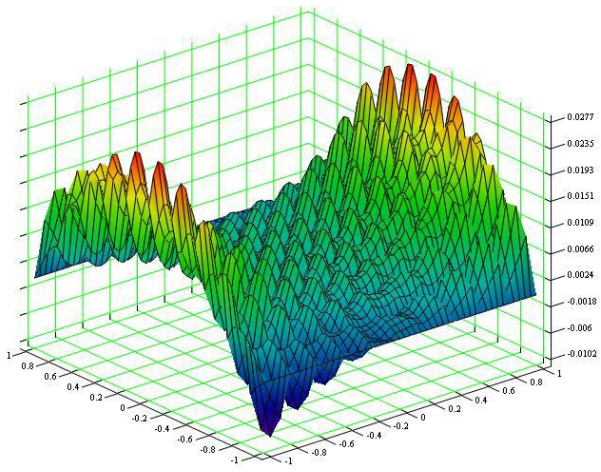
Національний університет "Львівська політехніка"

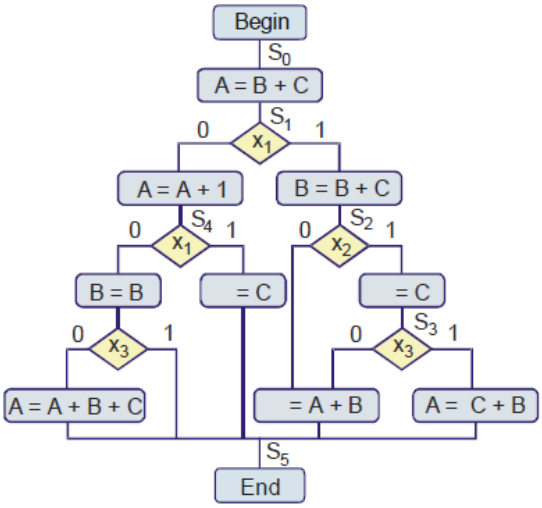
Харківський національний університет радіоелектроніки

Яворський Н.Б., Теслюк В.М., Литвинова Є.І.

Комп'ютерні методи в інженерії мікроелектромеханічних систем

Навчальний посібник





№530785-TEMPUS-1-2012-1-PL-TEMPUS-JPCR

Tempus

Навчальний посібник "**Технології тестування мікроситем"** створено для допомоги вищим навчальним закладам України впровадити нову магістерську навчальну програму "Проектування мікросистем".

Посібник "**Технології тестування мікроситем"** створено при підтримці Європейського Союзу за Спільним Європейським Проектом "Curricula Development for New Specialization: Master of Engineering in Microsystems Design" (MastMST), ідентифікаційний номер 530785-TEMPUS-1-2012-1-PL-TEMPUS-JPCR.

Координатор проекту проф. Збігнєв Лісік, Технічний університет м.Лодзь, Польща.

Учасники проекту:

- Національний університет "Львівська політехніка", м. Львів, Україна, координатор проф. Михайло Лобур.
- Київський Національний університет ім. Тараса Шевченка,м. Київ, Україна, координатор проф. Валерій Скришевський.
- Харківський національний університет радіоелектроніки, м. Харків, Україна, координатор проф. Володимир Хаханов.
- Донецький національний технічний університет, м. Красноармійськ, Україна, координатор проф. Володимир Святний.
- Технічний університет м.Ільменау, Німеччина, координатор проф. Іво Рангелов.
- Ліонський Національний інститут прикладних наук, Франція, координатор проф. Александра Апотолюк
- Університет Павії, м.Павії, Італія, координатор проф. Паоло Ді Барба

Посібник схвалено редакційним комітетом (проф. Паоло Ді Барба (Університет Павії) - співголова, проф. Александра Апостолюк(Ліонський Національний інститут прикладних наук) – співголова, члени: проф. Збігнєв Лісік(Технічний університет м.Лодзь), д-р Яцек Подгурські (Технічний університет м.Лодзь), Д-р Януш Возний (Технічний університет м.Лодзь), Д-р Валентин Іщук (Технічний університет м.Ільменау), Д-р Марія-Евеліна Могначі (Університет Павії), Д-р Роберто Галді (Університет Павії)) 6 травня 2016, м. Павія, Італія

Автори висловлюють глибоку вдячність керівництву вищеназваних університетів за всебічну підтримку Проекту.

Textbook "**Computer Methods in Microsystems Engineering**" developed to help higher education institutions in Ukraine to introduce new master's educational program "Designing microsystems".

Textbook "**Computer Methods in Microsystems Engineering**" was created with the support of the European Union within the Joint European Project "Curricula Development for New Specialization: Master of Engineering in Microsystems Design" (MastMST), identification number 530785-TEMPUS-1-2012-1-PL-TEMPUS-JPCR.

Project Coordinator prof. Zbigniew Lisik, Lodz University of Technology, Lodz, Poland.

Учасники проекту:

- Lviv Politechnical National University, Lviv, Ukraine ,
  Coordinator prof. Mykhailo Lobur.
- Taras Shevchenko National University of Kyiv, Ukraine,
  Coordinator prof. Valeriy Skryshevsky.
- Kharkiv National University of Radioelectronics, Ukraine ,
  Coordinator prof. Vladimir Hahanov.
- Donetsk National Technical University, Krasnoarmiysk,
  Coordinator prof. Volodymyr Sviatny.
- Ilmenau University of Technology, Germany,
  Coordinator prof. Ivo Rangelow.
- Lyon Institute of Applied Sciences, France,
  Coordinator prof. Alexandra Apostoluk.
- University of Pavia, Italy,
  Coordinator prof. Paolo Di Barba.

The Handbook was approved by Editorial Committee (prof. Paolo Di Barba (University of Pavia) - Co-Chair, prof. Alexandra Apostoluk (Lyon Institute of Applied Sciences) – Co-Chair, members: prof. Zbigniew Lisik (Lodz University of Technology), Dr Jacek Podgorski (Lodz University of Technology), Dr Janusz Wozny (Lodz University of Technology), Dr Valentyn lshchuk (Ilmenau University of Technology), Dr Maria Evelina Mognaschi (University of Pavia), Dr Roberto Galdi (University of Pavia) May 6, 2016, Pavia, Italy.

The authors express their deep gratitude to the aforementioned universities for full support of the project.

Назарій Яворський
Василь Теслюк
Євгенія Литвинова

Комп'ютерні методи в інженерії мікроелектро-механічних систем

Навчальний посібник

Львів – 2015

Яворський Н.Б., Теслюк В.М., Литвинова Є.І. Комп'ютерні методи в інженерії мікроелектромеханічних систем: Навчальний посібник. – Львів: Видавництво Національного університету "Львівська політехніка",2015. – 280 с.

Робота виконана в рамках проекту Curricula Development for New Specialization: Master of Engineering in Microsystems Design / MastMST, Ідентифікаційний номер 530785-TEMPUS-1-2012-1-PL-TEMPUS-JPCR. Програма фінансування ЄК: Міжрегіональна програма Європейського сусідства і партнерства. Цільова група: студенти, випускники, викладачі та адміністрація університетів, керівники промислових підприємств, міністерство освіти і науки України.

Основною метою проекту є створення умов в українських технічних університетах для наскрізного 3-х рівневого навчання по спеціальності Проектування та Інженерія Мікросистем відповідно до регіональних потреб ринку праці. Згідно поставленої мети, дана робота орієнтована на вирішення завдання підготовки Бакалаврського курсу лекцій з дисципліни "Computer Methods in Microsystem Engineering / Комп'ютерні методи в інженерії мікроелектромеханічних систем".

Робота підготовлена спільно: кафедрою систем автоматизованого проектування Інституту комп'ютерних наук та інформаційних технологій Національного університету "Львівська політехніка" в особі аспіранта Яворського Н.Б., д.т.н, проф. Теслюка В.М.; та кафедрою автоматизації проектування обчислювальної техніки факультету комп'ютерної інженерії й управління Харківського національного університету радіоелектроніки в особі д.т.н., проф. Литвинової Є.І.

Рецензенти: **Березький О.М.**, д-т тех. наук, зав. кафедри комп'ютерної інженерії Тернопільського національного економічного університету, професор;

Рак Т.Є., д-т тех. наук, проректор з науково-дослідної роботи львівського державного університету безпеки життєдіяльності, доцент;

Цмоць І.Г., д-т тех. наук, зав.кафедри автоматизованих систем управління Національного університету "Львівська політехніка", професор.

Рекомендувала Вчена рада
Національного університету "Львівська політехніка"
(протокол № 12 від 25.09.2015 р.)

ВСТУП

Протягом останніх років індустрія мікроелектроніки активно розвивається як у напрямку мініатюризації мікроелектронних пристроїв, так і у напрямку інтеграції в єдине ціле різних за фізичними принципами дії функціональних пристроїв. Називають такі інтегральні пристрої – мікроелектромеханічні системи (МЕМС). Процес розроблення таких мікросистем відбувається з використанням досвіду, знань, технічних прийомів і методів з різних галузей науки і техніки, що зумовлює необхідність функціональної інтеграції неоднорідних комп'ютерних систем або розроблення принципово нових інформаційних технологій (ІТ) проектування МЕМС. Центральне місце таких інформаційних технологій займає математичне забезпечення.

У галузі МЕМС за останні 30 років відбулися суттєві зміни, зокрема: значно вдосконалені технології їх виготовлення, відбувся стрімкий розвиток інфраструктури МЕМС, розпочато виробництво великої кількості різних конструкцій елементів МЕМС (давачі тиску, акселерометри, струменеві друкуючі головки, цифрові дзеркальні дисплеї) та в цілому МЕМС. Освоєність та домінуюча роль промисловості ІС разом з новими технологіями МЕМС відкриває нові можливості для мікроелектромеханічних систем.

У дані роботі розглядаються комп'ютерні методи в моделлюванні та проектуванні мікроелектромеханічних систем. Основна увага приділяється компонентному рівню проектування, що передбачає аналіз протікання фізичних процесів в МЕМС. Для цього детально описуються особливості використання чисельного методу рішення задач математичної фізики – методу скінченних елементів. Крім того, для повноти висвітлення матеріалу, у роботі розглядаються питання валідації та забезпечення якості при проектуванні МЕМС. Кожен розділ роботи супроводжується детальними ілюстративними прикладами, що спрямовані на краще розуміння та засвоєння матеріалу.

Перший розділ даної роботи присвячений основам курсу комп'ютерних методів в інженерії мікросистем та зокрема методам ієрархічного проектування мікроелектромеханічних систем. Тут описуються види електромеханічних систем та параметри росту їх світового виробництва. Порівнюються технології виготовлення МЕМС, що зазвичай визначають економічну та військову незалежність держави, забезпечують розвиток космічної області та конкурентноздатності продукції на світовому ринку. Наводяться структура та схема роботи мікродавачів та мікроактюаторів. Розділ описує застосування блочно-ієрархічного підходу до проектування МЕМС, що включає системний, схемотехнічний та компонентний рівні відповідно до яких застосовуються методи проектування "зверху-вниз", "знизу-вгору" та їх комбінації. Наведено динаміку тенденцій проектування на різних рівнях за останні роки. Розділ розглядає методи автоматизованого проектування МЕМС та існуючі системи проектування МЕМС на компонентному рівні, даються їх порівняльні характеристики.

Другий розділ даної роботи має за мету ознайомити читача з основами

формалізації задач компонентного рівня проектування МЕМС. Зокрема, тут розглядаються питання опису систем диференціальними рівняннями з частинними похідними еліптичного, гіперболічного та параболічного типу, що включають: класифікацію рівнянь; використання операторних форм запису; визначення початкових і крайових умов, а також коректність постановки відповідних задач.

У третьому розділі описуються основи методу скінченних елементів – найрозвиненішого чисельного методу наближеного рішення задач моделювання фізичних процесів в неперервних середовищах, що активно використовується на компонентному рівні проектування мікроелектромеханічних систем. Для кращого розуміння, дається коротка історична довідка основних етапів розвитку задач моделювання та методу скінченних елементів у контексті більш загальних методів зважених нев'язок, з допомогою яких можна розв'язати практично будь-яку задачу, визначену диференціальними рівняннями з частинними похідними. Піднімаються питання виведення слабких форм визначальних рівнянь з допомогою визначення головних та природних крайових умов задач. Розглядаються найпростіші симплекс елементи та їх геометричних зміст. Наводяться теоретичні властивості чисельних методів наближеного рішення задач, що включають апріорні та апостеріорні оцінки точності, стійкості та збіжності. Описуються вимоги кускової визначеності інтерполяційних функцій скінченних елементів, їх лінійна незалежність повнота та допустимість використання.

У четвертому розділі розглядаються особливості застосування методу скінченних елементів на компонентному рівні проектування МЕМС. У такому контексті, на основі теорії подібності описано фізичні аналогії скінченно-елементних моделей та дискретних систем загалом. Наведено способи рішення мультифізичних задач та систем диференціальних рівнянь. Піднято питання моделювання нелінійних та нестаціонарних задач.

П'ятий розділ присвячений особливостям апроксимації методом скінченних елементів. Тут детально описується формулювання інтерполяцій високих порядків точності. Наводяться відмінності між симплекс, комплекс, мультиплекс та криволінійними скінченними елементами. Детально описано методи чисельного інтегрування при побудові скінченно-елементних моделей. Наведено методи побудови криволінійних елементів, зокрема з використанням змішувальних процесів, а також методи побудови, так званих, нескінченних елементів. Піднято питання узгодженості інтерполяційного базису.

Шостий розділ описує методи декомпозиції обчислень на компонентному рівні проектування МЕМС. Тут наводяться основи доменної декомпозиції та розпаралелювання обчислень. Розділ присвячений яскравому прикладу таких методів – відносно молодому методу скінченних елементів розривів та з'єднань. Основний акцент зроблено на висвітлення геометричної інтерпретації методу, шляхом розгляду взаємозв’язків просторів лінійних операторів та векторів, що в них лежать. Для цього піднімаються питання знаходження псевдообернених матриць та їх геометричного змісту.

У сьомому розділі представлені основні напрямки технологій проектування, валідації та забезпечення якості мікроелектромеханічних систем. Пропонується технологія діагностування моделей систем на кристалах, яка базується на використанні транзакційних графів. Описується метод діагностування, спрямований на зменшення часу виявлення несправностей і пам'яті для зберігання діагностичної матриці за рахунок формування тернарних відносин між тестом, монітором і функціональним компонентом. Вирішуються: завдання розробки моделей цифрової системи у вигляді транзакційного графа і мультидерева таблиць несправностей, а також тернарні матриці активації функціональних компонентів обраного набору моніторів за допомогою тестових послідовностей; завдання розробки методів аналізу матриці активації з метою виявлення несправних блоків із заданою глибиною і синтезу логічних функцій для подальшого вбудованого апаратного діагностування несправностей.

1. Основи курсу та методи ієрархічного проектування МЕМС

1.1. Особливості та перспективи розвитку МЕМС

Рубіж XX–XXI століть характеризується інтенсивним розвитком існуючих та появою нових міждисциплінарних науково-прикладних областей. Однією з них є область мікроелектромеханічних систем (МЕМС) [1], [2], [3], які об'єднують в собі досягнення механіки, мікроелектроніки, оптики, електротехніки та інших науково-практичних областей [4]. Інтегральні пристрої даного типу володіють рядом переваг у порівнянні з макропристроями, вони:

- надійніші,
- дешевші,
- легші,
- інтеграція наукових областей носить синергетичний характер, виготовляють їх за груповою технологією тощо.

В загальному випадку, всі об'єкти проектування можна розділити згідно їх лінійних розмірів, приклад відповідного поділу наведено на *Рис. 1.1*. Як можна визначити з *Рис. 1.1*, пристрої з лінійними розмірами від декількох сантиметрів до міліметра називають мініпристроями, пристрої з лінійними розмірами від кількох міліметрів до мікрона – мікропристрої, а пристрої з лінійними розмірами меншими 1 – 0,1 мікрона – нанопристроями (наноелектромеханічні системи (НЕМС)) [5], [6], [7], [8]. Мікроелектромеханічні системи, як правило, відносяться до міні- та мікропристроїв і виготовляють їх за інтегральними груповими мікроелектронними та мікромеханічними технологіями [9].

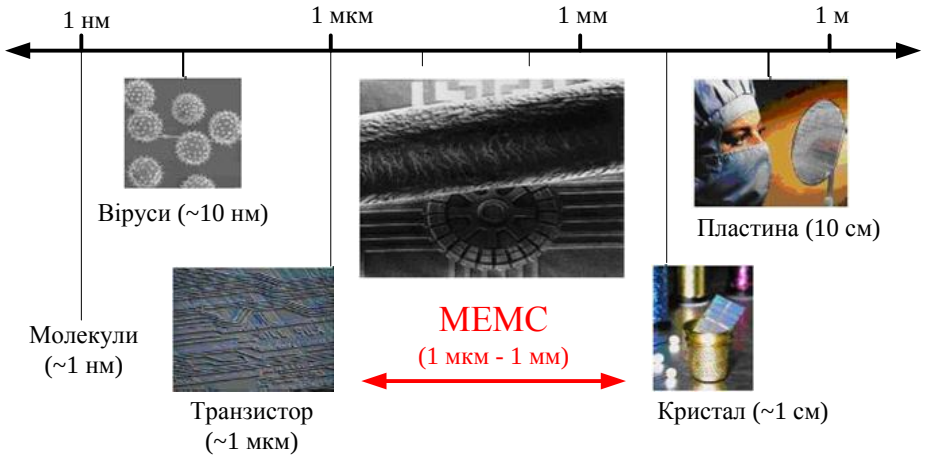


Рис. 1.1 Лінійні розміри об'єктів розробки

Разом з тим, існує область і традиційних електромеханічних систем (ЕМС) розміри яких більші за один сантиметр. Хоча, зрозуміло, що названі границі є чисто умовні та розмиті. Отже, згідно розмірного фактору існують ЕМС, МЕМС та НЕМС (див. *Рис. 1.2*). Особливістю цих трьох великих груп електромеханічних систем є те, що для опису принципів функціонування для ЕМС і МЕМС можна використати класичну теорію механіки,

електромагнетизму та ін., а для пристроїв НЕМС – квантову теорію наноелектротехніки. Особливі проблеми виникають, з точки зору теоретичних основ опису роботи МЕМС, при розмірах близьких та менших від одного мікрометра, де не завжди існуюча класична теорія коректно описує фізичні процеси, які відбуваються в конструкціях цих інтегральних пристроїв.

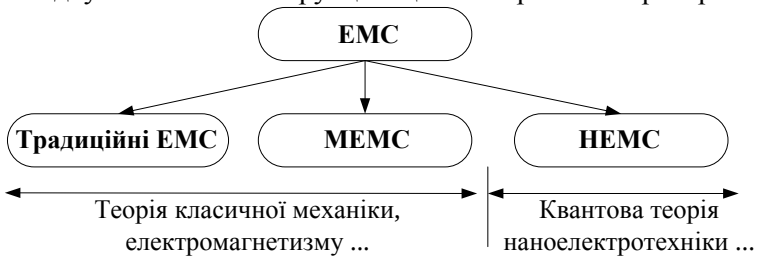


Рис. 1.2 Види електромеханічних систем

Переваги МЕМС над традиційними технічними пристроями, обумовили їх широке і масштабне використання. Тому сотні фірм світу займаються виготовленням МЕМС та використовують їх в технічних системах.

До найбільш відомих фірм [1], [2], [3], [4], [5], які займаються розробкою та виготовленням МЕМС, відносяться Analog Devices (США), Tanner Reaserchs (США), Berkeley Sensor & Actuator Center (BSAC), University of California (США), Tima-CMP (США), Sandia National Laboratories (США), Texas Instruments, Inc. (США), Московський інститут електронної техніки (Росія), Центр мікротехнологій та діагностики Санкт-Петербургського державного електротехнічного університету (Росія) та інші.

Згідно повідомлень міжнародної групи виробників МЕМС [10], ринок цих інтегральних пристроїв постійно зростає на 12 – 15 % кожного року (*Рис. 1.3*).

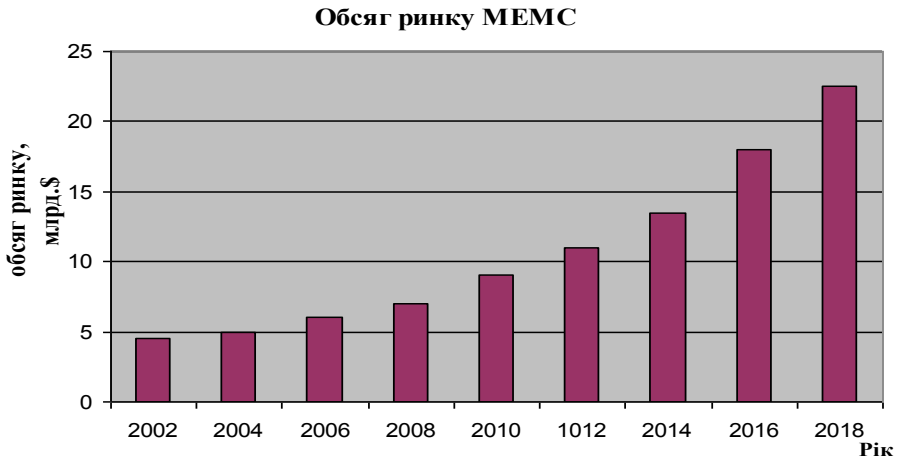


Рис. 1.3 Параметри росту світового ринку МЕМС

Для прикладу, в 2002 р. ринок МЕМС складав близько 4 млрд. US$, а в 2012 році ця цифра перевищила 11 млрд. US$. Тобто, за 10 років об'єм вартості ринку майже потроївся, а за довгостроковими прогнозами він буде розвиватись ще швидшими темпами (в 2018 буде складати близько 22,5 млрд. $), що свідчить про широкомасштабне впровадження МЕМС в сучасні промислові розробки та вироби [11], [12].

В процесі розвитку галузі МЕМС відбувається зміна асортименту, що відображено на *Рис. 1.4, Рис. 1.5* та *Рис. 1.6* [13]. Відсоткове співвідношення видів МЕМС, які знаходять масове використання в промислових виробах, зростає. Хоча відсоткове зменшення певних видів МЕМС не свідчить про зменшення до них уваги науковців та об'єму виробництва на фоні різкого росту сумарної кількості інтегральних пристроїв цього типу.

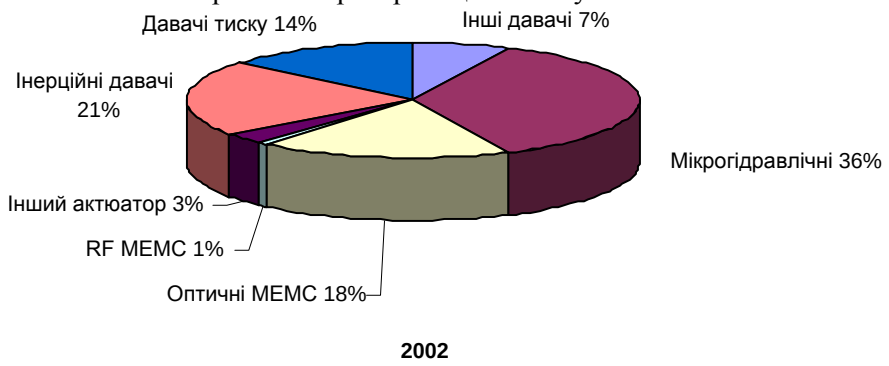


Рис. 1.4 Відсоткове співвідношення пристроїв МЕМС по видах за 2002 р.

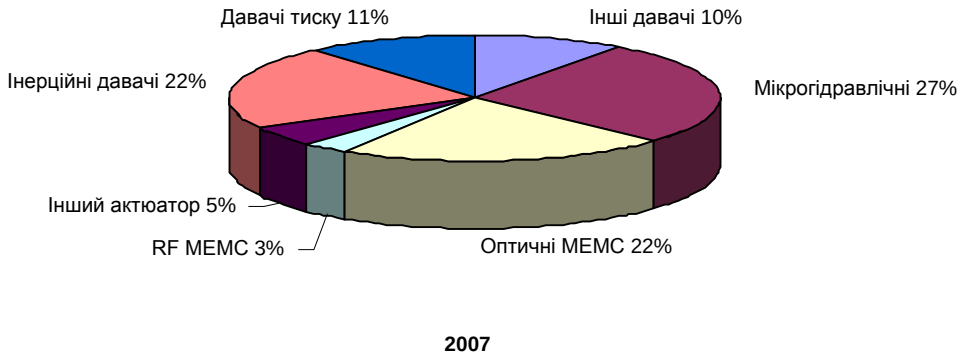


Рис. 1.5 Відсоткове співвідношення пристроїв МЕМС по видах на 2007 р.

Технології виготовлення МЕМС належать до так званих "критичних" технологій та технологій подвійного призначення (*Рис. 1.7*). Тому, в більшості випадків, дані технології визначають економічну та військову незалежність держави, забезпечують розвиток космічної області та конкурентноздатності продукції на світовому ринку. Разом з тим, вони базуються на відомих технологіях, зокрема:

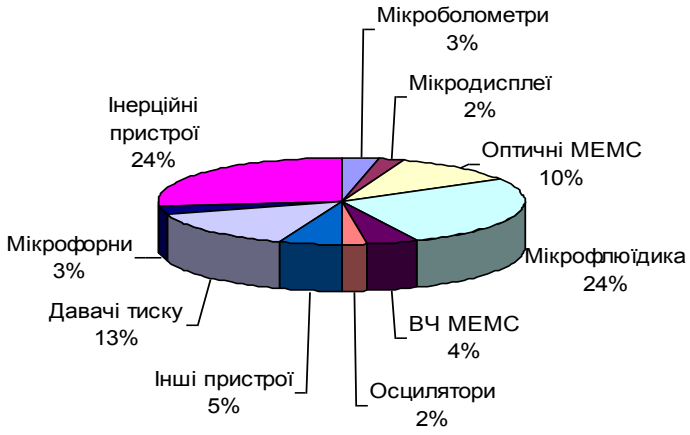


Рис. 1.6 Відсоткове співвідношення пристроїв МЕМС по видах на 2016 р. (прогноз Yole Development)

- мікромеханіки,
- мікроелектроніки,
- оптоелектроніки,
- акустоелектроніки,
- мехатроніки,
- мікроробототехніки,
- прецизійної механіки,
- матеріалознавства та інші.

Початковий етап розвитку будь-якого нового науково-прикладного напряму пов'язаний з труднощами в області термінології, стандартизації, тощо. Відповідні проблеми притаманні і МЕМС на даному етапі розвитку.

Історично склалося так, що перші МЕМС, які включали електронну складову (інтегральні схеми) та електромеханічні пристрої з наступним їх розміщенням на одному напівпровідниковому кристалі та використовували для виготовлення мікротехнології, були розроблені в США та отримали назву мікроелектромеханічні системи. Ця назва пішла від фізичного принципу роботи першого додаткового електромеханічного інтегрального пристрою (мікроелектромеханічні системи). Тому, до цього часу, в США ці пристрої називають мікроелектромеханічними системами, хоча вони можуть включати інтегральні елементи, які використовують інші принципи роботи. В Європі та Росії їх називають пристроями мікросистемної техніки, або мікросистеми, а в Японії – мікромашинами.

Відповідно, в США використовують термін "мікроелектромеханічні системи" і притримуються наступного визначення: "МЕМС – це інтегральні мікропроцесорні системи, які комбінують електричні та механічні компоненти виготовлені за технологіями сумісними з технологіями ІС з розмірами від мікрометрів до декількох міліметрів, а наявність зв’язків між актюаторами,

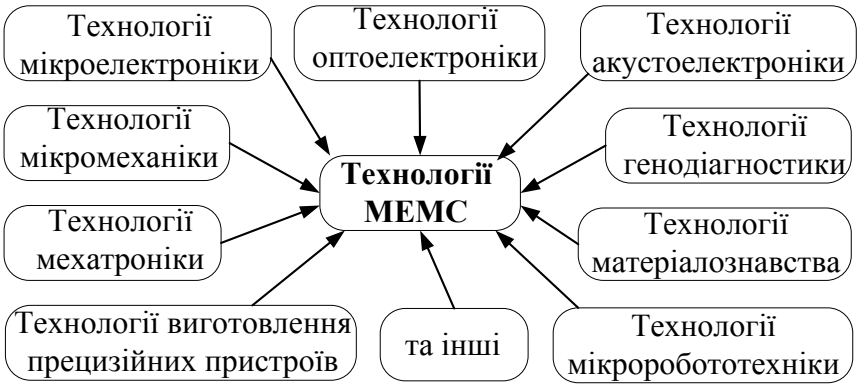


Рис. 1.7 Базові технології подвійного призначення

мікродавачів та системою обробки дає можливість відчувати та контролювати навколишнє середовище".

В Європі та Росії притримуються терміну "мікросистема" [14]. Мікросистема – інтелектуальна мінімізована система, яка володіє сенсорними, процесорними і/чи актюаторними функціями та використовує комбінацію двох чи більше пристроїв, що діють на основі використання електричних, механічних, оптичних, хімічних, біологічних, магнітних чи інших властивостей і інтегрованих на одному чіпі чи мультичіповій платі.

Починаючи з 1995 року дана область надзвичайно активно починає розвиватися в Японії та азійських країнах, де досить часто використовують термін "мехатроніка" або "мікромашини" та визначення: мікромашини (мехатроніка) складаються з функціональних елементів розміром у кілька міліметрів і здатних утворити комплексний мікроскопічний пристрій.

Слід зауважити, що всі визначення передбачають наявність таких основних елементів, як: розмірність, використання мікротехнологій для виготовлення, наявність інтерфейсу з оточуючим середовищем та засобів впливу на нього тощо.

Отже, МЕМС – науково-технічний напрямок, метою якого є створення в обмеженому об’ємі твердого тіла, або на його поверхні, мікросистем.

МЕМС контролює зміни в навколишньому середовищі за допомогою мікродавачів. Отже – це є пристрої, які реєструють зміни в оточуючому середовищі, або реагують на фізичні впливи. В загальному випадку принцип дії мікродавачів наведено на *Рис. 1.8*. Зміни в зовнішньому середовищі діють на чутливий елемент мікродавача. Чутливий елемент мікродавача (перетворювач, трандюсер), в свою чергу, перетворює зміну енергії зовнішнього середовища в зміну вихідного контрольованого параметра, який надалі, як правило, опрацьовує електрична схема. В якості зовнішнього впливу можна розглядати тиск, температуру, напруженість магнітного та електричного полів, силу будь-якої природи, деформацію тощо. Конструкція та особливості чутливого елемента залежать від контрольованого середовища, зміну якого необхідно реєструвати. Для прикладу, якщо в якості параметра зовнішнього середовища

розглядати тиск *P*, то чутливим елементом в мікродавачах МЕМС використовується, як правило, тонка кремнієва (полікремнієва) мембрана, на яку нанесено провідний для електричного струму матеріал (алюміній, золото та ін.) і яка виступає в ролі однієї з обкладок електричного конденсатора. В даному випадку вихідним контрольованим параметром є зміна електричної ємності ΔC, що реєструється, підсилюється і обробляється електричною схемою в інтегральному виконанні.

В даному випадку маємо наступну послідовність зміни параметрів ΔP→ΔL→ΔC. Тобто, зміна тиску призводить до змін переміщень в тонкій пластині, а зміни переміщень призводять до змін електричної ємності. Такий мікродавач називають мікродавачем тиску ємнісного типу.

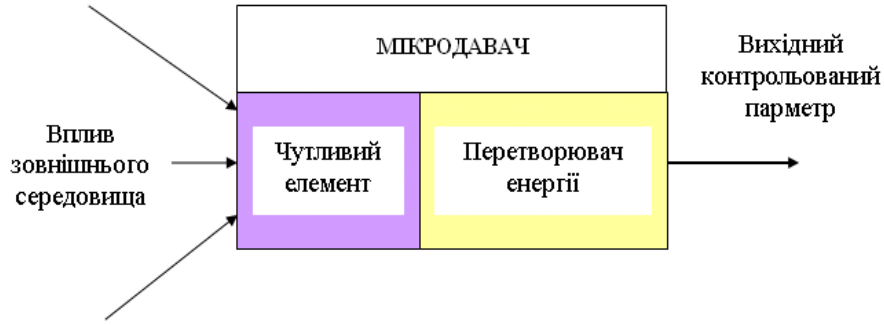


Рис. 1.8 Основні елементи мікродавача МЕМС

Реєструвати зміни тиску в зовнішньому середовищі можна і з допомогою дещо іншої схеми, а саме: ΔP→ΔL→ΔG→ΔR. В цьому випадку конструкція мікродавача відрізняється лише тим, що присутні п'єзорезистори на краях тонкої кремнієвої пластини. Принцип роботи мікродавача включає такі зміни параметрів. Зміна тиску навколишнього середовища призводить до переміщень пружного елемента. Згенеровані переміщення створюють зміну напружень ΔG на краях жорстко защемленої пластини, які, в свою чергу, на основі п'єзоефекту призводять до змін опору п'єзорезисторів. Вихідними контрольованими параметрами даного мікродавача є зміна опору п'єзорезистивних опорів. Такий мікродавач називають давачем тиску п'єзорезистивного типу.

Мікроактюатор [15] – це мікромеханічний пристрій, який перетворює енергію (електричну, магнітну, хімічну тощо) в механічну роботу, нагрівання, випромінювання світла, тощо.

Принцип роботи мікроактюаторів, в більшості випадків залежать від виду вхідної енергії та сил, які згенеровані цією енергією. В загальному випадку принцип роботи мікроактюатора можна зобразити схемою, яка наведена на *Рис. 1.9*. Згідно цієї схеми до мікроактюатора підводиться енергія, яка генерує сили, які, в свою чергу, обумовлюють механічне переміщення.

Досить часто в термінології МЕМС використовують термін перетворювач та трансдюсер. Отже під перетворювачем (трансдюсером) будемо розуміти пристрій, який виконує перетворення енергії одного виду в інший.

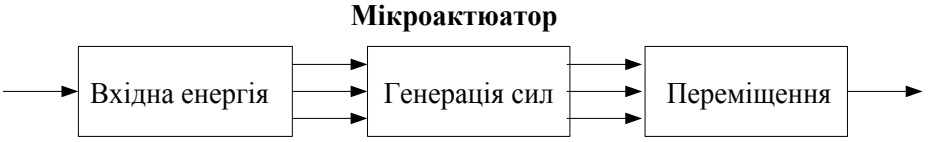


Рис. 1.9 Загальна схема роботи мікроактюатора МЕМС

Загальна структура МЕМС наведена на *Рис. 1.10*. Вона включає вхідний перетворювач, мікропроцесор (пристрій для обробки, збереження, передачі інформації) та вихідний перетворювач.

Отже, вхідний перетворювач (надалі мікродавач) призначений для визначення змін чи впливу оточуючого середовища на інтегральний пристрій. В багатьох мікропроцесорах, в якості вхідного електричного параметра, можуть виступати зміна опору, ємності, частоти, напруги, струму тощо.

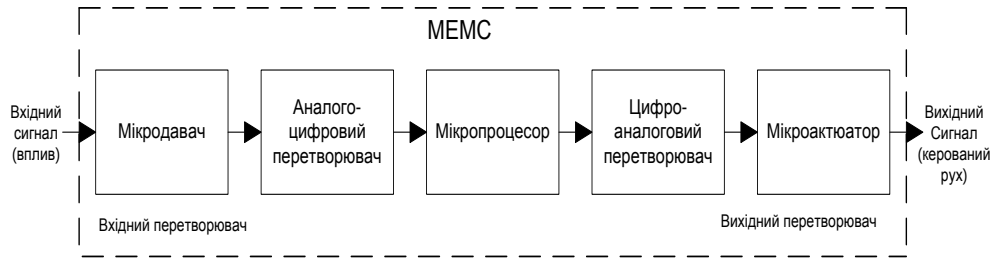


Рис. 1.10 Загальна структура МЕМС

Оскільки, безпосередньо, аналогову величину напруги чи струму мікропроцесор обробляти не може, то після мікродавача використано аналого-цифровий перетворювач (АЦП), з якого вже цифровий сигнал поступає на мікропроцесор. Мікропроцесор обробляє отримані дані згідно попередньо визначеного алгоритму, а результат обробки, у формі цифрового сигналу, видає на цифро-аналоговий перетворювач (ЦАП). ЦАП перетворює код в аналоговий сигнал, який безпосередньо подається на вихідний перетворювач. В якості вихідного перетворювача виступають актюатори.

Разом з тим, можлива й інша структура, яка наведена на *Рис. 1.11*. Її особливістю є те, що вона обробляє лише аналоговий сигнал і включає, відповідно, мікродавач, схему керування та обробки аналогового сигналу і мікроактюатор. Принцип дії такої МЕМС аналогічний до попередньої.

Наведені структури МЕМС на *Рис. 1.10* та *Рис. 1.11* відносяться до найпростіших. Особливістю їх є те, що реалізувати таку МЕМС можна за допомогою єдиної технології виготовлення, хоча можуть виникнути проблеми виготовлення мікродавача та мікроактюатора за єдиною технологією та розміщення їх на одному кристалі. Таким структурам притаманна найвища швидкодія (з двох вищенаведених структур кращі параметри швидкодії має структура з аналоговою схемою керування та обробки сигналу). Живлення в таких пристроях подається від макросистеми. В більшості випадків для реалізації мікродавача та мікроактюатора використовують технології

поверхневої чи об'ємної обробок, або їх похідні та КМОН технологія для виготовлення мікропроцесора, яка є сумісною з двома вище перерахованими. Прикладом МЕМС, які використовують такі структури, є підсистема викидання подушок безпеки в автомобілі, системи контролю цукру в крові людини тощо.

Зрозуміло, що на практиці реальні структури значно складніші включаючи також живлення, засоби зв'язку з іншими МЕМС і основною системою та інші.

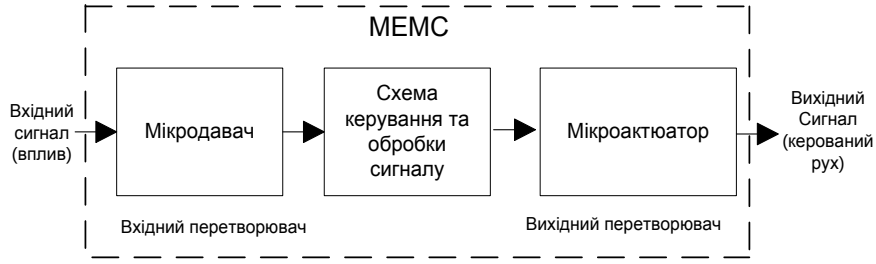


Рис. 1.11 Структура МЕМС для обробки аналогового сигналу

1.2. Застосування блочно-ієрархічного підходу до проектування МЕМС

При розв'язанні задач проектування МЕМС використано блочно-ієрархічний підхід, який передбачає використання принципу ієрархічності для структурування представлень про об'єкти по степені деталізації описів та принцип декомпозиції (блочності, модульності) для розбиття представлень кожного рівня на ряд складових (довершених блоків) з можливістю їх поблочного проектування [16], [17], [18].

Застосуємо теорію множин для формалізації процесу розроблення МЕМС. На верхньому рівні МЕМС позначимо, як S^1_{MEMS}, де одиниця означає перший рівень деталізації. Оскільки МЕМС є складною системою і її можна розбити на блоки нижчого рівня з ціллю зручності розв'язання задач проектування, то введемо рівень 2, який буде включати n блоків. Відповідно, кожний блок другого рівня позначимо через $S^{2,j}_{MEMS}$, де j – номер блока другого рівня розбиття ($j = 1,2,...n$). В даному випадку МЕМС можна описати, як

$$S^1_{MEMS} = \bigcup_{j=1}^{n} S^{2,j}_{MEMS}. \qquad (1.1)$$

Оскільки блоки другого рівня також є складними об'єктами і їх можна розглядати як системи по відношенню до блоків третього рівня та доцільно, з технічної сторони, розбити на простіші блоки, то кожний блок (система по відношенню блоків третього рівня) другого рівня можна описати як об'єднання блоків третього рівня:

$$S^{2,j}_{MEMS} = \bigcup_{l=1}^{K_j} S^{3,l}_{MEMS}, \qquad (1.2)$$

де K_j – кількість блоків третього рівня в j–му блоці (системі) другого рівня, l

– номер блока третього рівня розбиття ($l = 1,2,...,K_j$).

При технічній доцільності блоків четвертого рівня блоки третього рівня можна описати наступним чином:

$$S_{MEMS}^{3,l} = \bigcup_{z=1}^{Z_j} S_{MEMS}^{4,z}, \qquad (1.3)$$

де Z_l – кількість блоків четвертого рівня в l-му блоці (системі) третього рівня, z – номер блока четвертого рівня розбиття ($z = 1,2,...,Z_l$).

Таким чином процес продовжується доти, поки блоки m-го рівня вже недоцільно, з певних міркувань, піддавати декомпозиції на простіші. Блоки найнижчого рівня, як правило, називають базовими елементами.

Припустивши, що інформаційна технологія проектування МЕМС потребує чотири рівні ієрархії, їх можна описати з допомогою наступного виразу:

$$S_{MEMS}^{1} = \bigcup_{j=1}^{n} S_{MEMS}^{2,j} \bigcup_{l=1}^{K_j} S_{MEMS}^{3,l} \bigcup_{z=1}^{Z_j} S_{MEMS}^{4,z}. \qquad (1.4)$$

Слід зауважити, що поділ на блоки виконується, як правило, за функціональною ознакою. Тобто, у випадку побудови елементів МЕМС, де при розробці використовується три рівні на відміну від розробки інтегральних схем (розробка підсистеми обробки, збереження та передачі даних), перший рівень – МЕМС з набором функцій зазначених в технічному завданні, блоки другого рівня – це є пристрої для контролю стану навколишнього середовища, пристрої для збору, обробки, збереження та видачі керуючих сигналів, пристрої для впливу на оточуюче середовище та ін., а блоки третього рівня – балки, пружини, інерційні маси, інтегральні транзистори, резистори, конденсатори тощо.

В процесі розробки МЕМС, в більшості випадків, використовують класичне багаторівневе ієрархічне проектування [19], [20] "зверху-вниз", "знизу-вгору", паралельне, їх поєднання, наскрізне тощо. Процес розробки МЕМС з врахуванням особливостей проектування підсистеми збору, обробки, збереження та видачі керуючих сигналів у формі інтегральної схеми включає чотири рівні, які мають класичні назви: перший – системний, другий – функціональний, третій – схемотехнічний, а четвертий – компонентний.

На сучасному етапі розробки саме електромеханічних, електромагнітних, п’єзоелектричних, електротеплових та інших елементів МЕМС використовується три рівні, тобто:

- системний;
- схемотехнічний;
- компонентний.

Відповідно, відсутній функціональний рівень. Хоча складність фізичних процесів, які проходять в цих пристроях з мікронними розмірами, жорсткі вимоги до точності їх виготовлення (допуски на конструктивні параметри елементів МЕМС жорсткіші ніж на елементи електронних схем),

багатофункціональність елементів конструкції, необхідність оцінки принципової можливості функціонування пристрою та можливості його реалізації з допомогою наявних технологій особливо при розробці нових елементів МЕМС потребують розв'язання ряду задач, що знаходяться між системним та схемотехнічним рівнями (задачі кінематики в механіці, які можна віднести до функціонального рівня проектування, задачі пов'язані з розробкою алгоритмічного забезпечення МЕМС та інші). Надалі в роботі будемо притримуватися трирівневого підходу до розробки МЕМС. Пам'ятаючи, при цьому, про можливість включання функціонального рівня та розбиття компонентного на компонентний і елементний рівні. В більшості випадків кількість рівнів визначається технічною доцільністю, здоровим глуздом, наявними програмними системами для проектування МЕМС тощо.

Отже, у випадку використання розробки "зверху-вниз" (*Рис. 1.12*) пристрій МЕМС розбивається на функціонально довершені модулі: блоки живлення, мікродавачі, модулі обробки, передачі та збереження інформації, виконуючі пристроїв тощо. В даному випадку маємо справу з схемотехнічним рівнем автоматизованого проектування.

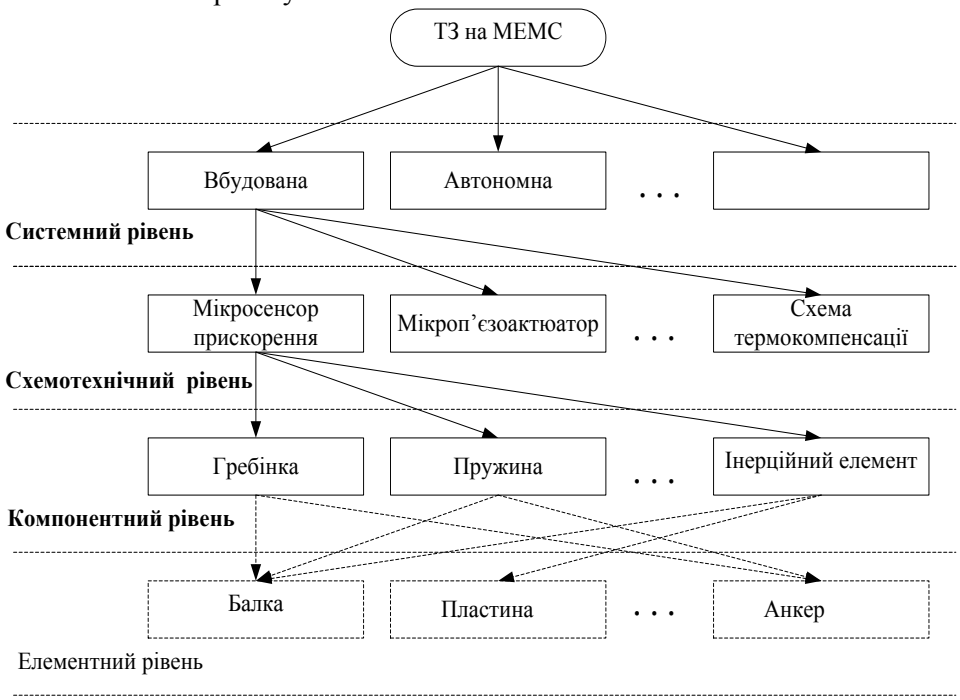


Рис. 1.12 Рівні розробки МЕМС "зверху-вниз"

Далі задача розробки мікродавачів та мікроактюаторів розбивається на задачі проектування компонент (елементів), які є задачами компонентного рівня і пов'язані, для прикладу, з розробкою пружини чи джерела живлення електростатичного мікроактюатора тощо. При потребі, компоненти МЕМС можна піддати подальшій декомпозиції на елементи. В цьому випадку будемо

мати справу з елементним рівнем, на якому виконуємо розробку балок, анкерів, пластин тощо.

