Дж.* Теория сплайнов и ее приложения. – М.: Мир, 1972. – 316 с.
37. *Стечкин С. Б., Субботин Ю. Н.* Сплайны в вычислительной математике. – М.: Наука, 1976. – 248 с.
38. *Зав'ялов Ю. С., Квасов В. І., Мірошниченко В. Д.* Методи сплайн функцій. – М.: Наука, 1980. – 382 с.
39. *Reinsh C. N.* Smoothing by spline function // Numer. math., 1967. Vol. 10, p. 177-183.
40. *Гребенников А. И.* Алгоритмы и программы аппроксимации функций одной переменной сплайнов и приложения. – М.: Изд-во Моск. ун-та, 1987. – 80 с.
41. *Соболев С. Л.* Некоторые применения функционального анализа в математической физике. – Новосибирск: Узд. СО АН СССР, 1962. – 255 с.
42. *Страхов В. Н.* Теория приближенного решения линейных некорректных задач в гильбертовом пространстве и ее использование в разведочной геофизике. I, II // Изв. АН УССР, Физика Земли. 1969, №8, с. 50-53; №9.– С. 64–96.
43. *Хромов Г. В.* О задаче восстановления производной // Вычислительные методы и программирование. – Саратов: Изд-во Сарат. ун-та , 1970.– Т.4.– С. 3–13.
44. *Карманов В. Г.* Математическое программирование: Учебное пособие. – М.: Наука, 1980. – 256 с.
45. *Licchetti R., Paizone F.* Hademard and Tyhonov wellpossedness of a certain class of convex functions // j.Math. Anal. and Appl., 1962. Vol. 88, p. 204-215.
46. *Люстерник Л. А., Соболев В. И.* Элементы функционального анализа. – М.: Наука, 1965. – 519с.
47. *Войникко Г. М.* Аналіз дискретизационных методов. – Тарту: Изд-во Тарт. университета, 1976. – 161 с.
48. *Stummel F.* Diskrete Konvergenz Linearer Operatoren I,II Math. Ann, 1970. Vol. 190, №1. s. 45-92; Math. Z. 1971. Vol. 120. s. 231-264.

49. *Daniel J. W.* On the approximate minimization of functionale // Math. Comput. 1969. Vol. 23,№107, p. 573-581.
50. *Васин В. В.* Дискретная аппроксимация и устойчивость в экстремальных задачах // Журн. вычисл. матем. и мат. физики. 1982, е. 22. №4,– С. 824-839.
51. *Ageev A. L., Babanov Y. A., Vasin V. V. et al*. Amorphous problem in EXAFS data analysis // Phys. Stat. Sol. 1983. Vol. 177, P. 343-350.
52. *Ахиезер Н. И.* Вариационное исчисление. – Харьков: Высшая школа, 1981. – 168 с.
53. *Бакушинский А. Б.* Об одном численном методе решения интегральных уравнений Фредгольма I рода // Журн. вычисл. мат. и мат. физики, 1965, т.5, №4.– С. 744-749.
54. *Тихонов А. Н., Гласко В. Б.* О приближенном решении интегральных уравнений Фредгольма I рода // Журн. вычисл. мат. и мат. физики, 1964, т.4, №3.– С. 564-571.
55. *Васин В. В.* Проксимальный алгоритм с проектированием в задачах выпуклого программирования. – Свердловск, 1982. 47 с./ Препринт / АН СССР, Урал. науч. центр, ин-т математики и механики.
56. *Vasin V. V.* Iterative methode for approximate solution of illposed problems with a priori information and their applications // Inverse and ill-posed problems. Boston: Acad. Press, 1987. P.211-229. (Notes and Reports in Math., in Sci and Eng.; Vol.4).
57. *Форсайт Дж., Малькольм М., Маулер К.* Машинные методы математических вычислений. – М.: Мир, 1980. – 279 стр.
58. *Охрименко М. Г.* Анализ точности вычисления сумм,Сб. Математическое обеспечение и организация вычислительного процесса, 1974, Киев.– С. 161–168.
59. *Охрименко М. Г., Савчак О. Н.* Замечание к вопросу о вычислении целых степеней на ЭВМ, Сб. Математическое обеспечение и организация вычислительного процесса, 1974, Киев. ИК АН УССР – С. 182–187.
60. *Охрименко М. Г.* Анализ точности вычисления сумм и произведений на ЭВМ. Сб. Оптимизация и организация вычислений, изд. ИК АН УССР, Киев, 1972.– С. 74–81.
61. *Витенко И. В.* Оптимальные алгоритмы сложения и умножения на машинах с плавающей запятой. ЖВМ и МФ, т.8, №5, 1968.– С. 108-114.
62. *J. H. Wilkinson.* Rounding Errors in Algebraic Processes, London,Her Majesty's stat.Offic. 1963, p 564.

63. *Джумаев С. О.* О приближенном вычислении псевдорешения // Докл. АН Тадж. ССР, 1982. Т.25, №10.– С. 584-587.
64. *Licchetti R., Patrone F.* Hadamard and Tyhonov Well-possedness of a Certain of Convex Functions // J. Math. Anal. and Appl. 1982. Vol. 88. p. 204–215 c.

НАВЧАЛЬНЕ ВИДАННЯ

М.Г. Охріменко, О.А. Жуковська, О.О. Купка

МЕТОДИ РОЗВ'ЯЗУВАННЯ НЕКОРЕКТНО ПОСТАВЛЕНИХ ЗАДАЧ

Підручник

Керівник видавничих проектів – *Б.А.Сладкевич*
Друкується в авторській редакції
Дизайн обкладинки – *Б.В. Борисов*

Підписано до друку 22.09.2007. Формат 60x84 1/16.
Друк офсетний. Гарнітура PetersburgC.
Умовн. друк. арк. 10,5.
Наклад 1000 прим.

Видавництво "Центр учбової літератури"
вул. Електриків, 23
м. Київ, 04176
тел./факс 425-01-34, тел. 451-65-95, 425-04-47, 425-20-63
8-800-501-68-00 (безкоштовно в межах України)
e-mail: office@uabook.com
сайт: WWW.CUL.COM.UA

Свідоцтво ДК №2458 від 30.03.2006