Використання блочно-ієрархічного підходу до проектування МЕМС має низку переваг над іншими, тобто:

- з простішими об'єктами розробки зручніше працювати;
- побудувати математичну модель даного об'єкта проектування;
- провести моделювання його роботи;
- виконати верифікацію та тестування результатів розробки тощо.

До того ж слід звернути увагу на особливість МЕМС, яка полягає тому, що функціональні пристрої можуть належати до різних наукових областей і розробнику, практично, не можливо бути спеціалістом в усіх галузях науки та техніки.

Розробку "знизу-вгору" (*Рис. 1.13*) використовують в тому випадку, коли необхідно побудувати подібний інтегральний пристрій і значна частина складових МЕМС частково чи повністю вже є спроектована. Цей вид розробки МЕМС з кожним роком все частіше використовується, оскільки бібліотека розроблених елементів з часом зростає. На перших етапах це стане можливим для окремих елементів МЕМС, потім компонент і підсистем.

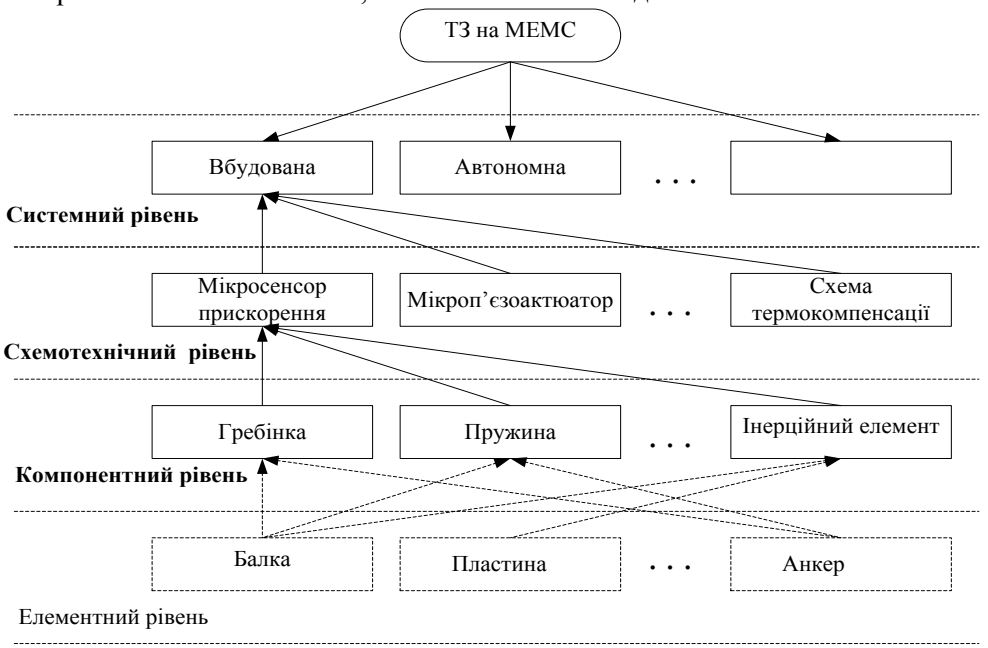


Рис. 1.13 Приклад розробки МЕМС "знизу-вгору"

Разом з тим, розробку ряду елементів мікроелектромеханічної системи доцільно проводити паралельно (*Рис. 1.14*), що дасть змогу на певних етапах значно прискорити процес проектування. Для прикладу, розробку мікродавачів, мікроактюаторів та системи керування обробки і передачі даних можна проводити паралельно.

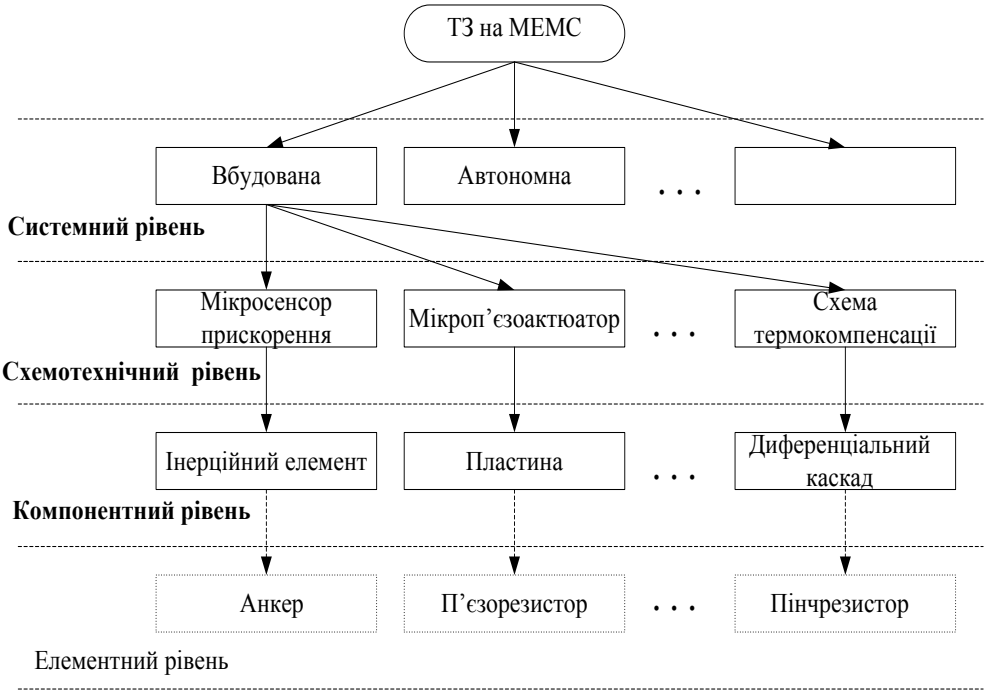


Рис. 1.14 Паралельна розробка елементів МЕМС

Отже, при побудові мікроелектромеханічних систем необхідно використовувати методи розробки "знизу-вгору", "зверху-вниз", їх поєднання і особливу увагу слід приділити паралельному проектуванню, що обумовлено особливостями мікросистем, тобто роботою складових за різними фізичними принципами.

Як було зазначено вище, розробка МЕМС ґрунтується на технологіях проектування ІС, які включали три аспекти проектування: функціонально-логічний, конструкторський та технологічний.

Оскільки розробка МЕМС є тісно пов'язано з технологією їх виготовлення, то особливого значення набуває відповідний аспект, де відбувається розробка процесу технологічного виготовлення МЕМС або використовується одна з базових мікротехнологій. Враховуючи те, що МЕМС технології, в основному, базуються на технологіях виготовлення ІС та мікромеханічних пристроїв, які, як правило, вже є відлагодженими, для використання базового технологічного процесу необхідно внести лише незначні корективи. Хоча впровадження нових конструкцій пристроїв МЕМС, використання нових матеріалів, зміни в базовому технологічному процесі виготовлення, впровадження нових мікротехнологій та дослідження впливу технологічних процесів на вихідні параметри інтегральних пристроїв МЕМС потребує розробки нових моделей, методів та програмних засобів для моделювання на технологічному рівні проектування технологічного маршруту.

Так склалося історично, що найбільші центри розробки та виготовлення МЕМС використовують різні технології. Технології, які використовують в США базуються на технологіях виготовлення ІС, а технології, які використовуються в Європі та Японії – на технологіях виготовлення мікромеханічних пристроїв (LIGA, SIGA та ін.).

Разом з тим, слід зауважити, як випливає із аналізу стану і перспектив світового розвитку проектних робіт на різних рівнях абстракції (*Рис. 1.15*), якщо в 1990 році реалізація проекту (починаючи з логічного рівня) займала 90% у всьому об'ємі проектних робіт, то в 2000 році ця доля скоротилась до 55%, а в 2010 році проектування на архітектурному і функціональному рівнях складає 70% у загальному об'ємі робіт, і тільки 30% припадає на конкретну реалізацію проекту в вибраному елементному (бібліотечному) базисі. В найближчому майбутньому відсоток робіт на архітектурному і функціональному рівнях буде лишень зростати.

В будь-якому випадку, розробка систем на кристалі, яке складається із кількох мільйонів вентилів, є непростою задачею, і засоби проектування інтегральних схем, безперервно ускладнюються, еволюціонують в сторону системного рівня проектування. В кінцевому рахунку, щоб використати множину базових блоків і об'єми інтегральних схем в кілька мільйонів вентилів, необхідні відповідні засоби розробки, які дають змогу використовувати всі ці можливості в розробках.

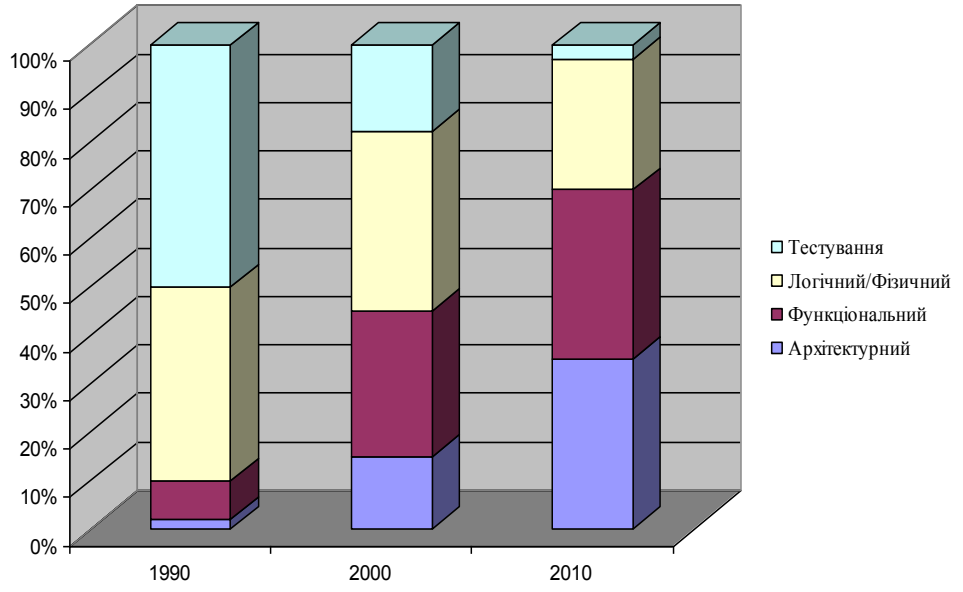


Рис. 1.15 Тенденції проектування на різних рівнях абстракції

1.3. Методи автоматизованого проектування МЕМС

Розроблення МЕМС, на відміну від розроблення стандартних інтегральних схем (ІС), має свою специфіку та ряд особливостей [21], [22]. Для визначення цих особливостей проведемо порівняння автоматизованого проектування МЕМС з стандартним процесом розроблення ІС [23].

На сьогодні найбільш відлагодженими є інформаційні технології автоматизованого проектування цифрових ІС. В цьому випадку відомий та розроблений набір програмних засобів, методологій, методів та математичних моделей з використанням підходу на основі базових елементів, стан яких добре відомий і які виготовляють за типовою мікротехнологією. Зрозуміло, що виникає ряд проблем з впровадженням нових інтегральних пристроїв при зменшенні їх лінійних розмірів, використанні нових фізичних принципів роботи елементів тощо.

Інформаційні технології проектування інтегральних пристроїв, де одночасно використовують аналогові та цифрові сигнали (змішаний сигнал) є менш автоматизовані, але добре зрозумілий підхід, який має використовувати розробник через широку різноманітність базових елементів в галузі проектування ІС змішаного сигналу, і зрозуміло, що методи, математичні моделі та підходи є значно складніші.

В галузі інформаційних технологій автоматизованого проектування МЕМС, де елементи функціонують за різними фізичними законами та принципами (змішана природа елементів), є найменш автоматизовані у порівнянні з технологіями розроблення ІС, і характеризується наявністю багатьох ітераційних циклів. Процес проектування МЕМС є складніший через розроблення різноманітних підсистем, які належать до різних наукових галузей та функціонують за різними фізичними законами. Особливо гостро стоять ці проблеми на етапах узгодження роботи елементів МЕМС між собою, їх виготовлення, комплексного проектування об’єкта розроблення тощо.

Існує два основні підходи для організації механізму проектування будь-якого складного технічного об’єкта: висхідний (знизу догори) і спадаючий (згори донизу). На початковій стадії розробки більшість проектувальних підходів є висхідними. Але якщо застосовувати інформаційні технології проектування більш обдумано із відомостями про параметри технології виготовлення, то ми змушені звернутись до розроблення згори донизу, оскільки цей підхід більш придатний до розроблення МЕМС з врахуванням технологічного аспекту.

Відповідні підходи широко використовуються при сучасному розроблені МЕМС. Слід зауважити, що з накопиченням досвіду та проектних рішень в галузі розробки та виготовлення МЕМС все частіше використовуються елементи розроблення знизу догори і частка цього підходу з часом буде лише зростати. Реальний процес автоматизованого проектування МЕМС використовує змішане розроблення, яке включає як проектування знизу догори, так і проектування зверху донизу. Разом з тим, в ряді робіт запропоновано використовувати підходи, пов’язані з паралельним та наскрізним розробленням

МЕМС, що обумовлено специфікою галузі МЕМС та дає змогу значно прискорити процес розробки інтегрального виробу. На сьогодні, паралельне проектування МЕМС можна зреалізувати лише на окремих ієрархічних рівнях з етапами синхронізації проектних робіт на початку та в кінці розроблення кожного з рівнів.

Наскрізне проектування ефективно використовується при проектуванні ІС, тому все ширше і масштабніше застосовується до розроблення мікроелектромеханічних систем.

Кожний з вищенаведених методів використовує блочно-ієрархічний підхід до автоматизованого проектування МЕМС, який передбачає розбиття процесу проектування на ієрархічні рівні.

Верхній рівень абстракції називають системним. На цьому рівні розроблення МЕМС розв'язують задачі, пов'язані з синтезом структури та визначенням її основних параметрів (задача аналізу) [24].

На сучасному етапі розвитку автоматизованого розроблення при розв'язанні задач системного рівня, зокрема, структурному синтезі, досить часто використовують експертні системи. В склад такої системи входить база даних, де розміщена інформація про елементи структури, база знань та монітор. Відповідну інформаційну технологію синтезу структури використати для розроблення МЕМС, на сьогодні, неможливо, оскільки відсутнє наповнення бази знань. Пояснити дану особливість можна з тих позицій, що науково-прикладна галузь МЕМС з'явилась не так давно і в цей час триває активний її розвиток та наукові дослідження МЕМС і продовжується процес накопичення та збору знань про об'єкти розробки.

Ряд інших інформаційних технологій синтезу структури об'єкта проектування базуються на інтерактивній роботі інженера-розробника з програмною системою, яка дає змогу визначити вихідні параметри сформованої структури. В цьому випадку інженер-розробник, на основі досвіду та попередньо спроектованих об'єктів подібного типу, генерує структуру та з допомогою системи визначає основні параметри і комплексний показник якості об'єкта проектування на основі розв'язання оптимізаційної багатокритеріальної задачі. Відповідна інформаційна технологія в більшості випадків використовують при сучасному проектуванні МЕМС на системному рівні.

Для визначення основних параметрів будь-якого об'єкта проектування на системному рівні використовують теорію масового обслуговування та теорію мереж Петрі [25], які варто використати і при розв'язанні задач аналізу МЕМС.

Інформаційні технології проектування систем сенсорних та актюаторних пристроїв і, зокрема, для аналізу перехідних процесів використовують методи, які дають змогу побудувати схематичні моделі (макромоделі) та макрооб'єкти. Називають цей рівень схемотехнічний на якому фізика процесів описується системою звичайних алгебричних чи диференціальних рівнянь. На відміну від системного рівня абстракції, моделювання компонентів близьке до фізики, через те, що ми можемо розпізнати фізичні частини, для прикладу резистори і транзистори, або маси чи пружини. При проектуванні МЕМС, як правило, на

цьому рівні використовується метод аналогій та теорія коливних процесів. Стосовно МЕМС, то на схемотехнічному рівні проектування використовують, як правило, методи та алгоритми, які дають змогу побудувати VHDL-AMS моделі елементів мікроелектромеханічних систем [26]. Наступний метод синтезу моделей елементів МЕМС ґрунтується на використанні інформації з компонентного рівня проектування і зменшення кількості рівнянь. На цей час відповідний метод активно використовується та розвивається.

Найнижчий рівень абстракції автоматизованого проектування МЕМС відзначений як компонентний рівень. Фізика процесів в елементах описується диференційними рівняннями в частинних похідних з відповідними початковими та краєвими умовами. На цьому рівні використовують такі методи, як метод скінченних різниць, метод скінченних елементів та граничних елементів, які дозволяють врахувати нелінійні та нестаціонарні процеси в конструкціях елементів МЕМС. Разом з тим, застосування вищенаведених методів потребує значних затрат ресурсу комп'ютера та найбільш підходить для визначення розподілу напружень, деформацій, власних частот в конструкції МЕМС тощо. Разом з тим, необхідно додати, що існують проблеми при синтезі ММ компонентного рівня для елементів МЕМС з розмірами співмірними з зерном матеріалу його конструкції.

1.4. Системи проектування МЕМС на компонентному рівні

Сучасні тенденції автоматизованого проектування складних об'єктів і систем дають можливість стверджувати, що частка вартості проектувальних робіт в загальній вартості виробу з кожним роком різко зростає. Особливо дана тенденція притаманна інтегральним пристроям, які виготовляють з допомогою мікротехнологій. Дещо сповільнити таку тенденцію можна з допомогою широкомасштабного використання програмних засобів проектування МЕМС.

Появі перших програмних систем для проектування МЕМС сприяло швидке зростання інтеграції МЕМС (моделююча програма має допомагати розробляти мікропристрої, які включають елементи, що діють за різними фізичними принципами: електричною, механічною, тепловою тощо), а також збільшення точності моделювання з ціллю зменшення витрат на розробку та виготовлення.

Швидкий прогрес в області проектування, моделювання та виготовлення МЕМС тісно пов'язаний з використанням програмних засобів, які дають змогу значно підвищити конкурентоздатність пристроїв цього типу. Найчастіше при проектуванні МЕМС використовуються наступні системи: SUGAR [27], IntelliSuite [28], NODASv1.4, Coventor [29], Tanner Research [30], ANSYS [31], CFD-ACE [32], Abaqus [33], MEMCAD [34], Solidis [35], Simulink [36], Saber [37], ALLTED [38], [39], [40].

Проаналізуємо також найвідоміші сучасні програмні системи MEMCAD [41], Cadence та Coventor, що стосуються проектування та моделювання МЕМС – пристроїв.

Програмна система SUGAR базується на методах вузлового аналізу, який

широко використовується при проектуванні та моделюванні інтегральних схем. Підкладки, компоненти інтегральних схем, електростатичні пристрої тощо моделюють з використанням систем диференціальних рівнянь. SUGAR успадковувала свою назву і філософію від програми SPICE [42].

До переваг цієї системи можна віднести наступне: розробник описує пристрій в компактному форматі таблиці з'єднань; не є складно моделювати поведінку пристрою, легко знаходити недоліки в конструкції елемента МЕМС або випробовувати нові ідеї.

Недоліком цієї системи є те, що для використовування програмного засобу SUGAR потрібен також MATLAB версії 5.0 або новішої версії; відсутні бібліотеки математичних моделей базових елементів МЕМС.

Мікроелектромеханічна дослідницька група (MEM Research) випустила програмний продукт під назвою EM3DS 4.2, який призначений для електромагнітного аналізу МЕМ-перемикачів і конденсаторів. Цей програмний інструмент використовує рівняння повної хвилі та новий підхід: узагальненої поперечної резонансної дифракції (Generalised Transverse Resonance Diffraction (GTRD)), призначений для розв’язання квазі-планарних багатошарових схем. На відміну від інших програмних продуктів для планарних структур, ця система дає змогу побудувати тривимірні пристрої з розрахованою кінцевою товщиною і кінцевою провідністю реальних провідників.

До переваг цієї програмної системи можна віднести наступне: має зручний інтерфейс; запропонований новий підхід до розроблення та моделювання забезпечив високу точність вихідних результатів для МЕМ – ємнісних перемикачів у порівнянні з аналогічними системами.

Недоліком програмної системи EM3DS 4.2 є: відсутність можливості конвертації отриманих результатів для використання в інших програмних системах; вузька область моделювання роботи МЕМС пристроїв (електромагнітні двигуни).

Однією з перших комерційних систем автоматизації проектування МЕМС є IntelliSuite (CorningIntelliSence Corporation). Вона призначена для проектування, моделювання і оптимізації МЕМ-пристроїв. Основною особливістю системи IntelliSuite є те, що процес конструювання починається не від геометрії пристрою, а від параметрів виготовлення інтегрального пристрою. IntelliSuite оптимізує конструкції МЕМС до виготовлення, скорочуючи цикл часу розвитку дослідного зразка і скорочуючи виробничі витрати. Об'єднуючи в собі шаблони технологічного процесу виготовлення, дані про матеріали, топологію фотошаблону і аналіз пристрою, IntelliSuite надає проектним групам інструментальний комплекс для розробки технологічних пристроїв з більш високим коефіцієнтом корисної дії (ККД). Недоліком цього програмного продукту є обмежена кількість технологій виготовлення МЕМ-пристроїв та неможливість зміни технологічного процесу розробником.

NODASv1.4 [43], [44] (Nodal Design of Actuators and Sensors) – це бібліотека параметризованих компонентів для використання в програмі моделювання вузловим методом SABER МЕМ-пристроїв, для виготовлення

яких використовується поверхнева мікрообробка. Бібліотека складена з підкладок, фотошаблонів, анкерів, електростатичних гребеневих мікродвигунів (горизонтальних і вертикальних) та електростатичних пристроїв. Особливістю проектування є те, що користувач спочатку синтезує схемне рішення, використовуючи позначення компонент, використовує статичні моделі для розрахунку розміщення кожного компонента і потім генерує новий опис поведінки, використовуючи набір динамічних моделей, в яких розміщення автоматично визначено як статичний параметр компонента. Запропонований підхід зменшує кількість вузлів і змінних для множини моделей, таким чином збільшуючи загальну швидкість моделювання, а також допомагає усувати помилки розміщення, зроблені користувачем.

До переваг цієї системи можна віднести наступне: компоненти можуть бути з'єднані для представлення складних систем; електричні властивості враховані в моделях компонентів, надають можливість одночасного електричного і електромеханічного аналізу.

Недоліком системи NODASv1.4 є відсутність: можливості моделювання гідравлічних та інших процесів в мікродавачах та актюаторах; засобів для розв'язання задач синтезу на системному рівні та ін.

Система для проектування МЕМС Coventor [45] дає змогу провести розробку з використанням підходу до розроблення "згори донизу" і "знизу догори". Вона включає чотири основні модулі та декількох додаткових, які забезпечують розробника усіма необхідними засобами для реалізації вищенаведених підходів.

Для аналізу елементів МЕМС на компонентному рівні проектування, як правило, використовують програмний комплекс Ansys. Основним недоліком цієї системи є значні затрати ресурсу ПК для розв'язання задач аналізу МЕМС та неточність вихідних результатів для пристроїв співмірних з розмірами зерна матеріалів конструкції елементів мікроелектромеханічних систем.

Основний недолік цієї системи полягає в низькому рівні автоматизації робіт на системному рівні проектування МЕМС.

При побудові макромоделей елементів МЕМС, які функціонують на основі різних фізичних процесів та законів, можна використати мову VHDL-AMS (Very High Speed Integrated Circuits Hardware Description Language Analog-Mixed Signals) [46], [47], [48], [49]. В даному випадку можна скористатися програмними продуктами при відлагодженні VHDL-AMS – моделей елементів МЕМС [50], [51], [52], зокрема: AMSWizard [47], hAMSter [53] та інші.

1.5. Список використаної літератури до розділу 1

[1] Лысенко И. Е. Проектирование сенсорных и актюаторных элементов микросистемной техники / И. Е. Лысенко. – Таганрог : ТРТУ, 2005. – 103 с.

[2] Теслюк В. М. Моделі та інформаційні технології синтезу мікроелектромеханічних систем: Монографія. – Львів: "Вежа і Ко", 2008 – 192 с.

[3] K. Petersen. A new age for MEMS. The 13 th International Conferences on Solid-State Sensors, Actuators and Microsystems, 2005. Digest of Technical Papers.

TRANSDUCER'05. On pages: 1-4 Vol.1.
[4] Maluf N., Williams K. 2004. An Introduction to Microelectromechanical Systems Engineering (second edition)/Artech House Inc. 305 p.
[5] Белявский В.И. Физические основы полупроводниковой нанотехнологии // Соросовский общеобразовательный Журнал. - 1998. - № 10. - С. 92 – 98.
[6] Нанотехнология в ближайшем десятилетии / Под ред. М.К.Роко, Р.С.Уильямса, П.Аливисатоса. - М., 2002.
[7] Головин Ю.И. Введение в нанотехнологию. - М., 2003.
[8] Zheng C., Changzhi G. Nanofabrication challenges for NEMS // 1st IEEE International Conference on Nano/Micro Engineered and Molecular Systems. – 2006. - Jan. - P. 607-610.
[9] Малышева И.А. Технология производства интегральных микросхем.: Учебник для техникумов. - 2-е изд., перераб. и доп. - М.: Радио и связь, 1991. - 344 с.
[10] Mounier E., Robin L., Steady 10-12% growth will double the MEMS market over next six years. – www.cowin4u.eu/analystcorner_memstrends_april2013
[11] Status of MEMS Industry. Explorung new growth opportunities/ – www/semiconwest/org|sites|semiconwest/org|files|docs|SW2013_JC%20Eloy_Yole%20 Developpement/pdf
[12] Mounier E., Bonnabel A. Driven by smartphones & microfluidics, emerging MEMS will account for 10%of the value of the total MEMS business by 2018", announces Yole Developpement. – www.yole.fr/iso_upload/News/2013/PR_EmergingMEMS_August 2013.pdf
[13] Васильев А., Борисов Е. Производство МЭМС. Перспективы и решения // ЭЛЕКТРОНИКА. Наука. Технология. Бізнес. – 2012, №3 (00117). – С. 60 – 64.
[14] Климов Д.М., Лучинин В.В., Васильев А.А., Мальцев П.П. Перспективы развития микросистемной техники в XXI веке // Микросистемная техника. - 1999. - № 1. - С. 3-6.
[15] Теслюк В. М., Денисюк П.Ю. Автоматизація проектування мікроелектромеханічних систем на компонентному рівні: Монографія. – Львів: Видавництво "Львівської політехніки", 2011 – 192 с.
[16] Петренко А. И., Семенков А. И. Основы построения систем автоматизированого проэктирования. К.: Вища школа., 1984. – 296 с.
[17] Teslyuk V., Pereyma M., Karkulyovskyy V., Lobur M. Features of microelectromechanical systems design // Proc. of the 2nd Inter. Conf. of Young Scientists "Perspective Technologies and Methods in MEMS Design" (MEMSTECH 2006).- Lviv–Polyana, Ukraine, 2006. – P. 67-70.
[18] Vasyl Teslyuk, Mykhaylo Lobur, Pavlo Denysyuk, Konstantin Kolesnyk. Methodology of the Automated MEMS Design. //Proc.of the IIId International Conference of Young Scientists MEMSTECH'2007, May, 23-26, Lviv, Polyana, 84-85.
[19] Петренко А. И. Основы автоматизации проэктирования. – К.: Техніка, 1982. 295 с.
[20] Коваль В. О., Лобур М, В. Автоматизация технологического моделирования полупроводниковых ИС. Учебное пособие, Львов, ЛПИ, 1987, с. 84
[21] Napieralski A., Napieralska M., Szermer M., Maj C. 2012. The evolution of MEMS and modeling methodologies, COMPEL: The International Journal for computation and Mathematics in Electrical and Electronic Engineering, vol.31, pp.1458 – 1469.
[22] Петренко А. І. Мережний пакет для комп'ютерного проектування мікроелектромеханічних систем (МЕМС) / А. І. Петренко // Развитие информационно–коммуникационных технологий и построение информационного

общества в Украине (CeBIT – 2007) : труды междун. науч. конф. – Київ, 2007. – С.143 – 156.
[23] Бургер Р. Основы технологии кремниевых интегральных схем. Окисление, диффузия, эпитаксия. / Р. Бургер, Р. Донован: пер. с англ. – М. : Мир. – 1969. 451 с.
[24] Teslyuk Vasyl, Tarik Al Omari, Hamza Alshavabkekh, Pavlo Denysyuk, Mykhaylo Melnyk Computer-Aided Design of MEMS at System Level // Journal Machine Dynamics Problems. - Poland, Warsaw University of Technology. - 2007., Vol. 31, No. 3 – P. 92 – 104.
[25] Lobur Mykhaylo, Teslyuk Vasyl, Zaharyuk Roman, Volodymyr Antonyuk Using Petri Nets In MEMS Design // Journal Machine Dynamics Problems. - Poland, Warsaw University of Technology. - 2006., Vol. 30, No. 4 – P. 29 – 36.
[26] Теслюк В.М., Тарік (Мох'д Тайсір) Алі Аль Омарі, Каркульовський В.І. VHDL-AMS модель для автоматизації схемотехнічного рівня розробки п'єзоелектричного мікрофона // Збірник наукових праць інституту проблем моделювання в енергетиці ім.Г.Є.Пухова НАН України. – Київ, 2008, Вип. 49. – С.206 - 212.
[27] Working Model Motion ver. 5.0, MSC.Working Knowledge [Електронний ресурс] . – San Mateo, CA. – Режим доступу: http://www.workingmodel.com.
[28] IntelliSuite, IntelliSense Corp., Wilmington, MA. – Режим доступу: http://www.intellisense.com.
[29] Потапов Ю. В. Программное обеспечение Coventor [Електронний ресурс]/ Ю. В. Потапов // Chip News. – 2002. – № 2, (EDA EXPERT № 1). – Режим доступу: (http://www.chip-news.ru/archive/chipnews/200202/index.html).
[30] L - Edit, Tanner Research Inc. 180 North Vinedo Avenue Pasadena, California 91107 USA. [Електронний ресурс]. – Режим доступу: http://www.mems.louisville.edu/lutz/resources/ledit/intro.html
[31] ANSYS/Multiphysics ver. 5.5, Ansys, Inc., Canonsburg, PA. [Електронний ресурс]. – Режим доступу: http://www.ansys.com.
[32] CFD – ACE+ and add – on modules [Електронний ресурс]. – CFD Research Corporation, Huntsville, AL. – Режим доступу: http://www.cfdrc.com.
[33] Abaqus ver. 5.7, Hibbitt, Karlsson & Sorensen, Inc., Pawtucket, RI. [Електронний ресурс]. – Режим доступу: http://www.hks.com.
[34] MEMCAD ver. 4.5, Microcosm Technologies, Inc., Research Triangle, NC. [Електронний ресурс]. – Режим доступу: http://www.memcad.com.
[35] SOLIDIS: A tool for microactuator simulation in 3 – D / J. M. Funk, J. G. Korvink, J. Buhler [et al.] // J. Microelectromechanical Systems. – 1997. – Vol. 6, No. 1. P. 70 – 82.
[36] Simulink ver. 3.0, The Mathworks, Inc., Natick, MA. [Електронний ресурс]. – Режим доступу: http://www.mathworks.com.
[37] SaberDesigner, Analogy, Inc., Beaverton, OR. [Електронний ресурс]. – Режим доступу: http://www.analogy.com.
[38] Чкалов А. В. Математическая модель микромеханического ультразвукового преобразователя для САПР ALLTED / А. В. Чкалов, А. В. Крамар // Электроника и связь. Тематичний випуск „Проблеми електроніки". – К. : НТУУ "КПИ", 2005. – Ч.1. – С.117 – 120.
[39] Капшук О. А. Моделирование тонкопленочных микроболометров с помощью пакета ALLTED / О. А. Капшук, А. В. Крамар, В. А. Рабышко // Электроника и связь. Тематичний випуск „Проблеми електроніки". – К. : НТУУ "КПИ", 2005. – Ч.1. – С.113 – 116.
[40] Ладогубец В. В. Состояние и перспективы развития автоматизированного

схемотехнического проектирования / В. В. Ладогубец // Электроника и связь. Тематичний випуск „Проблеми електроніки”. – К. : НТУУ “КПИ”, 2005. – Ч.1. – С.121 – 127.

[41] A computer aided design system for microelectromechanical systems / S. D. Senturia, R. Harris, B. Johnson [et al.] // Journal of Microelectromechanical Systems. – 1992. – Vol. 1, №.1. – P. 3 – 13.

[42] Star – HSPICE, Avant. Corp., Fremont, CA. [Електронний ресурс]. – Режим доступу: http://www.avanticorp.com.

[43] Jing Qi Schematic-based lumped parameterized behavioral modeling for suspended MEMS / Qi Jing, Tamal Mukherjee, Gary K. Fedder// Computer Aided Design : Proc. of the IEEE/ACM intern. Conf. – 2002, New York, NY, USA. – P. 367 – 373.

[44] Zhou N. Nodal Analysis for MEMS Design Using SUGAR v. 0.5. / N. Zhou, J. V. Clark, K. S. J. Pister – Santa Clara CA. – 1998. – P. 6 – 8.

[45] Chen R. T. Leveraging mainstream design and analysis tools for MEMS / R. T. Chen, I. Mirman // Electronics Manufacturing Technology Symposium: Proc. IEEE/CPMT/SEMI 29-th International, July 14-16. – 2004. – P. 332 – 337

[46] Авдеев Е. В. Аналоговые и смешанно - сигнальные расширения VHDL : [учебное пособие] / Е. В. Авдеев. – М. : МИЭТ. – 2000.– С. 92.

[47] Dewey A. VHDL-AMS modeling considerations and styles for composite systems. Version 2.0 [Електронний ресурс] / A. Dewey, J. H. Hillman, B. Hillman [et al.]. – Режим доступу: http://www.hamster.com.

[48] Ивченко В. Г. Применение языка VHDL при проектировании специализированных СБИС : [учебное пособие] / В. Г. Ивченко. – Таганрог. : ТРТУ, 1999. – 80 с.

[49] Kazmierski T. A formal description of VHDL-AMS analogue systems / T. Kazmierski // Design, Automation, and Test in Europe : Proc. of the conf. – 1998, IEEE Computer Society Washington, DC, USA. – P. 916 – 920.

[50] Golovatyj A., Teslyuk V., Kryvyy R. VHDL-Ams Model of Integrated Membrane-Type Micro-Accelerometer with Delta-Sigma (Δσ) Analog-To-Digital Converter for Schematic Design Level // ECONTECHMOD. – 2015, vol. 4, no. 2. – P. 65 – 70.

[51] Holovatyy A., Teslyuk V., Lobur M. Verilog-AMS model of comb-drive sensitive element of integrated capacitive microaccelerometer for behavioral level of computer-aided design // ECONTECHMOD. – 2014, vol. 3, no. 4. – P. 49 – 53.

[52] Zaharyuk R. VHDL-AMS – Model for Capacitive Interdigital Accelerometer / Roman Zaharyuk, Vasyl Teslyuk, Ihor Farmaga, Hamza Ali Yousef AlShawabkeh // Proc.of the IVth International Conference of Young Scientists (MEMSTECH’2008) – Lviv - Polyana, 2008. – P.134 – 137.

[53] Лысенко И. Е. Моделирование сенсорных и актюаторных элементов микросистемной техники с использованием языка VHDL – AMS / И. Е. Лысенко, Е. А. Рындин. – Таганрог : ТРТУ, 2003. – 26 с.

2. Формалізація задач компонентного рівня проектування МЕМС

"...Науки не пробують пояснити, навряд чи вони навіть стараються інтерпретувати – вони в основному створюють моделі. Під моделлю розуміється математична конструкція, яка при додаванні деяких словесних пояснень описує феномен, що вивчається. Виправданням для такої математичної конструкції служить єдина обставина: очікується, що вона спрацює...", – Джон фон Нейман[1].

2.1. Моделювання на основі диференціальних рівнянь

Математичні моделі на компонентному рівні проектування для багатьох фізичних процесів описуються диференціальними рівняннями з частинними похідними (ДРЧП). Більш загально, в інженерних задачах, будь-яке фізичне явище, що містить в собі предмет моделювання, зазвичай описується системою диференціальних рівнянь з частинними похідними, які беруться з розділів теоретичної фізики. Рівняння розглядаються у деякій області, що представляє собою об'єкт моделювання, починаючи від елементарних конструкцій і закінчуючи складними системами типу космічних апаратів, прискорювачів елементарних частинок, клімату цілої планети, еволюції зоряних кластерів, галактик чи навіть всього всесвіту. Але, дотримуючись теми даної роботи, в цьому розділі основну увагу буде приділено моделям, що описують саме МЕМС.

На систему обраних диференціальних рівнянь накладаються необхідні початкові та крайові умови і з цього моменту математична модель, що є аналітичною, вважається повною, а для практичного використання залишається лише знайти рішення для конкретної множини вхідних числових даних.

Аналітичні методи моделювання призначені для отримання функціональних залежностей шляхом послідовного застосування математичних формул та правил, коли модель записана у вигляді рівнянь, наприклад диференціальних. При використанні аналітичних методів моделювання часто виникають труднощі, пов'язані з неможливістю отримання розв'язку в такій формі, що значно обмежує їх застосування. Тобто, зазвичай рішення в аналітичному вигляді можна знайти лише для найпростіших рівнянь, які розглядаються в об'єктах тривіальної геометричної форми.

На практиці аналітичні рішення є зазвичай не придатними до використання, і не потрібні, особливо коли одна формула займає кілька сторінок. Натомість, завжди можна використати деяке наближене рішення, отримане не складними математичними операціями з застосуванням обчислювальної техніки, що цілком задовольняє потреби інженерних розрахунків. Саме для цього призначені чисельні або дискретні моделі.

Чисельні методи моделювання ґрунтуються на побудові скінченної послідовності дій над числами, яка призводить до бажаного результату.

[1] Цитата взята з: Gleick J. – Хаос: Создание новой науки, СПБ: Амфора, 2001, *ст.349*.

Дослідження аналітичних моделей за допомогою чисельних методів полягає в заміні математичних операцій та співвідношень на відповідні дискретні аналоги, цей процес називається дискретизацією. Результатом застосування чисельних методів завжди є набори чисел, які потім можна зручно подати у вигляді таблиць чи графіків.

Диференціальне рівняння з частинними похідними (скорочено ДРЧП) – це рівняння, що містить частинні похідні. На відміну від звичайних диференціальних рівнянь, де невідома функція залежить тільки від однієї змінної, в рівняннях з частинними похідними невідома функція залежить від кількох змінних. Наприклад, розподіл температури в задачах теплопровідності залежить від просторових координат і часу.

Розв'язком як звичайних диференціальних рівнянь так і рівнянь з частинними похідними завжди є деяка функція, тобто польова величина (*Рис. 2.1*). Класичною польовою величиною є потенціал. Наприклад, потенціал гравітаційного поля, або потенціал електричного поля. Поняття польових величин та потенціалів прийшло з, так званої, теорії поля – одного з способів опису фізичних феноменів. Згідно з нею [1], у кожній точці простору "фізично існує число", як абсурдно це б не звучало. Якщо деякі об'єкти помістити в такий простір, то всі числа певним чином зміняться, а сили взаємодії між об’єктами будуть діяти в напрямку найшвидшої зміни цих чисел. Кожне число можна визначити "локально" на основі сусідніх чисел. Саме ці числа і називають потенціалами (від французького "*potentiel*" – "*такий, що може бути*", що в свою чергу прийшло з латинського "*potentia*" – "*сила*" або "*міць*").

Тут і в подальшому абстрактну польову величину будемо позначати символом u, або функцією $u(x,y,...,\tau)=u(x_1,x_2,...,\tau)=u(\mathbf{r},\tau)$, де $\mathbf{r}$ – радіус-вектор точки з координатами $x,y,...$, або $x_1,x_2,...$ (три крапки (еліпсис) означають, що кількість вимірів наперед не визначено).

Наступні розділи будуть присвячені постановкам та чисельному рішенню диференціальних рівнянь з частинними похідними, що описують фізичні процеси на компонентному рівні проектування МЕМС.

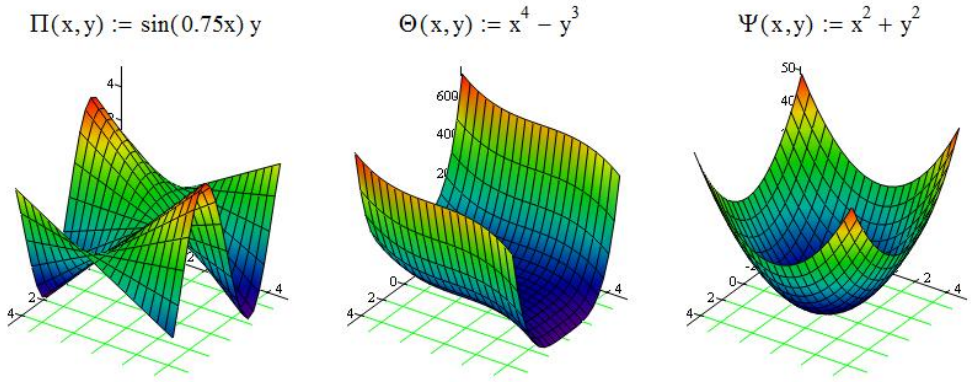


Рис. 2.1 Приклад двовимірних польових величин

2.2. Класифікація диференціальних рівнянь

Рівняння з частинними похідними можна класифікувати за багатьма ознаками. Класифікація важлива тому, що для кожного класу рівнянь існує своя загальна теорія і методи рішення. ДРЧП класифікуються за [2]:

- *Порядком рівнянь*. Порядком рівняння називається найвищий порядок похідних, що входять в рівняння.
- *Кількістю змінних*. Числом змінних називається число незалежних змінних рівняння, наприклад координатні осі і часова координата.
- *Лінійністю*. Рівняння з частинними похідними бувають лінійними та нелінійними. В лінійних рівняннях шукана функція і всі її частинні похідні входять лінійно, зокрема вони не множаться одна на одну, не підносяться до квадрату, ітераційно не залежать одна від одної та самі від себе і т.д. Більш точно, наприклад для рівнянь другого порядку, лінійним рівнянням називається рівняння виду:

$$A\frac{\partial^2 u}{\partial x^2} + B\frac{\partial^2 u}{\partial x \partial y} + C\frac{\partial^2 u}{\partial y^2} + \ldots = Q(x, y), \qquad (2.1)$$

  де A, B, C – константи або задані функції від незалежних змінних x та y (аналогічно і для більшої кількості вимірів).
- *Однорідністю*. Рівняння називається однорідним, якщо права частина рівняння рівна нулю або деякій константі, якщо права частина містить деякий вираз від незалежних змінних $Q(x, y)$, то рівняння називається неоднорідним.
- *Видом коефіцієнтів*. Якщо коефіцієнти A, B, C біля похідних є константами, то рівняння називається рівнянням з постійними коефіцієнтами, в іншому випадку – рівнянням зі змінними коефіцієнтами.
- *Типом*. Ця класифікація відноситься до рівнянь другого порядку виду (2.1). Річ в тому що це рівняння подібне до рівняння конічного перерізу. Так само, як конічні перерізи розділяють на еліпси, параболи та гіперболи, в залежності від знаку дискримінанту $B^2 - 4AC$, рівняння можна розділити на:
  - *Параболічний тип*. Рівняння цього типу описують процеси теплопровідності та дифузії і визначаються умовою $B^2 - 4AC = 0$, наприклад:

$$\frac{\partial u}{\partial \tau} = \frac{\partial^2 u}{\partial x^2}, \quad B^2 - 4AC = 0; \qquad (2.2)$$

  - *Гіперболічний тип*. Рівняння цього типу описують коливальні системи і хвильові процеси та визначаються умовою $B^2 - 4AC > 0$, наприклад:

$$\frac{\partial^2 u}{\partial \tau^2} = \frac{\partial^2 u}{\partial x^2}, \quad B^2 - 4AC = 4 > 0; \qquad (2.3)$$

- *Еліптичний тип*. Рівняння цього типу описують стаціонарні процеси і визначаються умовою $B^2 - 4AC < 0$, наприклад:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad B^2 - 4AC = -4 < 0. \tag{2.4}$$

У випадку рівняння зі змінними коефіцієнтами, тип рівняння може мінятися від точки до точки.

2.3. Операторна форма запису

Для дослідження диференціальних рівнянь з частинними похідними зручно використовувати операторну форму запису. У функціональному аналізі поняття оператору є розширенням поняття відображення з розділів лінійної алгебри, і не вдаючись в деталі, означає відображення, що ставить у відповідність функції іншу функцію [3]. Наприклад, лінійне диференціальне рівняння еліптичного типу в операторній формі записується як:

$$\begin{gathered} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0, \\ \mathcal{L}(.) = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}, \\ \mathcal{L}(u(x,y,z)) = \mathcal{L}u = 0, \end{gathered} \tag{2.5}$$

де $\mathcal{L}(.)$ – лінійний диференціальний оператор еліптичного типу. Цей оператор часто зустрічається у векторному та тензорному численні під назвою оператора Лапласа, або лапласіана і позначається як $\Delta(.)$ (дельта) або $\nabla^2(.)$, де останній вираз означає дивергенцію від градієнта скалярного поля. Оператор $\nabla(.)$ (набла) називається оператором Гамільтона або гамільтоніаном і позначає градієнт скалярного поля [4]:

$$\begin{gathered} \nabla = \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}, \\ \operatorname{grad}(u) = \nabla u = \frac{\partial u}{\partial \mathbf{r}} = \hat{\mathbf{i}}\frac{\partial u}{\partial x} + \hat{\mathbf{j}}\frac{\partial u}{\partial y} + \hat{\mathbf{k}}\frac{\partial u}{\partial z}, \end{gathered} \tag{2.6}$$

де $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, $\hat{\mathbf{k}}$ – одиничні ортогональні вектори, що утворюють базис простору, $\mathbf{r}$ – радіус-вектор точки з координатами x, y, z. Або в матричному вигляді:

$$\begin{gathered} \nabla = \left[\frac{\partial}{\partial x} \quad \frac{\partial}{\partial y} \quad \frac{\partial}{\partial z}\right]^{\mathbf{T}}, \\ \operatorname{grad}(u) = \nabla u = \left[\frac{\partial}{\partial x} \quad \frac{\partial}{\partial y} \quad \frac{\partial}{\partial z}\right]^{\mathbf{T}} [u] = \left[\frac{\partial u}{\partial x} \quad \frac{\partial u}{\partial y} \quad \frac{\partial u}{\partial z}\right]^{\mathbf{T}}. \end{gathered} \tag{2.7}$$

Застосувавши оператор Гамільтона до деякого векторного поля $\mathbf{J}$, отримаємо вираз дивергенції [4]:

$$\operatorname{div}(\mathbf{J}) = \nabla \cdot \mathbf{J} = \frac{\partial \mathbf{J}_x}{\partial x} + \frac{\partial \mathbf{J}_y}{\partial y} + \frac{\partial \mathbf{J}_z}{\partial z}, \tag{2.8}$$

або в матричному вигляді (скалярний добуток):

$$\begin{aligned} \operatorname{div}(\mathbf{J}) &= \nabla \cdot \mathbf{J} = \langle \nabla, \mathbf{J} \rangle = \nabla^{\mathbf{T}} \mathbf{J} = \\ &= \begin{bmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \end{bmatrix} \begin{Bmatrix} \mathbf{J}_x \\ \mathbf{J}_y \\ \mathbf{J}_z \end{Bmatrix} = \\ &= \frac{\partial \mathbf{J}_x}{\partial x} + \frac{\partial \mathbf{J}_y}{\partial y} + \frac{\partial \mathbf{J}_z}{\partial z}. \end{aligned} \tag{2.9}$$

Відповідно, оператор Лапласа виражається як:

$$\begin{aligned} \operatorname{div}(\operatorname{grad}(u)) &= \nabla \cdot \nabla u = \nabla^2 u = \\ &= \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial y}\right) + \frac{\partial}{\partial z}\left(\frac{\partial u}{\partial z}\right) = \\ &= \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}. \end{aligned} \tag{2.10}$$

Рівняння (2.5), часто називають рівнянням Лапласа. Якщо лінійне диференціальне рівняння еліптичного типу є неоднорідним, тобто права частина (2.5) рівна не нулю, чи довільній константі, а деякому виразу від незалежних змінних типу $Q(x, y, z)$, то таке рівняння називають рівнянням Пуассона [3].

У векторному і тензорному численні, та, як наслідок, в широкому колі задач, що описуються диференціальними рівняннями частинних похідних, також часто зустрічається оператор над векторним полем, що прийнято називати ротором. Ротор можна знайти як векторний добуток гамільтоніана ∇ на задане векторне поле $\mathbf{J}$ [4]:

$$\begin{aligned} \operatorname{rot}(\mathbf{J}) &= \nabla \times \mathbf{J} = \left\| \begin{bmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ \mathbf{J}_x & \mathbf{J}_y & \mathbf{J}_z \end{bmatrix} \right\| = \\ &= \hat{\mathbf{i}}\left(\frac{\partial \mathbf{J}_z}{\partial y} - \frac{\partial \mathbf{J}_y}{\partial z}\right) + \hat{\mathbf{j}}\left(\frac{\partial \mathbf{J}_x}{\partial z} - \frac{\partial \mathbf{J}_z}{\partial x}\right) + \hat{\mathbf{k}}\left(\frac{\partial \mathbf{J}_y}{\partial x} - \frac{\partial \mathbf{J}_x}{\partial y}\right) = \\ &= \left\{ \frac{\partial \mathbf{J}_z}{\partial y} - \frac{\partial \mathbf{J}_y}{\partial z} \quad \frac{\partial \mathbf{J}_x}{\partial z} - \frac{\partial \mathbf{J}_z}{\partial x} \quad \frac{\partial \mathbf{J}_y}{\partial x} - \frac{\partial \mathbf{J}_x}{\partial y} \right\}^{\mathbf{T}}. \end{aligned} \tag{2.11}$$

2.4. Початкові та крайові умови

Для однозначного розв'язку задачі, що описується диференціальним рівнянням, необхідно вказати початкові та крайові умови, тобто поставити, так звану, крайову задачу. Початкові умови задачі визначають значення потенціалу і його похідних у деякий початковий момент часу τ_0. Крайові умови, аналогічно до початкових, визначають значення потенціалу і його похідних на деяких границях Γ області моделювання Ω. Ці два типи умов нічим не відрізняються, крім фізичного змісту, проте форма границь зазвичай є значно складнішою від поняття моменту часу (винятком можуть послужити напевно тільки деякі, суто абстрактні, рівняння або певні задачі квантової механіки). Очевидно, що початкові та крайові умови задаються у вигляді функцій або відповідних операторних рівнянь.

Існують три основні типи крайових умов [3], [5]:

- Крайові умови першого роду, або крайові *умови Діріхле*, що задають значення потенціалу на границі та в загальному випадку мають вигляд:

$$u(\mathbf{r},\tau)\big|_{\Gamma} = f(\mathbf{r},\tau), \tag{2.12}$$

де $u(\mathbf{r},\tau)\big|_{\Gamma}$ – шукана польова величина на границі тіла Γ, $f(\mathbf{r},\tau)$ – деяка функція.

- Крайові умови другого роду, або крайові *умови Неймана*, що задають так звану густину потоку на границі, тобто першу похідну, та в загальному випадку мають вигляд:

$$J(\mathbf{r},\tau)\big|_{\Gamma} = \left.\frac{\partial u(\mathbf{r},\tau)}{\partial \mathbf{n}}\right|_{\Gamma} = f(\mathbf{r},\tau), \tag{2.13}$$

де $J(\mathbf{r},\tau)\big|_{\Gamma}$ – густина потоку на границі Γ, $\mathbf{n}$ – одинична нормаль до границі Γ.

- Крайові умови третього роду, або крайові *умови Робіна* (задача Робіна розглядається в механіці, натомість у задачах теплопровідності їх називають умовами Ньютона-Ріхмана [6]), що задають так званий потенціальний напір, тобто змішані крайові умови, які в загальному випадку мають вигляд:

$$\left(J(\mathbf{r},\tau)+k\left(u(\mathbf{r},\tau)-u_{\infty}\right)\right)\Big|_{\Gamma} = \left.\left(\frac{\partial u(\mathbf{r},\tau)}{\partial \mathbf{n}}+k\left(u(\mathbf{r},\tau)-u_{\infty}\right)\right)\right|_{\Gamma} = f(\mathbf{r},\tau), \tag{2.14}$$

де k – коефіцієнт пропорційності (в задачах теплопровідності це коефіцієнт тепловіддачі [7]), u_{∞} – потенціал навколишнього середовища.

Визначення початкових та крайових умов, також зручно робити в операторній формі, тут і в подальшому будемо позначати початкові умови оператором $\mathcal{T}(.)$, а крайові умови оператором $\ell(.)$, наприклад крайові умови Робіна можна записати як $\ell(.) = \partial/\partial\mathbf{n} + k - ku_{\infty}$, $\ell(u(\mathbf{r},\tau))\big|_{\Gamma} = f(\mathbf{r},\tau)$.

2.5. Поняття коректності формалізації крайових задач

Визначення початкових і крайових умов задачі визначає коректність чи некоректність її постановки. Задача поставлена коректно тоді і тільки тоді, коли рішення:

- існує;
- єдине;
- неперервно залежить від даних задачі (початкових та граничних умов, коефіцієнтів рівняння, тощо).

З точки зору функціонального аналізу, описані вимоги гарантують існування оберненого оператору $\mathcal{L}(.)^{-1}$, застосування якого дає однозначно визначений та відмінний від безмежності результат.

Вимога неперервної залежності розв'язку крайової задачі обумовлена тим, що фізичні дані, як правило, отримуються з експерименту наближено. Тому потрібно гарантувати, що розв'язок задачі в рамках вибраної математичної моделі не буде суттєво залежати від похибок вимірювання.

Задача, розв'язок якої задовольняє перераховані вище вимоги, називається коректно поставленою. Формально, доведення коректності вимагає конкретних постановок задач, оскільки для різних типів рівнянь та відповідних крайових умов розроблено теореми про існування та єдиність рішення (за Адамаром [3], [5], [8], [9]). На практиці, для коректної постановки задачі слід дотримуватися правила: кількість різних крайових умов, для шуканої польової величини, повинна бути рівна максимальному порядку похідних по часовим і просторовим координатам диференціального рівняння. Для рівнянь першого порядку – одна крайова умова, для рівнянь другого порядку – дві крайові умови, для третього порядку – три, і т.д.

2.6. Список використаної літератури до розділу 2

[1] Feynmann R. – The Character of Physical Law / Характеристики физических законов / пер. с англ. Наппельбаум Э., Голышева В. // Москва: АСТ, 2014.

[2] Farlow S. – Partial Differential Equations for Scientists and Engineers / Уравнения с частными производными для научных работников и инженеров / пер. с англ. Плис А., под ред. Похожаев С. // Москва: Мир, 1985.

[3] Михлин С. – Вариационные методы в математической физике. 2-е изд. перераб. и доп. // Москва: Наука, 1970.

[4] Кочин Н. – Векторное исчисление и начала тензорного исчисления. 9-е изд. // Москва: Наука 1965.

[5] Ладыженская О. – Краевые задачи математической физики // Москва: Наука, 1973.

[6] Zienkiewicz O., Morgan K. – Finite elements and approximation // New-York: Wiley, 1983.

[7] Лыков А. – Теория теплопроводности // Москва: Высшая школа, 1967.

[8] Тихонов А., Самарский А. – Уравнения математической физики: Учебное пособие, 6-е изд. испр. и доп. // Москва: МГУ, 1999.

[9] Тихонов А., Арсенин В. – Методы решения некорректных задач, 2-е изд. // Москва: Наука, 1979.

3. Основи методу скінченних елементів

3.1. Коротка історична довідка

Завдяки значному прогресу в області комп'ютерних наук, з появи перших ЕОМ і до сьогодні, чисельні методи стали основним інструментом математичного моделювання [1]. При цьому, за рядом причин найбільшого поширення набули *проекційно-сіткові* методи. Всі вони передбачають побудову в області, де вирішується задача, розрахункової сітки, тобто дискретизацію області на дрібні фрагменти (елементи) певного виду – трикутники, тетраедри, призми та ін., коли до розмірів і форм елементів також висуваються певні вимоги, так як вони суттєво впливають на похибки апроксимації та збіжність методів.

Одним з найбільш універсальних проекційно-сіткових методів розв'язку задач математичної фізики є *метод скінченних елементів* (скорочено МСЕ), основна ідея якого полягає в побудові дискретної моделі що апроксимує складну невідому функцію за допомогою скінченної множини простіших [2].

Вперше, метод скінченних елементів був запропонований інженерами, знайшов широке застосування на практиці, але значний період часу залишався поза полем зору математиків. Після детального математичного дослідження методу виявилося, що для більшості задач, метод скінченних елементів часто збігається до точного рішення швидше, ніж його основний конкурент – метод скінченних різниць [3].

Виникнення методу скінченних елементів пов'язано з рішенням задач космічних досліджень 50-их років XX століття. Вперше він був опублікований лише як чисельна процедура рішення, в роботі 1956 року[1], де описувалася задача теорії пружності з розв'язуванням в напруженнях. Ця робота спонукала до появи нових робіт, зокрема було опубліковано ряд статей з застосуванням методу скінченних елементів до задач будівельної механіки і механіки неперервних середовищ. Важливий внесок у теоретичну розробку методу було зроблено в 1965 році[2], коли було показано, що метод скінченних елементів можна розглядати як один з варіантів добре відомого в механіці методу Релея-Рітца, для якого вже була розвинута математична база варіаційного числення. Так в будівельній механіці метод скінченних елементів, завдяки процедурі мінімізації потенціальної енергії з методу Релея-Рітца, давав змогу звести задачу до системи лінійних рівнянь балансу.

Зв'язок методу скінченних елементів з процедурою мінімізації привів до широкого використання його при рішенні задач в інших областях інженерії. Метод застосовувався до задач, що описувалися рівняннями Лапласа або Пуассона. Рішення цих рівнянь також пов'язане з мінімізацією деякого функціоналу. В перших публікаціях, за допомогою методу скінченних

[1] Turner M., Clough R., Martin H., Topp L. – Stiffness and Deflection Analysis of Complex Structures // Jour. Aeronaut. Sci., 23:805-824, 1956.

[2] Melosh R. – Baisis for Derivation of Matrices for the Direct Stiffness Method // Jour. Am. Inst. for Aeron. and Astron. (NASA), 1:1631-1637, 1965.

елементів вирішувалися задачі поширення тепла, пізніше, метод був застосований до задач гідромеханіки, зокрема до задач протікання рідини в пористому середовищі.

Область застосування методу скінченних елементів значно розширилася, після того, як в 1969 році[1] було показано, що рівняння, які визначають елементи в задачах будівельної механіки, поширення тепла, гідромеханіки, можуть бути легко отримані за допомогою узагальнень – таких варіантів методу зважених нев'язок (до яких належить МСЕ), як метод Бубнова-Гальоркіна або спосіб найменших квадратів. Встановлення цього факту зіграло важливу роль в теоретичному обґрунтуванні методу скінченних елементів, так як дало змогу застосовувати його при рішенні будь-яких диференційних рівнянь. Слід підкреслити, що більш загальні теоретичні обґрунтування виключають необхідність варіаційного формулювання фізичних задач. Крім того, формулювання методу скінченних елементів, з допомогою методу зважених нев'язок, дає змогу виявити тісний взаємозв'язок з іншими поширеними чисельними методами, такими як метод скінченних різниць, метод граничних елементів, а також з спектральними методами Фур'є [4].

Вже з початку 1970-их років гальоркінський метод скінченних елементів став найбільш популярним методом зважених нев'язок, що застосовувалися з кусково-поліноміальними функціями малої степені. Ріст популярності формулювання Гальоркіна, як і одночасне зниження популярності варіаційного формулювання методу скінченних елементів, співпав з початком проникнення цього методу в області, далекі від механіки конструкцій, де він зародився.

Багато з вказаних областей застосування пов'язані з рухом – наприклад, всі різновиди механіки рідин і газів, а також теорії конвективної теплопередачі. Зазначимо, що переважна більшість "нових" областей важко піддаються опису з допомогою варіаційних формулювань.

Ера варіаційних методів, що почалася приблизно 1964 року з виходом зарубіжного видання [5][2], дала життя теорії скінченних елементів і забезпечила їй строге математичне підґрунтя. Закінчення ж цієї ери пов'язано з появою робіт Стренга і Фікса 1973 року [6], що дали дуже яскравий опис математичних досягнень стосовно методу скінченних елементів за весь період. Напевно, в історичному плані роботи Стренга і Фікса слід розглядати як надпис на надгробній плиті варіаційної ери.

Очевидно, що з математичної точки зору, найбільш цікавими виявилися шляхи подальшого розвитку методу скінченних елементів. Саме тому ми будемо розглядати метод скінченних елементів, як частковий випадок методів зважених нев'язок, а саме, як метод Бубнова-Гальоркіна з спеціальним вибором базисних функцій, кожна з яких має спеціальний мінімальний *скінченний носій*, тобто відмінна від нуля тільки в деякій невеликій підобласті всієї області задачі.

[1] Szabo B., Lee G. – Devariation of Stiffness Matrices for Problems in Plane Elasticity by Galerkin's Method // Intern. Jour. of Numerical Methods in Engineering, 1:301-310, 1969.

[2] Мається на увазі перше зарубіжне видання: Mikhlin S. – Variational methods in mathematical physics // Oxford: Pergamon, 1964.

Мінімальність полягає в тому, що в якості пробних функцій переважно вибираються поліноми низького порядку.

Метод скінченних елементів перетворився з чисельної процедури рішення задач будівельної механіки, в загальний метод чисельного рішення диференційних рівнянь чи їх систем, і не останню роль тут зіграло фінансування досліджень Американського національного комітету по дослідженню космічного простору (NASA). На початку XXI століття метод скінченних елементів став потужним засобом наближеного рішення диференційних рівнянь, що описують велике коло фізичних процесів. Різноманітні його застосування в техніці та наукових дослідженнях, і можна з повною впевненістю сказати, що без нього та його слуги ЕОМ, багато з задач не могли б бути вирішені взагалі. Важко уявити теперішню прикладну промислову систему автоматизованого моделювання, що не використовує метод скінченних елементів, який проник в усі інженерні галузі, і зокрема в галузь проектування мікроелектромеханічних систем.

3.2. Методи Бубнова-Гальоркіна

Перед тим, як розглянути метод скінченних елементів у контексті методів зважених нев'язок, для кращого розуміння, віддамо історичну данину частковим випадкам – методам Бубнова-Гальоркіна.

Виникнення методів Бубнова-Гальоркіна пов'язують з публікацією 1915-го року[1], що була присвячена пружній рівновазі стержнів та тонких пластин. Формулювання Гальоркіна дуже часто пов'язане з іменем Бубнова [5], який запропонував своє формулювання у зв'язку з варіаційним підходом до рішення задач на власні значення, тому в подальшому методи дістали назву методів Бубнова-Гальоркіна.

Методи до сьогоднішнього часу вже були застосовані при вирішенні численних задач механіки конструкцій, динаміки будівель, гідромеханіки, теорії гідродинамічної рівноваги, теорії тепло- і масообміну, акустики, теорії поширення мікрохвиль, теорії переносу нейтронів і т.д. З допомогою представлень Бубнова-Гальоркіна були проведені дослідження звичайних диференціальних рівнянь, рівнянь з частковими похідними та інтегральних рівнянь. Стаціонарні і нестаціонарні задачі, а також задачі на власні значення виявилися в однаковій мірі такими, що піддаються дослідженню на основі підходів Бубнова-Гальоркіна. Насправді, будь-яка задача, для якої можна вивести визначальне рівняння, може бути вирішена з допомогою одного з різновидів методів Бубнова-Гальоркіна [4].

Суть методів Бубнова-Гальоркіна полягає в апроксимації невідомої величини деякою сумою, так званих, лінійно незалежних *базисних* функцій, що переважно представляють собою прості в обчисленні аналітичні функції. Ці функції часто називають *пробними*. В термінах абстрактної алгебри і

[1] Галёркин Б. – Стержни и пластинки. Ряды в некоторых вопросах упругого равновесия стержней и пластинок // Вестник инженеров, 19:897-908, 1915.

функціонального аналізу методи Бубнова-Гальоркіна відносяться до класу *проекційних* методів, оскільки в них, шляхом апроксимації, будується проекція шуканого рішення в простір, утворений вибраними базисними функціями. Узагальнене формулювання цих методів отримало назву – метод Петрова-Гальоркіна [4], [7], [8], [9].

Розглянемо схему рішення крайової задачі методом Бубнова-Гальоркіна, на прикладі звичайного диференційного рівняння:

$$\frac{dy(x)}{dx} - y(x) = 0, \quad y(0) = 1, \quad 0 \le x \le 1, \tag{3.1}$$

або в операторній формі запису:

$$\mathcal{L}(.) = \frac{d}{dx} - 1, \quad \mathcal{L}(y(x)) = 0, \quad \ell(.) = 1, \quad \ell(y(x))\Big|_0 = 1, \quad x \in \Omega = [0;1]. \tag{3.2}$$

Щоб мати можливість порівняти результати, спочатку знайдемо аналітичне рішення задачі:

$$\frac{dy(x)}{dx} - y(x) = 0 \quad \Rightarrow \quad \frac{dy(x)}{dx} = y(x) \quad \Rightarrow \quad \frac{dy(x)}{dx} \cdot \frac{1}{y(x)} = 1,$$

$$\int \left(\frac{dy(x)}{dx} \cdot \frac{1}{y(x)} \right) dx = \int 1 dx \quad \Rightarrow \quad \ln(y(x)) = x + C, \tag{3.3}$$

$$y(x) = e^{x+C} \quad \Rightarrow \quad e^{0+C} = 1 \quad \Rightarrow \quad C = 0 \quad \Rightarrow \quad y(x) = e^x.$$

У методі Бубнова-Гальоркіна припускають, що невідома функція може бути достатньо точно апроксимована з допомогою *наближеного рішення* виду:

$$y(x) \approx \tilde{y}(x) = y_0 + \sum_{j=1}^{M} a_j \varphi_j(x), \tag{3.4}$$

де $\varphi_j(x)$ – відомі аналітичні, лінійно незалежні, базисні функції, що прийнято називати пробними, a_j – коефіцієнти, які необхідно знайти. Очевидно, що система пробних функцій повинна бути вибрана таким чином, щоб гарантувати збільшення точності рішення при збільшенні кількості M пробних функцій, тобто $\tilde{y}(x) \to y(x)$ при $M \to \infty$.

Для конкретного прикладу, виберемо у якості пробних функцій вираз x^j, таким чином апроксимація буде здійснюватися поліномом степеня M. Значення y_0, у методах Бубнова-Гальоркіна, зазвичай вибирається так, щоб в сукупності з сумою добутків $a_j \varphi_j(x)$ задовольнити крайові умови. В даному випадку $y_0 = y(0) = 1$, і при будь-яких a_j, $\tilde{y}(0) = 1$.

Якщо підставити в операторний вираз (3.2), замість точного рішення $y(x)$, його апроксимацію $\tilde{y}(x)$, то в загальному випадку, отримаємо відмінну від нуля нев'язку[1]:

[1] Не вдаючись в деталі, *нев'язкою* називають різницю правих частин, що утворюється між апроксимаційним та оригінальним рівняннями.

$$R(x) = \mathcal{L}(\tilde{y}(x)) = \mathcal{L}(1) + \sum_{j=1}^{M} a_j \mathcal{L}(x^j) =$$
$$= \left(\frac{d}{dx} - 1\right) \cdot 1 + \sum_{j=1}^{M} \left[a_j \left(\frac{d}{dx} - 1\right) \cdot x^j \right] = \tag{3.5}$$
$$= -1 + \sum_{j=1}^{M} a_j (j x^{j-1} - x^j).$$

Розширенням операції множення для функціональних залежностей чи польових величин є поняття *скалярного добутку*, що характерне для простору, в якому розглядається задача. Домовимося, що всі задачі розглядаються в Евклідовому просторі, де скалярний добуток двох функцій $\langle u(\mathbf{r}), v(\mathbf{r}) \rangle$ можна визначити як інтеграл [5], [10]:

$$\langle u(\mathbf{r}), v(\mathbf{r}) \rangle \equiv \int_{\Omega} u(\mathbf{r}) v(\mathbf{r}) d\Omega. \tag{3.6}$$

Якщо аргументами є не функції, а векторні величини, то скалярний добуток прийме звичну форму скалярного добутку, відомого з курсу лінійної алгебри.

За допомогою операції скалярного добутку, диференціальне рівняння, що описує фізичні явища локально в нескінченно малих межах відносно довільної точки, переноситься на конкретний об'єкт моделювання, що має свої специфічні форми та відповідні границі, після чого задача вже розглядається глобально відносно цього об'єкту. Це пояснюється тим, що скалярний добуток тісно пов’язаний з *ортогональною проекцією* точного рішення задачі в підпростір базисних функцій, які утворюють апроксимацію при використанні проекційного чи проекційно-сіткового методу. Дві функції є ортогональними в деякій області, якщо їх скалярний добуток є рівний нулю [5], тобто:

$$\langle u(\mathbf{r}), v(\mathbf{r}) \rangle = \int_{\Omega} u(\mathbf{r}) v(\mathbf{r}) d\Omega = 0. \tag{3.7}$$

Важливою особливістю, що визначає простір, є функціональна залежність, яка ставить кожній точці простору деяке число – абстрактну "відстань" чи "довжину". Цю залежність прийнято називати *нормою* і позначати як $\|.\|$. Норми бувають різні, в основному ми будемо використовувати норми сімейства лінійних диференціальних операторів $\mathcal{L}_2(\Omega)$, що визначаються формулою [10]:

$$\|u(\mathbf{r})\|_{\mathcal{L}_p(\Omega)} \equiv \left(\int_{\Omega} |u(\mathbf{r})|^p \, d\Omega \right)^{\frac{1}{p}}, \tag{3.8}$$

або її дискретний аналог:

$$\| u \|_{\mathcal{L}_p(\Omega^d),d} \equiv \left(\sum_{l=1}^{L} |u_l|^p \right)^{\frac{1}{p}}. \tag{3.9}$$

Так для $p = 2$, дискретна $\mathcal{L}_{2,d}$-норма, це класична Евклідова норма, за допомогою якої можна визначити відстань між двома точками:

$$\|a-b\|_2 \equiv \sqrt{\sum_{l=1}^{N}(a_l-b_l)^2}. \qquad (3.10)$$

Щоб знайти значення коефіцієнтів y_j з рівнянь (3.4) та (3.5), потрібно поставити умову ортогональності нев'язки до обраного базису, тобто розв'язати систему рівнянь:

$$\langle R(x),\varphi_i(x)\rangle = 0, \quad i=1,2,\ldots,M, \qquad (3.11)$$

де $\varphi_i(x)$ – ті самі відомі аналітичні, лінійно незалежні, базисні функції, що розглядалися в (3.4). В даному випадку:

$$\left\langle \left(-1+\sum_{j=1}^{M} a_j\,(jx^{j-1}-x^j)\right), x^{i-1}\right\rangle = 0,$$
$$\left\langle \left(\sum_{j=1}^{M} a_j\,(jx^{j-1}-x^j)\right), x^{i-1}\right\rangle = \langle 1, x^{i-1}\rangle. \qquad (3.12)$$

Враховуючи, що $i, j = 1,2,\ldots,M$, та $0 \le x \le 1$, то:

$$\int_0^1 (jx^{j-1}-x^j)x^{i-1}dx = \frac{j}{i+j-1}-\frac{1}{i+j}, \quad \int_0^1 x^{i-1}dx = \frac{1}{i}, \qquad (3.13)$$

або в матричній формі:

$$[\mathbf{K}]\{\mathbf{a}\} = \{\mathbf{f}\}, \quad [\mathbf{K}]_{i,j} = \frac{j}{i+j-1}-\frac{1}{i+j}, \quad \{\mathbf{f}\}_i = \frac{1}{i}. \qquad (3.14)$$

Розв'язавши матричну систему рівнянь (3.14), отримаємо поліном, що апроксимує рішення. Так для $M=1$, поліном виглядає як $y(x)=1+2x$, для $M=2$ $y(x)=1+0{,}857143x+0{,}857143x^2$, для $M=3$ $y(x)=1+1{,}014085x+$ $+0{,}422535x^2+0{,}281690x^3$, і т.д.

Таблиця 3.1

Рішення рівняння $dy(x)/dx - y(x) = 0$
з допомогою традиційного методу Гальоркіна

x	Апроксимоване рішення $\tilde{y}(x)$				Точне рішення $y(x)=e^x$
	$M=1$	$M=2$	$M=3$	$M=4$	
0	1,000000	1,000000	1,000000	1,000000	1,000000
0,2	1,400000	1,205714	1,221972	1,221411	1,221403
0,4	1,800000	1,480000	1,491268	1,491860	1,491825
0,6	2,200000	1,822857	1,821408	1,822090	1,822119
0,8	2,600000	2,234286	2,225915	2,225526	2,225541
1	3,000000	2,714286	2,718310	2,718282	2,718282
$\|y(x)-\tilde{y}(x)\|_{2,d}$	0,690066	0,009788	0,000069	$2{,}698782\times10^{-7}$	0,000000
$\|R(x)\|_{2,d}$	2,449490	0,349927	0,034500	0,002447	0,000000

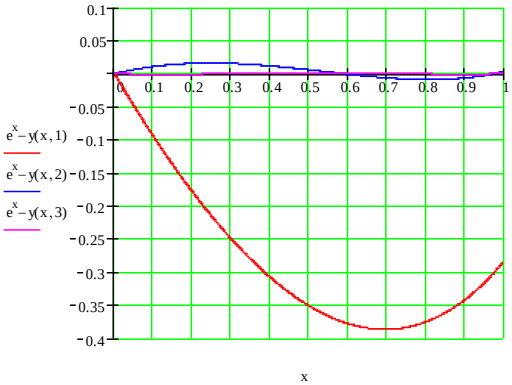


Рис. 3.1 Розподіл похибки $y(x)-\tilde{y}(x)$ *для рішення рівняння* $dy(x)/dx-y(x)=0$ *з параметром* M

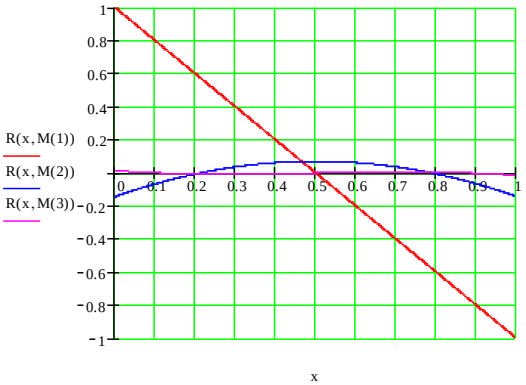


Рис. 3.2 Розподіл нев'язки $R(x)$ *для рішення рівняння* $dy(x)/dx-y(x)=0$ *з параметром* M

Для оцінки точності отриманого апроксимованого рішення, використаємо $\mathcal{L}_{2,d}$-норму (3.8), (3.9), таким чином $\left\|y(x)-\tilde{y}(x)\right\|_{2,d}$ буде оцінювати похибку отриманого результату, або іншими словами, показуватиме відстань в функціональному просторі між точним і апроксимованим рішенням. Чим менша відстань, тим точніше апроксимоване рішення.

Як видно з *Таблиця 3.1*, та *Рис. 3.2*, дискретна норма нев'язки $\left\|R(x)\right\|_{2,d}$ також швидко зменшується зі збільшенням кількості M пробних функцій. Якщо врахувати, що крайова умова задовольняється точно, можна очікувати, що $\left\|R(x)\right\|_2\to 0$ при $\left\|y(x)-\tilde{y}(x)\right\|_2\to 0$. В практичних розрахунках точне рішення зазвичай невідоме, і значення $\left\|y(x)-\tilde{y}(x)\right\|_2$ вирахувати неможливо, однак завжди можна визначити значення $\left\|R(x)\right\|_2$.

3.3. Різновиди методів зважених нев'язок

Методи Бубнова-Гальоркіна можна трактувати як часткові випадки більш загального класу методів, під назвою *методи зважених нев'язок* (скорочено МЗН). Назва методів зважених нев'язок, швидше всього була введена в роботі 1956 року[1], але аналогічна ідея розглядалася ще в 1953 році[2], під назвою "принцип розподілу похибок". Основна ідея методів зважених нев'язок полягає у введенні, так званих, *вагових* функцій, що прийнято називати *повірочними*, за допомогою яких, при збільшенні кількості пробних функцій, прямує до нуля нев'язка між точним і апроксимованим рішенням задачі. Річ у тому, що функції $\varphi_i(x)$ з рівняння (3.11) виступають у якості вагових функцій, що зважують нев'язки $R(x)$. У загальному випадку методів зважених нев'язок, повірочні

[1] Crandall S. – Engineering analysis // New York: McGraw-Hill, 1956.
[2] Collatz L. – The numerical treatment of differential equations // Berlin: Springer-Verlag, 1953.

функції не обов'язково співпадають з пробними, щоб їх розрізняти, в літературі повірочні функції часто позначають як $W_i(\mathbf{r})$, або $\omega_i(\mathbf{r})$ [3], [4]. Таким чином, методи Бубнова-Гальоркіна є частковими випадками методів зважених нев'язок, де пробні і повірочні функції співпадають.

Формально, методи зважених нев'язок можна описати наступним чином [4]: Нехай в деякій області Ω, з границями Γ, задано диференціальне рівняння:

$$\mathcal{L}(u(\mathbf{r},\tau)) = 0, \tag{3.15}$$

яке повинно бути вирішене при початкових умовах $\mathcal{T}(u(\mathbf{r},\tau))\big|_{\tau=\tau_0} = 0$ і крайових умовах $\ell(u(\mathbf{r},\tau))\big|_{\Gamma} = 0$. Вводиться наближене рішення $\tilde{u}(\mathbf{r},\tau)$, таке що:

$$\mathcal{L}(\tilde{u}(\mathbf{r},\tau)) = R^{\Omega}(\mathbf{r},\tau), \quad \mathcal{T}(\tilde{u}(\mathbf{r},\tau))\big|_{\tau=\tau_0} = R^{\mathcal{T}}(\mathbf{r},\tau), \quad \ell(\tilde{u}(\mathbf{r},\tau))\big|_{\Gamma} = R^{\Gamma}(\mathbf{r},\tau). \tag{3.16}$$

При побудові наближеного рішення $\tilde{u}(\mathbf{r},\tau)$, можна йти по одному з наступних шляхів:

- Диференціальне рівняння задовольняється точно, тобто $R^{\Omega}(\mathbf{r},\tau) = 0$. Такі методи відносяться до підкласу *граничних методів*.
- Крайові умови задовольняються точно, тобто $R^{\Gamma}(\mathbf{r},\tau) = 0$. Такі методи відносяться до підкласу *внутрішніх методів*.
- Ні диференціальне рівняння, ні крайові умови не задовольняються точно. Такі методи відносяться до підкласу *змішаних методів*.

Наближене рішення, аналогічно до (3.4), представляється у вигляді:

$$u(\mathbf{r},\tau) \approx \tilde{u}(\mathbf{r},\tau) = u_0(\mathbf{r},\tau) + \sum_{j=1}^{M} a_j(\tau)\varphi_j(\mathbf{r}), \tag{3.17}$$

де $a_j(\tau)$ – коефіцієнти, що необхідно знайти. Для цього, аналогічно до (3.11), отримані нев'язки R^{Ω}, $R^{\mathcal{T}}$ та R^{Γ} прирівнюють до нуля, за допомогою скалярного добутку з системою повірочних функцій $\omega_i(\mathbf{r})$. Тобто, ставиться вимога ортогональності нев'язки до обраних вагових функцій:

$$\langle R(\mathbf{r}), \omega_i(\mathbf{r}) \rangle = 0, \quad i = 1,2,\ldots,M, \tag{3.18}$$

і в залежності від того, як визначений скалярний добуток, тобто чи простір де розглядається задача є неперервним або дискретним, отримаємо класичний або дискретний метод зважених нев'язок. Останнє рівняння часто називають *рівнянням методу зважених нев'язок*.

Якщо задача, що розглядається, описується еліптичним рівнянням, то скалярний добуток завжди можна розписати, аналогічно до (3.14), як систему лінійних рівнянь в матричному вигляді:

$$[\mathbf{K}]\{\mathbf{a}\} = \{\mathbf{f}\}. \tag{3.19}$$

Вектор $\{\mathbf{a}\}$ містить невідомі коефіцієнти a_j. Починаючи від задач теорії пружності, матрицю $[\mathbf{K}]$ прийнято називати *матрицею жорсткості*, а вектор $\{\mathbf{f}\}$ – *вектором навантажень* [2], хоча назва та позначення не принципові і

можуть відрізнятися в кожній окремій задачі, залежно від фізичного змісту, що в них закладається.

Наведемо приклади методів зважених нев'язок, що найчастіше використовуються, їх порівняння можна знайти в *Таблиця 3.2*.

- **Метод найменших квадратів**

Найстаріший з методів, що відносяться до методів зважених нев'язок. Вперше запропонований Гаусом у 1795 році[1]. Ідея методу полягає у мінімізації інтегралу від квадрату нев'язки:

$$I(a_1, a_2, \ldots, a_M) = \int_\Omega R(\mathbf{r})^2 d\Omega, \tag{3.20}$$

для чого припускають, що:

$$\frac{\partial I}{\partial a_i} = 0, \quad i = 1, 2, \ldots, M. \tag{3.21}$$

Це еквівалентно тому, що:

$$\omega_i(\mathbf{r}) = \frac{\partial R(\mathbf{r})}{\partial a_i}. \tag{3.22}$$

Оскільки в даному випадку $\partial R(\mathbf{r})/\partial a_i = \varphi_i(\mathbf{r})$, то $\omega_i(\mathbf{r}) = \varphi_i(\mathbf{r})$, тому можна показати, що I досягає мінімуму при:

$$\left\langle R(\mathbf{r}), \varphi_i(\mathbf{r}) \right\rangle = \int_\Omega R(\mathbf{r})\varphi_i(\mathbf{r}) d\Omega = 0. \tag{3.23}$$

Останнє рівняння точно співпадає з стандартним рівнянням методів зважених нев'язок (3.18), крім того, в даному випадку рівняння співпадає з рівнянням методів Бубнова-Гальоркіна.

- **Метод підобластей (коллокацій по підобластях)**

Вперше з'явився в 1923 році[2]. У цьому методі, система вагових функцій ставиться в залежність від деяких підобластей Ω_i, загальної області Ω, і записується у вигляді:

$$\omega_i(\mathbf{r}) = \begin{cases} 1, & \mathbf{r} \in \Omega_i, \\ 0, & \mathbf{r} \notin \Omega_i. \end{cases} \tag{3.24}$$

Вибір такої системи вагових функцій еквівалентний тому, що рівна нулю нев'язка по кожній з підобластей $R^{\Omega_i}(\mathbf{r}) = 0$. Таким чином скалярний добуток (3.18) запишеться у формі матричної системи рівнянь (3.19) де елементи системи рівні:

$$[\mathbf{K}]_{i,j} = \int_{\Omega_i} \mathcal{L}(\varphi_j(\mathbf{r})) d\Omega_i, \quad [\mathbf{f}]_i = \int_{\Omega_i} \ell(u(\mathbf{r})) d\Omega_i. \tag{3.25}$$

Метод підобластей, є напевно першим з, так званих, *локальних* методів, де пробні і повірочні функції не поширюються на всю область рішення, а визначені на підобластях. Інші ж методи є *глобальними*.

[1] Crandall S. – Engineering analysis // New York: McGraw-Hill, 1956.
[2] Biezeno C., Koch J. // Jour. Ingenieur., 38:25-36, 1923.

- **Метод коллокацій (поточкових коллокацій)**

Вперше запропонований в 1937 році[1]. У цьому методі, система вагових функцій записується у вигляді:

$$\omega_i(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{r}_i), \tag{3.26}$$

де δ – дельта-функція Дірака, що за визначенням має властивості:

$$\delta(\mathbf{r} - \mathbf{r}_i) = \begin{cases} \infty, & \mathbf{r} = \mathbf{r}_i, \\ 0, & \mathbf{r} \neq \mathbf{r}_i, \end{cases} \quad \int_{-\infty}^{+\infty} G(\mathbf{r})\delta(\mathbf{r} - \mathbf{r}_i)d\Omega = G(\mathbf{r}_i). \tag{3.27}$$

Вибір такої системи вагових функцій еквівалентний тому, що нев'язка $R^{\Omega}(\mathbf{r}_i) = 0$. Таким чином скалярний добуток (3.18), запишеться у формі матричної системи рівнянь (3.19), де елементи системи рівні:

$$[\mathbf{K}]_{i,j} = \mathcal{L}(\phi_j(\mathbf{r}))\Big|_{\mathbf{r}=\mathbf{r}_i}, \quad [\mathbf{f}]_i = \ell(u(\mathbf{r}))\Big|_{\mathbf{r}=\mathbf{r}_j}. \tag{3.28}$$

- **Метод моментів**

Вперше запропонований в 1947 році[2]. У цьому методі, система вагових функцій записується у вигляді:

$$\omega_i(x) = x^{i-1}. \tag{3.29}$$

- **Метод Гальоркіна (метод Бубнова-Гальоркіна)**

Як вже було сказано, у методах Бубнова-Гальоркіна, вагові і базисні функції вибираються з одного і того ж сімейства функцій:

$$\omega_i(\mathbf{r}) = \varphi_i(\mathbf{r}), \quad i = 1, 2, \ldots, M. \tag{3.30}$$

Відповідно, аналогічно до (3.12)-(3.14), скалярний добуток (3.18) запишеться у формі матричної системи рівнянь (3.19) де елементи системи рівні:

$$[\mathbf{K}]_{i,j} = \int_{\Omega} \varphi_i(\mathbf{r})\mathcal{L}(\varphi_j(\mathbf{r}))d\Omega, \quad [\mathbf{f}]_i = \int_{\Gamma} \varphi_i(\mathbf{r})\ell(u(\mathbf{r}))d\Gamma. \tag{3.31}$$

Наведемо основні вимоги, що повинні виконуватися при використанні традиційного методу Гальоркіна:

- Повірочні функції $\omega_i(\mathbf{r})$ вибираються з того ж сімейства, що і пробні $\varphi_j(\mathbf{r})$;
- Пробні і повірочні функції повинні бути лінійно незалежними.

Крім того, сюди можна додати ще кілька умов, що в першу чергу пов'язані з ефективністю використання методу, зокрема:

- Пробні і повірочні функції повинні бути ортогональними одні до одного (умова Бубнова);
- Пробні і повірочні функції повинні представляти собою M перших елементів повної системи функцій (застосування функцій, починаючи одразу з високих порядків, приведе до погіршення збіжності методу);

[1] Frazar R., Jones W., Skan S. // ARC R&M 1799, 1937.
[2] Yamada H. // Rept. Res. Inst. Fluid Eng. Kyushu Univ., 3:29, 1947.

- Пробні функції повинні точно задовольняти початкові та крайові умови (побудова внутрішнього методу зважених нев'язок дає змогу значно спростити обчислення).

- **Узагальнений метод Гальоркіна (метод Петрова-Гальоркіна)**

Вперше був запропонований у 1940 році[1] для задач конвективно-дифузійного протікання рідини з переважним конвективним вкладом, оскільки класичний метод в подібних випадках мав небажані характеристики стійкості. У цьому методі, система вагових функцій записується у вигляді:

$$\omega_i(\mathbf{r}) = P_i(\mathbf{r}), \tag{3.32}$$

де $P_i(\mathbf{r})$ – аналітична функція, аналогічна до повірочної функції $\varphi_i(\mathbf{r})$, що використовується при застосуванні методів Бубнова-Гальоркіна, але містить додаткові члени або множники, що необхідні для виконання деяких додаткових вимог до рішення задачі. Іншими словами, пробне рішення будується по одному базису, а ортогональність нев'язок вимагається до іншого.

Таблиця 3.2

Порівняння основних методів зважених нев'язок

Метод зважених нев'язок	Методи Бубнова-Гальоркіна	Метод найменших квадратів	Метод підобластей	Метод коллокацій
Точність	Дуже висока	Дуже висока	Висока	Помірна
Простота формулювання	Помірна	Низька	Висока	Дуже висока
Примітки	Еквівалентні методу Релея-Рітца, якщо його можна застосувати до даного рівняння	Непридатний до часозалежних задачі та задач на власні значення	Еквівалентний методу скінченних об'ємів, підходить для законів збереження	Ортогональна коллокація дає високу точність

- **Спектральні методи зважених нев'язок**

Спектральні методи використовуються при рішенні задач з багатьох областей, але найбільш широко вони застосовувались до двох класів проблем: глобальне атмосферне моделювання (вперше в 1954 році[2]) і фундаментальні дослідження турбулентності (вперше в 1968 році[3]).

Методи, що відносяться до цього підкласу, подібно до традиційних методів Бубнова-Гальоркіна є глобальними методами, тобто обрані пробні і повірочні функції охоплюють всю область рішення. Основною вимогою спектральних методів є ортогональність пробних і повірочних функцій:

$$\left\langle \varphi_j(\mathbf{r}), \omega_i(\mathbf{r}) \right\rangle \begin{cases} \neq 0, & i = j, \\ = 0, & i \neq j, \end{cases} \tag{3.33}$$

[1] Петров Г. – Применение метода Галеркина к задаче об устойчивости вязкой жидкости // ПММ, т. 4(3):3-11, 1940.
[2] Siberman I., J. Meteorol. // 11:27-34, 1954.
[3] Orszag S., Kruskal M. // Physics of Fluids, 11:43-60, 1968.

завдяки чому в розрахунках, майже завжди, приймають участь тільки діагональні елементи матриці (3.19), що в свою чергу веде до майже лінійної складності матричних обчислень.

У спектральних методах прийнято використовувати пробні і повірочні функції з сімейства ортогональних функцій наведених в *Таблиця 3.3*.

Таблиця 3.3

Ієрархія сімейства пробних функцій, що застосовуються в спектральних методах зважених нев'язок

Пробна функція	Примітки
Розклад за власними функціями	Показується рішенням подібної задачі
Ряди Фур'є	Періодичні крайові умови, нескінченна диференційованість рішення
Ряди за поліномами Лежандра	Хороша роздільна здатність на довжину хвилі, неперіодичність
Ряди за поліномами Чебишова	Дуже ефективні, неперіодичність, наявність мінімаксу

3.4. Використання методів зважених нев'язок при рішенні задач

Для того, щоб зрозуміти, як апроксимувати крайові умови, будуючи розв'язок на основі методів зважених нев'язок, спочатку детально розглянемо процес апроксимації в підкласі внутрішніх методів, тобто методів де крайові умови задовольняються точно $R^{\Gamma}(\mathbf{r},\tau)=0$. А пізніше перенесемо результати на підкласи змішаних методів, де по ряду причин, виникає необхідність апроксимації крайових умов, та граничних методів.

- **Внутрішні методи зважених нев'язок**

Розглянемо однорідне диференціальне рівняння в області Ω з границями Γ, що описується лінійним еліптичним оператором:

$$\mathcal{L}(u(\mathbf{r}))=k, \tag{3.34}$$

де k – константа. Рішення повинно задовольняти однорідні крайові умови:

$$\ell(u(\mathbf{r}))\big|_{\Gamma}=f. \tag{3.35}$$

Наприклад, це можуть бути крайові умови Діріхле і Неймана:

$$\begin{aligned} &\ell_1(u(\mathbf{r}))=u(\mathbf{r}), \quad f_1=u_\infty \quad \mathbf{r}\in\Gamma_u \quad \Rightarrow \quad u(\mathbf{r})\big|_{\Gamma_u}=u_\infty, \\ &\ell_2(u(\mathbf{r}))=\frac{\partial u(\mathbf{r})}{\partial \mathbf{n}}, \quad f_2=q \quad \mathbf{r}\in\Gamma_q \quad \Rightarrow \quad \frac{\partial u(\mathbf{r})}{\partial \mathbf{n}}\bigg|_{\Gamma_q}=q. \end{aligned} \tag{3.36}$$

Побудуємо наближене рішення $\tilde{u}(\mathbf{r})$ методом зважених нев'язок відповідно до (3.17). При цьому, як вже говорилося, зробимо це так, щоб задовольнити крайові умови:

$$\ell_g\left(u_0(\mathbf{r})+\sum_{j=1}^{M}a_j\varphi_j(\mathbf{r})\right)=f_g, \quad \mathbf{r}\in\Gamma_g, \quad g=1,2,\ldots,G, \tag{3.37}$$

де G – порядок диференціального рівняння.

Дійсно, якщо здійснити безпосереднє диференціювання апроксимацій (3.17), то можна отримати апроксимації похідних від $u(\mathbf{r})$. Як наслідок, якщо пробні функції $\varphi_j(\mathbf{r})$ є неперервними в області Ω і всі їх похідні існують, то:

$$\begin{aligned} u(\mathbf{r}) \quad &\approx \tilde{u}(\mathbf{r}) \quad &&= u_0(\mathbf{r}) + \sum_{j=1}^{M} a_j \varphi_j(\mathbf{r}), \\ \frac{\partial u(\mathbf{r})}{\partial \mathbf{r}} \quad &\approx \frac{\partial \tilde{u}(\mathbf{r})}{\partial \mathbf{r}} \quad &&= \frac{\partial u_0(\mathbf{r})}{\partial \mathbf{r}} + \sum_{j=1}^{M} a_j \frac{\partial \varphi_j(\mathbf{r})}{\partial \mathbf{r}}, \\ \frac{\partial^2 u(\mathbf{r})}{\partial \mathbf{r}^2} \quad &\approx \frac{\partial^2 \tilde{u}(\mathbf{r})}{\partial \mathbf{r}^2} \quad &&= \frac{\partial^2 u_0(\mathbf{r})}{\partial \mathbf{r}^2} + \sum_{j=1}^{M} a_j \frac{\partial^2 \varphi_j(\mathbf{r})}{\partial \mathbf{r}^2}, \\ &\dots \end{aligned} \tag{3.38}$$

Так як побудований розклад задовольняє крайові умови, то для отримання апроксимації шуканого потенціалу $u(\mathbf{r})$, потрібно гарантувати, щоб $\tilde{u}(\mathbf{r})$ було наближеним рішенням диференціального рівняння. Підставляючи $\tilde{u}(\mathbf{r})$ в це рівняння отримаємо нев'язку по області R^{Ω} (3.16):

$$R^{\Omega}(\mathbf{r}) = \mathcal{L}(\tilde{u}(\mathbf{r})) - k = \mathcal{L}(u_0(\mathbf{r})) + \sum_{j=1}^{M} a_j \mathcal{L}(\varphi_j(\mathbf{r})) - k. \tag{3.39}$$

Щоб отримати наближену рівність $R^{\Omega} = 0$ по всій області Ω, використаємо скалярний добуток (3.18) з системою вагових функцій $\omega_i(\mathbf{r})$:

$$\int_{\Omega} R^{\Omega}(\mathbf{r}) \omega_i(\mathbf{r}) d\Omega = \int_{\Omega} \omega_i(\mathbf{r}) \left[\mathcal{L}(u_0(\mathbf{r})) + \sum_{j=1}^{M} a_j \mathcal{L}(\varphi_j(\mathbf{r})) - k \right] d\Omega = 0. \tag{3.40}$$

Вибираючи $i, j = 1, 2, \dots, M$, отримаємо систему лінійних алгебраїчних рівнянь (3.19), де:

$$\begin{aligned} [\mathbf{K}]_{i,j} &= \int_{\Omega} \omega_i(\mathbf{r}) \mathcal{L}(\varphi_j(\mathbf{r})) d\Omega, \quad 1 \le i, j \le M, \\ [\mathbf{f}]_i &= \int_{\Omega} \omega_i(\mathbf{r}) k d\Omega - \int_{\Omega} \omega_i(\mathbf{r}) \mathcal{L}(u_0(\mathbf{r})) d\Omega, \quad 1 \le i \le M. \end{aligned} \tag{3.41}$$

- **Змішані методи зважених нев'язок**

Необхідність у виборі пробних функцій в підкласі внутрішніх методів, що задовольняють крайові умови, суттєво звужує кількість можливих видів цих функцій. Уникнути цього недоліку можна за допомогою використання підкласу змішаних методів зважених нев'язок.

Відповідно до цього, будемо тепер вважати, що розклад наближеного рішення (3.17) не обов'язково задовольняє одну чи всі крайові умови задачі. Тобто, виключимо доданок початкового наближення $u_0(\mathbf{r})$ і знімемо певні обмеження на вибір пробних функцій:

$$u(\mathbf{r}) \approx \tilde{u}(\mathbf{r}) = \sum_{j=1}^{M} a_j \varphi_j(\mathbf{r}). \tag{3.42}$$

Щоб виконати крайові умови задачі, в такому випадку, їх потрібно апроксимувати аналогічно до шуканого рішення, використовуючи нев'язку по крайовим умовам $\ell(\tilde{u}(\mathbf{r}))\big|_{\Gamma} = R^{\Gamma}(\mathbf{r})$ (і якщо необхідно, нев'язки по початковим умовам $\mathcal{T}(\tilde{u}(\mathbf{r},\tau))\big|_{\tau=\tau_0} = R^{\mathcal{T}}(\mathbf{r},\tau)$). Так ми отримаємо змішаний метод зважених нев'язок, коли $R^{\Omega}, R^{\Gamma} \neq 0$:

$$R^{\Gamma}(\mathbf{r}) = \ell(\tilde{u}(\mathbf{r})) - f = \sum_{j=1}^{M} a_j \ell(\varphi_j(\mathbf{r})) - f. \tag{3.43}$$

Система рівнянь зважених нев'язок для визначення коефіцієнтів u_j будується на основі суми скалярних добутків по всіх нев'язках:

$$\left\langle R^{\Omega}(\mathbf{r}), \omega_i^{\Omega}(\mathbf{r}) \right\rangle + \left\langle R^{\Gamma}(\mathbf{r}), \omega_i^{\Gamma}(\mathbf{r}) \right\rangle = 0, \quad i = 1,2,\ldots,M, \tag{3.44}$$

де, вагові функції ω_i^{Ω} та ω_i^{Γ} в принципі можуть бути вибрані незалежно. Тобто, якщо система рівнянь (3.44) виконується для великої кількості довільних ω_i^{Ω} та ω_i^{Γ}, то апроксимація $\tilde{u}(\mathbf{r})$ повинна наближувати точне рішення $u(\mathbf{r})$ при умові, що розклад (3.42) взагалі здатний це зробити. Це твердження не змінюється, якщо ω_i^{Ω} та ω_i^{Γ} якимось чином пов'язані.

Виведемо формули для елементів матричної системи (3.19):

$$\begin{aligned} \int_{\Omega} R^{\Omega}(\mathbf{r})\omega_i^{\Omega}(\mathbf{r})d\Omega &= \int_{\Omega} \omega_i^{\Omega}(\mathbf{r})\left[\sum_{j=1}^{M} a_j \mathcal{L}(\varphi_j(\mathbf{r})) - k\right] d\Omega, \\ \int_{\Gamma} R^{\Gamma}(\mathbf{r})\omega_i^{\Gamma}(\mathbf{r})d\Gamma &= \int_{\Gamma} \omega_i^{\Gamma}(\mathbf{r})\left[\sum_{j=1}^{M} a_j \ell(\varphi_j(\mathbf{r})) - f\right] d\Gamma, \end{aligned} \tag{3.45}$$

та:

$$\begin{aligned} [\mathbf{K}]_{i,j} &= \int_{\Omega} \omega_i^{\Omega}(\mathbf{r})\mathcal{L}(\varphi_j(\mathbf{r}))d\Omega + \int_{\Gamma} \omega_i^{\Gamma}(\mathbf{r})\ell(\varphi_j(\mathbf{r}))d\Gamma, \quad 1 \le i, j \le M, \\ [\mathbf{f}]_i &= \int_{\Omega} \omega_i^{\Omega}(\mathbf{r})k d\Omega + \int_{\Gamma} \omega_i^{\Gamma}(\mathbf{r})\, f d\Gamma, \quad 1 \le i \le M. \end{aligned} \tag{3.46}$$

Зауважимо, що подібний підхід можна застосовувати і для неоднорідних рівнянь де в правій частині замість констант k та f присутній деякий вираз типу $Q(\mathbf{r})$.

Процес апроксимації в змішаних методах зважених нев'язок зазвичай є практично набагато складнішим, ніж у внутрішніх методах. Основною проблемою є необхідність обчислення інтегралів по границям області, які можуть мати складні криволінійні форми чи інші ускладнюючі фактори. Проте, існує ряд задач для яких описану проблему можна значно спростити, а інколи і усунути повністю.

Для цього використовується, так звана, *слабка форма* рівняння [5], [10], при якій вихідний диференціальний оператор розбивається на кілька операторів з меншим порядком диференціювання. Наприклад скалярний добуток нев'язки

по області, зазвичай можна записати у вигляді:

$$\left\langle R(\mathbf{r}), \omega_i(\mathbf{r}) \right\rangle = \int_{\Omega} \omega_i(\mathbf{r}) \mathcal{L}(\varphi_j(\mathbf{r})) d\Omega = \\ = \int_{\Omega} \mathcal{E}(\omega_i(\mathbf{r})) \mathcal{D}(\varphi_j(\mathbf{r})) d\Omega + \int_{\Gamma} \omega_i(\mathbf{r}) \mathcal{Q}(\varphi_j(\mathbf{r})) d\Gamma. \tag{3.47}$$

де $\mathcal{E}$, $\mathcal{D}$ та $\mathcal{Q}$ – диференціальні оператори більш низького порядку, ніж вихідний диференціальний оператор $\mathcal{L}$. Після такого перетворення, при належним чином вибраних вагових функціях ω_i^{Ω} та ω_i^{Γ}, можна досягти того, що останній доданок з (3.47) та частина останнього доданку з (3.44) взаємно знищуються, завдяки чому, буде виключено інтеграл який містить пробну функцію чи її похідні вздовж границі області.

Така процедура припустима тільки для деяких крайових умов, що називають *природними* для даного рівняння (решту крайових умов називають *головними*) [4], [5]. У загальному випадку, застосування процедури для крайових умов, що включають тільки значення шуканого потенціалу на границях, не принесе корисних результатів, але подібний підхід може бути вигідним, коли на границі задані похідні шуканої функції, тобто умови Неймана.

Щоб отримати рівняння з більш низьким порядком диференціювання і фактично перенести частину диференціювання з відповідного оператору на повірочну функцію застосовують правило *інтегрування за частинами* [11] з подальшим застосуванням теореми Стокса чи її часткових випадків на відповідну кількість вимірів [2], [3]:

$$\int_a^b u dv = uv\Big|_a^b - \int_a^b v du, \tag{3.48}$$

або, правило *диференціювання добутку* [11] з подальшим застосуванням тієї ж теореми Стокса чи її часткових випадків:

$$(uv)' = u'v + uv'. \tag{3.49}$$

Наприклад, для однорідного еліптичного рівняння, що можна розписати як:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0. \tag{3.50}$$

Скалярний добуток повірочних функцій і нев'язки по області (3.45) можна розписати як:

$$a_j \iiint_{\Omega} \omega_i^{\Omega} \left[\frac{\partial^2 \varphi_j}{\partial x^2} + \frac{\partial^2 \varphi_j}{\partial y^2} + \frac{\partial^2 \varphi_j}{\partial z^2} \right] dxdydz. \tag{3.51}$$

Приймаючи $u = \omega_i^{\Omega}$, $v = d\varphi_j = \frac{\partial \varphi_j}{\partial x} + \frac{\partial \varphi_j}{\partial y} + \frac{\partial \varphi_j}{\partial z}$, $du = \frac{\partial \omega_i^{\Omega}}{\partial x} + \frac{\partial \omega_i^{\Omega}}{\partial y} + \frac{\partial \omega_i^{\Omega}}{\partial z}$,

$dv = \frac{\partial^2 \varphi_j}{\partial x^2} + \frac{\partial^2 \varphi_j}{\partial y^2} + \frac{\partial^2 \varphi_j}{\partial z^2}$, відповідно до правила інтегрування за частинами (3.48)

та застосувавши теорему Гріна, що є частковим випадком теореми Стокса, останній вираз можна розписати як:

$$\begin{aligned}&\iiint_{\Omega}\omega_i^{\Omega}\left[\frac{\partial^2\varphi_j}{\partial x^2}+\frac{\partial^2\varphi_j}{\partial y^2}+\frac{\partial^2\varphi_j}{\partial z^2}\right]dxdydz=\\&=\int_{\Gamma}\omega_i^{\Omega}\left[l_x\frac{\partial\varphi_j}{\partial x}+l_y\frac{\partial\varphi_j}{\partial y}+l_y\frac{\partial\varphi_j}{\partial z}\right]d\Gamma-\\&-\iiint_{\Omega}\left[\frac{\partial\omega_i^{\Omega}}{\partial x}\frac{\partial\varphi_j}{\partial x}+\frac{\partial\omega_i^{\Omega}}{\partial y}\frac{\partial\varphi_j}{\partial y}+\frac{\partial\omega_i^{\Omega}}{\partial z}\frac{\partial\varphi_j}{\partial z}\right]dxdydz,\end{aligned}\tag{3.52}$$

де, l_x, l_y, l_z – направляючі косинуси нормалі до границі Γ. Враховуючи, що:

$$\int_{\Gamma}\omega_i^{\Omega}\left[l_x\frac{\partial\varphi_j}{\partial x}+l_y\frac{\partial\varphi_j}{\partial y}+l_y\frac{\partial\varphi_j}{\partial z}\right]d\Gamma=\int_{\Gamma}\omega_i^{\Omega}\frac{\partial\varphi_j}{\partial\mathbf{n}}d\Gamma,\tag{3.53}$$

отримаємо:

$$\begin{aligned}&\iiint_{\Omega}\omega_i^{\Omega}\left[\frac{\partial^2\varphi_j}{\partial x^2}+\frac{\partial^2\varphi_j}{\partial y^2}+\frac{\partial^2\varphi_j}{\partial z^2}\right]dxdydz=\int_{\Gamma}\omega_i^{\Omega}\frac{\partial\varphi_j}{\partial\mathbf{n}}d\Gamma-\\&-\iiint_{\Omega}\left[\frac{\partial\omega_i^{\Omega}}{\partial x}\frac{\partial\varphi_j}{\partial x}+\frac{\partial\omega_i^{\Omega}}{\partial y}\frac{\partial\varphi_j}{\partial y}+\frac{\partial\omega_i^{\Omega}}{\partial z}\frac{\partial\varphi_j}{\partial z}\right]dxdydz.\end{aligned}\tag{3.54}$$

Для описаного рівняння, природними крайовими умовами є умови Неймана (3.36), для яких, розписавши скалярний добуток вагових функцій і нев'язки по границі отримаємо:

$$\int_{\Gamma}\omega_i^{\Gamma}\left[a_j\frac{\partial\varphi_j}{\partial\mathbf{n}}-q\right]d\Gamma=a_j\int_{\Gamma}\omega_i^{\Gamma}\frac{\partial\varphi_j}{\partial\mathbf{n}}d\Gamma-\int_{\Gamma}q\omega_i^{\Gamma}d\Gamma.\tag{3.55}$$

Приймаючи вагові функцій для нев'язок по області і по границі як $\omega_i^{\Omega}=-\omega_i^{\Gamma}$, суму скалярних добутків по всіх нев'язках (3.44) можна розписати як:

$$\begin{aligned}&a_j\iiint_{\Omega}\omega_i^{\Omega}\left[\frac{\partial^2\varphi_j}{\partial x^2}+\frac{\partial^2\varphi_j}{\partial y^2}+\frac{\partial^2\varphi_j}{\partial z^2}\right]dxdydz+a_j\int_{\Gamma}\omega_i^{\Gamma}\frac{\partial\varphi_j}{\partial\mathbf{n}}d\Gamma-\int_{\Gamma}q\omega_i^{\Gamma}d\Gamma=0\\&a_j\int_{\Gamma}\omega_i^{\Omega}\frac{\partial\varphi_j}{\partial\mathbf{n}}d\Gamma-a_j\iiint_{\Omega}\left[\frac{\partial\omega_i^{\Omega}}{\partial x}\frac{\partial\varphi_j}{\partial x}+\frac{\partial\omega_i^{\Omega}}{\partial y}\frac{\partial\varphi_j}{\partial y}+\frac{\partial\omega_i^{\Omega}}{\partial z}\frac{\partial\varphi_j}{\partial z}\right]dxdydz+\\&+a_j\int_{\Gamma}\omega_i^{\Gamma}\frac{\partial\varphi_j}{\partial\mathbf{n}}d\Gamma-\int_{\Gamma}q\omega_i^{\Gamma}d\Gamma=0,\\&a_j\int_{\Gamma}\omega_i^{\Omega}\frac{\partial\varphi_j}{\partial\mathbf{n}}d\Gamma-a_j\iiint_{\Omega}\left[\frac{\partial\omega_i^{\Omega}}{\partial x}\frac{\partial\varphi_j}{\partial x}+\frac{\partial\omega_i^{\Omega}}{\partial y}\frac{\partial\varphi_j}{\partial y}+\frac{\partial\omega_i^{\Omega}}{\partial z}\frac{\partial\varphi_j}{\partial z}\right]dxdydz-\\&-a_j\int_{\Gamma}\omega_i^{\Omega}\frac{\partial\varphi_j}{\partial\mathbf{n}}d\Gamma+\int_{\Gamma}q\omega_i^{\Omega}d\Gamma=0,\end{aligned}$$

$$a_j \iiint\limits_{\Omega} \left[\frac{\partial \omega_i^{\Omega}}{\partial x}\frac{\partial \varphi_j}{\partial x} + \frac{\partial \omega_i^{\Omega}}{\partial y}\frac{\partial \varphi_j}{\partial y} + \frac{\partial \omega_i^{\Omega}}{\partial z}\frac{\partial \varphi_j}{\partial z} \right] dxdydz = -\int\limits_{\Gamma} q\omega_i^{\Omega} d\Gamma. \qquad (3.56)$$

Таким чином, ми позбулися інтегралу, що включає похідну від шуканої функції по границі та понизили вимоги до порядку пробних і повірочних функцій.

Наведемо приклад використання описаних процедур для задачі стаціонарної теплопровідності, в двовимірному випадку. Нехай коефіцієнт теплопровідності матеріалу $\lambda = 1$ Вт/м°С, матеріал займає квадратну область $-1 \le x \le 1$м, $-1 \le y \le 1$м. На сторонах $y = \pm 1$ підтримується постійна температура 0°С, тоді як через сторони $x = \pm 1$ подається тепло з швидкістю $\cos(\pi y/2)$ Вт/м^2°С на одиницю довжини (*Рис. 3.3*). Запишемо відповідну крайову задачу:

$$\begin{cases} \lambda \nabla^2 T(x,y) = \lambda \dfrac{\partial^2 T(x,y)}{\partial x^2} + \lambda \dfrac{\partial^2 T(x,y)}{\partial y^2} = 0, \\ \dfrac{\partial T(-1,y)}{\partial \mathbf{n}} = \dfrac{\partial T(1,y)}{\partial \mathbf{n}} = \cos\left(\dfrac{\pi y}{2}\right), \\ T(x,-1) = T(x,1) = 0, \quad -1 \le x, y \le 1, \end{cases} \qquad (3.57)$$

або в операторній формі запису:

$$\begin{aligned} &\mathcal{L}(.) = \lambda \nabla^2 = \lambda \frac{\partial^2}{\partial x^2} + \lambda \frac{\partial^2}{\partial y^2}, \quad \mathcal{L}(T(x,y)) = 0, \\ &\ell_1(.) = \frac{\partial}{\partial \mathbf{n}}, \quad \ell_1(T(x,y))\big|_{\Gamma_q \equiv [x=\pm 1]} = \cos(\pi y/2), \\ &\ell_2(.) = 1, \qquad \ell_2(T(x,y))\big|_{\Gamma_T \equiv [y=\pm 1]} = 0, \\ &(x,y) \in \Omega \equiv [-1;1] \times [-1;1]. \end{aligned} \qquad (3.58)$$

Виберемо пробні функції так, щоб задовольнити крайову умову на Γ_T. Для цього використаємо систему функцій $\varphi_1 = 1 - y^2$, $\varphi_2 = (1-y^2)x^2$, $\varphi_3 = (1-y^2)y^2$, $\varphi_4 = (1-y^2)x^2y^2$, $\varphi_5 = (1-y^2)x^4$, і так далі. Очевидно, що при обраній системі, наприклад 5-ти елементна апроксимація:

$$\tilde{T}(x,y) = (1-y^2)(a_1 + a_2x^2 + a_3y^2 + a_4x^2y^2 + a_5x^4), \qquad (3.59)$$

буде задовольняти крайову умову на Γ_T, тобто $\tilde{T}(x=\pm 1) = 0$. Тоді згідно (3.47) отримаємо:

$$\int\limits_{-1}^{1}\int\limits_{-1}^{1} \omega_i^{\Omega}\left(\frac{\partial^2 \varphi_j}{\partial x^2} + \frac{\partial^2 \varphi_j}{\partial y^2}\right) dxdy + \int\limits_{\Gamma_q} \omega_i^{\Gamma}\left(\frac{\partial T}{\partial \mathbf{n}} - \cos\left(\frac{\pi y}{2}\right)\right) d\Gamma = 0,$$

$$\int\limits_{-1}^{1}\int\limits_{-1}^{1}\left(\frac{\partial \omega_i^{\Omega}}{\partial x}\frac{\partial \varphi_j}{\partial x} + \frac{\partial \omega_i^{\Omega}}{\partial y}\frac{\partial \varphi_j}{\partial y}\right) dxdy - \left[\int\limits_{\Gamma_T} \omega_i^{\Omega}\frac{\partial T}{\partial \mathbf{n}} d\Gamma + \int\limits_{\Gamma_q} \omega_i^{\Omega}\frac{\partial T}{\partial \mathbf{n}} d\Gamma\right] -$$

$$-\int_{\Gamma_q}\omega_i^{\Gamma}\left(\frac{\partial T}{\partial \mathbf{n}}-\cos\left(\frac{\pi y}{2}\right)\right)d\Gamma=0. \qquad (3.60)$$

Оскільки $\omega_i^{\Omega}=\varphi_j$ та $\varphi_j\big|_{\Gamma_T}=0$, то інтеграл по Γ_T перетворюється в 0. Отримане рівняння можна переписати:

$$\int_{-1}^{1}\int_{-1}^{1}\left(\frac{\partial\varphi_i}{\partial x}\frac{\partial\varphi_j}{\partial x}+\frac{\partial\varphi_i}{\partial y}\frac{\partial\varphi_j}{\partial y}\right)dxdy-\int_{\Gamma_q}\varphi_i\frac{\partial T}{\partial \mathbf{n}}d\Gamma- \\ -\int_{\Gamma_q}\omega_i^{\Gamma}\frac{\partial T}{\partial \mathbf{n}}d\Gamma+\int_{\Gamma_q}\omega_i^{\Gamma}\cos\left(\frac{\pi y}{2}\right)d\Gamma=0. \qquad (3.61)$$

Знову приймемо $\omega_i^{\Omega}=\varphi_i=-\omega_i^{\Gamma}$, звідки випливає, що крайова умова на Γ_q є природною для даного рівняння:

$$\int_{-1}^{1}\int_{-1}^{1}\left(\frac{\partial\varphi_i}{\partial x}\frac{\partial\varphi_j}{\partial x}+\frac{\partial\varphi_i}{\partial y}\frac{\partial\varphi_j}{\partial y}\right)dxdy=\int_{\Gamma_q}\varphi_i\cos\left(\frac{\pi y}{2}\right)d\Gamma. \qquad (3.62)$$

Підставляючи сюди обрану систему базисних функцій, отримаємо симетричну систему лінійних рівнянь:

$$[\mathbf{K}]_{i,j}=\int_{-1}^{1}\int_{-1}^{1}\left(\frac{\partial\varphi_i}{\partial x}\frac{\partial\varphi_j}{\partial x}+\frac{\partial\varphi_i}{\partial y}\frac{\partial\varphi_j}{\partial y}\right)dxdy,\quad 1\le i,j\le M, \\ [\mathbf{f}]_i=\int_{-1}^{1}\left[\varphi_i\cos\left(\frac{\pi y}{2}\right)\right]_{x=-1}dy+\int_{-1}^{1}\left[\varphi_i\cos\left(\frac{\pi y}{2}\right)\right]_{x=1}dy,\quad 1\le i\le M. \qquad (3.63)$$

Обчисливши елементи матричної системи для $M=5$, отримаємо:

$$[\mathbf{K}]=\begin{bmatrix}5{,}333333 & 1{,}777778 & 1{,}066667 & 0{,}355556 & 1{,}066667\\ 1{,}777778 & 3{,}911111 & 0{,}355556 & 0{,}619683 & 4{,}175238\\ 1{,}066667 & 0{,}355556 & 1{,}676190 & 0{,}558730 & 0{,}213333\\ 0{,}355556 & 0{,}619683 & 0{,}558730 & 0{,}470688 & 0{,}640000\\ 1{,}066667 & 4{,}175238 & 0{,}213333 & 0{,}640000 & 5{,}468783\end{bmatrix},\quad \{\mathbf{f}\}=\begin{Bmatrix}2{,}064098\\ 2{,}064098\\ 0{,}281921\\ 0{,}281921\\ 2{,}064098\end{Bmatrix}. \qquad (3.64)$$

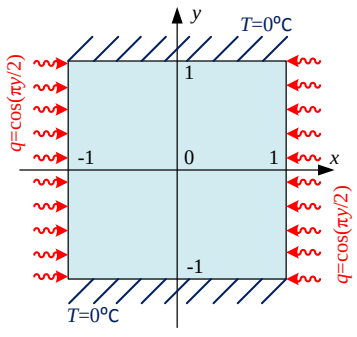


Рис. 3.3 Зображення умов двовимірної задачі стаціонарної теплопровідності

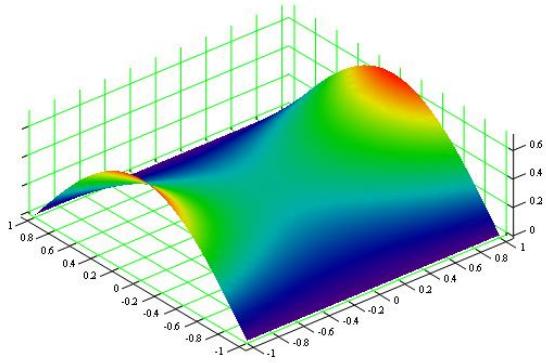


Рис. 3.4 Апроксимоване рішення задачі з допомогою методу Бубнова-Гальоркіна

Розв'язавши систему рівнянь, отримаємо вектор шуканих коефіцієнтів a_j:

$$\{\mathbf{a}\} = \{0{,}276308 \quad 0{,}339251 \quad -0{,}058746 \quad -0{,}092205 \quad 0{,}077615\}^{\mathbf{T}}. \qquad (3.65)$$

Апроксимоване рішення задачі показано на *Рис. 3.4*. На *Рис. 3.5* показано поступову збіжність отриманих апроксимованих результатів на прямих $x = \pm 1$ до природних крайових умов задачі.

Рис. 3.5 Порівняння значень похідних від температури по нормалі до границь $x = \pm 1$ для точного і апроксимованого рішення двовимірної задачі стаціонарної теплопровідності

- **Граничні методи зважених нев'язок**

У попередніх підрозділах було описано способи формулювання наближеного розв'язку крайових задач внутрішніми і змішаними методами зважених нев'язок, для яких система базисних функцій обиралася спираючись на визначені крайові умови. Очевидно що існує варіант вибору таких базисних функцій, що задовольняють не крайові умови, а саме диференціальне рівняння, тобто нев'язка по області $R^{\Omega}(\mathbf{r}, \tau) = 0$. Як вже було сказано, такий варіант вибору базисних функцій розглядається у підкласі граничних методів зважених нев'язок, що часто називають *методами граничних рішень*, чи *методами граничних елементів*.

Якщо вихідне диференціальне рівняння є лінійним, то описаний варіант може бути реалізований вибором базисних функцій, що самі є рішенням диференціального рівняння. Вибираючи систему функцій таким чином, припустимо:

$$u(\mathbf{r}) \approx \tilde{u}(\mathbf{r}) = \sum_{j=1}^{M} a_j \varphi_j(\mathbf{r}). \qquad (3.66)$$

Тоді рівняння методів зважених нев'язок (3.44) зводиться до відношення:

$$\left\langle R^{\Gamma}(\mathbf{r}), \omega_i^{\Gamma}(\mathbf{r}) \right\rangle = \int_{\Gamma} R^{\Gamma}(\mathbf{r}) \omega_i^{\Gamma}(\mathbf{r}) d\Gamma = 0, \quad i = 1, 2, \ldots, M, \qquad (3.67)$$

оскільки:

$$R^{\Omega}(\mathbf{r}) = \mathcal{L}(\tilde{u}(\mathbf{r})) = \sum_{j=1}^{M} a_j \mathcal{L}(\varphi_j(\mathbf{r})) = 0. \qquad (3.68)$$

Тепер необхідно визначити тільки систему вагових функцій ω_i^{Γ}, і при чому фактично тільки на границі Γ.

Основною перевагою граничних методів зважених нев'язок є надзвичайно швидка збіжність до точного рішення, основним недоліком – необхідність використання функцій, що відповідають рішенню вихідного рівняння, що можливо далеко не для всіх випадків [4].

Для диференціальних рівнянь більш загального виду вибір системи базисних і вагових функцій в граничних методах зважених нев'язок є менш очевидним. В загальному випадку можуть бути використані сингулярні функції типу функції Гріна[1], і тоді результуюча апроксимація записується у вигляді системи інтегральних рівнянь. До методів такого типу відносять так звані методи граничних інтегральних рівнянь.

3.5. Формулювання методу скінченних елементів

Реалізація попередньо описаних методів зважених нев'язок, за допомогою обчислювальної техніки, супроводжується рядом проблем, зокрема [4]:

- Для досягнення великої точності слід використовувати апроксимації з великою кількістю базисних функцій. Збільшення числа M базисних функцій веде до того, що елементи результуючих матриць систем лінійних рівнянь будуть мало відрізнятися один від одного, а враховуючи обмеженість розрядності чисел, якими оперує обчислювальна машина, така різниця взагалі може губитися в обчислювальній похибці. Це веде до того, що при великій кількості базисних функцій, в межах похибки може не здійснюватися умова лінійної незалежності системи базисних функцій, і як наслідок, неможливо буде отримати апроксимоване рішення задачі.
- При застосуванні таких методів зважених нев'язок, як методи Бубнова-Гальоркіна, навіть при тому, що результуючі матриці системи лінійних рівнянь будуть симетричними, вони будуть повністю заповнені коефіцієнтами. Знову ж таки, при великому числі M базисних функцій ми отримаємо систему, рішення якої шукається зі складністю $O(M^3)$, а застосування наближених рішень може бути ускладненим, оскільки матриця є повністю заповнена. Подібна складність обчислень стає критичною при рішенні нестаціонарних чи нелінійних задач.
- Попередня проблема автоматично веде до проблем з розміщенням елементів матриць в пам'яті обчислювальної машини та їх опрацюванням, що значно ускладнює програми, які реалізують обчислення.

[1] Функція Гріна $\mathcal{L}^{-1}$ це обернений оператор до диференціального оператора $\mathcal{L}$, що використовується для знаходження рішення диференціального рівняння [12].

- Щоб зменшити кількість обчислень і швидше отримати результати з задовільною точністю, бажано підбирати систему базисних функцій так, щоб вона автоматично задовольняла головні крайові умови задачі. Проте, такі процедури є очевидними тільки для простих просторових областей з границями, що є паралельні координатним осям. Задачі де фізичне явище розглядається в області складної форми, у таких випадках є на порядок складнішими.
- Нерідко задача описує фізичний процес, що характеризується великими градієнтами в малій частині області рішення та малими градієнтами в усіх інших її частинах, тому тут виникає питання ефективного застосування чи навіть доцільності використання великої кількості базисних функцій.

У попередньо описаних методах зважених нев'язок неявно передбачалося, що базисні функції $\varphi_j(\mathbf{r})$, які входять в розклад (3.17), були визначенні єдиним виразом на всій області задачі Ω, а інтеграли скалярних добутків типу (3.18) обчислювалися одразу по всій цій області, тобто шукана апроксимація була *глобальною*. Частковим винятком служив тільки метод підобластей, який був *локальним*, тобто передбачав пошук наближеного рішення на основі розбиття області Ω на ряд підобластей простої форми Ω_i, де були специфічним чином визначені повірочні функції $\omega_i(\mathbf{r})$.

В принципі, ідея пошуку рішення складної задачі на основі розбиття області на деякі підобласті, що при тому могли перекривати одна одну, вперше була запропонована ще до появи методу підобластей, а саме в 1870 році під назвою альтернуючий метод Шварца[1] [13], пізніше ми ще повернемося до неї, при розгляді методів декомпозиції обчислень.

Подібним до методу підобластей є і метод коллокацій, де рішення шукається глобально, але на основі визначення в області Ω ряду *вузлів* де відбуваються коллокації. В одновимірному випадку система цих вузлів фактично ділить область Ω на міжвузлові підобласті Ω_i. Аналогічно можна поступити і в багатовимірних випадках, тобто розбити складну область Ω на скінченний ряд простих Ω_i, що не перетинаються, при чому кожна з таких Ω_i будується як комбінація скінченної кількості вузлів. Такі підобласті, разом з визначеними для них пробними функціями, прийнято називати *елементами*, а процес розбиття неперервної області на елементи – *дискретизацією*.

Якщо застосовувати подібне формулювання до попередньо описаних методів зважених нев'язок, то їх можна розглядати відносно єдиного *суперелементу*, що охоплює всю область задачі Ω.

При застосуванні методів зважених нев'язок, до цього моменту, невідомими величинами були абстрактні коефіцієнти a_j розкладу наближеного

[1] Schwartz, H. – Über einen Grenzübergang durch alternierendes Verfahren // Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürichб 15:272–286, 1870.

рішення. Ці коефіцієнти не мають ніякого очевидного фізичного змісту. Проте використовуючи варіант поелементного розбиття області, пробне рішення типу (3.17) може бути задано як:

$$u(\mathbf{r},\tau) \approx \tilde{u}(\mathbf{r},\tau) = u_0(\mathbf{r},\tau) + \sum_{j=1}^{M} u_j(\tau)\varphi_j(\mathbf{r}), \tag{3.69}$$

де, $u_j(\tau)$ – так зване, вузлове значення шуканого потенціалу $u(\mathbf{r},\tau)$. Очевидно, що $u_j(\tau)$ тепер мають прямий фізичний зміст. Більше того, у такому формулюванні пробні функції $\varphi_j(\mathbf{r})$ тепер мають обов'язкову інтерполяційну природу та відповідну *інтерполяційну похибку*. Це означає, що аналогічно до вагових функцій у методі підобластей (3.24), кожна базисна функція $\varphi_j(\mathbf{r}) = 1$ у вузлі під номером j, та $\varphi_j(\mathbf{r}) = 0$ в інших вузлах, але не між вузлами, тобто:

$$\varphi_j(\mathbf{r}) = \begin{cases} 1, & \mathbf{r} = \mathbf{r}_j, \\ 0, & \mathbf{r} = \mathbf{r}_i, \end{cases} \quad i \neq j. \tag{3.70}$$

Для кожного з елементів, це означає, що:

$$\sum_{j=1}^{M} \varphi_j(\mathbf{r}) = 1, \quad \mathbf{r} \in \Omega_i, \tag{3.71}$$

де, тепер M крім кількості пробних функцій позначає кількість вузлів для кожного елементу Ω_i. Тобто пробні функції є *кусково-визначеними* і відмінними від нуля тільки в деякій невеликій підобласті всієї області визначення задачі. Це еквівалентно тому, що система функцій для елементу має, так званий, *скінченний носій*, а відповідні елементи називають *скінченними елементами*. Також зауважимо, що навіть при тому, що значення шуканого потенціалу у вузлах повинно прямувати до точного, ця умова ніяк не поширюється на значення похідних від шуканого потенціалу.

Скінченність носія дає велику перевагу при розв'язку результуючих систем лінійних рівнянь, оскільки завдяки тому, що функції визначені тільки в невеликій підобласті, матриці цих систем стають сильно розрідженими, тобто містять велику кількість нульових коефіцієнтів. Після деяких нескладних маніпуляцій такі матриці можна звести до стрічкового виду, коли ненульові коефіцієнти будуть розміщуватися недалеко від діагоналі. Це дає змогу значно скоротити розміри машинної пам'яті, необхідної для зберігання коефіцієнтів, а також можливість розв'язку систем рівнянь наближеними методами, зі складністю, що менша за $O(M^3)$.

Крім того, завдяки використанню розбиття області визначення задачі на множину підобластей, відкривається широке коло можливостей локального контролю деталізації апроксимації, наприклад в зонах де присутній великий градієнт шуканої функції, кількість елементів можна збільшити, а в зонах де градієнт відсутній – навпаки зменшити.

Іншою важливою можливістю, що відкривається при використанні розбиття на скінченні елементи, є можливість розглядати рівняння зі змінними

коефіцієнтами, тобто середовища, що мають різні властивості в залежності від координат, наприклад об'єднання різних матеріалів.

Мінімальність скінченного носія функцій, тобто переважне використання поліномів низького порядку в сукупності з використанням скінченних елементів примітивної форми, наприклад прямокутників з осями паралельними координатним, дає значну перевагу в складності та часі обчислення інтегралів з рівняння методу зважених нев'язок (3.18). Особливо це відчутно при розв'язку складних задач, де скалярні добутки нев'язок і вагових функцій повинні визначатися з допомогою чисельного інтегрування.

Для прикладу розглянемо ситуацію, що зображена на *Рис. 3.6*, де невідома функція $u(x)$ апроксимується за допомогою набору лінійних кусково-визначених базисних функцій у вигляді "пірамідок", для кожної з яких справедливо:

$$\varphi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}}, & x_{j-1} \le x \le x_j, \\ \dfrac{x_{j+1} - x}{x_{j+1} - x_j}, & x_j \le x \le x_{j+1}, \\ 0, & x < x_{j-1} \vee x > x_{j+1}. \end{cases} \tag{3.72}$$

Тобто, кожна з функцій $\varphi_j(x)$ визначена тільки на ділянках $[x_{j-1}, x_{j+1}]$ і рівна нулю для всіх інших значень x. Легко перевірити, що $\varphi_j(x)$ відповідає критеріям (3.70) та (3.71). Наближене рішення $\tilde{u}(x)$ шукається на основі розкладу:

$$\tilde{u}(x)\big|_{x \in [x_j, x_{j+1}]} = \sum_{k=j}^{j+1} u_k \varphi_k(x) = u_j \varphi_j(x) + u_{j+1} \varphi_{j+1}(x). \tag{3.73}$$

Сума таких розкладів по кожному з елементів лінійно апроксимує невідому функцію $\tilde{u}(x)$ по всій області її визначення, за умови, що обрана система базисних функцій взагалі здатна це зробити.

Оскільки метод скінченних елементів базується на методах Бубнова-Гальоркіна, і пробні і повірочні функції тут вибираються з одного і того ж сімейства поліномів низького порядку. В літературі ці функції прийнято позначати як $N(\mathbf{r})$ і називати *функціями форми* (скінченного елементу що є підобластю дискретизації) або *інтерполяційними функціями* [2], [3], [4], [14], [15], [16]. Як і раніше, функції форми повинні бути лінійно незалежними і по можливості задовольняти початкові та крайові умови задачі.

Розглянемо приклад використання методу скінченних елементів для еліптичних рівнянь, що визначені у багатовимірному просторі. Координати будемо позначати не як $x, y, \ldots$, а як $x_1, x_2, \ldots$. Нехай в деякій області Ω, з межами Γ, необхідно вирішити крайову задачу

$$\mathcal{L}(u(\mathbf{r})) = k, \quad \mathbf{r} \in \Omega, \quad \ell(u(\mathbf{r}))\big|_{\Gamma} = f, \tag{3.74}$$

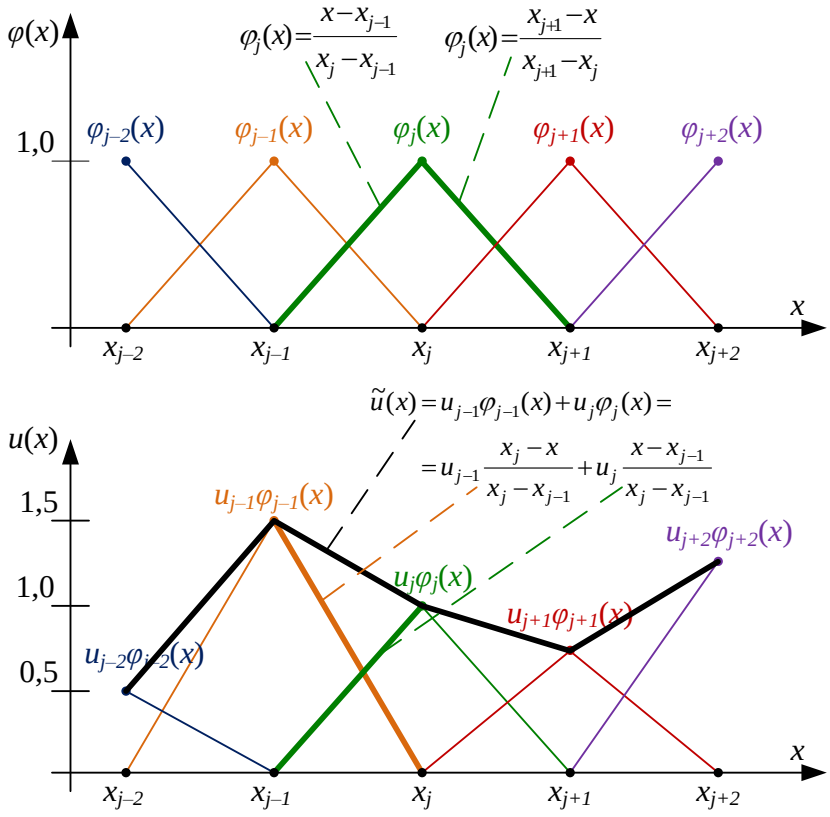


Рис. 3.6 Приклад апроксимації невідомої функції $u(x)$ *за допомогою набору лінійних кусково-визначених базисних функцій*

де, u, k, f – шукана, та задані функції, $\mathcal{L}$, ℓ – диференційні оператори, що визначають вхідне рівняння та крайові умови. Наприклад, це можуть бути головні крайові умови Діріхле і природні крайові умови Неймана:

$$\ell_1(u(\mathbf{r}))=u(\mathbf{r}),\quad f_1=u_\infty\quad \mathbf{r}\in\Gamma_u\quad\Rightarrow\quad u(\mathbf{r})\big|_{\Gamma_u}=u_\infty,$$
$$\ell_2(u(\mathbf{r}))=\frac{\partial u(\mathbf{r})}{\partial \mathbf{n}},\quad f_2=q\quad \mathbf{r}\in\Gamma_q\quad\Rightarrow\quad \frac{\partial u(\mathbf{r})}{\partial \mathbf{n}}\bigg|_{\Gamma_q}=q. \tag{3.75}$$

Розіб'ємо Ω на P підобластей, що не перетинаються: $\Omega=\bigcup_{i=1}^{P}\Omega_i$. Межі кожної з підобластей позначимо як $\Gamma_{i,i'}$. Вхідній задачі (3.74) поставимо у відповідність сукупність допоміжних крайових задач в підобластях, якщо такі необхідні:

$$\mathcal{L}(u_i(\mathbf{r}))=k_i(\mathbf{r}),\quad \mathbf{r}\in\Omega_i,\quad \ell_{i,i'}(u_i(\mathbf{r}))\big|_{\Gamma_{i,i'}}=f_{i,i'}(\mathbf{r})\equiv\ell_{i',i}(u_{i'}(\mathbf{r}))\big|_{\Gamma_{i',i}},$$
$$i'\in\psi_i,\quad i=1,2,...,P, \tag{3.76}$$

де, ψ_i – сукупність номерів підобластей $\Omega_{i'}$. На зовнішній межі Γ_i ставляться задані крайові умови вхідної задачі. Вважаємо що рішення задач (3.74) та (3.76) існують, єдині та співпадають.

Припустимо, що кожна невідома функція u_i може бути достатньо точно апроксимована з допомогою наближеного рішення:

$$u_i(\mathbf{r}) \approx \tilde{u}_i(\mathbf{r}) = u_{i,0} + \sum_{j=1}^{M} u_{i,j} N_{i,j}(\mathbf{r}), \tag{3.77}$$

де, $N_{i,j}(\mathbf{r})$ – відомі аналітичні базисні функції (функції форми), $u_{i,j}$ – вузлові коефіцієнти, які необхідно знайти. Початкове значення $u_{i,0}$ приймемо рівним нулю, процес включення головних крайових умов буде показано окремо. Підставивши (3.77) в (3.74) отримаємо відмінні від нуля нев'язки:

$$\begin{aligned} R_i^{\Omega_i}(\mathbf{r}) &= \mathcal{L}(\tilde{u}_i(\mathbf{r})) = \sum_{j=1}^{M} \mathcal{L}(N_{i,j}(\mathbf{r})) u_{i,j} \neq 0, \\ R_i^{\Gamma_i}(\mathbf{r}) &= \ell(\tilde{u}_i(\mathbf{r})) = \sum_{j=1}^{M} \ell(N_{i,j}(\mathbf{r})) u_{i,j} \neq 0. \end{aligned} \tag{3.78}$$

Щоб здійснити апроксимацію, ставимо умову ортогональності нев'язки до вагових функцій, де $N_i(\mathbf{r})$ – вагова функція для нев'язки по області та $-N_i(\mathbf{r})$ – вагова функція для нев'язки по границі:

$$\begin{aligned} &\left\langle R_i^{\Omega_i}(\mathbf{r}), N_i(\mathbf{r}) \right\rangle + \left\langle R_i^{\Gamma_i}(\mathbf{r}), -N_i(\mathbf{r}) \right\rangle = \\ &= \int_{\Omega_i} R_i^{\Omega_i}(\mathbf{r}), N_i(\mathbf{r}) d\Omega - \int_{\Gamma_i} R_i^{\Gamma_i}(\mathbf{r}), N_i(\mathbf{r}) d\Gamma = 0. \end{aligned} \tag{3.79}$$

Загальна задача у такому випадку отримується від суми інтегралів кожної з підобластей:

$$\begin{aligned} \int_{\Omega} R(\mathbf{r}) N(\mathbf{r}) d\Omega &= \sum_{i=1}^{P} \int_{\Omega_i} R_i^{\Omega_i}(\mathbf{r}) N_i(\mathbf{r}) d\Omega, \\ \int_{\Gamma} R(\mathbf{r}) N(\mathbf{r}) d\Gamma &= \sum_{i=1}^{P} \int_{\Gamma_i} R_i^{\Gamma_i}(\mathbf{r}) N_i(\mathbf{r}) d\Gamma. \end{aligned} \tag{3.80}$$

Враховуючи природні крайові умови задачі, та те, що $\nabla \tilde{u}(\mathbf{r}) = \nabla \left(\sum_{j=1}^{M} u_{i,j} N_{i,j}(\mathbf{r}) \right)$, запишемо слабку форму рівняння:

$$\left(\int_{\Omega_i} \nabla N_g(\mathbf{r}) \nabla N_j(\mathbf{r}) d\Omega \right) u_{i,j} = \int_{\Gamma_i} f N_g(\mathbf{r}), d\Gamma, \quad g, j = 1, 2, \ldots M. \tag{3.81}$$

Перепишемо результати в матричну форму і сформуємо систему лінійних алгебраїчних рівнянь. Апроксимацію (3.77) запишемо як:

$$u_i(\mathbf{r}) \approx \tilde{u}_i(\mathbf{r}) = \sum_{j=1}^{M} u_{i,j} N_{i,j}(\mathbf{r}) =$$

$$= \begin{bmatrix} N_{i,1}(\mathbf{r}) & N_{i,2}(\mathbf{r}) & \cdots & N_{i,M}(\mathbf{r}) \end{bmatrix} \begin{Bmatrix} u_{i,1} \\ u_{i,2} \\ \vdots \\ u_{i,M} \end{Bmatrix} = [\mathbf{N}]_i \{\mathbf{u}\}_i. \qquad (3.82)$$

Знайдемо першу похідну від наближеного розв'язку $\tilde{u}(\mathbf{r})$ по всім просторовим координатам, тобто градієнт $\nabla[\mathbf{N}]\{\mathbf{u}\}$, для спрощення матрицю градієнтів позначимо як $[\mathbf{B}] = \nabla[\mathbf{N}]$:

$$\nabla \tilde{u}_i = \begin{Bmatrix} \frac{\partial \tilde{u}_i}{\partial x_1} \\ \frac{\partial \tilde{u}_i}{\partial x_2} \\ \frac{\partial \tilde{u}_i}{\partial x_3} \end{Bmatrix} = \begin{Bmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \frac{\partial}{\partial x_3} \end{Bmatrix} \cdot \begin{bmatrix} N_{i,1} & N_{i,2} & \cdots & N_{i,M} \end{bmatrix} \cdot \begin{Bmatrix} u_{i,1} \\ u_{i,2} \\ \vdots \\ u_{i,M} \end{Bmatrix} =$$

$$= \begin{bmatrix} \frac{\partial N_{i,1}}{\partial x_1} & \frac{\partial N_{i,2}}{\partial x_1} & \cdots & \frac{\partial N_{i,M}}{\partial x_1} \\ \frac{\partial N_{i,1}}{\partial x_2} & \frac{\partial N_{i,2}}{\partial x_2} & \cdots & \frac{\partial N_{i,M}}{\partial x_2} \\ \frac{\partial N_{i,1}}{\partial x_3} & \frac{\partial N_{i,2}}{\partial x_3} & \cdots & \frac{\partial N_{i,M}}{\partial x_3} \end{bmatrix} \cdot \begin{Bmatrix} u_{i,1} \\ u_{i,2} \\ \vdots \\ u_{i,M} \end{Bmatrix} = [\mathbf{B}]_i \{\mathbf{u}\}_i. \qquad (3.83)$$

Еліптичний оператор $\mathcal{L}(.)$ може містити коефіцієнт пропорційності, наприклад коефіцієнт теплопровідності λ, чи діяти з більш складним варіантом у вигляді тензору $[\mathbf{D}]$ для лінійного тензору напружень з задач теорії пружності. Тому, в загальному випадку ми будемо його враховувати при обчисленні інтегралів по кожному з елементів. Таким чином, рівняння (3.81) у матричній формі можна переписати як:

$$\left(\int_{\Omega_i} \nabla N_g(\mathbf{r}) \nabla N_j(\mathbf{r}) d\Omega \right) u_{i,j} = \left(\int_{\Omega_i} [\mathbf{B}]_i^{\mathrm{T}} [\mathbf{D}]_i [\mathbf{B}]_i \, d\Omega \right) \{\mathbf{u}\}_i = [\mathbf{K}]_i \{\mathbf{u}\}_i$$
$$\int_{\Gamma_i} f N_g(\mathbf{r}), d\Gamma = \int_{\Gamma_i} [\mathbf{N}]_i^{\mathrm{T}} f_i d\Gamma = \{\mathbf{f}\}_i. \qquad (3.84)$$

Суму інтегралів кожного з елементів (3.80) у такому випадку можна записати як систему лінійних алгебраїчних рівнянь типу (3.19):

$$[\mathbf{K}]\{\mathbf{u}\} = \{\mathbf{f}\}, \qquad (3.85)$$

де:

$$[\mathbf{K}] = \sum_{i=1}^{P} [\mathbf{K}]_i = \sum_{i=1}^{P} \int_{\Omega_i} [\mathbf{B}]_i^{\mathrm{T}} [\mathbf{D}]_i [\mathbf{B}]_i \, d\Omega,$$

$$\{\mathbf{f}\}=\sum_{i=1}^{P}\{\mathbf{f}\}_i=\sum_{i=1}^{P}\int_{\Gamma_i}[\mathbf{N}]_i^{\mathrm{T}}\,f_i d\Gamma. \tag{3.86}$$

Глобальний вектор шуканих вузлових значень $\{\mathbf{u}\}$ будується на основі об'єднання локальних векторів $\{\mathbf{u}\}_i$, з врахуванням тієї особливості, що сусідні скінченні елементи мають спільні вузли. Процес знаходження суми інтегралів у (3.86), тобто процес побудови глобальної матриці жорсткості $[\mathbf{K}]$, вектору навантажень $\{\mathbf{f}\}$ та вектору шуканих вузлових значень $\{\mathbf{u}\}$ називається *ансамблюванням* [2], [3], [15]. Розмір матриці та векторів відповідає кількості вузлів. Якщо, для прикладу, в якості функцій форми обрано функції типу (3.72), то кожен елемент, що відповідає відрізку $[x_j, x_{j+1}]$, обмеженому двома вузлами j та $j+1$, буде вносити вклад у глобальну матрицю так, як це показано на *Рис. 3.7*. Опустивши всі нульові коефіцієнти отримаємо локальну матрицю 2×2:

$$[\mathbf{K}]_i=\int_{x_i}^{x_{i+1}}\begin{bmatrix}\dfrac{\partial N_{i,1}}{\partial x}\\[2mm]\dfrac{\partial N_{i,2}}{\partial x}\end{bmatrix}[D]_i\begin{bmatrix}\dfrac{\partial N_{i,1}}{\partial x} & \dfrac{\partial N_{i,2}}{\partial x}\end{bmatrix}dx=\begin{bmatrix}\displaystyle\int_{x_i}^{x_{i+1}}\frac{\partial N_{i,1}}{\partial x}D_i\frac{\partial N_{i,1}}{\partial x}dx & \displaystyle\int_{x_i}^{x_{i+1}}\frac{\partial N_{i,1}}{\partial x}D_i\frac{\partial N_{i,2}}{\partial x}dx\\[2mm]\displaystyle\int_{x_i}^{x_2}\frac{\partial N_{i,2}}{\partial x}D_i\frac{\partial N_{i,1}}{\partial x}dx & \displaystyle\int_{x_i}^{x_2}\frac{\partial N_{i,2}}{\partial x}D_i\frac{\partial N_{i,2}}{\partial x}dx\end{bmatrix}. \tag{3.87}$$

а глобальна матриця $[\mathbf{K}]$, наприклад для трьох елементів будується у відповідності з глобальною індексацією вузлів як:

$$[\mathbf{K}]=\begin{bmatrix}\displaystyle\int_{x_1}^{x_2}\frac{\partial N_{1,1}}{\partial x}D_1\frac{\partial N_{1,1}}{\partial x}dx & \displaystyle\int_{x_1}^{x_2}\frac{\partial N_{1,1}}{\partial x}D_1\frac{\partial N_{1,2}}{\partial x}dx & 0 & 0\\[2mm]\displaystyle\int_{x_1}^{x_2}\frac{\partial N_{1,2}}{\partial x}D_1\frac{\partial N_{1,1}}{\partial x}dx & \displaystyle\int_{x_1}^{x_2}\frac{\partial N_{1,2}}{\partial x}D_1\frac{\partial N_{1,2}}{\partial x}dx & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}+$$

$$+\begin{bmatrix}0 & 0 & 0 & 0\\ 0 & \displaystyle\int_{x_2}^{x_3}\frac{\partial N_{2,1}}{\partial x}D_2\frac{\partial N_{2,1}}{\partial x}dx & \displaystyle\int_{x_2}^{x_3}\frac{\partial N_{2,1}}{\partial x}D_2\frac{\partial N_{2,2}}{\partial x}dx & 0\\ 0 & \displaystyle\int_{x_2}^{x_3}\frac{\partial N_{2,2}}{\partial x}D_2\frac{\partial N_{2,1}}{\partial x}dx & \displaystyle\int_{x_2}^{x_3}\frac{\partial N_{2,2}}{\partial x}D_2\frac{\partial N_{2,2}}{\partial x}dx & 0\\ 0 & 0 & 0 & 0\end{bmatrix}+$$

$$+\begin{bmatrix}0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & \displaystyle\int_{x_3}^{x_4}\frac{\partial N_{3,1}}{\partial x}D_3\frac{\partial N_{3,1}}{\partial x}dx & \displaystyle\int_{x_3}^{x_4}\frac{\partial N_{3,1}}{\partial x}D_3\frac{\partial N_{3,2}}{\partial x}dx\\ 0 & 0 & \displaystyle\int_{x_3}^{x_4}\frac{\partial N_{3,2}}{\partial x}D_3\frac{\partial N_{3,1}}{\partial x}dx & \displaystyle\int_{x_3}^{x_4}\frac{\partial N_{3,2}}{\partial x}D_3\frac{\partial N_{3,2}}{\partial x}dx\end{bmatrix}=$$

$$= \begin{bmatrix} \int_{x_1}^{x_2} \frac{\partial N_{1,1}}{\partial x} D_1 \frac{\partial N_{1,1}}{\partial x} dx & \int_{x_1}^{x_2} \frac{\partial N_{1,1}}{\partial x} D_1 \frac{\partial N_{1,2}}{\partial x} dx & 0 & 0 \\ \int_{x_1}^{x_2} \frac{\partial N_{1,2}}{\partial x} D_1 \frac{\partial N_{1,1}}{\partial x} dx & \left(\int_{x_1}^{x_2} \frac{\partial N_{1,2}}{\partial x} D_1 \frac{\partial N_{1,2}}{\partial x} dx + \int_{x_2}^{x_3} \frac{\partial N_{2,1}}{\partial x} D_2 \frac{\partial N_{2,1}}{\partial x} dx \right) & \int_{x_2}^{x_3} \frac{\partial N_{2,1}}{\partial x} D_2 \frac{\partial N_{2,2}}{\partial x} dx & 0 \\ 0 & \int_{x_2}^{x_3} \frac{\partial N_{2,2}}{\partial x} D_2 \frac{\partial N_{2,1}}{\partial x} dx & \left(\int_{x_2}^{x_3} \frac{\partial N_{2,2}}{\partial x} D_2 \frac{\partial N_{2,2}}{\partial x} dx + \int_{x_3}^{x_4} \frac{\partial N_{3,1}}{\partial x} D_3 \frac{\partial N_{3,1}}{\partial x} dx \right) & \int_{x_3}^{x_4} \frac{\partial N_{3,1}}{\partial x} D_3 \frac{\partial N_{3,2}}{\partial x} dx \\ 0 & 0 & \int_{x_3}^{x_4} \frac{\partial N_{3,2}}{\partial x} D_3 \frac{\partial N_{3,1}}{\partial x} dx & \int_{x_3}^{x_4} \frac{\partial N_{3,2}}{\partial x} D_3 \frac{\partial N_{3,2}}{\partial x} dx \end{bmatrix}. \quad (3.88)$$

І відповідно вектор навантажень $\{\mathbf{f}\}$:

$$\{\mathbf{f}\} = \begin{Bmatrix} \int_{x_1}^{x_2} N_{1,1} f_1 dx \\ \int_{x_1}^{x_2} N_{1,2} f_1 dx \\ 0 \\ 0 \end{Bmatrix} + \begin{Bmatrix} 0 \\ \int_{x_2}^{x_3} N_{2,1} f_2 dx \\ \int_{x_2}^{x_3} N_{2,2} f_2 dx \\ 0 \end{Bmatrix} + \begin{Bmatrix} 0 \\ 0 \\ \int_{x_3}^{x_4} N_{3,1} f_3 dx \\ \int_{x_3}^{x_4} N_{3,2} f_3 dx \end{Bmatrix} = \begin{Bmatrix} \int_{x_1}^{x_2} N_{1,1} f_1 dx \\ \int_{x_1}^{x_2} N_{1,2} f_1 dx + \int_{x_2}^{x_3} N_{2,1} f_2 dx \\ \int_{x_2}^{x_3} N_{2,2} f_2 dx + \int_{x_3}^{x_4} N_{3,1} f_3 dx \\ \int_{x_3}^{x_4} N_{3,2} f_3 dx \end{Bmatrix}. \quad (3.89)$$

Як видно, отримана матриця є симетричною і стрічковою.

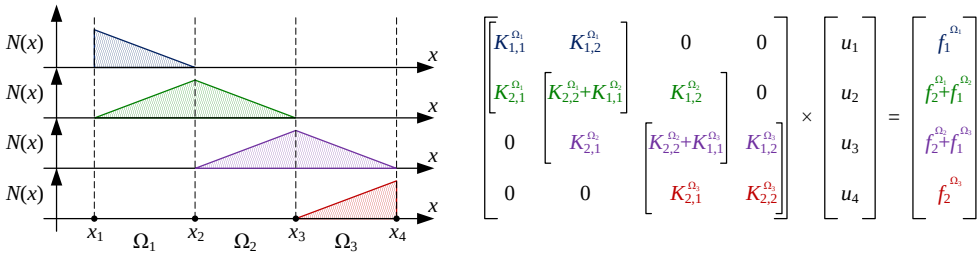


Рис. 3.7 Вклад кожного скінченного елементу в глобальну матрицю жорсткості

Значення вектору навантажень з виразу (3.89) в конкретних задачах відмінні від нуля тільки для елементів, вузли яких розміщені на границях, тобто для елементів де дійсно вказані природні крайові умови, оскільки згідно визначення $\partial u(\mathbf{r})/\partial \mathbf{n} = f$ тоді і тільки тоді, коли $\mathbf{r} \in \Gamma_q$, в усіх інших відповідні інтеграли обертаються в нуль.

Залишається врахувати крайові умови Діріхле, і ми отримаємо придатну для розв'язку систему рівнянь. Нагадаємо, що крайові умови Діріхле вказують значення потенціалу на границі області. Їх можна було б врахувати як і раніше, при побудові розкладу наближеного рішення (3.69) за допомогою $u_{i,0}$, але

оскільки згідно розкладу невідомими коефіцієнтами є значення потенціалу у вузлах, задання крайових умов Діріхле фактично відповідає заданню значень у вузлах, що розміщені на відповідній границі, а міжвузлові значення потенціалу інтерполюються функціями форми скінченних елементів. Таким чином, стають відомими деякі невідомі вектору вузлових значень $\{\mathbf{u}\}$, і щоб їх врахувати необхідно лише певним чином модифікувати матрицю жорсткості та вектор навантаження.

Існує кілька способів це зробити, найбільш популярним з яких є наступна процедура модифікації локальної матриці жорсткості та вектору навантажень [2], [15], [16]:

- Припустимо, нам відомо значення $u_j = u_\infty$. Віднімемо від кожного елементу локального вектору навантаження, добуток відомого значення потенціалу та коефіцієнту локальної матриці жорсткості у відповідному рядку і j-му стовбці, тобто:
$$\{\mathbf{f}\}_{i,k} = \{\mathbf{f}\}_{i,k} - u_{i,j}[\mathbf{K}]_{i,k,j}; \tag{3.90}$$
- Прирівняємо всі елементи j-го рядка локальної матриці жорсткості $[\mathbf{K}]_i$ до нуля, крім діагонального елементу, тобто елементу в j-му стовпці;
- Прирівняємо всі елементи j-го стовпця локальної матриці жорсткості $[\mathbf{K}]_i$ до нуля, крім діагонального елементу, тобто елементу в j-му рядку;
- Присвоїмо j-му елементу локального вектору навантаження значення $u_{i,j}[\mathbf{K}]_{i,j,j}$.

Наприклад, деяка локальна система рівнянь задана як:

$$\begin{bmatrix} 175 & -75 & -150 & 0 & 50 \\ -75 & 75 & 150 & -150 & 0 \\ -150 & 150 & 300 & 150 & -450 \\ 0 & -150 & 150 & 0 & 0 \\ 50 & 0 & -450 & 0 & 400 \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{Bmatrix} = \begin{Bmatrix} 250 \\ 0 \\ 420 \\ 280 \\ 0 \end{Bmatrix}. \tag{3.91}$$

Оскільки крайові умови ще не враховані, не виконується умова існування та єдності рішення. Можна перевірити, що на даному етапі матриця є виродженою і система немає розв'язків. Нехай відомо, що $u_2 = 2$ та $u_5 = 6$. Знайдемо значення (3.90) для u_2:

$$\begin{Bmatrix} 250 \\ 0 \\ 420 \\ 280 \\ 0 \end{Bmatrix} - 2 \begin{Bmatrix} -75 \\ 75 \\ 150 \\ -150 \\ 0 \end{Bmatrix} = \begin{Bmatrix} 400 \\ -150 \\ 120 \\ 580 \\ 0 \end{Bmatrix}. \tag{3.92}$$

Модифікуємо матрицю жорсткості і вектор навантаження:

$$\left[\begin{array}{c|c|ccc} 175 & 0 & -150 & 0 & 50 \\ \hline 0 & 75 & 0 & 0 & 0 \\ \hline -150 & 0 & 300 & 150 & -450 \\ 0 & 0 & 150 & 0 & 0 \\ 50 & 0 & -450 & 0 & 400 \end{array}\right]\left\{\begin{array}{c} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{array}\right\}=\left\{\begin{array}{c} 400 \\ \hline 150 \\ \hline 120 \\ 580 \\ 0 \end{array}\right\}. \qquad (3.93)$$

Виконаємо аналогічні дії для u_5:

$$\left\{\begin{array}{c} 400 \\ 150 \\ 120 \\ 580 \\ 0 \end{array}\right\}-6\left\{\begin{array}{r} 50 \\ 0 \\ -450 \\ 0 \\ 400 \end{array}\right\}=\left\{\begin{array}{r} 100 \\ 150 \\ 2820 \\ 580 \\ -2400 \end{array}\right\}, \qquad (3.94)$$

$$\left[\begin{array}{rrrr|c} 175 & 0 & -150 & 0 & 0 \\ 0 & 75 & 0 & 0 & 0 \\ -150 & 0 & 300 & 150 & 0 \\ 0 & 0 & 150 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 400 \end{array}\right]\left\{\begin{array}{c} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{array}\right\}=\left\{\begin{array}{r} 100 \\ 150 \\ 2820 \\ 300 \\ \hline 2400 \end{array}\right\}. \qquad (3.95)$$

Тепер, коли всі крайові умови враховані, система має розв'язок:

$$\{\mathbf{u}\}=\{3{,}885714 \quad 2{,}000000 \quad 3{,}866667 \quad 14{,}952381 \quad 6{,}000000\}^{\mathbf{T}}. \qquad (3.96)$$

Як і у випадку природних крайових умов, описану процедуру необхідно застосовувати тільки для елементів, вузли яких розміщені на границі де задані крайові умови Діріхле. Після завершення процедури, локальна матриця та вектори можуть бути використана в процесі ансамблювання.

Конкретний приклад застосування описаного методу скінченних елементів до еліптичних рівнянь розглянемо пізніше. А поки що, щоб краще зрозуміти процес ансамблювання, застосуємо описаний метод до розв'язку звичайного однорідного диференціального рівняння:

$$d^2y(x)/dx^2 - y(x) = 0, \quad y(0) = 0, \quad y(1) = 1, \quad 0 \le x \le 1. \qquad (3.97)$$

Щоб мати можливість порівняти результати, спочатку знайдемо аналітичне рішення задачі:

$$d^2y(x)/dx^2 - y(x) = 0 \;\Rightarrow\; \lambda^2 - 1 = 0, \quad D = B^2 - 4AC = 0 - 4\cdot 1\cdot(-1) = 4,$$

$$\lambda_1 = (-B + \sqrt{D})/2A = (0+2)/2 = 1 \quad \lambda_2 = (-B - \sqrt{D})/2A = (0-2)/2 = -1,$$

$$y(x) = C_1 e^{\lambda_1 x} + C_2 e^{\lambda_2 x} = C_1 e^{x} + C_2 e^{-x},$$

$$\begin{cases} y(0) = 0, \\ y(1) = 1, \end{cases} \Rightarrow \begin{cases} C_1 e^0 + C_2 e^0 = 0, \\ C_1 e^1 + C_2 e^{-1} = 1, \end{cases}$$

$$\begin{bmatrix} 1 & 1 \\ e & 1/e \end{bmatrix}\begin{Bmatrix} C_1 \\ C_2 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} \;\Rightarrow\; \begin{Bmatrix} C_1 \\ C_2 \end{Bmatrix} = \begin{Bmatrix} e/(e^2-1) \\ -e/(e^2-1) \end{Bmatrix},$$

$$y(x) = e \cdot e^{x} / (e^{2} - 1) - e \cdot e^{-x} / (e^{2} - 1) = (e^{1+x} - e^{1-x}) / (e^{2} - 1). \qquad (3.98)$$

Розіб'ємо відрізок $0 \le x \le 1$ на три елементи різної довжини, наприклад $[0, 1/4]$, $[1/4, 3/5]$ та $[3/5, 1]$, пронумеруємо їх як $\Omega_i = [X_{i,1}, X_{i,2}]$, де $i = 1, 2, 3$. Для кожного елементу побудуємо розклад наближеного рішення як:

$$y_i(x) \approx \tilde{y}_i(x) = \begin{bmatrix} N_{i,1}(x) & N_{i,2}(x) \end{bmatrix} \begin{Bmatrix} y_{i,1} \\ y_{i,2} \end{Bmatrix} = [\mathbf{N}]_i \{\mathbf{y}\}_i, \qquad (3.99)$$

де у якості функцій форми елементів виберемо кусково-визначені лінійні функції (3.72):

$$N_{i,1}(x) = (X_{i,2} - x) / (X_{i,2} - X_{i,1}),$$
$$N_{i,1}(x) = (x - X_{i,1}) / (X_{i,2} - X_{i,1}), \qquad (3.100)$$
$$X_{i,1} \le x \le X_{i,2}.$$

Запишемо рівняння методу зважених нев'язок:

$$\int_{X_{i,1}}^{X_{i,2}} [\mathbf{N}]_i^{\mathbf{T}} \left(\frac{d^2 [\mathbf{N}]_i}{dx^2} \{\mathbf{y}\}_i - [\mathbf{N}]_i \{\mathbf{y}\}_i \right) dx = 0. \qquad (3.101)$$

Зведемо рівняння до слабкої форми:

$$\left(-\int_{X_{i,1}}^{X_{i,2}} \frac{d[\mathbf{N}]_i^{\mathbf{T}}}{dx} \frac{d[\mathbf{N}]_i}{dx} dx + \left[[\mathbf{N}]_i^{\mathbf{T}} \frac{d[\mathbf{N}]_i}{dx} \right]_{X_{i,1}}^{X_{i,2}} - \int_{X_{i,1}}^{X_{i,2}} [\mathbf{N}]_i^{\mathbf{T}} [\mathbf{N}]_i dx \right) \{\mathbf{y}\}_i = 0. \qquad (3.102)$$

Враховуючи те, що середній доданок це:

$$\left[[\mathbf{N}]_i^{\mathbf{T}} \frac{d[\mathbf{N}]_i}{dx} \right]_{X_{i,1}}^{X_{i,2}} \{\mathbf{y}\}_i = \begin{bmatrix} N_{i,1}(x) \dfrac{d\tilde{y}_i(x)}{dx} \\ N_{i,2}(x) \dfrac{d\tilde{y}_i(x)}{dx} \end{bmatrix}_{X_{i,1}}^{X_{i,2}} = \begin{bmatrix} N_{i,1}(x) \dfrac{d}{dx} \left(y_{i,1} N_{i,1}(x) + y_{i,2} N_{i,2}(x) \right) \\ N_{i,2}(x) \dfrac{d}{dx} \left(y_{i,1} N_{i,1}(x) + y_{i,2} N_{i,2}(x) \right) \end{bmatrix}_{X_{i,1}}^{X_{i,2}}, \qquad (3.103)$$

з обраними базисними функціями він завжди буде обертатися в нуль. Звідки отримаємо формулу для локальної системи рівнянь:

$$[\mathbf{K}]_i = \int_{X_{i,1}}^{X_{i,2}} \frac{d[\mathbf{N}]_i^{\mathbf{T}}}{dx} \frac{d[\mathbf{N}]_i}{dx} dx + \int_{X_{i,1}}^{X_{i,2}} [\mathbf{N}]_i^{\mathbf{T}} [\mathbf{N}]_i dx, \quad \{\mathbf{f}\}_i = \{\mathbf{0}\}. \qquad (3.104)$$

Позначимо довжину елементу як $h_i = X_{i,2} - X_{i,1}$ і знайдемо локальні матриці жорсткості для кожного з елементів:

$$[\mathbf{K}]_i = \int_{X_{i,1}}^{X_{i,2}} \left(\frac{d}{dx} \begin{bmatrix} \dfrac{X_{i,2} - x}{h_i} \\ \dfrac{x - X_{i,1}}{h_i} \end{bmatrix} \frac{d}{dx} \begin{bmatrix} \dfrac{X_{i,2} - x}{h_i} & \dfrac{x - X_{i,1}}{h_i} \end{bmatrix} + \begin{bmatrix} \dfrac{X_{i,2} - x}{h_i} \\ \dfrac{x - X_{i,1}}{h_i} \end{bmatrix} \begin{bmatrix} \dfrac{X_{i,2} - x}{h_i} & \dfrac{x - X_{i,1}}{h_i} \end{bmatrix} \right) dx =$$

$$= \int_{X_{i,1}}^{X_{i,2}} \left(\frac{1}{h_i^2} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \begin{bmatrix} -1 & 1 \end{bmatrix} + \frac{1}{h_i^2} \begin{bmatrix} X_{i,2} - x \\ x - X_{i,1} \end{bmatrix} \begin{bmatrix} X_{i,2} - x & x - X_{i,1} \end{bmatrix} \right) dx =$$

$$= \frac{1}{h_i^2} \int_{X_{i,1}}^{X_{i,2}} \left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \begin{bmatrix} (X_{i,2} - x)^2 & (X_{i,2} - x)(x - X_{i,1}) \\ (X_{i,2} - x)(x - X_{i,1}) & (x - X_{i,1})^2 \end{bmatrix} \right) dx =$$

$$= \frac{1}{h_i^2}\left(\begin{bmatrix} h_i & -h_i \\ -h_i & h_i \end{bmatrix} + \begin{bmatrix} \frac{1}{3}h_i^3 & \frac{1}{6}h_i^3 \\ \frac{1}{6}h_i^3 & \frac{1}{3}h_i^3 \end{bmatrix} \right) = \begin{bmatrix} \frac{1}{h_i}+\frac{h_i}{3} & -\frac{1}{h_i}+\frac{h_i}{6} \\ -\frac{1}{h_i}+\frac{h_i}{6} & \frac{1}{h_i}+\frac{h_i}{3} \end{bmatrix}. \qquad (3.105)$$

Зберемо глобальну матрицю жорсткості:

$$[\mathbf{K}] = \begin{bmatrix} \frac{1}{h_1}+\frac{h_1}{3} & -\frac{1}{h_1}+\frac{h_1}{6} & 0 & 0 \\ -\frac{1}{h_1}+\frac{h_1}{6} & \frac{1}{h_1}+\frac{h_1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{h_2}+\frac{h_2}{3} & -\frac{1}{h_2}+\frac{h_2}{6} & 0 \\ 0 & -\frac{1}{h_2}+\frac{h_2}{6} & \frac{1}{h_2}+\frac{h_2}{3} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{h_3}+\frac{h_3}{3} & -\frac{1}{h_3}+\frac{h_3}{6} \\ 0 & 0 & -\frac{1}{h_3}+\frac{h_3}{6} & \frac{1}{h_3}+\frac{h_3}{3} \end{bmatrix} = \qquad (3.106)$$

$$= \begin{bmatrix} \frac{1}{h_1}+\frac{h_1}{3} & -\frac{1}{h_1}+\frac{h_1}{6} & 0 & 0 \\ -\frac{1}{h_1}+\frac{h_1}{6} & \frac{1}{h_1}+\frac{h_1}{3}+\frac{1}{h_2}+\frac{h_2}{3} & -\frac{1}{h_2}+\frac{h_2}{6} & 0 \\ 0 & -\frac{1}{h_2}+\frac{h_2}{6} & \frac{1}{h_2}+\frac{h_2}{3}+\frac{1}{h_3}+\frac{h_3}{3} & -\frac{1}{h_3}+\frac{h_3}{6} \\ 0 & 0 & -\frac{1}{h_3}+\frac{h_3}{6} & \frac{1}{h_3}+\frac{h_3}{3} \end{bmatrix}.$$

Враховуючи що $h_1 = 1/4$, $h_2 = 7/20$ та $h_3 = 2/5$, отримаємо глобальну систему рівнянь:

$$\begin{bmatrix} 49/12 & -95/24 & 0 & 0 \\ -95/24 & 49/12+1249/420 & -2351/840 & 0 \\ 0 & -2351/840 & 1249/420+70/30 & -73/30 \\ 0 & 0 & -73/30 & 70/30 \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{Bmatrix}. \qquad (3.107)$$

або:

$$\begin{bmatrix} 4{,}083333 & -3{,}958333 & 0 & 0 \\ -3{,}958333 & 7{,}057143 & -2{,}798810 & 0 \\ 0 & -2{,}798810 & 5{,}607143 & -2{,}433333 \\ 0 & 0 & -2{,}433333 & 2{,}633333 \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{Bmatrix}. \qquad (3.108)$$

Враховуючи початкові умови, тобто відомі y_1 та y_4, систему слід модифікувати:

$$\begin{bmatrix} 4{,}083333 & 0 & 0 & 0 \\ 0 & 7{,}057143 & -2{,}798810 & 0 \\ 0 & -2{,}798810 & 5{,}607143 & 0 \\ 0 & 0 & 0 & 2{,}633333 \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 2{,}433333 \\ 2{,}633333 \end{Bmatrix}. \qquad (3.109)$$

Розв'язком цієї системи є вектор:

$$\{\mathbf{y}\} = \{0{,}000000 \quad 0{,}214589 \quad 0{,}541083 \quad 1{,}000000\}^{\mathbf{T}}. \qquad (3.110)$$

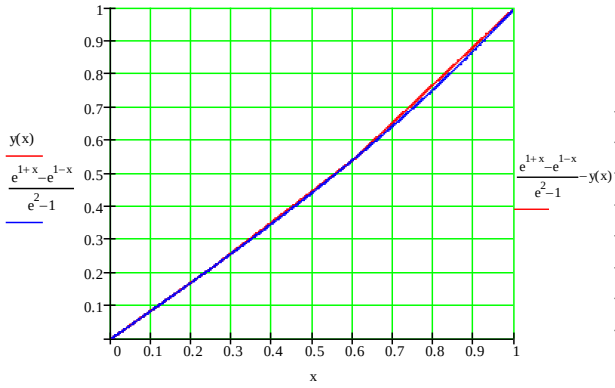


Рис. 3.8 Точне та наближене рішення рівняння $d^2y(x)/dx^2 - y(x) = 0$*, отримане скінченно-елементною апроксимацією*

Рис. 3.9 Похибка між точним та отриманим наближеним рішенням рівняння $d^2y(x)/dx^2 - y(x) = 0$

На *Рис. 3.8* показано графік точного рішення та його отриманої апроксимації. На *Рис. 3.9* показано похибку між точним та отриманим наближеним рішенням. Як і слід було очікувати, похибка є мінімальною у вузлах дискретизації.

Отже, метод скінченних елементів є *чисельним* методом рішення диференційних рівнянь, що зустрічаються при розв'язку інженерних задач. Коротко алгоритм методу можна описати наступним чином [2]:

- в області, що розглядається, вибирається скінченна кількість вузлів, значення неперервної величини в кожному з цих вузлів – це змінна яку потрібно знайти;
- між вузлами вибираються підобласті (скінченні елементи), їх сукупність апроксимує форму області;
- неперервна величина апроксимується на кожному елементі (переважно поліномом), завдяки вузловим значенням.

Широке використання методу скінченних елементів зумовлене його перевагами над іншими чисельними методами. Зокрема до них можна віднести:

- властивості суміжних елементів не обов'язково мають бути однаковими, це дозволяє використовувати метод для тіл, складених з різних матеріалів;
- криволінійна область може бути апроксимована за допомогою прямолінійних елементів або описана точно за допомогою криволінійних елементів, отже метод можна використовувати не лише для областей з хорошою формою границь;
- розміри елементів можуть бути змінними, це дозволяє збільшувати або зменшувати сітку розбиття області на елементи, якщо в цьому є необхідність;
- при використанні методу скінченних елементів не виникає проблем при розгляді змішаних крайових умов.

До недоліків методу можна віднести дуже громіздкі розрахунки. Навіть у випадках простих задач, необхідність використовувати швидкодіючу ЕОМ з досить великим об'ємом оперативної пам'яті.

3.6. Симплекс елементи та лінійна інтерполяція

Спробуємо розібратися, яким чином будуються лінійні кусково-визначені функції форми, що використовувалися в попередніх прикладах, а також, як їх застосовувати до задач в багатовимірних просторах.

Вибір функцій форми $N(\mathbf{r})$ у загальному випадку залежить від задачі що розглядається і необхідної точності розв'язку. Трішки забігаючи вперед, розглянемо одну з класифікацій скінченних елементів за кількістю їх вузлів і відповідних функцій форм, де розрізняють:

- *симплекс* елементи, у яких кількість вузлів на одиницю більша за розмірність задачі, що розв'язується і відповідні інтерполяційні функції є лінійними [17];
- *комплекс* і *мультиплекс* скінченні елементи, у яких кількість вузлів більша за одиницю від розмірності задачі і відповідні функції інтерполяції можуть бути поліномами вищих порядків [2];
- крім того, у деяких задачах, де розглядаються криволінійні границі об'єктів, застосовують *криволінійні* скінченні елементи [3], [4], [15].

Від вибору типу елементів залежить похибка інтерполяції шуканої величини в межах елементу, і як наслідок, в межах всієї області, тобто порядок точності скінченно-елементної моделі. З іншої сторони, вибір поліномів високого порядку призводить до збільшення кількості обчислень при інтегруванні. Перевага симплекс елементів полягає по-перше, в простій і досконалій математичній базі для задач будь-якої розмірності, і як наслідок по-друге, в наявності ефективних методів автоматичної побудови скінченно-елементних сіток для тіл практично будь-якої складності. Окремим пунктом сюди також можна приписати зв'язок математичної бази та відповідних алгоритмів з методами геометричного моделювання, які відіграють не останню роль в комп'ютерних системах моделювання і проектування.

Симплексом[1] T^N, S^N або в нашій нотації Ω^N, називають частину N-вимірного простору $\Re^N$, що обмежена випуклою оболонкою з $M = N+1$ геометрично незалежних точок заданих радіус-векторами $\mathbf{r}_i$, $i = 1,2,\ldots M$:

$$\Omega = \left(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_M\right) \subset \Re^N. \qquad (3.111)$$

Кожна з граней симплексу Ω_i^{N-1} також є симплексом в $(N-1)$- вимірному просторі.

Розглянемо лінійну функцію, що є поліномом першого порядку:

$$u(x) = \alpha + \beta x. \qquad (3.112)$$

[1] Більш строге математичне визначення симплексу дається на основі барицентричних координат [17], [18].

Відповідним симплекс елементом є одновимірний відрізок, що використовувався в попередніх прикладах. Позначимо його вузли як X_1 та X_2, отримаємо систему рівнянь:

$$\begin{aligned} u_1 &= \alpha + \beta X_1, \\ u_2 &= \alpha + \beta X_2. \end{aligned} \tag{3.113}$$

Розв'язавши систему отримаємо коефіцієнти:

$$\begin{aligned} \alpha &= \frac{u_1 X_2 - u_2 X_1}{X_2 - X_1}, \\ \beta &= \frac{u_1 - u_2}{X_2 - X_1}. \end{aligned} \tag{3.114}$$

Підставивши (3.114) в (3.112) отримаємо вираз:

$$u(x) = \frac{u_1 X_2 - u_2 X_1}{X_2 - X_1} + \frac{u_1 - u_2}{X_2 - X_1} x = \left(\frac{X_2 - x}{X_2 - X_1} \right) u_1 + \left(\frac{x - X_1}{X_2 - X_1} \right) u_2. \tag{3.115}$$

Множники у виразі (3.115) є лінійними інтерполяційними функціями, тобто функціями форми симплекс елементу $N(\mathbf{r})$:

$$N_1(x) = \frac{X_2 - x}{X_2 - X_1}, \quad N_2(x) = \frac{x - X_1}{X_2 - X_1}. \tag{3.116}$$

Щоб отримати загальну формулу знаходження функцій форми симплекс елементів будь-якої розмірності [19], перепишемо (3.112) у матричну форму [2], [3], [15]:

$$u(x) = \alpha + \beta x = [1 \quad x] \begin{Bmatrix} \alpha \\ \beta \end{Bmatrix} = [\mathbf{P}] \begin{Bmatrix} \alpha \\ \beta \end{Bmatrix}. \tag{3.117}$$

Систему рівнянь (3.113) можна переписати в матричну форму як:

$$\{\mathbf{u}\} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \end{bmatrix} \begin{Bmatrix} \alpha \\ \beta \end{Bmatrix} = [\mathbf{C}] \begin{Bmatrix} \alpha \\ \beta \end{Bmatrix}. \tag{3.118}$$

Розв'язок (3.114) у матричній формі записується як:

$$\begin{Bmatrix} \alpha \\ \beta \end{Bmatrix} = [\mathbf{C}]^{-1} \{\mathbf{u}\}. \tag{3.119}$$

Підставивши (3.119) у (3.117) отримаємо загальну матричну форму:

$$u(\mathbf{r}) = [\mathbf{P}][\mathbf{C}]^{-1} \{\mathbf{u}\}, \tag{3.120}$$

звідки загальна матрична форма для функцій форми симплекс елементу:

$$[\mathbf{N}] = [\mathbf{P}][\mathbf{C}]^{-1}. \tag{3.121}$$

Щоб застосувати формулу для будь-якої розмірності потрібно розширити відповідні матриці. Так матриця $[\mathbf{P}]$ у загальному випадку будується як:

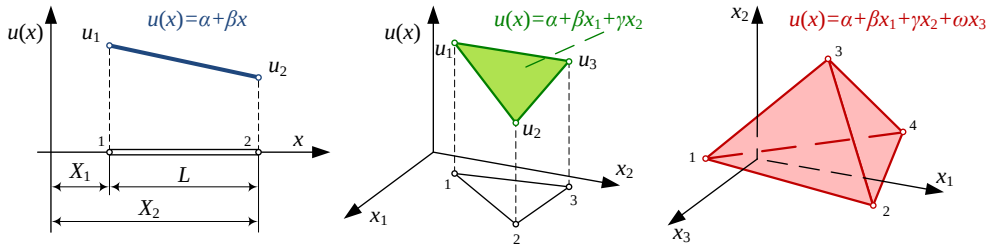
$$[\mathbf{P}] = \begin{bmatrix} 1 & x_1 & x_2 & \dots & x_N \end{bmatrix}. \tag{3.122}$$

Позначивши координати вузла симплекс елементу як $X_{i,j}$, де перший індекс, це

індекс вузла $i=1,2,\ldots,M$, а другий – це індекс відповідної координати $j=1,2,\ldots,N,\quad N=M-1$, квадратна матриця $[\mathbf{C}]$ у загальному випадку будується як:

$$[\mathbf{C}]=\begin{bmatrix}1 & X_{1,1} & X_{1,2} & \cdots & X_{1,N}\\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,N}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & X_{M,1} & X_{M,2} & \cdots & X_{M,N}\end{bmatrix}. \qquad (3.123)$$

На *Рис. 3.10* показано приклади симплекс елементів та їх інтерполяційних поліномів в одно-, дво- і тривимірному просторах.



a) *b)* *c)*

Рис. 3.10 Одно- (a), дво- (b) і тривимірний (c) симплекс елементи та їх інтерполяційні поліноми

Розглядаючи матрицю $[\mathbf{N}]$ можна зауважити, що оскільки похідна $\partial x_i/\partial x_i=1$, то матриця похідних функцій форми $[\mathbf{B}]=\nabla[\mathbf{N}]$ (3.83) є фактично матрицею $[\mathbf{C}]^{-1}$ без першого рядка:

$$[\mathbf{B}]=\nabla[\mathbf{N}]=\nabla\left([\mathbf{P}][\mathbf{C}]^{-1}\right)=\begin{Bmatrix}\frac{\partial}{\partial x_1}\\ \frac{\partial}{\partial x_2}\\ \vdots\\ \frac{\partial}{\partial x_N}\end{Bmatrix}\cdot\begin{bmatrix}1\cdot[\mathbf{C}]^{-1}_{1,1} & 1\cdot[\mathbf{C}]^{-1}_{1,2} & \cdots & 1\cdot[\mathbf{C}]^{-1}_{1,M}\\ + & + & & +\\ x_1\cdot[\mathbf{C}]^{-1}_{2,1} & x_1\cdot[\mathbf{C}]^{-1}_{2,2} & \cdots & x_1\cdot[\mathbf{C}]^{-1}_{2,M}\\ + & + & & +\\ x_2\cdot[\mathbf{C}]^{-1}_{3,1} & x_2\cdot[\mathbf{C}]^{-1}_{3,2} & \cdots & x_2\cdot[\mathbf{C}]^{-1}_{3,M}\\ + & + & & +\\ \vdots & \vdots & \ddots & \vdots\\ + & + & & +\\ x_N\cdot[\mathbf{C}]^{-1}_{M,1} & x_N\cdot[\mathbf{C}]^{-1}_{M,2} & \cdots & x_N\cdot[\mathbf{C}]^{-1}_{M,M}\end{bmatrix}= \qquad (3.124)$$

$$=\begin{bmatrix}[\mathbf{C}]^{-1}_{2,1} & [\mathbf{C}]^{-1}_{2,2} & \cdots & [\mathbf{C}]^{-1}_{2,M}\\ [\mathbf{C}]^{-1}_{3,1} & [\mathbf{C}]^{-1}_{3,2} & \cdots & [\mathbf{C}]^{-1}_{3,M}\\ \vdots & \vdots & \ddots & \vdots\\ [\mathbf{C}]^{-1}_{M,1} & [\mathbf{C}]^{-1}_{M,2} & \cdots & [\mathbf{C}]^{-1}_{M,M}\end{bmatrix}.$$

Отримана матриця не містить змінних, всі її коефіцієнти залежать тільки від координат вузлів симплекс елементу, що є наперед визначеними, тому

підставляючи отримані результати в (3.84) можна знайти локальну матрицю жорсткості $[\mathbf{K}]_i$:

$$[\mathbf{K}]_i = \int_{\Omega_i} [\mathbf{B}]_i^{\mathrm{T}} [\mathbf{D}]_i [\mathbf{B}]_i \, d\Omega_i = [\mathbf{B}]_i^{\mathrm{T}} [\mathbf{D}]_i [\mathbf{B}]_i \int_{\Omega_i} d\Omega_i = [\mathbf{B}]_i^{\mathrm{T}} [\mathbf{D}]_i [\mathbf{B}]_i \Omega_i, \quad (3.125)$$

де, Ω_i – це об'єм симплекс елементу, який у загальному випадку, для будь-якої розмірності, можна знайти за допомогою формули орієнтованого об'єму (знак отриманого результату залежить від нумерації вузлів, для правильного результату він має бути додатнім):

$$\Omega^N = \frac{1}{N!}\left|[\mathbf{C}]\right| = \frac{1}{N!}\left\|\begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,N} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{M,1} & X_{M,2} & \cdots & X_{M,N} \end{bmatrix}\right\| = \\ = \frac{1}{N!}\left\|\begin{bmatrix} X_{1,1} - X_{M,1} & X_{1,2} - X_{M,2} & \cdots & X_{1,N} - X_{M,N} \\ X_{2,1} - X_{M,1} & X_{2,2} - X_{M,2} & \cdots & X_{2,N} - X_{M,N} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N,1} - X_{M,1} & X_{N,2} - X_{M,2} & \cdots & X_{N,N} - X_{M,N} \end{bmatrix}\right\|, \quad (3.126)$$

або за допомогою формули з використанням визначника Кейлі-Менгера [20], [21], [22], що не залежить від розмірності простору і може бути застосована також до граней симплекс елементу:

$$\Omega^N = \sqrt{\frac{(-1)^{N-1}}{2^N (N!)^2}\begin{bmatrix} 0 & 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & 0 & d_{1,2}^2 & d_{1,3}^2 & \cdots & d_{1,N}^2 & d_{1,N+1}^2 \\ 1 & d_{1,2}^2 & 0 & d_{2,3}^2 & \cdots & d_{2,N}^2 & d_{2,N+1}^2 \\ 1 & d_{1,3}^2 & d_{2,3}^2 & 0 & \cdots & d_{3,N}^2 & d_{3,N+1}^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & d_{1,N}^2 & d_{2,N}^2 & d_{3,N}^2 & \cdots & 0 & d_{N,N+1}^2 \\ 1 & d_{1,N+1}^2 & d_{2,N+1}^2 & d_{3,N+1}^2 & \cdots & d_{N,N+1}^2 & 0 \end{bmatrix}}, \quad (3.127)$$

$$d_{i,j} = d(X_i, X_j) = \left\| X_i - X_j \right\|_2.$$

Підставивши результати в (3.86) отримаємо глобальну матрицю жорсткості:

$$[\mathbf{K}] = \sum_{i=1}^{P} [\mathbf{K}]_i = \sum_{i=1}^{P} [\mathbf{B}]_i^{\mathrm{T}} [\mathbf{D}]_i [\mathbf{B}]_i \frac{1}{N!}\left|[\mathbf{C}]_i\right|. \quad (3.128)$$

Щоб зрозуміти геометричний зміст, розглянемо двовимірний симплекс елемент, тобто трикутник (*Рис. 3.10.b*). Інтерполяційний поліном записується у вигляді:

$$u(x_1, x_2) = \alpha + \beta x_1 + \gamma x_2. \quad (3.129)$$

Для вузлових значень можна записати систему рівнянь:

$$\begin{aligned} u_1 &= \alpha + \beta X_{1,1} + \gamma X_{1,2}, \\ u_2 &= \alpha + \beta X_{2,1} + \gamma X_{2,2}, \\ u_3 &= \alpha + \beta X_{3,1} + \gamma X_{3,2}. \end{aligned} \tag{3.130}$$

Розв'язавши систему отримаємо коефіцієнти:

$$\begin{aligned} \alpha &= \frac{1}{|[\mathbf{C}]|}\left((X_{2,1}X_{3,2} - X_{3,1}X_{2,2})u_1 + (X_{3,1}X_{1,2} - X_{1,1}X_{3,2})u_2 + (X_{1,1}X_{2,2} - X_{2,1}X_{1,2})u_3\right), \\ \beta &= \frac{1}{|[\mathbf{C}]|}\left((X_{2,2} - X_{3,2})u_1 + (X_{3,2} - X_{1,2})u_2 + (X_{1,2} - X_{2,2})u_3\right), \\ \gamma &= \frac{1}{|[\mathbf{C}]|}\left((X_{3,1} - X_{2,1})u_1 + (X_{1,1} - X_{3,1})u_2 + (X_{2,1} - X_{1,1})u_3\right), \end{aligned} \tag{3.131}$$

де:

$$|[\mathbf{C}]| = \left\| \begin{bmatrix} 1 & X_{1,1} & X_{1,2} \\ 1 & X_{2,1} & X_{2,2} \\ 1 & X_{3,1} & X_{3,2} \end{bmatrix} \right\| = 2\Omega, \tag{3.132}$$

є подвійною площею трикутника.

Підставляючи знайдені значення α, β і γ в рівняння (3.129) отримаємо вираз для функцій форми:

$$u(x_1, x_2) = N_1(x_1, x_2)u_1 + N_2(x_1, x_2)u_2 + N_3(x_1, x_2)u_3, \tag{3.133}$$

де:

$$N_1(x_1, x_2) = [\mathbf{C}]^{-1}_{1,1} + [\mathbf{C}]^{-1}_{2,1}x_1 + [\mathbf{C}]^{-1}_{3,1}x_2, \quad \begin{cases} [\mathbf{C}]^{-1}_{1,1}|[\mathbf{C}]| = X_{2,1}X_{3,2} - X_{3,1}X_{2,2}, \\ [\mathbf{C}]^{-1}_{2,1}|[\mathbf{C}]| = X_{2,2} - X_{3,2}, \\ [\mathbf{C}]^{-1}_{3,1}|[\mathbf{C}]| = X_{3,1} - X_{2,1}, \end{cases}$$

$$N_2(x_1, x_2) = [\mathbf{C}]^{-1}_{1,2} + [\mathbf{C}]^{-1}_{2,2}x_1 + [\mathbf{C}]^{-1}_{3,2}x_2, \quad \begin{cases} [\mathbf{C}]^{-1}_{1,2}|[\mathbf{C}]| = X_{3,1}X_{1,2} - X_{1,1}X_{3,2}, \\ [\mathbf{C}]^{-1}_{2,2}|[\mathbf{C}]| = X_{3,2} - X_{1,2}, \\ [\mathbf{C}]^{-1}_{3,2}|[\mathbf{C}]| = X_{1,1} - X_{3,1}, \end{cases} \tag{3.134}$$

$$N_3(x_1, x_2) = [\mathbf{C}]^{-1}_{1,3} + [\mathbf{C}]^{-1}_{2,3}x_1 + [\mathbf{C}]^{-1}_{3,3}x_2, \quad \begin{cases} [\mathbf{C}]^{-1}_{1,3}|[\mathbf{C}]| = X_{1,1}X_{2,2} - X_{2,1}X_{1,2}, \\ [\mathbf{C}]^{-1}_{2,3}|[\mathbf{C}]| = X_{1,2} - X_{2,2}, \\ [\mathbf{C}]^{-1}_{3,3}|[\mathbf{C}]| = X_{2,1} - X_{1,1}. \end{cases}$$

Оскільки інтерполяційна функція є лінійною, градієнти шуканого потенціалу в межах елементів завжди будуть постійними, це легко перевірити:

$$\begin{aligned} \frac{\partial u(x_1, x_2)}{\partial x_1} &= \frac{N_1(x_1, x_2)}{\partial x_1}u_1 + \frac{N_2(x_1, x_2)}{\partial x_1}u_2 + \frac{N_3(x_1, x_2)}{\partial x_1}u_3, \\ \frac{\partial u(x_1, x_2)}{\partial x_2} &= \frac{N_1(x_1, x_2)}{\partial x_2}u_1 + \frac{N_2(x_1, x_2)}{\partial x_2}u_2 + \frac{N_3(x_1, x_2)}{\partial x_2}u_3. \end{aligned} \tag{3.135}$$

Враховуючи вираз (3.134) отримаємо:

$$\begin{aligned}\frac{\partial u(x_1,x_2)}{\partial x_1} &= [\mathbf{C}]^{-1}_{2,1}u_1 + [\mathbf{C}]^{-1}_{2,2}u_2 + [\mathbf{C}]^{-1}_{2,3}u_3 = \text{const},\\ \frac{\partial u(x_1,x_2)}{\partial x_2} &= [\mathbf{C}]^{-1}_{3,1}u_1 + [\mathbf{C}]^{-1}_{3,2}u_2 + [\mathbf{C}]^{-1}_{3,3}u_3 = \text{const},\end{aligned} \tag{3.136}$$

або:

$$\begin{bmatrix}[\mathbf{C}]^{-1}_{2,1} & [\mathbf{C}]^{-1}_{2,2} & [\mathbf{C}]^{-1}_{2,3}\\ [\mathbf{C}]^{-1}_{3,1} & [\mathbf{C}]^{-1}_{3,2} & [\mathbf{C}]^{-1}_{3,3}\end{bmatrix}\{\mathbf{u}\} = \begin{bmatrix}\frac{\partial N_1}{\partial x} & \frac{\partial N_2}{\partial x} & \frac{\partial N_3}{\partial x}\\ \frac{\partial N_1}{\partial y} & \frac{\partial N_2}{\partial y} & \frac{\partial N_3}{\partial y}\end{bmatrix}\{\mathbf{u}\} = \nabla[\mathbf{N}]\{\mathbf{u}\} = [\mathbf{B}]\{\mathbf{u}\}. \tag{3.137}$$

Звідси можна зробити висновок – через те, що градієнти в межах елементу є постійними, необхідно використовувати достатньо малі за величиною елементи, щоб апроксимувати потенціали, значення яких швидко змінюються в залежності від координат.

На *Рис. 3.11* зображено геометричний зміст матриці градієнтів [**B**] для двовимірного симплекс елементу, на основі виразів, отриманих в (3.134). За необхідності, аналогічні співвідношення проекцій сторін можна вивести для симплекс елементів будь-якої розмірності.

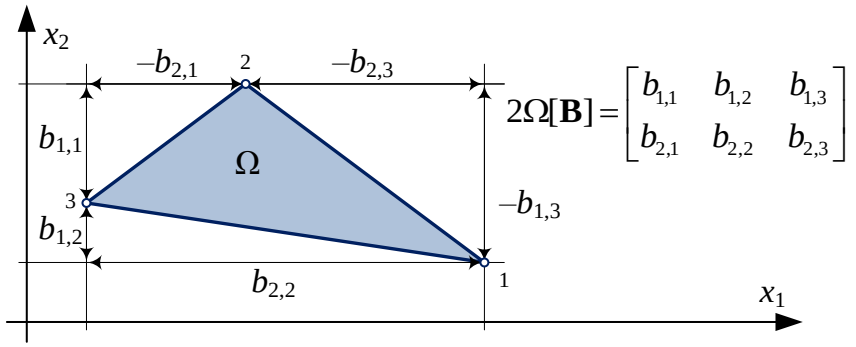


Рис. 3.11 Геометричний зміст матриці градієнтів [**B**] *для двовимірного симплекс елементу*

Для знаходження вектору навантажень $\{\mathbf{f}\}$ розглядають *барицентричні координати* [17] симплекс елементу (у літературі по МСЕ також можна зустріти назви L- координати [2], природні координати [16], [23], симплекс координати або однорідні координати [24]).

Нехай симплекс $\Omega \subset \Re^N$ заданий вузлами з радіус-векторами $\mathbf{r}_i$, $i = 1,2,\ldots M$, $M = N+1$. Барицентричними координатами деякої точки, що задана радіус-вектором $\mathbf{r}$ є набір коефіцієнтів $L_1, L_2, \ldots, L_M$, таких що [17]:

$$\begin{aligned} L_1 + L_2 + \ldots + L_M &= 1,\\ \mathbf{r} &= L_1\mathbf{r}_1 + L_2\mathbf{r}_2 + \ldots + L_M\mathbf{r}_M, \end{aligned} \tag{3.138}$$

або у матричній формі:

$$\begin{Bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_N \end{Bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{1,1} & X_{2,1} & \cdots & X_{M,1} \\ X_{1,2} & X_{2,2} & \cdots & X_{M,2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1,N} & X_{2,N} & \cdots & X_{M,N} \end{bmatrix} \begin{Bmatrix} L_1 \\ L_2 \\ \vdots \\ L_N \\ L_M \end{Bmatrix}. \qquad (3.139)$$

Підставивши в останній вираз (3.122) та (3.123) отримаємо [7]:

$$[\mathbf{P}]^{\mathbf{T}} = [\mathbf{C}]^{\mathbf{T}}[\mathbf{L}]^{\mathbf{T}}, \quad [\mathbf{L}]^{\mathbf{T}} = \left([\mathbf{C}]^{\mathbf{T}}\right)^{-1}[\mathbf{P}]^{\mathbf{T}},$$

$$[\mathbf{L}]^{\mathbf{T}} = \left([\mathbf{C}]^{-1}\right)^{\mathbf{T}}[\mathbf{P}]^{\mathbf{T}}, \quad [\mathbf{L}]^{\mathbf{T}} = \left([\mathbf{P}][\mathbf{C}]^{-1}\right)^{\mathbf{T}}, \qquad (3.140)$$

$$[\mathbf{L}] = [\mathbf{P}][\mathbf{C}]^{-1} = [\mathbf{N}].$$

Тобто, у випадку використання симплекс елементів функції їх форми є їх барицентричними координатами.

Геометрично кожна барицентрична координата L_i це відношення об'єму симплексу, що утворений заданою точкою і гранню, протилежною до i-го вузла базового симплексу, до об'єму базового симплексу (*Рис. 3.12*), тобто:

$$L_i = \frac{\Omega_i}{\Omega} = \frac{\left|[\mathbf{C}]^i\right|}{\left|[\mathbf{C}]\right|} = \frac{\left\|\begin{bmatrix} 1 & x_1 & x_2 & \cdots & x_N \\ 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{i-1,1} & X_{i-1,2} & \cdots & X_{i-1,N} \\ 1 & X_{i+1,1} & X_{i+1,2} & \cdots & X_{i+1,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{M,1} & X_{M,2} & \cdots & X_{M,N} \end{bmatrix}\right\|}{\left\|\begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,N} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{M,1} & X_{M,2} & \cdots & X_{M,N} \end{bmatrix}\right\|}. \qquad (3.141)$$

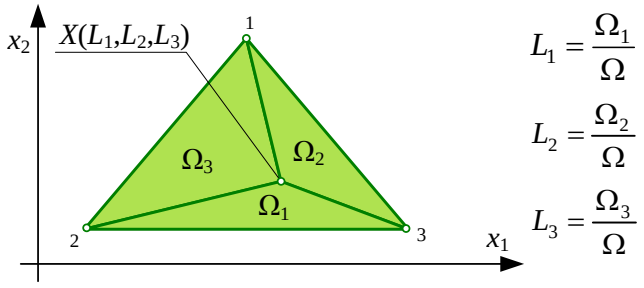


Рис. 3.12 Геометричний зміст барицентричних координат на прикладі двовимірного симплекс елементу.

Якщо якась з барицентричних координат L_i рівна нулю, то задана точка знаходиться на грані, що протилежна до i-го вузла. Якщо якась з барицентричних координат L_i менша нуля, то задана точка не належить симплексу:

$$\begin{gathered} i = 1, 2, \ldots, N, \\ X \in \Omega^N \quad \Leftrightarrow \quad \forall i : L_i \geq 0, \\ X \in \Omega_i^{N-1} \quad \Leftrightarrow \quad \exists i : L_i = 0, \\ X \notin \Omega^N \quad \Leftrightarrow \quad \exists i : L_i < 0. \end{gathered} \tag{3.142}$$

Обчислимо інтеграл для вектору навантажень $\{\mathbf{f}\}$ (3.84), що береться по границі Γ, переводячи функції форми симплекс елементу з глобальних в барицентричні координати. Для цього застосуємо процедуру деформації (сукупності ізопараметричних афінних перетворень – перенесення, стискування та обертання) довільного симплекс елементу в універсальний елемент одиничної довжини, що лежить на координатних осях.

Відомо [11], що будь-яка деформація може бути описана лише однозначними перетвореннями, що мають неперервні похідні необхідного порядку. Математично, такі перетворення описуються взаємно-однозначним, неперервним *відображенням* (що також називають *бієкцією*), в нашому випадку з простору $\Re^N$ де заданий довільний симплекс елемент, в простір $\left(\Re^N\right)^*$ де заданий універсальний елемент одиничної довжини, що лежить на координатних осях, тобто $\mathbf{F} : \Re \to \Re^*$.

Нехай задано точку $P \in \Re^N$, з радіус-вектором $\mathbf{R} \equiv (X_1, X_2, \ldots, X_N)$, взаємно-однозначне відображення $\mathbf{F} : \Re \to \Re^*$ переводить цю точку в точку $p \in \left(\Re^N\right)^*$, з радіус-вектором $\mathbf{r} \equiv (x_1, x_2, \ldots, x_N)$, тобто кожна з координат точки P є функцією координат точки p: $X_i = (x_1, x_2, \ldots, x_N)$ (*Рис. 3.13*). Таке відображення можна побудувати на основі розкладу в ряд Тейлора:

$$\mathbf{F}(\mathbf{r}) = \mathbf{F}(\mathbf{R}) + \mathbf{Jac}_{\mathbf{r}}(\mathbf{R})(\mathbf{r} - \mathbf{R}) + \theta(\|\mathbf{r} - \mathbf{R}\|), \tag{3.143}$$

де, $\mathbf{Jac}_{\mathbf{r}}(\mathbf{R})$ – це матриця Якобі:

$$[\mathbf{Jac}_{\mathbf{r}}\mathbf{R}] = \begin{bmatrix} \frac{\partial X_1}{\partial x_1} & \frac{\partial X_1}{\partial x_2} & \cdots & \frac{\partial X_1}{\partial x_N} \\ \frac{\partial X_2}{\partial x_1} & \frac{\partial X_2}{\partial x_2} & \cdots & \frac{\partial X_2}{\partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial X_N}{\partial x_1} & \frac{\partial X_N}{\partial x_2} & \cdots & \frac{\partial X_N}{\partial x_N} \end{bmatrix} = \frac{\partial (X_1, X_2, \ldots, X_N)}{\partial (x_1, x_2, \ldots, x_N)} = \frac{\partial X_\alpha}{\partial x_\beta}. \tag{3.144}$$

Визначник матриці Якобі, що часто називають функціональним визначником або просто *Якобіаном*, має безпосередній фізичний зміст – він

показує відношення елементарних об'ємів (площ чи довжин в залежності від кількості вимірів) тіла при деформації. Якщо тіло описується деякою функцією $f(X,Y,Z)$, і після деформації функцією від нових координат $f(X(x,y,z),Y(x,y,z),Z(x,y,z))$, то об'єм можна визначити як:

$$\iiint\limits_{\Omega\subset\Re} f(X,Y,Z)dXdYdZ = \iiint\limits_{\Omega^*\subset\Re^*} f(X(x,y,z),Y(x,y,z),Z(x,y,z))\left|[\mathbf{Jac_r R}]\right|dxdydz, \quad (3.145)$$

або в загальному випадку:

$$\int\limits_{\Omega} f(\mathbf{R})d\Omega = \int\limits_{\Omega^*} f^*(\mathbf{r})\left|[\mathbf{Jac_r R}]\right|d\Omega^*. \quad (3.146)$$

Щоб перевести функції форми симплекс елементу з глобальних координат в барицентричні, необхідно визначити Якобіан $\left|[\mathbf{Jac_L r}]\right|$. Для цього використаємо відношення (3.139), позначивши $L_M = 1-\sum_{i=1}^{N} L_i$. Виконавши множення, для кожної координати отримаємо залежність:

$$x_j = X_{1,j}L_1 + X_{2,j}L_2 + \ldots + X_{N,j}L_N + X_{M,j}\left(1-\sum_{i=1}^{N} L_i\right), \quad j=1,2,\ldots N,$$
$$\frac{\partial x_j}{\partial L_k} = X_{k,j} - X_{M,j}, \quad k=1,2,\ldots N. \quad (3.147)$$

Звідси, використовуючи (3.126), Якобіан $\left|[\mathbf{Jac_L r}]\right|$ рівний:

$$\left|[\mathbf{Jac_L r}]\right| = \left|\begin{bmatrix} X_{1,1}-X_{M,1} & X_{1,2}-X_{M,2} & \ldots & X_{1,N}-X_{M,N} \\ X_{2,1}-X_{M,1} & X_{2,2}-X_{M,2} & \ldots & X_{2,N}-X_{M,N} \\ \vdots & \vdots & \ddots & \ldots \\ X_{N,1}-X_{M,1} & X_{N,2}-X_{M,2} & \ldots & X_{N,N}-X_{M,N} \end{bmatrix}\right| = N!\Omega. \quad (3.148)$$

Отримане співвідношення є очевидним – оскільки об'єм гіперкубу в якому розташований універсальний симплекс елемент одиничної довжини, що лежить на координатних осях, рівний одиниці, а відповідний об'єм елементу рівний $1/N!$, то Якобіан $\left|[\mathbf{Jac_L r}]\right|$ показує відношення об'ємів при деформації з Ω в $1/N!$(*Рис. 3.13*).

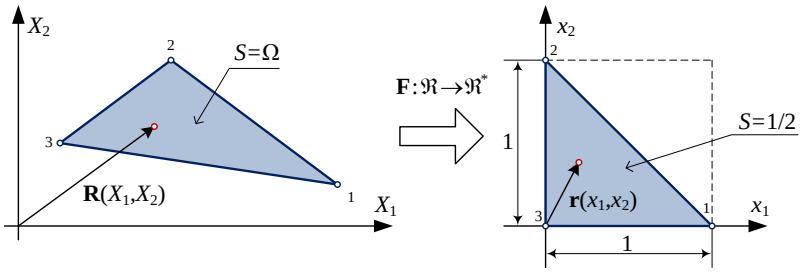


Рис. 3.13 Приклад взаємно-однозначного відображення двовимірного симплекс елементу

Границя Γ в інтегралі для вектору навантажень $\{\mathbf{f}\}$ (3.84) є гранню симплекс елементу, що також є симплексом нижчої розмірності, тому функція форми N_i у вузлі симплексу i, що протилежний до даної грані, рівна нулю. Для прикладу, нехай це буде останній вузол, в іншому випадку вузли елементу завжди можна перенумерувати. Записуючи інтегрування в барицентричних координатах отримаємо:

$$\{\mathbf{f}\} = \int_{\Gamma} [\mathbf{N}]^{\mathbf{T}} f d\Gamma = \int_{\Gamma} \operatorname{diag} \begin{Bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \\ 0 \end{Bmatrix} \begin{bmatrix} N_1 = L_1 \\ N_2 = L_2 \\ \vdots \\ N_N = L_N \\ 0 \end{bmatrix} d\Gamma = \tag{3.149}$$

$$= \int_0^1 \int_0^{1-L_1} \dots \int_0^{1-L_1-\dots-L_{N-1}} \operatorname{diag} \begin{Bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \\ 0 \end{Bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_N \\ 0 \end{bmatrix} \left| [\mathbf{Jac}_{\mathbf{L}} \mathbf{r}]^{\Gamma} \right| dL_N \dots dL_2 dL_1.$$

Для спрощення розглянемо тривимірний симплекс елемент з двовимірною границею:

$$\{\mathbf{f}\} = \int_0^1 \int_0^{1-L_1} \operatorname{diag} \begin{Bmatrix} f_1 \\ f_2 \\ f_3 \\ 0 \end{Bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ 1-L_1-L_2 \\ 0 \end{bmatrix} \left| [\mathbf{Jac}_{\mathbf{L}} \mathbf{r}]^{\Gamma} \right| dL_2 dL_1 = (3-1)! \Omega^{\Gamma} \int_0^1 \operatorname{diag} \begin{Bmatrix} f_1 \\ f_2 \\ f_3 \\ 0 \end{Bmatrix} \begin{bmatrix} L_1 L_2 \\ \frac{L_2^2}{2} \\ L_2 - L_1 L_2 - \frac{L_2^2}{2} \\ 0 \end{bmatrix}_{L_2=0}^{L_2=1-L_1} dL_1 = \tag{3.150}$$

$$= \Omega^{\Gamma} \int_0^1 \operatorname{diag} \begin{Bmatrix} f_1 \\ f_2 \\ f_3 \\ 0 \end{Bmatrix} \begin{bmatrix} 2L_1 - 2L_1^2 \\ 1-2L_1+L_1^2 \\ 1-2L_1+L_1^2 \\ 0 \end{bmatrix} dL_1 = \Omega^{\Gamma} \operatorname{diag} \begin{Bmatrix} f_1 \\ f_2 \\ f_3 \\ 0 \end{Bmatrix} \begin{bmatrix} L_1^2 - \frac{2L_1^3}{3} \\ L_1 - L_1^2 + \frac{L_1^3}{3} \\ L_1 - L_1^2 + \frac{L_1^3}{3} \\ 0 \end{bmatrix}_{L_1=0}^{L_1=1} = \Omega^{\Gamma} \operatorname{diag} \begin{Bmatrix} f_1 \\ f_2 \\ f_3 \\ 0 \end{Bmatrix} \begin{bmatrix} 1-\frac{2}{3} \\ 1-1+\frac{1}{3} \\ 1-1+\frac{1}{3} \\ 0 \end{bmatrix} = \frac{\Omega^{\Gamma}}{3} \begin{Bmatrix} f_1 \\ f_2 \\ f_3 \\ 0 \end{Bmatrix}.$$

Знаменник в дробі $\Omega^{\Gamma}/3$, це кількість вузлів грані, по якій проводиться інтегрування, а також розмірність задачі. Тобто, в загальному випадку, щоб знайти вектор навантажень де задані природні крайові умови, для симплекс елементів справедлива формула:

$$\{\mathbf{f}\} = \frac{\Omega^{\Gamma}}{N!} \begin{Bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{Bmatrix}. \tag{3.151}$$

Застосуємо описаний метод на практиці. Для цього повторно розглянемо приклад (3.57). Нагадаємо умови задачі: нехай коефіцієнт теплопровідності матеріалу $\lambda = 1$ Вт/м°С, матеріал займає квадратну область $-1 \le x \le 1$ м, $-1 \le y \le 1$ м. На сторонах $y = \pm 1$ підтримується постійна температура 0°С, тоді як через сторони $x = \pm 1$ подається тепло з швидкістю $\cos(\pi y/2)$ Вт/м^2°С на одиницю довжини (*Рис. 3.3*).

Розіб'ємо область регулярною сіткою з 200 трикутників так, як це показано на *Рис. 3.14*. Кожну вершину всередині елементу локально пронумеруємо проти годинникової стрілки так, як це показано на *Рис. 3.15*.

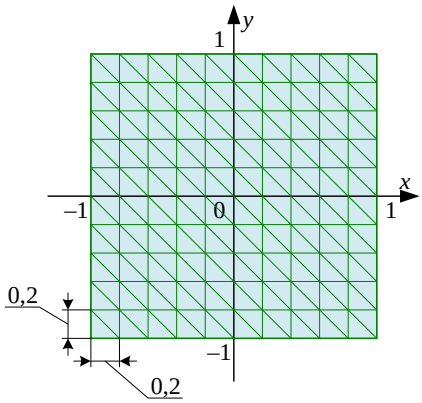


Рис. 3.14 Дискретизація пластини регулярною сіткою з 200 трикутних елементів

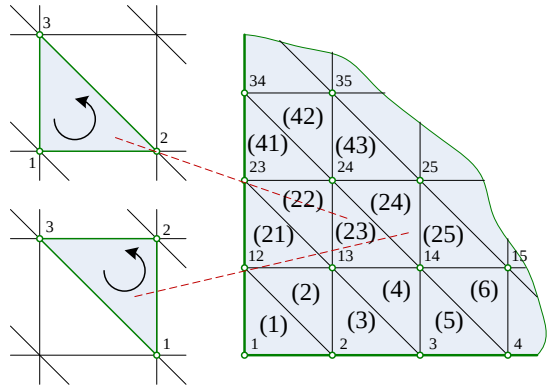


Рис. 3.15 Локальна та глобальна нумерація елементів і вузлів дискретизації

Оскільки сітка регулярна і всі елементи є однаковими за розмірами, кожна локальна матриця жорсткості буде однаковою для парних і непарних елементів. Щоб їх знайти, спочатку запишемо матрицю координат симплексу (3.123), наприклад для першого (тобто непарного) елементу:

$$[\mathbf{C}]_1 = \begin{bmatrix} 1 & -1 & -1 \\ 1 & -0,8 & -1 \\ 1 & -1 & -0,8 \end{bmatrix}. \qquad (3.152)$$

Визначник матриці, тобто подвійна площа трикутника рівна:

$$\left|[\mathbf{C}]_1\right| = \left|\begin{bmatrix} 1 & -1 & -1 \\ 1 & -0,8 & -1 \\ 1 & -1 & -0,8 \end{bmatrix}\right| = 0,04. \qquad (3.153)$$

Обернена матриця рівна:

$$[\mathbf{C}]_1^{-1} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & -0,8 & -1 \\ 1 & -1 & -0,8 \end{bmatrix}^{-1} = \begin{bmatrix} -9 & 5 & 5 \\ -5 & 5 & 0 \\ -5 & 0 & 5 \end{bmatrix}. \qquad (3.154)$$

Матриця градієнтів $[\mathbf{B}]$, це матриця $[\mathbf{C}]^{-1}$ без першого рядка:

$$[\mathbf{B}]_1 = \begin{bmatrix} -5 & 5 & 0 \\ -5 & 0 & 5 \end{bmatrix}. \tag{3.155}$$

Оскільки матеріал пластини є ізотропний, то коефіцієнт теплопровідності однаковий в усіх напрямках, запишемо його як:

$$[\mathbf{D}]_1 = \begin{bmatrix} \lambda_x & 0 \\ 0 & \lambda_y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.156}$$

Тепер можна знайти локальну матрицю жорсткості:

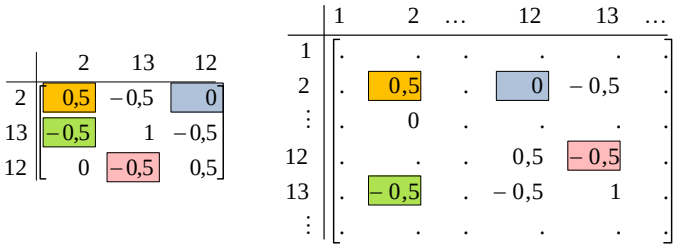
$$[\mathbf{K}]_{2i+1} = [\mathbf{B}]^{\mathrm{T}}_{2i+1}[\mathbf{D}]_{2i+1}[\mathbf{B}]_{2i+1}\frac{1}{2}\left|[\mathbf{C}]_{2i+1}\right| =$$

$$= \begin{bmatrix} -5 & -5 \\ 5 & 0 \\ 0 & 5 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} -5 & 5 & 0 \\ -5 & 0 & 5 \end{bmatrix}\frac{1}{2}0.04 = \begin{bmatrix} 1 & -0,5 & -0,5 \\ -0,5 & 0,5 & 0 \\ -0,5 & 0 & 0,5 \end{bmatrix}. \tag{3.157}$$

Для другого елементу (тобто парного), отримаємо трішки іншу локальну матрицю жорсткості:

$$[\mathbf{K}]_{2i} = [\mathbf{B}]^{\mathrm{T}}_{2i}[\mathbf{D}]_{2i}[\mathbf{B}]_{2i}\frac{1}{2}\left|[\mathbf{C}]_{2i}\right| = \begin{bmatrix} 0,5 & -0,5 & 0 \\ -0,5 & 1 & -0,5 \\ 0 & -0,5 & 0,5 \end{bmatrix}. \tag{3.158}$$

Зберемо глобальну матрицю жорсткості відповідно до нумерації вузлів. Це робиться аналогічно до того, як це робилося для одновимірних елементів (див. наприклад (3.88) або (3.106)) так, як показано на *Рис. 3.16*. Тобто, не потрібно розширювати і додавати всі розширені матриці з великою кількістю нульових коефіцієнтів. Достатньо враховувати тільки вклади ненульових коефіцієнтів у відповідності до глобальної нумерації вузлів.



a) *b*)

Рис. 3.16 Розширення та переформування локальної матриці жорсткості елементу, при побудові глобальної матриці жорсткості; a) – локальна матриця; b) – вклад кожного коефіцієнту локальної матриці в глобальну

Після врахування вкладу коефіцієнтів всіх локальних матриць отримаємо стрічкову глобальну матрицю жорсткості, з шириною стрічки $k = 12$. Тобто, всі елементи, індекс яких починаючи від індексу діагонального елементу +12 є рівними нулю. Таке число з'явилося внаслідок того, що на один ряд сітки припадає 11 вузлів. Фрагмент матриці наведено нижче:

$$[\mathbf{K}]=\begin{bmatrix}
1 & -0,5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0,5 & 0 & \cdots \\
-0,5 & 2 & -0,5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & \cdots \\
0 & -0,5 & 2 & -0,5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & -0,5 & 2 & -0,5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & -0,5 & 2 & -0,5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & -0,5 & 2 & -0,5 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & -0,5 & 2 & -0,5 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & -0,5 & 2 & -0,5 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -0,5 & 2 & -0,5 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0,5 & 2 & -0,5 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0,5 & 1 & 0 & 0 & \cdots \\
-0,5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & -1 & \cdots \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 4 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}. \quad (3.159)$$

Природну крайову умову $\cos(\pi y/2)$ на $x=\pm 1$, для спрощення будемо інтерполювати як середнє значення між вузлами на границі, тобто:

$$f_i = f_{i+1} = \frac{\cos(\pi y_i/2)+\cos(\pi y_{i+1}/2)}{2}\cdot(y_{i+1}-y_i). \quad (3.160)$$

Це еквівалентно тому, що потік тепла по нормалі до границі є сталим в межах кожного елементу, що допустимо при використанні великої кількості елементів.

Звідси, на основі (3.151), знаходимо локальний вектор навантаження для елементів границі, для сторони $x=-1$ це елементи з непарними індексами:

$$\{\mathbf{f}\}_{2k+1}=\frac{\Omega^{\Gamma}}{N!}\begin{Bmatrix} f_1 \\ f_2 \\ f_3 \end{Bmatrix}=0.5\begin{Bmatrix} 0,5\cdot\left(\cos(\pi y_i/2)+\cos(\pi y_{i+1}/2)\right)\cdot(y_{i+1}-y_i) \\ 0 \\ 0,5\cdot\left(\cos(\pi y_i/2)+\cos(\pi y_{i+1}/2)\right)\cdot(y_{i+1}-y_i) \end{Bmatrix}, \quad (3.161)$$

а для сторони $x=1$ це елементи з парними індексами (див. *Рис. 3.15*):

$$\{\mathbf{f}\}_{2k}=\frac{\Omega^{\Gamma}}{N!}\begin{Bmatrix} f_1 \\ f_2 \\ f_3 \end{Bmatrix}=0.5\begin{Bmatrix} 0,5\cdot\left(\cos(\pi y_i/2)+\cos(\pi y_{i+1}/2)\right)\cdot(y_{i+1}-y_i) \\ 0,5\cdot\left(\cos(\pi y_i/2)+\cos(\pi y_{i+1}/2)\right)\cdot(y_{i+1}-y_i) \\ 0 \end{Bmatrix}. \quad (3.162)$$

Глобальний вектор навантаження будується аналогічно до глобальної матриці жорсткості, його фрагмент наведено нижче:

$$\{\mathbf{f}\}=\{0{,}015451 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0{,}015451 \quad 0{,}060291 \quad 0 \quad \ldots\}^{\mathbf{T}}. \quad (3.163)$$

Після врахування головних крайових умов, тобто $T(x,\pm 1)=0$, глобальна матриця жорсткості, та вектор навантаження приймуть вигляд (3.164).

Оскільки процедура включення крайових умов Діріхле змінює крім матриці жорсткості, ще й вектор навантаження, важливо щоб вона застосовувалася в останню чергу, після включення всіх інших крайових умов.

Після розв'язку системи отримаємо вузлові значення шуканої температури. Між вузлами температура інтерполюється функціями форми скінченних елементів, що є барицентричними координатами. Тому, щоб знайти апроксимовану температуру в довільній точці поверхні, потрібно:

- знайти елемент, куди входить задана точка, за допомогою (3.142);
- інтерполювати значення температури як (3.82).

$$[\mathbf{K}]=\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & -1 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 4 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}
\quad
\{\mathbf{f}\}=\begin{Bmatrix}
0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0,060291 \\ 0 \\ \vdots
\end{Bmatrix}
\qquad (3.164)$$

На *Рис. 3.17* зображено отримане апроксимоване поле температури. Різниця між результатами апроксимації методом Гальоркіна при п'яти базисних функціях, що були отримані в попередньому прикладі та результатами скінченно-елементної апроксимації, наведено на *Рис. 3.18*.

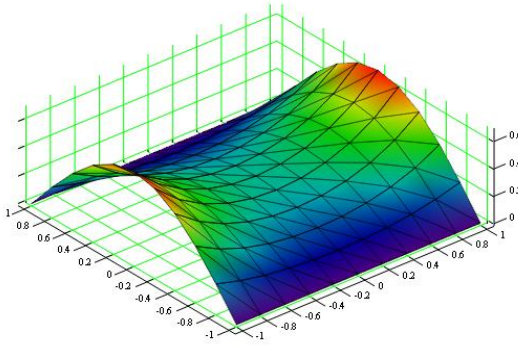


Рис. 3.17 Апроксимоване рішення задачі теплопровідності з допомогою методу скінченних елементів, при використанні регулярної сітки 200 симплекс елементів

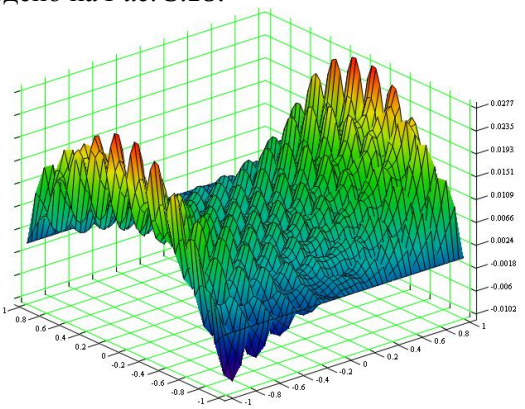


Рис. 3.18 Різниця між апроксимованим рішенням, отриманим з допомогою методу Бубнова-Гальоркіна, при $M=5$, та рішенням скінченно-елементної апроксимації

На *Рис. 3.19* та *Рис. 3.20* зображено похідні від отриманого рішення, їх сукупність показує градієнт тепла в пластині. Як вже було сказано, значення шуканої температури у вузлах повинно прямувати до точного. Ця умова ніяк не поширюється на значення її похідних. З останніх рисунків видно, що похідні мають розриви першого роду в міжелементних зонах, навіть при тому, що рішення прямує до точного.

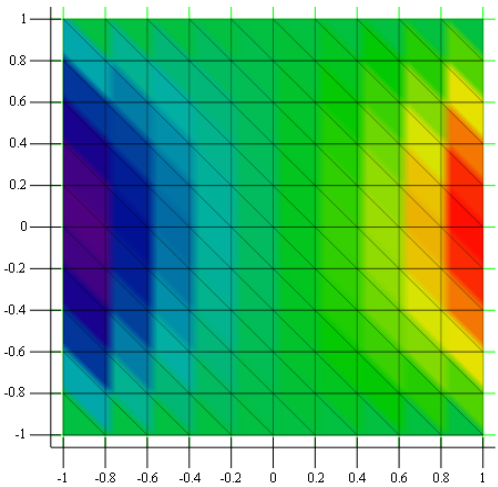


Рис. 3.19 Значення похідної $\partial\tilde{T}/\partial x$ від отриманого апроксимованого рішення

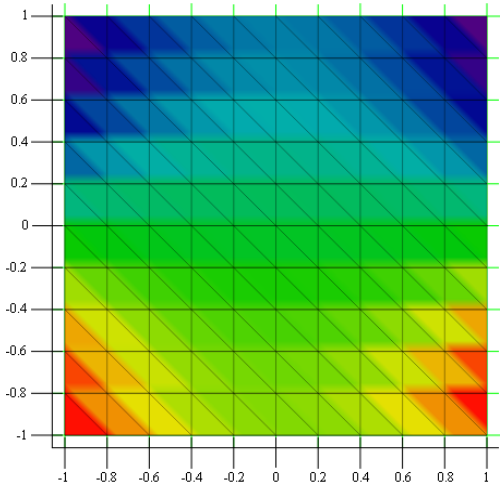


Рис. 3.20 Значення похідної $\partial\tilde{T}/\partial y$ від отриманого апроксимованого рішення

Така поведінка зумовлена природою використаної скінченно-елементної моделі – похідні в межах симплекс елементу завжди є константами (див. (3.136)). Тому, знову ж таки, щоб отримати достатньо точне рішення в зонах, де присутні великі градієнти, потрібно використовувати багато малих за розмірами скінченних елементів.

3.7. Теоретичні властивості

Не знаючи точного рішення крайової задачі, **неможливо** в загальному випадку **обчислити точність** отриманого апроксимованого рішення [16]. У таких випадках *оцінку* рішення, тобто межі в яких розміщена похибка, шукають спираючись на *апріорну оцінку*[1] точності, при якій аналізується функція, що апроксимується, та сам метод апроксимації, або спираючись на *апостеріорну оцінку*[2] точності, при якій порівнюються результати отримані з використанням різних методів апроксимації чи результати апроксимації, отримані одним і тим ж методом при різних обчислювальних параметрах.

Обидві оцінки точності, для будь-якого чисельного методу апроксимації, вимагають проведення аналізу *стійкості* та *збіжності* обчислень в моделі, що відповідає задачі.

Всі дослідження фізичних процесів з застосуванням чисельних методів містять в собі похибки, спричинені трьома обставинами:

- шукане рішення заміняється деяким наближенням, похибка такого наближення називається *похибкою апроксимації*;

[1] Від латинського "*a priori*" – буквально "*від попереднього*", тобто знання, отримані до досвіду і незалежно від нього, іншими словами те, що наперед відомо.

[2] Від латинського "*a posteriori*" – буквально "*від наступного*", тобто знання, що випливають з досвіду, антонім до "*a priori*".

- обчислення здійснюються засобами, що здатні оперувати числами скінченної розрядності внаслідок чого виникає *обчислювальна похибка*;
- фізико-математична модель лише приблизно описує реальний фізичний процес.

Останній пункт зазвичай не розглядається в літературі по чисельним методам, оскільки це компетенція іншої наукової дисципліни і при фундаментальних дослідженнях такою дисципліною може стати навіть філософія.

Стійкість чисельного методу визначається ростом помилок при виконанні окремих обчислювальних операцій. Нестійкі обчислення є результатом заокруглення чи інших помилок, які необмежено накопичуються, внаслідок чого точне рішення швидко тоне в помилках.

Збіжність чисельного методу це поступове наближення послідовно обчислених результатів до гранично-точного результату, по мірі того, як уточнюються деякі обчислювальні параметри. В обчислювальному процесі, що збігається, різниця результатів між ітераціями поступово зменшується і в границі прямує до нуля. З *Рис. 3.21* видно, що по мірі уточнення деяких обчислювальних параметрів точність росте, якщо процес обчислень збігається і падає якщо процес обчислень є незбіжним.

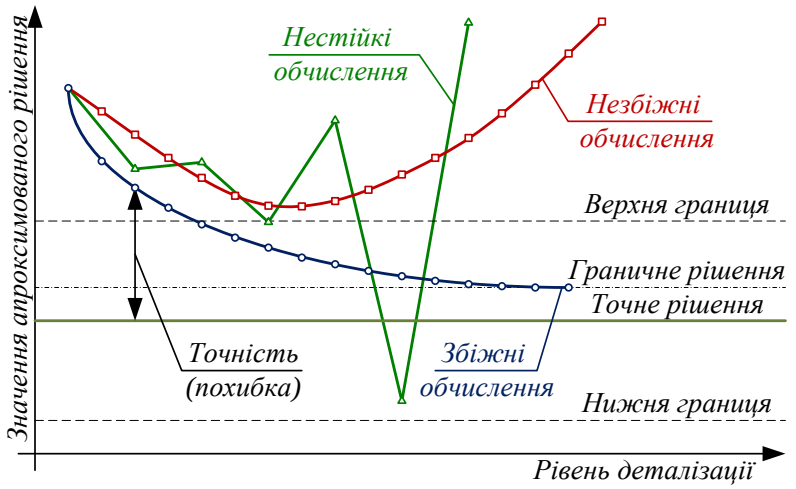


Рис. 3.21 Точність стійкість та збіжність чисельних методів

При рішенні задач, що можуть бути описані еквівалентною варіаційною постановкою [5], властивості збіжності, що відповідають методу Релея-Рітца, поширюються і на методи Бубнова-Гальоркіна, і як наслідок, на метод скінченних елементів. Як вже було сказано, для методу Релея-Рітца, що ефективно застосовується для задач механіки, вже є добре розвинута математична база, на яку ми і будемо опиратися.

Припустимо, що нам необхідно розв'язати операторне рівняння:

$$\mathcal{A}(u) = f, \qquad (3.165)$$

де оператор є симетричним, тобто для довільних елементів одного і того ж простору u та v:

$$\langle \mathcal{A}(u), v \rangle = \langle u, \mathcal{A}(v) \rangle, \tag{3.166}$$

та позитивно визначений, тобто для довільного елементу u:

$$\langle \mathcal{A}(u), u \rangle > 0. \tag{3.167}$$

Можна показати [5], [10], що рівняння (3.165) має єдине рішення (теорема про існування та єдиність рішення за Адамаром). Крім того, задача рішення цього рівняння може бути замінена задачею знаходження функції u, що мінімізує функціонал[1]:

$$\mathcal{F}(u) = \langle \mathcal{A}(u), u \rangle - 2\langle u, f \rangle. \tag{3.168}$$

По аналогії зі скалярним добутком (3.6) вводиться поняття енергетичного добутку, що зв'язаний з оператором $\mathcal{A}$, і визначається як:

$$[u, v] \equiv \langle \mathcal{A}(u), v \rangle. \tag{3.169}$$

У літературі такий скалярний добуток також часто позначають як:

$$\langle \mathcal{A}(u), v \rangle = \langle u, v \rangle_{\mathcal{A}}. \tag{3.170}$$

Маючи апроксимоване рішення рівняння $\tilde{u}$, вираз (3.168) можна переписати у вигляді:

$$\begin{aligned} \mathcal{F}(\tilde{u}) &= \langle \tilde{u}, \tilde{u} \rangle_{\mathcal{A}} - 2\langle u, \tilde{u} \rangle_{\mathcal{A}} = \\ &= \langle u - \tilde{u}, u - \tilde{u} \rangle_{\mathcal{A}} - \langle u, u \rangle_{\mathcal{A}} = \\ &= \|u - \tilde{u}\|_{2,\mathcal{A}} - \|u\|_{2,\mathcal{A}}, \end{aligned} \tag{3.171}$$

де, $\|.\|_{2,\mathcal{A}}$ – позначає енергетичну норму оператора $\mathcal{A}$, що визначається аналогічно до (3.8), як:

$$\|u\|_{2,\mathcal{A}} = \langle \mathcal{A}(u), u \rangle^{\frac{1}{2}}. \tag{3.172}$$

Очевидно, коли пробне рішення $\tilde{u}$ є рівним точному рішенню u, функціонал $\mathcal{F}(\tilde{u})$ має мінімальне значення, при чому це значення пропорційне енергії системи.

За визначенням, енергетична норма $\|u\|_{2,\mathcal{A}}$ є скінченною, якщо оператор $\mathcal{A}$ є позитивно визначений і обмежений знизу, а також якщо вільний член f має скінченну норму [4], [5]. Це означає, що:

$$\langle \mathcal{A}(u), u \rangle \geq \gamma \|u\|_2^2, \tag{3.173}$$

де, γ – деяка додатня константа. У такому випадку, послідовність функцій $u_1, u_2, \ldots, u_k$ мінімізує функціонал, коли:

[1] Функціонал, на відміну від оператора, ставить у відповідність кожному елементу множини деякий, не обов'язково один і тільки один, елемент іншої множини, при чому кілька елементів першої можуть відповідати одному і тому ж елементу останньої. Таким чином оператор є частковим випадком функціоналу [5].

$$\lim_{k\to\infty}\mathcal{F}(u_k)=\inf(\mathcal{F}(u)), \tag{3.174}$$

де, $\inf(.)$ – найбільша нижня границя $\mathcal{F}(u)$. Будь-яка послідовність $u_1,u_2,\ldots,u_k$ що відповідає умові (3.174), збігається по енергії до рішення рівняння (3.165). Збіжність по енергії означає, що u_k збігається до точного рішення u, якщо:

$$\lim_{k\to\infty}\|u-u_k\|_{2,\mathcal{A}}\to\varepsilon, \tag{3.175}$$

де, ε – довільно вибрана мала додатня константа.

Доведено [5], що метод Релея-Рітца дозволяє отримати послідовність функцій $u_1,u_2,\ldots,u_k$, яка збігається по енергії до точного рішення u, при умові, що u – є рішенням зі скінченною енергією. При доведенні збіжності методу Релея-Рітца, виявляється що пробні функції в формулі (3.17) повинні задовольняти двом умовам:

- послідовність пробних функцій $\varphi_1,\varphi_2,\ldots,\varphi_j,\ldots,\varphi_M$ повинна бути *повною* по енергії;
- всі функції φ_j повинні бути лінійно незалежними;

Перша умова гарантує, що послідовність обраних пробних функцій взагалі здатна апроксимувати точне рішення. Вважається, що така послідовність є повною, коли лінійна комбінація:

$$\lim_{M\to\infty}\sum_{j=1}^{M}a_j\varphi_j\to u. \tag{3.176}$$

Тобто, в деякому сенсі збігається до точного рішення u, при кількості функцій, що прямує до безмежності[1] [16]. За допомогою теореми Стоуна-Вейєрштрасса можна довести, що поліноміальні ряди є повними, тобто здатні апроксимувати деяку неперервну функцію на визначеному відрізку. За деталями слід звернутися до літератури по функціональному аналізу.

---

[1] Таке твердження випливає теорії функціонального аналізу, а саме з поняття послідовності Коші, або фундаментальної послідовності, члени якої наближаються як завгодно близько один до одного зі збільшенням порядкових номерів. Формально послідовність точок $\{x_n\}$ в лінійному метричному просторі називається послідовністю Коші, якщо для будь-якого $\varepsilon>0$ знайдеться таке $N=N(\varepsilon)$, що при всіх $n,m\geq N$:

$$\|x_n-x_m\|<\varepsilon.$$

Послідовність називається збіжною послідовністю, якщо в цьому просторі існує така точка x, що для кожного $\varepsilon>0$ знайдеться таке $N=N(\varepsilon)$, що при всіх $n\geq N$:

$$\|x-x_n\|<\varepsilon.$$

Простір, для якого всі послідовності Коші є збіжними, називається повним (кожна послідовність збігається до елементу того ж простору). Повний лінійний простір, з визначеним в ньому скалярним добутком називається Гільбертовим простором. В проекційному методі апроксимація будується як ортогональна проекція шуканої функції в функціональний Гільбертовий простір, базис якого утворений з системи пробних функцій φ_j. Тому апроксимацію можна побудувати тоді, коли з обраної системи пробних функцій можна утворити Гільбертовий простір, або іншими словами, коли послідовність функцій є повною.

Очевидно, що поліноміальний ряд, у загальному випадку може дати точне рішення тільки тоді, коли він має безмежну степінь[1]. На практиці можливо використовувати тільки скінченну кількість доданків, тому рішення завжди буде наближеним. Проаналізуємо при яких умовах похибка такого рішення збігатиметься до нуля. Нехай диференціальний оператор $\mathcal{A}$ має порядок $2p$, тобто шуканими значеннями є потенціал u та всі його похідні до $\partial^{2p}u/\partial\mathbf{r}^{2p}$ включно. Щоб апроксимувати це рішення, необхідно використовувати поліном, як мінімум порядку $2p$, якщо похідна порядку $2p$ відмінна від нуля, наприклад:

$$\begin{aligned}
\tilde{u}(x) &= a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_{2p} x^{2p},\\
\frac{d\tilde{u}(x)}{dx} &= a_1 + 2a_2 x + 3a_3 x^2 + \ldots + 2m a_{2p} x^{2p-1},\\
\frac{d^2\tilde{u}(x)}{dx^2} &= 2a_2 x + 6a_3 x + \ldots + 2p(2p-1) p a_{2p} x^{2p-2},\\
&\ldots\\
\frac{d^{2p-1}\tilde{u}(x)}{dx^{2p-1}} &= (2p-1)!\, p a_{2p-1} + (2p)!\, a_{2p} x,\\
\frac{d^{2p}\tilde{u}(x)}{dx^{2p}} &= (2p)!\, a_{2p}.
\end{aligned} \qquad (3.177)$$

З останнього відношення видно, що обираючи для апроксимації поліноми степеня не нижчого $2p$, кожна з похідних починає прямувати до свого точного значення. Далі, при збільшенні степені поліному, слід очікувати зменшення похибки апроксимованого рішення та збіжності його до точного рішення, навіть при наявності обчислювальної похибки [6].

У ряді задач, де визначені природні крайові умови, за допомогою процедури пониження порядку в рівнянні методу зважених нев'язок (3.47), в загальному випадку можна перенести половину порядку похідних з пробних функцій на повірочні. Тобто, для апроксимації задач, що визначаються диференціальними рівняннями в слабкій формі є *допустимим* використання поліномів порядку не нижчого від p, де $2p$ – порядок рівняння.

Рішення Релея-Рітца, що мінімізує функціонал (3.168) співпадає з рішенням методів Бубнова-Гальоркіна рівняння (3.165). Як наслідок, для класу задач, що описуються даним рівнянням, властивості збіжності, що відповідають рішенню Релея-Рітца, відносяться також і до рішень Бубнова-Гальоркіна. А за необхідності, подібні судження можна розширити і на всі методи зважених нев'язок. Очевидно, що чим складніше диференціальне рівняння, тим важче визначити межі похибки його рішення. Проте нев'язку рівняння, що отримана шляхом підстановки в нього пробного рішення, визначити не складно. Як наслідок, виникає можливість пов'язати апостеріорну оцінку точності з

[1] Єдиним винятком є випадок, коли шукана функція сама є поліномом скінченного порядку – тоді можна отримати точне рішення задачі.

відповідною нормою по відношенню до нев'язки. Зауважимо, що така оцінка зазвичай є дуже заниженою, наприклад аналізуючи результати апроксимації з *Таблиця 3.1*, не важко помітити, що при збільшенні числа базисних функцій, норма похибки значно швидше збігається до нуля, ніж відповідна нев'язка.

3.8. Список використаної літератури до розділу 3

[1] Щеглов И. – Дискретизация сложных двумерных и трехмерных областей для решения задач математического моделирования / автореф. // Москва: МГТУ, 2010.

[2] Segerlind L. – Applied Finite Element Analysis / Применение метода конечных элементов / пер. с англ. Шестаков А., под. ред. Победри Б. // Москва: Мир, 1979.

[3] Zienkiewicz O., Morgan K. – Finite elements and approx. // New-York: Wiley, 1983.

[4] Fletcher C. – Computational Galerkin Methods / Численные методы на основе метода Галёркина / пер. с англ. под ред. Шидловский В. // Москва: Мир, 1988.

[5] Михлин С. – Вариационные методы в мат. физике // Москва: Наука, 1970.

[6] Strang G., Fix G. – An Analysis of the Finite Element Method. / Теория метода конечных элементов / пер с англ. под ред. Марчука Г. // Москва: Мир, 1977.

[7] Гантмахер Ф. – Теория матриц. 2-е изд., доп. // Москва: Наука, 1966.

[8] Винберг Э. – Курс Алгебры. 2-е изд. // Москва: Факториал Пресс, 2001.

[9] Гельфанд И. – Лекции по линейной алгебре. 4-е изд., доп. // Москва: Наука, 1971.

[10] Ладыженская О. – Краевые задачи математической физики // Москва: Наука, 1973.

[11] Banach S. – Rachunek Rozniczkowy i Calkowy / Дифференциальное и интегральное исчисление. 2-е изд. / пер. с польск. Зуховицкий С. // Москва: Наука, 1966.

[12] Тихонов А., Самарский А. – Уравнения математической физики: Учебное пособие, 6-е изд. испр. и доп. // Москва: МГУ, 1999.

[13] Годунов С. – Уравнения математической физики. 2-е изд. // Москва: Наука, 1979.

[14] Thomee V. – Galerkin Finite Element Methods for Parabolic Problems. 2-nd ed. // New-York: Springer, 2006.

[15] Knabner P., Angerman L. – Numerical Methods for Elliptic and Parabolic Partial Differential Equations // New-York: Springer, 2003.

[16] Norrie D., Vries G. – An Introduction to Finite Element Analysis // New-York: Academic press, 1978.

[17] Александров П., Пасынков Б. – Введение в теорию топологических пространств и общую теорию размерности // Москва: Наука, 1973.

[18] Александров П. – Введение в теор. множ. и общ. топ. // Москва: Наука, 1977.

[19] Hanson A. – Geom. for N-Dimensional Graphics // New-York: Academic Press, 1994.

[20] Liberti L., Lavor C., Maculan N., Mucherino A. – Euclidean distance geometry and applications // Tech. Report, arXiv.12050349, 2012.

[21] Д'Андреа К., Сомбра М. – Определитель Кэли-Менгера неприводим при $n \geq 3$ // Сибирский математический журнал, Том 46, №1, сс. 92-97, 2005.

[22] [Electronic resource] Math Pages – Simplex Volumes and the Cayley-Menger Determinant, http://www.mathpages.com/home/kmath664/kmath664.htm.

[23] Eisenberg M., Malvern L. – On finite element integration in natural coordinates // Int. Journal for Numerical Methods in Engineering, 7(4):574-575, 1973.

[24] Silvester P., Ferrari R. – Finite Elements for Electrical Engineers / Метод конечных элементов для радиоинженеров и инженеров-электриков / пер. с англ. Хотяинцева С., под ред. Дубровка Ф. // Москва: Мир, 1986.

4. Застосування МСЕ на компонентному рівні проектування МЕМС

4.1. Фізичні аналогії скінченно-елементної моделі

Як вже було сказано, метод скінченних елементів вперше з'явився в 50-их рр. XX століття лише як чисельна процедура рішення задачі пошуку плоских напружень. Метод був запропонований інженерами та завоював велику популярність, оскільки його початкові формулювання будувалися без зайвих, для практичних інженерних розрахунків, складних математичних викладок. Натомість, використовувалися безпосередні інтерпретації неперервних фізичних задач, як взаємозв'язок примітивних елементів аналогічних *дискретних систем*[1] [1], [2], [3], [4], методи дослідження яких добре відомі інженерам.

Знову повертаючись до витоків теорії методу скінченних елементів, далі буде показано взаємозв'язок його моделей з цими дискретними системами на основі методу аналогій та теорії подібності. Це дасть можливість зрозуміти безпосередній фізичний зміст скінченно-елементних моделей.

Не вдаючись в деталі процесу моделювання, дослідження дискретних систем складається з таких основних етапів [5]:

- ідеалізація системи: реальна система ідеалізується як сукупність окремих елементів;
- приведення балансу елементів: виведення залежностей, що описують рівняння балансу змінних стану[2] реальної системи в межах окремих елементів;
- ансамблювання: об'єднання певним чином всіх елементів, з метою отримання можливості описувати поведінку системи одночасним рішенням множини всіх рівнянь балансу;
- обчислення відгуку моделі: одночасне обчислення множини всіх рівнянь балансу та отримання значень змінних стану системи, як реакцію на зовнішні чинники.

В скінченно-елементній моделі невідомими є вузлові значення шуканих величин, таких як переміщення вздовж осей координат, температура, електричний потенціал тощо. Позначивши кількість усіх вузлів як G, отримаємо $\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_G$ невідомих.

У залежності від конкретної задачі, між сусідніми вузлами встановлюється взаємозв'язок максимально простим, переважно лінійним способом (використовується мінімальний носій), наприклад це може бути закон Гука, який описує поведінку пружини з заданою жорсткістю, закон Фур'є, який описує поведінку теплового потоку в матеріалі з заданим коефіцієнтом теплопровідності, чи закон Ома, який описує поведінку струму, що долає ділянку кола з заданим опором. Позначивши кількість вузлів скінченного елементу як M, абстрактно ці явища можна описати набором сил або потоків в

[1] В літературі також часто зустрічається назва "*системи з зосередженими параметрами*".

[2] Термін "*змінні стану*" прийшов з області системного аналізу та дуже часто використовується при описі термодинамічних, чи більш загально, фізичних систем.

межах елементу $\mathbf{J}_1, \mathbf{J}_2, \ldots \mathbf{J}_M$. Ці сили/потоки є лінійними функціями від вузлових значень та описуються залежностями [1]:

$$\mathbf{J}_j = \mathbf{K}_{j,g}\mathbf{u}_g - \mathbf{f}_j \quad 1 \le g \le G, \quad 1 \le j \le M. \tag{4.1}$$

Рівняння для всієї системи отримуються простим додаванням потоків $\mathbf{J}_1, \mathbf{J}_2, \ldots \mathbf{J}_M$ по елементах. Тому для лінійного прикладу всі P рівнянь будуть мати вигляд:

$$\sum_{i=1}^{P} \mathbf{J}_{i,j} = \sum_{i=1}^{P} \mathbf{K}_{i,j,g}\mathbf{u}_{i,g} - \sum_{i=1}^{P} \mathbf{f}_{i,j}, \tag{4.2}$$

і, як наслідок, рівняння повної системи можуть бути записані в стандартній формі:

$$[\mathbf{K}]\{\mathbf{u}\} = \{\mathbf{f}\}. \tag{4.3}$$

Напевно найбільш відомим прикладом таких моделей є задача дослідження складних механічних конструкцій, що складаються з простих елементів [2], [3], [5]. Таку систему можна інтерпретувати як дискретну систему взаємозв'язаних пружин (*Рис. 4.1*). Тоді, у нашому формулюванні невідомими є переміщення $\mathbf{u} = \{u, v\}^{\mathrm{T}}$, а $\mathbf{J}$ – це механічні сили, що діють на вузли зі сторони сусідніх елементів. Беручи за основу плоскі лінійні напруження, можна записати:

$$\begin{Bmatrix} \mathbf{J}_u \\ \mathbf{J}_v \end{Bmatrix}_i = [\mathbf{K}]_i \cdot \{\mathbf{u}\}_i + \begin{Bmatrix} \mathbf{f}_u \\ \mathbf{f}_v \end{Bmatrix}_i, \tag{4.4}$$

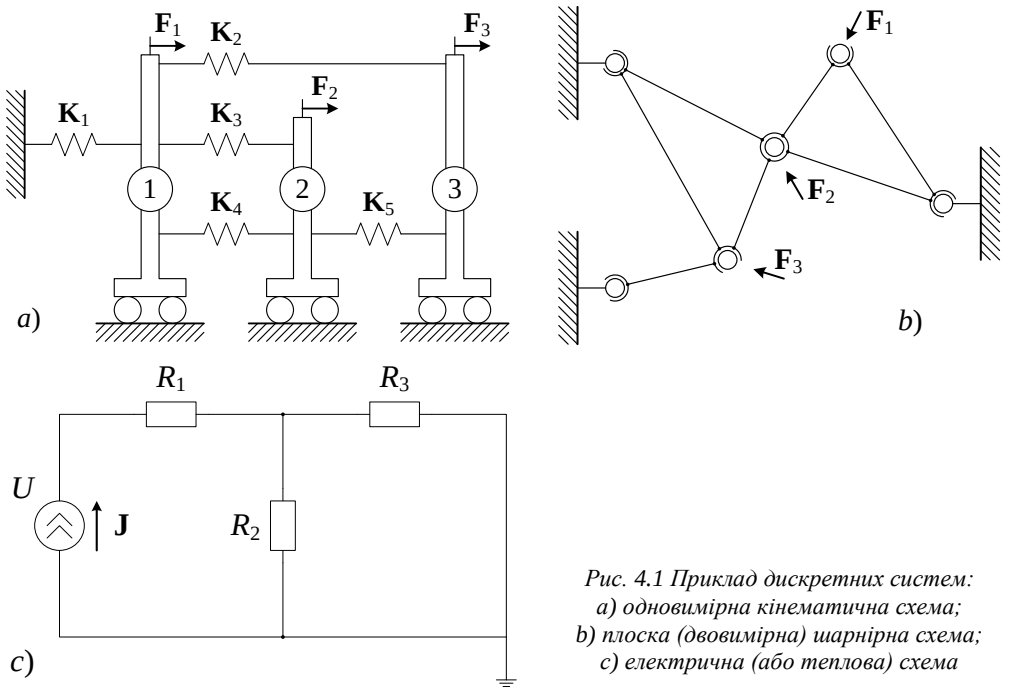


Рис. 4.1 Приклад дискретних систем:
a) одновимірна кінематична схема;
b) плоска (двовимірна) шарнірна схема;
c) електрична (або теплова) схема

де $[\mathbf{K}]$ – описує жорсткість пружин, тобто є матрицею жорсткості, $\{\mathbf{f}\}$ – вектор сил у вузлі, що необхідні для балансу деякого розподіленого на елементі напруження. Якщо конструкція знаходиться під дією деяких зовнішніх сил $\{\mathbf{F}\}$, що прикладені до одного з вузлів, то для балансу необхідно, щоб:

$$\mathbf{F}_g = \sum_{i=1}^{P} \mathbf{J}_{i,g}, \tag{4.5}$$

де не дорівнюють нулю тільки компоненти елементів, що включають вузол, який розглядається. Комбінуючи останні рівняння та ансамблюючи систему, отримаємо:

$$[\mathbf{K}]\{\mathbf{u}\} = \{\mathbf{f}\} + \{\mathbf{F}\}. \tag{4.6}$$

Аналогічними є судження й для інших дискретних систем.

Якщо говорити більш абстрактно, то застосування методу аналогій [6] є дуже корисним при аналізі в недосліджених областях. За допомогою аналогій невідома система може порівнюватися з раніше дослідженою системою. А в більш повно дослідженій системі, взаємодія її елементів є більш наглядною, і відомі методи досліджень застосовуються з більшим успіхом.

Аналогії, що існують між електричними, механічними, акустичними та іншими системами, давно та успішно використовуються фізиками та інженерами в дослідженнях та обчисленнях. Метод аналогій дає змогу значно спростити математичні викладки та робить більш зрозумілими як проміжні етапи досліджень, так і їх результати. Перевага цього методу проявляється, в першу чергу, при аналізі складних систем, що складаються з великої кількості елементів де одночасно протікають різні фізичні процеси [7].

Основна ідея методу аналогій пов'язана з введенням на основі математично записаних фізичних законів для аналогічних параметрів (в електричних, теплових, механічних та інших) фізичних систем, формальних позначень (наприклад потоків $\mathbf{J}$) що відрізняються тільки індексами. Це звичайно зумовлює вимушений відхід від стандартизованих понять з конкретних областей, але, з іншої сторони, дозволяє єдиним чином описувати предмет дослідження.

Формальну фізико-математичну теорію, що має за мету, з точки зору системного аналізу, об’єднати єдиним чином аналогічні явища та дати можливість їх досліджувати однаковим чином, не вникаючи при цьому у фундаментальні дослідження самих явищ, почали розвивати на основі термодинаміки на початку XX століття. Значний поштовх вперед було зроблено в 1930-их, коли в термодинаміці вперше було запропоновано принцип найменшого розсіювання (дисипації) енергії, та об’єднано різні лінійні фізичні процеси, за допомогою, так званих, кінетичних коефіцієнтів, або коефіцієнтів Онзагера[1,2]. Пізніше в 1960-1970-их з’явилися роботи, що на основі математичної теорії поля, якою тут користуємося, та методу аналогій,

[1] Onsager L. // Phys. Rev., 37:405; 38:2265, 1931.
[2] Onsager L., Fuoss J. // Journ. Phys. Chem., 36:2689, 1932.

формально описували фактично всю лінійну термодинаміку та відповідні похідні фізичні процеси[1]. Логічним продовженням розвитку такої теорії вже на початку XXI століття стали роботи[2], що максимально абстрактно, але все ще безвідривно від фізичної суті, описують предметні явища переносу або перетворення енергії, незалежно від тієї чи іншої області науки. Коротко, така теорія дістала назву Енергодинаміка.

Теоретично, метод аналогій базується на теорії *подібності* [8], перші строгі формулювання якої з'явилися завдяки Ньютону, ще в 1687 році[3]. Подібність аналогічних явищ полягає в однаковому характері протікання всіх процесів. Математично аналогічні явища описуються формально однаковими диференціальними рівняннями та умовами однозначності. Однак фізичний зміст і розмірність вхідних величин різні. Більш строго, *подібність* – це взаємно-однозначна відповідність між двома об'єктами, коли відомі функції переходу від параметрів одного об'єкта до параметрів іншого, а математичні описи цих об'єктів можуть бути тотожними.

Теорія подібності формулює властивості аналогічних систем, стверджуючи, що подібні явища мають однакові *критерії подібності*. Тобто безрозмірні набори величин, що характеризують середню міру відношення інтенсивності фізичних явищ, важливих для досліджуваного процесу. Ці критерії встановлюються з умов тотожності рівнянь для фізичних процесів, або на основі аналізу формальних розмірностей, що використовуються в моделях.

Для подібності властиві деякі загальні закономірності, які прийнято називати першою та другою теоремами подібності, а також додатковими положеннями до них. Ці додаткові положення необхідні при дослідженні подібності явищ в складних нелінійних, в тому чи іншому сенсі неоднорідних чи стохастичних системах. Обидві теореми встановлюють співвідношення між параметрами подібних явищ, не звертаючи уваги, при цьому, на реалізацію подібності при побудові моделей. Для останнього застосовується третя теорема подібності (або обернена теорема), що визначає необхідні і достатні умови для того, щоб явища виявилися подібними. Теорема вимагає подібності умов однозначності та такого підбору параметрів моделі, при якому критерії подібності, що містять крайові умови, стають однаковими.

Перед тим, як продовжити, розглянемо взаємозв'язок різних систем одиниць вимірювання фізичних величин. Виміряти деяку фізичну величину Θ означає порівняти її з іншою величиною θ тієї самої фізичної природи, тобто визначити, у скільки разів Θ більше або менше θ. Щоб уникнути непорозумінь, для Θ та θ прийнято певний зміст, чи більш конкретно, семантику, відповідно до тієї фізичної природи, де вони розглядаються. Цей умовний зміст називається одиницею вимірювання.

---

[1] Gyarmati I. – Non-Equilibrium Thermodynamics. Field Theory and Variational Principles // New-York: Springer, 1970.

[2] Эткин В. – Энергодинамика (синтез теорий переноса и преобразования энергии) // Санкт-Петербург: Наука, 2008.

[3] Newton Is. – Philosophiæ Naturalis Principia Mathematica // Londini, 1687.

При дослідженні різних явищ природи розвиваються формальні теорії, що можуть оперувати новими одиницями вимірювання. В залежності від явищ, ці одиниці бувають незалежними від інших одиниць, або ж утворюються на їх основі. У теперішньому світі існують різні системи таких одиниць вимірювання. Тут використовуємо міжнародну систему одиниць СІ (англ. SI), що є метричною. СІ побудована на основі семи базових одиниць вимірювання[1]: метр [м], кілограм [кг], секунда [с], ампер [А], кельвін [°К], моль [моль], кандела [кд].

Формула, за якою визначається залежність, між похідними та основними одиницями, називається розмірністю величини. У загальному випадку ця формула виражається як добуток степенів базових одиниць вимірювання, наприклад для другого закону Ньютона:

$$\mathbf{F} = m\mathbf{a} = m \cdot \frac{d\mathbf{v}}{d\tau} = \frac{d^2\mathbf{r}}{d\tau^2} \quad \Rightarrow \quad [m]^1 \cdot [\mathbf{r}]^1 \cdot [\tau]^{-2} = \frac{\text{кг} \cdot \text{м}}{\text{с}^2} = \text{Н}, \qquad (4.7)$$

де квадратні дужки означають взяття одиниці вимірювання конкретної фізичної величини. Слід зазначити, що залежності між одиницями вимірювання є справедливими на будь-яких (великих чи малих) масштабах. Тому, якщо відповідні формули містять інтегральні, диференціальні, трансцендентні або інші вирази, то, оскільки вони не мають розмірності, при виведенні похідних одиниць, а також встановлення умов подібності, їх можна опустити [8].

При розгляді конкретних задач, зазвичай неявно вибирається деяка підсистема одиниць вимірювання, і всі обчислення здійснюються на її основі. Наприклад, відштовхуючись від рівняння стаціонарної теплопровідності, в обчисленнях присутні тільки температура в градусах Кельвіна [°К], чи її похідна в градусах Цельсія [°С], відстань в метрах [м], а також, неявно, маса [кг] і час [с], що є базовими для коефіцієнту теплопровідності [Вт/м°С], де [Вт = кг м^2/с^3].

Якщо за основу взято деякий набір одиниць $P_1, P_2, \ldots, P_K$, як можна замість них вибрати стільки ж похідних одиниць $R_1, R_2, \ldots, R_K$, що утворили б нову систему одиниць вимірювання? Це можливо тоді і тільки тоді, коли розмірності $[R_1], [R_2], \ldots, [R_K]$ є лінійно незалежними функціями від розмірностей $[P_1], [P_2], \ldots, [P_K]$, і є можливість побудувати взаємно однозначне відображення між просторами P і R, тобто єдиним чином виразити $[R_1], [R_2], \ldots, [R_K]$ через $[P_1], [P_2], \ldots, [P_K]$. Запишемо розмірності похідних величин:

$$\begin{aligned} [R_1] &= [P_1]^{\alpha_1} \cdot [P_2]^{\beta_1} \cdot \ldots \cdot [P_K]^{\omega_1}, \\ [R_2] &= [P_1]^{\alpha_2} \cdot [P_2]^{\beta_2} \cdot \ldots \cdot [P_K]^{\omega_2}, \\ &\ldots \\ [R_N] &= [P_1]^{\alpha_N} \cdot [P_2]^{\beta_N} \cdot \ldots \cdot [P_K]^{\omega_N}. \end{aligned} \qquad (4.8)$$

[1] Див. міжнародне бюро з мір та ваг: http://www.bipm.org/en/measurement-units/base-units.html.

Виразимо з останньої системи степені:

$$\begin{aligned}\ln[R_1] &= \alpha_1 \ln[P_1]\cdot \beta_1 \ln[P_2]\cdot \ldots \cdot \omega_1 \ln[P_K],\\ \ln[R_2] &= \alpha_2 \ln[P_1]\cdot \beta_2 \ln[P_2]\cdot \ldots \cdot \omega_2 \ln[P_K],\\ &\ldots\\ \ln[R_K] &= \alpha_K \ln[P_1]\cdot \beta_K \ln[P_2]\cdot \ldots \cdot \omega_K \ln[P_K].\end{aligned} \tag{4.9}$$

Остання система має єдиний розв'язок тоді і тільки тоді, коли:

$$\left\| \begin{bmatrix} \alpha_1 & \beta_1 & \cdots & \omega_1 \\ \alpha_2 & \beta_2 & \cdots & \omega_2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_N & \beta_N & \cdots & \omega_N \end{bmatrix} \right\| \neq 0. \tag{4.10}$$

Тобто, в такому випадку $[P_1],[P_2],\ldots,[P_K]$ та $[R_1],[R_2],\ldots,[R_K]$ єдиним чином виражаються одні через одних.

Наприклад, для деякої механічної системи, де базовими одиницями є маса m, відстань $\mathbf{r}$ та час τ, у якості базових одиниць можна вибрати силу $\mathbf{F}$, відстань $\mathbf{r}$ та час τ:

$$\begin{aligned}[\mathbf{F}] &= [m]^1\cdot[\mathbf{r}]^1\cdot[\tau]^{-2}\\ [\mathbf{r}] &= [m]^0\cdot[\mathbf{r}]^1\cdot[\tau]^0\\ [\tau] &= [m]^0\cdot[\mathbf{r}]^0\cdot[\tau]^1\end{aligned} \quad \Rightarrow \quad \left\| \begin{bmatrix} 1 & 1 & -2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right\| = 1 \neq 0. \tag{4.11}$$

Повернемося до розгляду критеріїв подібності. Нехай деяка задача розглядає множину розмірних параметрів $P_1, P_2, \ldots, P_N$. З цих параметрів обрано $\Theta_1, \Theta_2, \ldots, \Theta_S$ незалежних, кожен з яких є функцією деяких початкових розмірних параметрів $P_1, P_2, \ldots, P_N$:

$$\Theta_i = f_i(P_1, P_2, \ldots, P_N). \tag{4.12}$$

Сукупність цих функцій однозначно описує задачу, що розглядається, або говорячи абстрактно, однозначно визначає положення системи в деякому функціональному просторі. Кожне з таких рівнянь утворює, так звану, *узагальнену координату* в побудованому функціональному просторі. Також приймемо, що $P_1, P_2, \ldots, P_K$ початкових параметрів утворюють базову систему одиниць вимірювання для цієї задачі, а $(N-K)$ з них є похідними.

Нехай задача розглядає деяку залежність між параметрами, для спрощення приймемо у якості такої, рівняння:

$$\Theta_1 + \Theta_2 + \ldots + \Theta_N = 0 \Rightarrow \sum_{i=1}^{N} \Theta_i = 0. \tag{4.13}$$

Для більш складних виразів судження не змінюються, оскільки залежності між одиницями вимірювання є справедливими на будь-яких ділянках, та не змінюють розмірностей. Приведемо рівняння до безрозмірного виду[1], поділивши параметри на один з довільно вибраних, нехай це буде останній:

[1] Такий метод відомий як метод аналізу рівнянь процесу або спосіб інтегральних аналогів.

$$\frac{\Theta_1}{\Theta_S}+\frac{\Theta_2}{\Theta_S}+\ldots+\frac{\Theta_{S-1}}{\Theta_S}+1=0\Rightarrow\sum_{i=1}^{S-1}\frac{\Theta_i}{\Theta_S}+1=0. \qquad (4.14)$$

Розглянемо іншу задачу, що оперує множиною незалежних розмірних параметрів $\Psi_1,\Psi_2,\ldots,\Psi_S$, кожен з яких є функцією деяких початкових розмірних параметрів:

$$\Psi_i=\Psi_i(R_1,R_2,\ldots,R_N). \qquad (4.15)$$

Приймемо, що задачі є подібними. При цьому формально вважаємо, що дві задачі (або більш загально системи) є *подібними*, якщо їх відповідні узагальнені координати є пропорційними відносно деякого масштабного множника. Першу задачу назвемо оригіналом або натурою, а другу задачу – моделлю. Для моделі відповідно отримаємо:

$$\frac{\Psi_1}{\Psi_S}+\frac{\Psi_2}{\Psi_S}+\ldots+\frac{\Psi_{S-1}}{\Psi_S}+1=0\Rightarrow\sum_{i=1}^{S-1}\frac{\Psi_i}{\Psi_S}+1=0. \qquad (4.16)$$

Оскільки модель є подібна оригіналу за визначенням, їх параметри описуються деяким лінійним масштабним відношенням:

$$P_1=m_1R_1,P_2=m_2R_2,\ldots,P_N=m_NR_N. \qquad (4.17)$$

Тобто:

$$\begin{aligned}\Theta_i=\Theta_i(P_1,P_2,\ldots,P_N)=\Theta_i(m_1R_1,m_2R_2,\ldots,m_NR_N)=\\ =M_i\Psi_i(R_1,R_2,\ldots,R_N)=M_i\Psi_i,\end{aligned} \qquad (4.18)$$

де M_i – деякі загальні масштабні множники. Оскільки задачі подібні, кожен з масштабних множників повинен бути рівним один одному, тобто:

$$\begin{gathered}M_1=M_2=\ldots=M_S,\\ \frac{M_1}{M_S}=\frac{M_2}{M_S}=\ldots=\frac{M_{S-1}}{M_S}=\frac{M_S}{M_S}=1.\end{gathered} \qquad (4.19)$$

Тепер можна записати відношення параметрів оригіналу і моделі як:

$$\frac{\Theta_1}{\Theta_S}=\frac{\Psi_1}{\Psi_S},\quad \frac{\Theta_2}{\Theta_S}=\frac{\Psi_2}{\Psi_S},\quad \ldots,\quad \frac{\Theta_{S-1}}{\Theta_S}=\frac{\Psi_{S-1}}{\Psi_S}. \qquad (4.20)$$

Кожне з таких відношень, якраз і є тією безрозмірною величиною, що називається *критерієм подібності*, та зазвичай позначається в літературі символом π. Якщо розглядається одразу декілька подібних систем, то можна записати:

$$\pi_i^{(1)}=\frac{\Theta_i^{(1)}}{\Theta_S^{(1)}}=\pi_i^{(2)}=\frac{\Theta_i^{(2)}}{\Theta_S^{(2)}}=\ldots=\pi_i^{(S)}=\frac{\Theta_i^{(G)}}{\Theta_S^{(G)}}=\text{idem}.^{1} \qquad (4.21)$$

Сформулюємо *першу теорему подібності* – у всіх подібних явищ критерії подібності однакові, або коротко: $\pi=\text{idem}$. Це достатні умови існування подібності.

[1] Від латинського "*identicus*" чи "*idem*" – буквально "*такий самий*", тобто однаковий, такий ж, ідентичний. В конкретному використанні означає "відповідно однаково для всіх".

Друга теорема подібності, також відома як π-теорема, стверджує, що будь-яке повне рівняння (узагальнена координата) фізичного процесу чи деякої системи, що записане у визначеній системі одиниць вимірювання, може бути записано у вигляді залежності між критеріями подібності, тобто рівнянням, що зв'язує безрозмірні величини, отримані на основі параметрів процесу. Іншими словами, завжди можна перетворити вираз типу (4.12) у вираз типу:

$$f(1,1,\ldots,1,\pi_1,\ldots,\pi_{N-K})=0. \tag{4.22}$$

Оскільки кожна узагальнена координата є комбінацією базових та похідних параметрів, то будь-який критерій подібності можна виразити як комбінацію базових параметрів[1]:

$$\begin{aligned}\pi_i &= \frac{\Theta_i(P_1,P_2,\ldots P_N)}{\Theta_S(P_1,P_2,\ldots P_N)} = P_1^{z_1}P_2^{z_2}\ldots P_N^{z_N} = c\cdot[P_1]^{z_1}[P_2]^{z_2}\ldots[P_N]^{z_N} = \\ &= c\cdot\left([P_1]^{\alpha_1}[P_2]^{\beta_1}\ldots[P_K]^{\omega_1}\right)^{z_1}\left([P_1]^{\alpha_2}[P_2]^{\beta_2}\ldots[P_K]^{\omega_2}\right)^{z_2}\ldots\left([P_1]^{\alpha_N}[P_2]^{\beta_N}\ldots[P_K]^{\omega_N}\right)^{z_N} = \\ &= c\cdot[P_1]^{\alpha_1 z_1+\alpha_2 z_2+\ldots+\alpha_N z_N}[P_2]^{\beta_1 z_1+\beta_2 z_2+\ldots+\beta_N z_N}\ldots[P_K]^{\omega_1 z_1+\omega_2 z_2+\ldots+\omega_N z_N}.\end{aligned} \tag{4.23}$$

Критерії подібності є безрозмірними, тому:

$$\begin{aligned}&\alpha_1 z_1+\alpha_2 z_2+\ldots+\alpha_N z_N=0,\\ &\beta_1 z_1+\beta_2 z_2+\ldots+\beta_N z_N=0,\\ &\ldots\\ &\omega_1 z_1+\omega_2 z_2+\ldots+\omega_N z_N=0.\end{aligned} \tag{4.24}$$

Останнє рівняння має N невідомих та K лінійно незалежних рівнянь, тобто існує тільки $(N-K)$ лінійно незалежних розв'язків, і, відповідно, тільки $(N-K)$ фундаментальних критерії подібності.

Третя теорема подібності встановлює необхідні і достатні умови для практичної реалізації подібності: щоб дві системи були подібними повинні бути відповідно однаковими базові критерії подібності та умови однозначності. При чому, під базовими критеріями розуміються ті критерії, що побудовані на базових, в конкретному випадку, параметрах. Умови однозначності визначають індивідуальні особливості процесу, виділяючи з множини всіх процесів даного класу конкретний. До них відносяться фактори і умови, що не залежать від механізму самого явища:

- геометричні властивості системи, де протікає процес;
- фізичні параметри середовища і тіл, що утворюють систему;
- початковий стан системи;
- крайові умови;
- взаємодія з зовнішнім середовищем.

У кожному конкретному випадку умови однозначності можуть бути різними, в залежності від роду задачі чи виду рівняння, яке її описує.

Наприклад, задача описує поведінку механічної системи, що складається з

[1] Такий метод відомий як метод визначення критерії подібності на основі аналізу розмірностей (π-теореми), або метод визначальних рівнянь.

вантажу масою m [кг] який коливається на пружині з жорсткістю k [кг с$^{-2}$] ($\mathbf{F} = -k\mathbf{r} \Rightarrow k = -\mathbf{F}/\mathbf{r}$, [$k$] = кг м с$^{-2}$ м$^{-1}$ = кг с$^{-2}$), під дією сили $\mathbf{F}$ [кг м с$^{-2}$], яка діє з частотою ω [с$^{-1}$], протягом часу τ [с]. Положення вантажу визначається як $\mathbf{r} = f(m,k,\mathbf{F},\omega,\tau)$ [м], що утворює узагальнену координату, яка однозначно описує положення системи. Базовими параметрами для цієї системи можна вибрати $\mathbf{r}$, m та τ, оскільки з комбінації їх розмірностей найпростіше вивести розмірності всіх інших параметрів. Запишемо систему лінійних рівнянь (4.24):

	m	k	$\mathbf{F}$	ω	τ	$\mathbf{r}$	
$\mathbf{r}$			z_3			$+z_6$	$=0$
m	z_1	$+z_2$	$+z_3$				$=0$
τ		$-2z_2$	$-2z_3$	$-z_4$	$+z_5$		$=0$

. (4.25)

Система має $N = 6$ невідомих і $K = 3$ лінійно незалежних рівнянь, отже існує $(N-K) = 3$ лінійно незалежні розв'язки. Для обчислення π_1 приймемо $z_4 = 2$ і $z_5 = z_6 = 0$, тоді $z_1 = 1$, $z_2 = -1$ та $z_3 = 0$. Для обчислення π_2 приймемо $z_5 = 2$ і $z_4 = z_6 = 0$, тоді $z_1 = -1$, $z_2 = 1$ та $z_3 = 0$. Для обчислення π_3 приймемо $z_6 = 1$ і $z_4 = z_5 = 0$, тоді $z_1 = 0$, $z_2 = 1$ та $z_3 = -1$. Звідки на основі (4.23):

$$\pi_1 = m^1 k^{-1} \mathbf{F}^0 \omega^2 \tau^0 \mathbf{r}^0 = \frac{m\omega^2}{k} \quad [\pi_1] = \text{кг}^1\text{с}^{-2}\text{кг}^{-1}\text{с}^2 = 1$$

$$\pi_2 = m^{-1} k^1 \mathbf{F}^0 \omega^0 \tau^2 \mathbf{r}^0 = \frac{k\tau^2}{m} \quad [\pi_2] = \text{кг}^1\text{с}^{-2}\text{с}^2\text{кг}^{-1} = 1 \qquad (4.26)$$

$$\pi_3 = m^0 k^1 \mathbf{F}^{-1} \omega^0 \tau^0 \mathbf{r}^1 = \frac{k\mathbf{r}}{\mathbf{F}} \quad [\pi_3] = \text{кг}^1\text{с}^{-2}\text{м}^1\text{с}^2\text{кг}^{-1}\text{м}^{-1} = 1$$

Розглянемо інший приклад – задачу нестаціонарної теплопровідності, що описується параболічним рівнянням:

$$c\rho \frac{\partial T}{\partial \tau} = \lambda \nabla^2 T. \qquad (4.27)$$

Тут параметрами задачі виступають: питома теплоємність c [Дж/кг°С], де [Дж = кг м^2 / с^2], густина ρ [кг/м^3], коефіцієнт теплопровідності λ [Вт/м°С], де [Вт = кг м^2/с^3], час τ [с], відстань x, y, z, або, відкинувши диференціальні оператори[1], деяка характеристична відстань l [м] та температура T [°С]. Знайдемо критерій подібності звівши рівняння до безрозмірного:

$$\pi_1 = \frac{\lambda \frac{T}{l^2}}{c\rho \frac{T}{\tau}} = c^{-1}\rho^{-1}\lambda\tau l^{-2} T^0 = \frac{\lambda\tau}{c\rho l^2},$$

[1] Опускаючи диференціальні чи інтегральні оператори отримуємо:
$d^n/dx^n \Rightarrow 1/x^n$, $\int xdy \Rightarrow xy$, $\nabla \Rightarrow 1/l$, $\nabla^2 \Rightarrow 1/l^2$.

$$[\pi_1] = \left(\text{кг}^{-1}\text{м}^{-2}\text{с}^{2}\text{кг}^{1}\text{C}^{1}\right)\left(\text{кг}^{-1}\text{м}^{3}\right)\left(\text{кг}^{1}\text{м}^{2}\text{с}^{-3}\text{м}^{-1}\text{C}^{-1}\right)\text{с}^{1}\text{м}^{-2} = \\ = \text{кг}^{-1+1-1+1}\text{м}^{-2+3+2-1-2}\text{с}^{2-3+1}\text{C}^{1-1} = \text{кг}^{0}\text{м}^{0}\text{с}^{0}\text{C}^{0} = 1. \tag{4.28}$$

Отриманий критерій відомий під назвою критерію Фур'є. Його, та інші загальноприйняті критерії подібності, прийнято позначати першими літерами відповідних прізвищ, в даному випадку "Fo" – Fourier. Критерій встановлює відповідність між швидкістю розвитку різних ефектів, що впливають на хід досліджуваного процесу. Такі критерії характерні для будь-яких нестаціонарних процесів. Їх часто називають критеріями гомохронності, тобто часової однорідності.

Аналізуючи рівняння (4.27), можна побачити, що шість параметрів системи є комбінацією чотирьох базових: маси [кг], відстані [м], часу [с] та температури [°C]. Тобто, будуючи систему рівнянь (4.24), ми б отримали $N = 6$ невідомих та $K = 4$ лінійно незалежних рівнянь, а відповідна кількість лінійно незалежних розв'язків, і значить – критеріїв подібності, рівна $(N - K) = 2$. Отже, критерій Фур'є є обов'язковим, але не єдиним для визначення подібності даної системи до інших.

Нагадаємо, для того, щоб система була повною необхідно вказати крайові умови. Початкові умови задачі, а також умови Діріхле не дадуть можливості для визначення наступного критерію подібності, тому, розглядаючи загальний випадок, використаємо крайові умови Робіна (Ньютона-Ріхмана), тобто температурний напір:

$$\lambda \left.\frac{\partial T}{\partial \mathbf{n}}\right|_{\Gamma} = \alpha \Delta T\big|_{\Gamma}, \tag{4.29}$$

де α – коефіцієнт тепловіддачі [Вт/м^2°C]. Зведемо останнє рівняння до безрозмірного:

$$\pi_2 = \frac{\alpha T}{\lambda \dfrac{T}{l}} = \alpha \lambda^{-1} l = \frac{\alpha l}{\lambda},$$

$$[\pi_2] = \left(\text{кг}^{1}\text{м}^{2}\text{с}^{-3}\text{м}^{-2}\text{C}^{-1}\right)\left(\text{кг}^{-1}\text{м}^{-2}\text{с}^{3}\text{м}^{1}\text{C}^{1}\right)\text{м}^{1} = \\ = \text{кг}^{1-1}\text{м}^{2-2-2+1+1}\text{с}^{-3+3}\text{C}^{-1+1} = \text{кг}^{0}\text{м}^{0}\text{с}^{0}\text{C}^{0} = 1. \tag{4.30}$$

Отриманий критерій відомий під назвою критерію Біо ("Bi"). Він є приблизною мірою відношення температурного перепаду в об'єкті до температурного напору між зовнішнім середовищем та об'єктом. Якщо значення критерію набагато більше одиниці, то температурним напором можна знехтувати, і крайові умови Робіна перетворюються в крайові умови Діріхле. Якщо навпаки, критерій набагато менший одиниці, можна розглядати тільки температурний напір, і крайові умови Робіна перетворюються в крайові умови Неймана [9].

Розглянемо тепер задачу електропровідності, що описує комутацію в деякому електричному пристрої. Відомо [10], [11], що комутація, або процес, що відбувається в перші моменти часу після замикання чи розмикання різних

ділянок електричного кола (абстрактної схеми заміщення реального об'єкту), є перехідним процесом. Явища електромагнетизму описуються системою рівнянь Максвелла [10], [11], [12], до якої в якості основних невідомих входять: **E** – напруженість електричного поля [В/м], **H** – напруженість магнітного поля [А/м], **D** – електрична індукція або електричне зміщення [Кл/м^2], **B** – магнітна індукція [Тл], **J** – густина електричного струму [А/м^2] та ρ – густина електричного заряду [Кл/м^3] . У диференціальній формі ці невідомі та відповідно система рівнянь Максвелла записується як:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial \tau}, \quad \nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial \tau}, \quad \nabla \cdot \mathbf{D} = \rho, \quad \nabla \cdot \mathbf{B} = 0. \tag{4.31}$$

До цих диференціальних співвідношень додаються, так звані, матеріальні рівняння:

$$\mathbf{D} = \varepsilon_0 \varepsilon \mathbf{E}, \quad \mathbf{B} = \mu_0 \mu \mathbf{H}, \quad \mathbf{J} = \sigma \mathbf{E}, \tag{4.32}$$

де ε_0 – діелектрична проникність вакууму [Ф/м], ε – діелектрична проникність середовища (безрозмірна), μ_0 – магнітна проникність вакууму [Гн/А^2], μ – магнітна проникність середовища (безрозмірна) та σ – питома електропровідність середовища [Ом$^{-1}$м$^{-1}$].

Візьмемо дивергенцію від другого рівняння з (4.31):

$$\nabla \cdot (\nabla \times \mathbf{H}) = \nabla \cdot \mathbf{J} + \frac{\partial \nabla \cdot \mathbf{D}}{\partial \tau}. \tag{4.33}$$

Можна перевірити, що дивергенція від ротору завжди рівна нулю, тому враховуючи, що $\nabla \cdot \mathbf{D} = \rho$, отримаємо:

$$\frac{\partial \rho}{\partial \tau} + \nabla \cdot \mathbf{J} = 0. \tag{4.34}$$

Зміна заряду за одиницю часу описує ємнісні характеристики середовища відносно його об'єму[1]. Позначивши цю ємнісну характеристику як c [Ф/м^3 = А^{2}с^{4}кг$^{-1}$м$^{-5}$], і враховуючи (4.32) та той факт, що напруженість електричного поля **E** залежить від зміни потенціального поля в просторі:

$$\mathbf{E} = -\nabla U, \tag{4.35}$$

отримаємо:

$$c \frac{\partial U}{\partial \tau} = \sigma \nabla^2 U. \tag{4.36}$$

Відповідними критеріями подібності для останнього рівняння будуть:

$$\pi_1 = \frac{\sigma \frac{U}{l^2}}{c \frac{U}{\tau}} = c^{-1} \sigma \tau l^{-2} U^0 = \frac{\sigma \tau}{c l^2}, \quad \pi_2 = \frac{\beta U}{\sigma \frac{U}{l}} = \beta \sigma^{-1} l = \frac{\beta l}{\sigma}. \tag{4.37}$$

При підборі таких параметрів моделі (4.36), при яких її критерії подібності

[1] Robertson A., Gross D. – An Electrical-Analog Method for Transient Heat-Flow Analysis // Journal of Research of the National Bureau of Standards, 61(2):105-115, 1958.

будуть відповідно однаковими для всіх критеріїв подібності оригіналу (4.27), задачі будуть подібними та аналогічними.

Звісно, що використання останньої неперервної моделі, що описує задачу електропровідності, є незручним на практиці. Набагато зручніше, використати дискретну систему. Продовжуючи розвивати аналогії, можна показати, що для задачі теплопровідності, так само, як і для задачі електропровідності, можна побудувати схему заміщення, тобто, так зване, теплове коло при відповідній теплоелектричній аналогії [9], [13], [14], (*Таблиця 4.1*).

Таблиця 4.1

Елементи аналогій теплового та електричного кіл

Елемент теплової схеми		Електрична аналогія		Графічне позначення
Ізотермічна поверхня чи об'єм	$T = \text{const}$ [°C]	Провідник	$U = \text{const}$ [В]	
Ідеальний тепловий зв'язок	$R_T = \frac{\Delta T}{q} = \frac{h}{\lambda}$ [м^2°C/Вт]	Резистор	$R_U = \frac{\Delta U}{i} = \frac{h}{\sigma}$ [Ом]	
Зосереджений тепловий опір/провідність	$T = 0$	Заземлення	$U = 0$	
Джерело температурного напору або конвекція (крайові умови Робіна = умови Діріхле + умови Неймана)	$\lambda \frac{\partial T}{\partial \mathbf{n}}\Big\|_{\Gamma} = \alpha \Delta T\Big\|_{\Gamma}$	Джерело напруги (електрорушійної сили)	$\sigma \frac{\partial U}{\partial \mathbf{n}}\Big\|_{\Gamma} = \beta \Delta U\Big\|_{\Gamma}$	
Джерело теплового потоку (крайові умови Неймана)	$\lambda \frac{\partial T}{\partial \mathbf{n}}\Big\|_{\Gamma} = q$	Джерело струму (потоку носіїв заряду)	$\sigma \frac{\partial U}{\partial \mathbf{n}}\Big\|_{\Gamma} = i$	
Теплова ємність	$c_T \rho = \frac{\Delta Q_T}{\Omega \Delta T}$ [Дж/кг°C · кг/м^3]	Конденсатор	$c_U = \frac{\Delta Q_U}{\Delta U}$ [Ф]	

Очевидно, що подібну методику можна застосовувати для будь-яких аналогічних явищ. Так на основі електричних кіл, крім теплових, можна також моделювати поведінку гідравлічних (гідродинамічних), акустичних, механічних і навіть квантових систем [15].

Нерозглянутим питанням залишається процес переходу від неперервної до дискретної системи. Річ у тому, що згідно третьої теореми подібності, дискретна система може розглядатися аналогом оригінальної неперервної системи, тільки при однакових умовах однозначності. Особливу роль тут відіграють геометричні властивості обох систем. Наприклад, для електричного поля в провідниках приймають ряд спрощень, що формально за допомогою інтегральних операторів переводять диференціальні рівняння Максвелла (4.31) в їх інтегральні аналоги [10], тобто дискретні системи – електричні кола. Тут зазвичай в ролі об’єктів моделювання виступають тривіальні об’єкти, простої

геометричної форми.

Більш складною стає ситуація, коли необхідно змоделювати поведінку полів в об'єкті складної форми. Для цього розглянемо двовимірні симплекс елементи, що були описані в попередньому розділі, а точніше їх геометричний зміст. Приймаючи для елементу деякий абстрактний коефіцієнт провідності (жорсткості) α, або у випадку анізотропії, тензор характеристик середовища $[\mathbf{D}]$, можна знайти локальну матрицю жорсткості $[\mathbf{K}]$, що визначає взаємозв'язок між вузлами елементу:

$$[\mathbf{K}] = \iint [\mathbf{B}]^{\mathrm{T}} [\mathbf{D}][\mathbf{B}] d\Omega = [\mathbf{B}]^{\mathrm{T}} \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix} [\mathbf{B}]\Omega. \tag{4.38}$$

Виходячи з того, що коефіцієнти матриці градієнтів $[\mathbf{B}]$ мають безпосередній геометричний зміст, тобто є проекціями сторін елементу на координатні осі, отримаємо:

$$\begin{aligned} [\mathbf{K}] &= \frac{1}{2\Omega} \begin{bmatrix} b_{1,1} & b_{2,1} \\ b_{1,2} & b_{2,2} \\ b_{1,3} & b_{2,3} \end{bmatrix} \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix} \frac{1}{2\Omega} \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,1} & b_{2,2} & b_{2,3} \end{bmatrix} \Omega = \\ &= \frac{\alpha}{4\Omega} \begin{bmatrix} b_{1,1}^2 + b_{2,1}^2 & b_{1,1}b_{1,2} + b_{2,1}b_{2,2} & b_{1,1}b_{1,3} + b_{2,1}b_{2,3} \\ b_{1,1}b_{1,2} + b_{2,1}b_{2,2} & b_{1,2}^2 + b_{2,2}^2 & b_{1,2}b_{1,3} + b_{2,2}b_{2,3} \\ b_{1,1}b_{1,3} + b_{2,1}b_{2,3} & b_{1,2}b_{1,3} + b_{2,2}b_{2,3} & b_{1,3}^2 + b_{2,3}^2 \end{bmatrix}. \end{aligned} \tag{4.39}$$

Нагадаємо, що площу трикутника можна записати через коефіцієнти матриці градієнтів $[\mathbf{B}]$. Наприклад віднявши перший рядок:

$$2\Omega = \left\| \begin{bmatrix} 1 & X_{1,1} & X_{1,2} \\ 1 & X_{2,1} & X_{2,2} \\ 1 & X_{3,1} & X_{3,2} \end{bmatrix} \right\| = \left\| \begin{bmatrix} X_{2,1} - X_{1,1} & X_{2,2} - X_{1,2} \\ X_{3,1} - X_{1,1} & X_{3,2} - X_{1,2} \end{bmatrix} \right\| = \left\| \begin{bmatrix} b_{2,3} & -b_{1,3} \\ -b_{2,2} & b_{1,2} \end{bmatrix} \right\|, \tag{4.40}$$

або, віднімаючи інші рядки:

$$2\Omega = b_{1,1}b_{2,2} - b_{1,2}b_{2,1} = b_{1,1}b_{2,3} - b_{1,3}b_{2,1} = b_{1,2}b_{2,3} - b_{1,3}b_{2,2}. \tag{4.41}$$

Враховуючи останню формулу, взаємозв'язок між вузлами елементу можна виразити за допомогою провідностей Y, або обернених до них величин – опорів R:

$$\begin{aligned} [\mathbf{K}]_{1,2} &= [\mathbf{K}]_{2,1} = Y_{1,2} = \frac{1}{R_{1,2}} = \frac{1}{2}\alpha \frac{b_{1,1}b_{1,2} + b_{2,1}b_{2,2}}{b_{1,1}b_{2,2} - b_{1,2}b_{2,1}}, \\ [\mathbf{K}]_{1,3} &= [\mathbf{K}]_{3,1} = Y_{1,3} = \frac{1}{R_{1,3}} = \frac{1}{2}\alpha \frac{b_{1,1}b_{1,3} + b_{2,1}b_{2,3}}{b_{1,1}b_{2,3} - b_{1,3}b_{2,1}}, \\ [\mathbf{K}]_{2,3} &= [\mathbf{K}]_{3,2} = Y_{2,3} = \frac{1}{R_{2,3}} = \frac{1}{2}\alpha \frac{b_{1,2}b_{1,3} + b_{2,2}b_{2,3}}{b_{1,2}b_{2,3} - b_{1,3}b_{2,2}}. \end{aligned} \tag{4.42}$$

У такому випадку, локальна матриця жорсткості (опорів/провідностей) $[\mathbf{K}]$

буде мати вигляд:

$$[\mathbf{K}]=\begin{bmatrix} -(Y_{1,2}+Y_{1,3}) & Y_{1,2} & Y_{1,3} \\ Y_{1,2} & -(Y_{1,2}+Y_{2,3}) & Y_{2,3} \\ Y_{1,3} & Y_{2,3} & -(Y_{1,3}+Y_{2,3}) \end{bmatrix}. \qquad (4.43)$$

Отримана матриця, є нічим іншим, ніж комбінація діагональної матриці провідностей та булевої матриці з'єднань з методу вузлових потенціалів – методу розрахунку електричних кіл шляхом запису системи лінійних алгебраїчних рівнянь, в якій невідомими є потенціали у вузлах кола [10]. У матричному вигляді система рівнянь для методу вузлових потенціалів виглядає наступним чином:

$$[\mathbf{A}][\mathbf{Y}][\mathbf{A}]^{\mathrm{T}}\{\mathbf{U}\}=-[\mathbf{A}](\{\mathbf{J}\}+[\mathbf{Y}]\{\mathbf{E}\}), \qquad (4.44)$$

де $[\mathbf{A}]$ – булева матриця з'єднань (матриця інцидентності вузлів до ребер), $[\mathbf{Y}]$ – діагональна матриця провідностей, $\{\mathbf{U}\}$ – шукані вузлові потенціали, $\{\mathbf{J}\}$ – джерела струму (потоків), $\{\mathbf{E}\}$ – джерела напруги. Використовуючи елементи аналогії (*Таблиця 4.1*), цю систему рівнянь можна звести до вже звичної:

$$[\mathbf{K}]=[\mathbf{A}][\mathbf{Y}][\mathbf{A}]^{\mathrm{T}},\quad \{\mathbf{f}\}=-[\mathbf{A}](\{\mathbf{J}\}+[\mathbf{Y}]\{\mathbf{E}\}),\quad [\mathbf{K}]\{\mathbf{U}\}=\{\mathbf{f}\}. \qquad (4.45)$$

Отже, використані функції форми симплекс елементу мають безпосередній фізичний зміст – вони відображають параметри опорів/провідностей аналогічної дискретної системи (*Рис. 4.2*). Якщо ж розглядати задачі механіки, як це було на початку розвитку методу скінченних елементів, матриця жорсткості $[\mathbf{K}]$ буде описувати поведінку симплекс елементу, кожне ребро якого є ідеалізованою пружиною з заданим коефіцієнтом жорсткості аналогічної дискретної механічної системи. І так далі, для інших аналогій. Також очевидно, що можна розглядати взаємозв'язок не тільки між сусідніми вузлами. Тоді ситуація буде аналогічною до використання комплекс елементів.

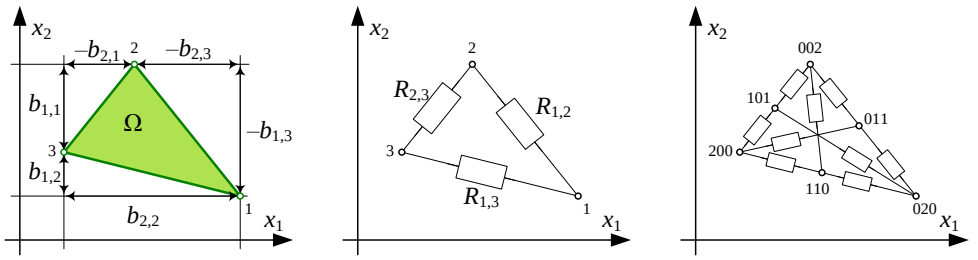


Рис. 4.2 Приклад фрагменту аналогічної дискретної системи для неперервних задач провідності на основі трикутних елементів різних порядків

При розгляді нестаціонарних задач, набувають безпосереднього фізичного змісту матриці демпфування $[\mathbf{C}]$, що описують ємнісні характеристики системи, та матриці маси $[\mathbf{M}]$, що описують індуктивні характеристики

аналогічних дискретних систем. На *Рис. 4.3* зображено процес побудови аналогічної дискретної системи – теплового кола, для задачі нестаціонарної теплопровідності. Зверніть увагу, що ансамблювання елементів приводить до утворення паралельного з'єднання опорів, по одному для кожного з сусідніх елементів.

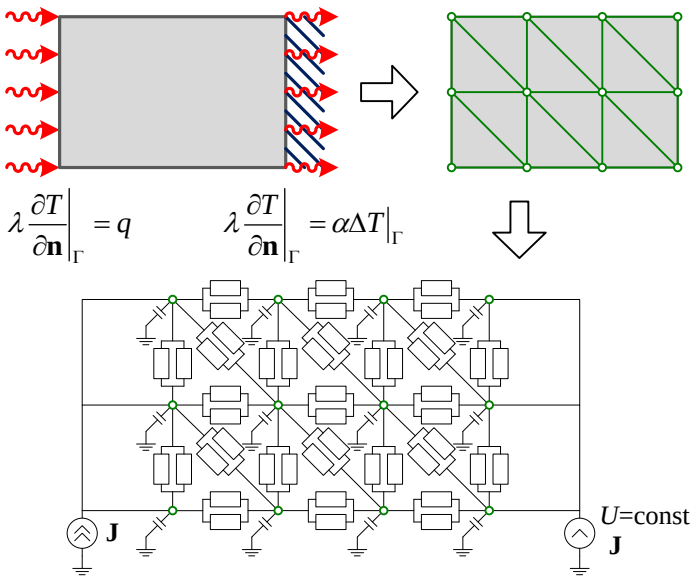


Рис. 4.3 Приклад аналогічної дискретної системи для задачі нестаціонарної теплопровідності

4.2. Рішення систем диференціальних рівнянь

Метод скінченних елементів, як і всі методи зважених нев'язок, може бути ефективно застосований і для рішення систем диференціальних рівнянь, що виникають, наприклад при розгляді задач механіки чи задач, що описують мультифізичні процеси. У формулюванні постановки таких задач, шукана польова величина виступає в ролі не скалярного, а векторного (чи тензорного) потенціалу:

$$\mathbf{u}(\mathbf{r}) = \{u_1(\mathbf{r}), u_2(\mathbf{r}), \ldots, u_D(\mathbf{r})\}^{\mathrm{T}}. \qquad (4.46)$$

Невідомий вектор $\mathbf{u}(\mathbf{r})$ в деякій області Ω задовольняє диференціальним рівнянням:

$$\mathcal{A}_1(\mathbf{u}(\mathbf{r})) = 0, \quad \mathcal{A}_2(\mathbf{u}(\mathbf{r})) = 0, \quad \ldots, \quad \mathcal{A}_D(\mathbf{u}(\mathbf{r})) = 0, \qquad (4.47)$$

або в матричній формі:

$$\mathbf{A}(\mathbf{u}) = \begin{bmatrix} \mathcal{A}_1(\mathbf{u}(\mathbf{r})) \\ \mathcal{A}_2(\mathbf{u}(\mathbf{r})) \\ \vdots \\ \mathcal{A}_D(\mathbf{u}(\mathbf{r})) \end{bmatrix}_{\Omega} = 0. \qquad (4.48)$$

Для коректності постановки задачі, на границі Γ області Ω задано

необхідну кількість крайових умов:

$$\mathcal{B}_1(\mathbf{u}(\mathbf{r}))=0,\quad \mathcal{B}_2(\mathbf{u}(\mathbf{r}))=0,\quad \ldots,\quad \mathcal{B}_D(\mathbf{u}(\mathbf{r}))=0, \tag{4.49}$$

або в матричній формі:

$$\mathbf{B}(\mathbf{u})=\begin{bmatrix}\mathcal{B}_1(\mathbf{u}(\mathbf{r}))\\ \mathcal{B}_2(\mathbf{u}(\mathbf{r}))\\ \vdots\\ \mathcal{B}_D(\mathbf{u}(\mathbf{r}))\end{bmatrix}_\Gamma=0. \tag{4.50}$$

Для кожної компоненти шуканого векторного потенціалу $\mathbf{u}(\mathbf{r})$ використовується розклад по базисним функціям:

$$\begin{gathered}u_1(\mathbf{r})\approx\tilde{u}_1(\mathbf{r})=(u_1)_0(\mathbf{r})+\sum_{j=1}^{M}a_{j,1}\varphi_{j,1}(\mathbf{r}),\\ u_2(\mathbf{r})\approx\tilde{u}_2(\mathbf{r})=(u_2)_0(\mathbf{r})+\sum_{j=1}^{M}a_{j,2}\varphi_{j,2}(\mathbf{r}),\\ \ldots\\ u_D(\mathbf{r})\approx\tilde{u}_D(\mathbf{r})=(u_D)_0(\mathbf{r})+\sum_{j=1}^{M}a_{j,D}\varphi_{j,D}(\mathbf{r}),\end{gathered} \tag{4.51}$$

або у векторній формі:

$$\mathbf{u}(\mathbf{r})\approx\tilde{\mathbf{u}}(\mathbf{r})=\{\mathbf{u}_0(\mathbf{r})\}+\sum_{j=1}^{M}[\boldsymbol{\varphi}_j(\mathbf{r})]\{\mathbf{a}\}_j, \tag{4.52}$$

де:

$$\begin{gathered}\{\mathbf{u}_0(\mathbf{r})\}=\left\{(u_1)_0(\mathbf{r}),(u_2)_0(\mathbf{r}),\ldots,(u_D)_0(\mathbf{r})\right\}^{\mathbf{T}},\\ \{\mathbf{a}\}_j=\left\{a_{j,1},a_{j,2},\ldots,a_{j,D}\right\}^{\mathbf{T}},\end{gathered} \tag{4.53}$$

та:

$$[\boldsymbol{\varphi}_j(\mathbf{r})]=\begin{bmatrix}\varphi_{j,1}(\mathbf{r}) & 0 & \cdots & 0\\ 0 & \varphi_{j,2}(\mathbf{r}) & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \varphi_{j,D}(\mathbf{r})\end{bmatrix}. \tag{4.54}$$

Очевидно, що в даному випадку можуть бути використані попередньо описані види апроксимацій методів зважених нев'язок. При цьому, розуміється що $[\boldsymbol{\varphi}_j(\mathbf{r})]$ – діагональна матриця, яка побудована з базисних функцій, а параметр $\{\mathbf{a}\}_j$ – вектор з числом компонент, рівним числу невідомих функцій у розкладі $\mathbf{u}(\mathbf{r})$.

Щоб отримати для задачі даного типу узагальнене рівняння методу зважених нев'язок, необхідно розглянути кожне з рівнянь (4.47) та відповідні їм крайові умови. Приписуючи до них вагові функції, отримаємо нову систему:

$$\begin{aligned}
&\int_{\Omega} \omega_{i,1}^{\Omega}(\mathbf{r})\mathcal{A}_1(\tilde{\mathbf{u}}(\mathbf{r}))d\Omega + \int_{\Gamma} \omega_{i,1}^{\Gamma}(\mathbf{r})\mathcal{B}_1(\tilde{\mathbf{u}}(\mathbf{r}))d\Gamma = 0,\\
&\int_{\Omega} \omega_{i,2}^{\Omega}(\mathbf{r})\mathcal{A}_2(\tilde{\mathbf{u}}(\mathbf{r}))d\Omega + \int_{\Gamma} \omega_{i,2}^{\Gamma}(\mathbf{r})\mathcal{B}_2(\tilde{\mathbf{u}}(\mathbf{r}))d\Gamma = 0,\\
&\cdots\\
&\int_{\Omega} \omega_{i,D}^{\Omega}(\mathbf{r})\mathcal{A}_D(\tilde{\mathbf{u}}(\mathbf{r}))d\Omega + \int_{\Gamma} \omega_{i,D}^{\Gamma}(\mathbf{r})\mathcal{B}_D(\tilde{\mathbf{u}}(\mathbf{r}))d\Gamma = 0.
\end{aligned} \tag{4.55}$$

Якщо ввести діагональні матриці вагових функцій $[\mathbf{W}_i^{\Omega}(\mathbf{r})]$ та $[\mathbf{W}_i^{\Gamma}(\mathbf{r})]$, де:

$$[\mathbf{W}]_i^{\Omega} = \begin{bmatrix} \omega_{i,1}^{\Omega}(\mathbf{r}) & 0 & \cdots & 0 \\ 0 & \omega_{i,2}^{\Omega}(\mathbf{r}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_{i,D}^{\Omega}(\mathbf{r}) \end{bmatrix}, [\mathbf{W}]_i^{\Gamma} = \begin{bmatrix} \omega_{i,1}^{\Gamma}(\mathbf{r}) & 0 & \cdots & 0 \\ 0 & \omega_{i,2}^{\Gamma}(\mathbf{r}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_{i,D}^{\Gamma}(\mathbf{r}) \end{bmatrix}, \tag{4.56}$$

то систему (4.55) можна записати в компактній формі:

$$\int_{\Omega} [\mathbf{W}]_i^{\Omega}\mathbf{A}(\tilde{\mathbf{u}})d\Omega + \int_{\Gamma} [\mathbf{W}]_i^{\Gamma}\mathbf{B}(\tilde{\mathbf{u}})d\Gamma = 0, \tag{4.57}$$

розв'язавши яку, отримаємо апроксимацію шуканого векторного потенціалу $\mathbf{u}(\mathbf{r})$.

Для прикладу розглянемо стаціонарну двовимірну задачу лінійної теорії пружності в плоских напруженнях [1]. Основними невідомими величинами є повздовжнє і поперечне переміщення кожної з точок тіла, що піддається деформації. Якщо позначити переміщення вздовж осі x, як функцію $u(x,y)$, а переміщення вздовж осі y, як функцію $v(x,y)$, то шуканий векторний потенціал можна записати як:

$$\mathbf{u}(x,y) = \begin{Bmatrix} u(x,y) \\ v(x,y) \end{Bmatrix}. \tag{4.58}$$

Не вдаючись в деталі теорії пружності [16], [17], деформації, і як наслідок напруження, можуть бути виражені через описані переміщення. Так деформації тіла записуються через лінійний тензор деформації, у вигляді системи диференціальних рівнянь:

$$[\boldsymbol{\varepsilon}] = \begin{Bmatrix} \varepsilon_x \\ \varepsilon_y \\ \gamma_{xy} \end{Bmatrix} = \begin{Bmatrix} \partial u/\partial x \\ \partial v/\partial y \\ \partial u/\partial y + \partial v/\partial x \end{Bmatrix} = \mathcal{L}(\mathbf{u}(x,y)) = [\mathcal{L}]\{\mathbf{u}\}, \tag{4.59}$$

де:

$$\mathcal{L}(.) = [\mathcal{L}] = \begin{bmatrix} \partial/\partial x & 0 \\ 0 & \partial/\partial y \\ \partial/\partial y & \partial/\partial x \end{bmatrix}. \tag{4.60}$$

Напруження, в описаній задачі, виражаються через лінійний тензор напружень [18]:

$$[\boldsymbol{\sigma}]=\begin{Bmatrix}\sigma_x\\ \sigma_y\\ \tau_{xy}\end{Bmatrix}=\frac{E}{1-\mu^2}\begin{bmatrix}1 & \mu & 0\\ \mu & 1 & 0\\ 0 & 0 & (1-\mu)/2\end{bmatrix}[\boldsymbol{\varepsilon}]=[\mathbf{D}][\boldsymbol{\varepsilon}]=[\mathbf{D}]\big([\mathcal{L}]\{\mathbf{u}\}\big),\quad (4.61)$$

де, E – модуль пружності матеріалу, також відомий як модуль Юнга, тобто величина, що показує здатність матеріалу чинити опір розтягуванню чи стискуванню при пружній деформації. μ – коефіцієнт Пуассона, тобто коефіцієнт, що показує міру зміни поперечних розмірів тіла при розтягуванні. Відповідно до теорії, коефіцієнт Пуассона та модуль Юнга повністю описують пружні властивості матеріалу в рамках даної задачі. Залишається лише записати систему диференціальних рівнянь балансу в деякій двовимірній області Ω:

$$\mathbf{A}(\mathbf{u})=\begin{bmatrix}\partial\sigma_x/\partial x+\partial\tau_{xy}/\partial y+X\\ \partial\tau_{xy}/\partial x+\partial\sigma_y/\partial y+Y\end{bmatrix}_\Omega=[\mathcal{L}]^{\mathbf{T}}[\mathbf{D}]\big([\mathcal{L}]\{\mathbf{u}\}\big)+\{\mathbf{X}\}=0.\quad (4.62)$$

Тут, X та Y – внутрішні сили, що діють на нескінченно малий об'єм в межах області Ω, тобто $\mathbf{X}=\{X,Y\}^{\mathbf{T}}$.

Крайові умови, для типових задач лінійної теорії пружності, можуть бути коректно поставлені шляхом вказання поверхневих навантажень, тобто сил, що діють на границях тіла (умови Неймана), та відомих переміщень на границях тіла (умови Діріхле). У такому випадку:

$$\mathbf{B}(\mathbf{u})=\begin{bmatrix}\sigma_x l_x+\tau_{xy}l_y-t_x\\ \tau_{xy}l_x+\sigma_y l_y-t_y\end{bmatrix}_{\Gamma_\sigma}=0,\quad (4.63)$$

та:

$$\mathbf{B}(\mathbf{u})=\begin{bmatrix}u-u_\infty\\ v-v_\infty\end{bmatrix}_{\Gamma_u}=0,\quad (4.64)$$

де, l_x та l_y – направляючі косинуси нормалі до границі Γ_σ, t_x та t_y – відомі напруження на границі Γ_σ, u_∞ та v_∞ – відомі переміщення на границі Γ_u.

Припустимо, що для описаної системи рівнянь можна побудувати систему базисних функцій, таку що:

$$u_0\big|_{\Gamma_u}=u_\infty,\quad v_0\big|_{\Gamma_u}=v_\infty.\quad (4.65)$$

Розклад (4.52) стає рівним нулю на Γ_u, тобто задовольняє крайові умови Діріхле. Цю систему можна записати у вигляді:

$$\mathbf{u}(x,y)\approx\begin{Bmatrix}\tilde{u}(x,y)\\ \tilde{v}(x,y)\end{Bmatrix}=\begin{Bmatrix}u_\infty\\ v_\infty\end{Bmatrix}+\sum_{j=1}^{M}\begin{bmatrix}\varphi_{j,u}(x,y) & 0\\ 0 & \varphi_{j,v}(x,y)\end{bmatrix}\begin{Bmatrix}a_{j,u}\\ a_{j,v}\end{Bmatrix}.\quad (4.66)$$

Система вагових функцій запишеться у вигляді:

$$[\mathbf{W}]_i^\Omega=\begin{bmatrix}\omega_{i,u}^\Omega(x,y) & 0\\ 0 & \omega_{i,v}^\Omega(x,y)\end{bmatrix},\quad [\mathbf{W}]_i^\Gamma=\begin{bmatrix}\omega_{i,u}^\Gamma(x,y) & 0\\ 0 & \omega_{i,v}^\Gamma(x,y)\end{bmatrix}.\quad (4.67)$$

Згідно методу зважених нев'язок, для рівняння балансу в напруженнях

отримаємо:

$$\int_{\Omega}\begin{Bmatrix}\omega_{i,u}^{\Omega}\left(\partial\tilde{\sigma}_x/\partial x+\partial\tilde{\tau}_{xy}/\partial y+X\right)\\ \omega_{i,v}^{\Omega}\left(\partial\tilde{\tau}_{xy}/\partial x+\partial\tilde{\sigma}_y/\partial y+Y\right)\end{Bmatrix}d\Omega+\int_{\Gamma_\sigma}\begin{Bmatrix}\omega_{i,u}^{\Gamma}\left(\tilde{\sigma}_x l_x+\tilde{\tau}_{xy}l_y-t_x\right)\\ \omega_{i,v}^{\Gamma}\left(\tilde{\tau}_{xy}l_x+\tilde{\sigma}_y l_y-t_y\right)\end{Bmatrix}d\Gamma=0, \quad (4.68)$$

де, $\tilde{\sigma},\tilde{\tau}=[\mathbf{D}]([\mathcal{L}]\{\tilde{\mathbf{u}}\})$ [1]. Використовуючи техніку пониження порядку, на основі теореми Стокса, ці співвідношення можна перетворити до виду:

$$-\int_{\Omega}\begin{Bmatrix}\tilde{\sigma}_x\dfrac{\omega_{i,u}^{\Omega}}{\partial x}+\tilde{\tau}_{xy}\dfrac{\omega_{i,u}^{\Omega}}{\partial y}-\omega_{i,u}^{\Omega}X\\ \tilde{\tau}_{xy}\dfrac{\omega_{i,v}^{\Omega}}{\partial x}+\tilde{\sigma}_y\dfrac{\omega_{i,v}^{\Omega}}{\partial y}-\omega_{i,v}^{\Omega}Y\end{Bmatrix}d\Omega+$$

$$+\int_{\Gamma_u+\Gamma_\sigma}\begin{Bmatrix}\omega_{i,u}^{\Omega}\left(\tilde{\sigma}_x l_x+\tilde{\tau}_{xy}l_y\right)\\ \omega_{i,v}^{\Omega}\left(\tilde{\tau}_{xy}l_x+\tilde{\sigma}_y l_y\right)\end{Bmatrix}d\Gamma+\int_{\Gamma_\sigma}\begin{Bmatrix}\omega_{i,u}^{\Gamma}\left(\tilde{\sigma}_x l_x+\tilde{\tau}_{xy}l_y-t_x\right)\\ \omega_{i,v}^{\Gamma}\left(\tilde{\tau}_{xy}l_x+\tilde{\sigma}_y l_y-t_y\right)\end{Bmatrix}d\Gamma=0. \quad (4.69)$$

Приймемо $\omega_{i,u}^{\Omega}=\omega_{i,v}^{\Omega}=\varphi_{i,u}=\varphi_{i,v}$, $\omega_{i,u}^{\Omega}=-\omega_{i,u}^{\Gamma}\big|_{\Gamma_\sigma}$ та $\omega_{i,v}^{\Omega}=-\omega_{i,v}^{\Gamma}\big|_{\Gamma_\sigma}$. Оскільки $\varphi_j\big|_{\Gamma_u}=0$, то і $\omega_i\big|_{\Gamma_u}=0$, наведене вище рівняння можна записати компактно в напруженнях:

$$\int_{\Omega}\left([\mathcal{L}][\mathbf{W}]_i\right)^{\mathbf{T}}[\tilde{\boldsymbol{\sigma}}]d\Omega=\int_{\Omega}[\mathbf{W}]_i\{\mathbf{X}\}d\Omega+\int_{\Gamma_\sigma}[\mathbf{W}]_i\{\mathbf{t}\}d\Gamma, \quad (4.70)$$

де, $\{\mathbf{t}\}^{\mathbf{T}}=(t_x,t_y)$. Або виразимо його в переміщеннях:

---

[1] Вважаємо за необхідне також навести матричні формули для розв'язку стаціонарної тривимірної задача лінійної теорії пружності в ізотропному матеріалі, де двовимірна задача в плоских напруженнях є частковим випадком [18]:

$$\mathbf{u}(x,y,z)=\begin{bmatrix}u_x(x,y,z)\\ u_y(x,y,z)\\ u_z(x,y,z)\end{bmatrix}\quad[\mathcal{L}]=\begin{bmatrix}\partial/\partial x & 0 & 0\\ 0 & \partial/\partial y & 0\\ 0 & 0 & \partial/\partial z\\ \partial/\partial y & \partial/\partial x & 0\\ \partial/\partial z & 0 & \partial/\partial x\\ 0 & \partial/\partial z & \partial/\partial y\end{bmatrix}\quad[\boldsymbol{\varepsilon}]=\begin{bmatrix}\varepsilon_x\\ \varepsilon_y\\ \varepsilon_z\\ \gamma_{xy}\\ \gamma_{xz}\\ \gamma_{yz}\end{bmatrix}=[\mathcal{L}]\{\mathbf{u}\}=\begin{bmatrix}\partial/\partial x & 0 & 0\\ 0 & \partial/\partial y & 0\\ 0 & 0 & \partial/\partial z\\ \partial/\partial y & \partial/\partial x & 0\\ \partial/\partial z & 0 & \partial/\partial x\\ 0 & \partial/\partial z & \partial/\partial y\end{bmatrix}\begin{Bmatrix}u_x\\ u_y\\ u_z\end{Bmatrix}=\begin{bmatrix}\partial u_x/\partial x\\ \partial u_y/\partial y\\ \partial u_z/\partial z\\ \partial u_x/\partial y+\partial u_y/\partial x\\ \partial u_x/\partial z+\partial u_z/\partial x\\ \partial u_y/\partial z+\partial u_z/\partial y\end{bmatrix}$$

$$[\mathbf{D}]=\frac{E(1-\mu)}{(1+\mu)(1-2\mu)}\begin{bmatrix}1 & \mu/(1-\mu) & \mu/(1-\mu) & 0 & 0 & 0\\ \mu/(1-\mu) & 1 & \mu/(1-\mu) & 0 & 0 & 0\\ \mu/(1-\mu) & \mu/(1-\mu) & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & (1-2\mu)/2(1-\mu) & 0 & 0\\ 0 & 0 & 0 & 0 & (1-2\mu)/2(1-\mu) & 0\\ 0 & 0 & 0 & 0 & 0 & (1-2\mu)/2(1-\mu)\end{bmatrix}$$

$$[\boldsymbol{\sigma}]=\begin{pmatrix}\sigma_x & \sigma_y & \sigma_z & \tau_{xy} & \tau_{xz} & \tau_{yz}\end{pmatrix}^{\mathbf{T}}=[\mathbf{D}][\boldsymbol{\varepsilon}]=[\mathbf{D}]([\mathcal{L}]\{\mathbf{u}\})$$

$$\mathbf{B}(\mathbf{u})=\begin{bmatrix}\sigma_x l_x+\tau_{xy}l_y+\tau_{xz}l_z-t_x\\ \tau_{xy}l_x+\sigma_y l_y+\tau_{yz}l_z-t_y\\ \tau_{xz}l_x+\tau_{yz}l_y+\sigma_z l_z-t_z\end{bmatrix}_{\Gamma_\sigma}=0\qquad \mathbf{B}(\mathbf{u})=\begin{bmatrix}u_x-u_{\infty,x}\\ u_y-u_{\infty,y}\\ u_z-u_{\infty,z}\end{bmatrix}_{\Gamma_u}=0$$

$$\mathbf{A}(\mathbf{u})=[\mathcal{L}]^{\mathbf{T}}[\mathbf{D}]([\mathcal{L}]\{\mathbf{u}\})+\{\mathbf{X}\}=$$

$$=\begin{bmatrix}\partial\sigma_x/\partial x+\partial\tau_{xy}/\partial y+\partial\tau_{xz}/\partial z+X\\ \partial\tau_{xy}/\partial x+\partial\sigma_y/\partial y+\partial\tau_{yz}/\partial z+Y\\ \partial\tau_{xz}/\partial x+\partial\tau_{yz}/\partial y+\partial\sigma_z/\partial z+Z\end{bmatrix}_{\Omega}=0$$

$$\int_{\Omega}\left([\mathcal{L}][\mathbf{W}]_i\right)^{\mathbf{T}}[\mathbf{D}]\left([\mathcal{L}]\{\tilde{\mathbf{u}}\}\right)d\Omega=\int_{\Omega}[\mathbf{W}]_i\{\mathbf{X}\}d\Omega+\int_{\Gamma_\sigma}[\mathbf{W}]_i\{\mathbf{t}\}d\Gamma. \qquad (4.71)$$

Тобто, отримаємо систему лінійних алгебраїчних рівнянь методу зважених нев'язок для рівняння балансу (4.62), де крайові умови задання поверхневого навантаження (умови Неймана) є природними крайовими умовами.

Залишається підставити розклад апроксимації шуканого векторного потенціалу (4.66) в останнє рівняння, після чого отримаємо систему лінійних рівнянь з симетричною матрицею жорсткості:

$$\begin{aligned}[\mathbf{K}]_{i,j}&=\int_{\Omega}\left([\mathcal{L}][\mathbf{W}]_i\right)^{\mathbf{T}}[\mathbf{D}]\left([\mathcal{L}][\mathbf{W}]_j\right)d\Omega,\quad 1\le i,j\le M,\\ [\mathbf{f}]_i&=\int_{\Omega}[\mathbf{W}]_i\{\mathbf{X}\}d\Omega+\int_{\Gamma_\sigma}[\mathbf{W}]_i\{\mathbf{t}\}d\Gamma-\int_{\Omega}\left([\mathcal{L}][\mathbf{W}]_i\right)^{\mathbf{T}}[\mathbf{D}]\left([\mathcal{L}]\{\mathbf{u}_0\}\right)d\Omega,\quad 1\le i\le M.\end{aligned} \qquad (4.72)$$

Слід зауважити, що оскільки шукана величина є вектором, в даному випадку 2×1, то кожен елемент матриці жорсткості $[\mathbf{K}]_{i,j}$ є також матрицею, в даному випадку 2×2, і елементи вектору навантажень $[\mathbf{f}]_i$ є також векторами, в даному випадку 2×1. Щоб побудувати загальну систему лінійних алгебраїчних рівнянь, достатньо просто послідовно об'єднати всі обчислені елементи відповідно до індексів i та j. Так наприклад для $M=2$, отримаємо систему з чотирьох рівнянь і чотирьох невідомих:

$$\begin{bmatrix}[\mathbf{K}]_{1,1} & [\mathbf{K}]_{1,2}\\ [\mathbf{K}]_{2,1} & [\mathbf{K}]_{2,2}\end{bmatrix}=\left[\begin{array}{cc|cc}[\mathbf{K}]_{1,1}^{i=1,j=1} & [\mathbf{K}]_{1,2}^{i=1,j=1} & [\mathbf{K}]_{1,1}^{i=1,j=2} & [\mathbf{K}]_{1,2}^{i=1,j=2}\\ [\mathbf{K}]_{2,1}^{i=1,j=1} & [\mathbf{K}]_{2,2}^{i=1,j=1} & [\mathbf{K}]_{2,1}^{i=1,j=2} & [\mathbf{K}]_{2,2}^{i=1,j=2}\\ \hline [\mathbf{K}]_{1,1}^{i=2,j=1} & [\mathbf{K}]_{1,2}^{i=2,j=1} & [\mathbf{K}]_{1,1}^{i=2,j=2} & [\mathbf{K}]_{1,2}^{i=2,j=2}\\ [\mathbf{K}]_{2,1}^{i=2,j=1} & [\mathbf{K}]_{2,2}^{i=2,j=1} & [\mathbf{K}]_{2,1}^{i=2,j=2} & [\mathbf{K}]_{2,2}^{i=2,j=2}\end{array}\right],$$

$$\begin{Bmatrix}\{\mathbf{a}\}_1\\ \{\mathbf{a}\}_2\end{Bmatrix}=\begin{Bmatrix}a_{1,u}\\ a_{1,v}\\ \hline a_{2,u}\\ a_{2,v}\end{Bmatrix},\quad \begin{Bmatrix}\{\mathbf{f}\}_1\\ \{\mathbf{f}\}_2\end{Bmatrix}=\begin{Bmatrix}\{\mathbf{f}\}_1^{i=1}\\ \{\mathbf{f}\}_2^{i=1}\\ \hline \{\mathbf{f}\}_1^{i=2}\\ \{\mathbf{f}\}_2^{i=2}\end{Bmatrix}, \qquad (4.73)$$

при чому $[\mathbf{K}]_{1,2}=\left([\mathbf{K}]_{2,1}\right)^{\mathbf{T}}$.

Описаний підхід є достатньо абстрактним і може бути використаний для будь-якої ситуації в лінійній теорії пружності. Крім того, рівняння (4.70) може бути легко отримане виходячи з варіаційної постановки у методі Релея-Рітца, на основі використання принципу віртуальної роботи [16], [17], згідно якого всі точки тіла знаходяться у стані рівноваги при умові рівності робіт, що здійснюються внутрішніми напруженнями та зовнішніми силами на довільному або "віртуальному" переміщенні тіла.

Розглянемо описаний підхід на конкретному прикладі. Нехай задано прямокутну алюмінієву пластину, що займає область $-2\le x\le 2$ м, $-1\le y\le 1$м. Пластина закріплена на сторонах $y=\pm1$ та знаходиться під дією навантаження $t_x=E(1-y^2)/(1+\mu)$ ГПа, $t_y=0$ ГПа на сторонах $x=\pm2$ (*Рис. 4.4*). Необхідно

знайти поля переміщень, що виникають під дією навантажень. Модуль пружності матеріалу пластини $E = 70$ ГПа, коефіцієнт Пуассона $\mu = 0{,}34$.

Для симетричності виберемо систему базисних функцій $\varphi_{1,u} = x(1-y^2)$, $\varphi_{2,u} = x^3(1-y^2)$, $\varphi_{3,u} = xy^2(1-y^2)$, …, для переміщень вздовж осі x. І аналогічно $\varphi_{1,v} = y(1-y^2)$, $\varphi_{2,v} = x^2y(1-y^2)$, $\varphi_{3,v} = y^3(1-y^2)$, …, для переміщень вздовж осі y. Так триелементна апроксимація переміщень запишеться у вигляді:

$$\tilde{\mathbf{u}}(x,y) = \begin{Bmatrix} \tilde{u}(x,y) \\ \tilde{v}(x,y) \end{Bmatrix} = \begin{Bmatrix} u_0 + a_{1,u}x(1-y^2) + a_{2,u}x^3(1-y^2) + a_{3,u}xy^2(1-y^2) \\ v_0 + a_{1,v}y(1-y^2) + a_{2,v}x^2y(1-y^2) + a_{3,v}y^3(1-y^2) \end{Bmatrix}. \quad (4.74)$$

Враховуючи те, що на сторонах $y = \pm 1$, за умовою задачі немає ніяких переміщень, приймаючи $u_0 = v_0 = 0$, можна побачити, що апроксимація автоматично задовольняє цю крайову умову. Тоді рівняння методу зважених нев'язок (4.70) буде мати вигляд:

$$\int_{-1}^{1}\int_{-2}^{2} \left([\mathcal{L}][\mathbf{W}]_i\right)^{\mathbf{T}} [\tilde{\boldsymbol{\sigma}}] dxdy = \int_{-1}^{1} [\mathbf{W}]_i\big|_{x=-2} \{\mathbf{t}\} dy - \int_{-1}^{1} [\mathbf{W}]_i\big|_{x=2} \{\mathbf{t}\} dy, \quad (4.75)$$

або:

$$[\mathbf{K}]_{i,j} = \frac{E}{1-\mu^2}\int_{-1}^{1}\int_{-2}^{2}\left(\begin{bmatrix} \varphi_{i,u} & 0 \\ 0 & \varphi_{i,v} \end{bmatrix}\begin{bmatrix} \partial/\partial x & 0 & \partial/\partial y \\ 0 & \partial/\partial y & \partial/\partial x \end{bmatrix}\right)\cdot$$
$$\cdot\begin{bmatrix} 1 & \mu & 0 \\ \mu & 1 & 0 \\ 0 & 0 & (1-\mu)/2 \end{bmatrix}\left(\begin{bmatrix} \partial/\partial x & 0 \\ 0 & \partial/\partial y \\ \partial/\partial y & \partial/\partial x \end{bmatrix}\begin{bmatrix} \varphi_{j,u} & 0 \\ 0 & \varphi_{j,v} \end{bmatrix}\right) dxdy, \quad (4.76)$$
$$[\mathbf{f}]_i = \frac{E}{1+\mu}\int_{-1}^{1}\begin{bmatrix} \varphi_{i,u} & 0 \\ 0 & \varphi_{i,v} \end{bmatrix}_{x=-2}\begin{bmatrix} 1-y^2 \\ 0 \end{bmatrix} dy - \frac{E}{1+\mu}\int_{-1}^{1}\begin{bmatrix} \varphi_{i,u} & 0 \\ 0 & \varphi_{i,v} \end{bmatrix}_{x=2}\begin{bmatrix} 1-y^2 \\ 0 \end{bmatrix} dy.$$

Здійснивши матричне множення для останнього виразу отримаємо:

$$[\mathbf{K}]_{i,j} = \frac{E}{1-\mu^2}\int_{-1}^{1}\int_{-2}^{2}\begin{bmatrix} \dfrac{\partial\varphi_{i,u}}{\partial x}\dfrac{\partial\varphi_{j,u}}{\partial x} + \dfrac{1-\mu}{2}\dfrac{\partial\varphi_{i,u}}{\partial y}\dfrac{\partial\varphi_{j,u}}{\partial y} & \mu\dfrac{\partial\varphi_{i,u}}{\partial x}\dfrac{\partial\varphi_{j,u}}{\partial y} + \dfrac{1-\mu}{2}\dfrac{\partial\varphi_{i,u}}{\partial y}\dfrac{\partial\varphi_{j,u}}{\partial x} \\ \mu\dfrac{\partial\varphi_{i,v}}{\partial y}\dfrac{\partial\varphi_{j,u}}{\partial x} + \dfrac{1-\mu}{2}\dfrac{\partial\varphi_{i,v}}{\partial x}\dfrac{\partial\varphi_{j,u}}{\partial y} & \dfrac{\partial\varphi_{i,v}}{\partial y}\dfrac{\partial\varphi_{j,u}}{\partial y} + \dfrac{1-\mu}{2}\dfrac{\partial\varphi_{i,v}}{\partial x}\dfrac{\partial\varphi_{j,u}}{\partial x} \end{bmatrix} dxdy, \quad (4.77)$$
$$[\mathbf{f}]_i = \frac{E}{1+\mu}\int_{-1}^{1}\begin{Bmatrix} \varphi_{i,u}\big|_{x=-2}(1-y^2) - \varphi_{i,u}\big|_{x=2}(1-y^2) \\ 0 \end{Bmatrix} dy.$$

Обчисливши дані інтеграли та об'єднавши отримані результати, отримаємо загальну систему рівнянь аналогічно до (4.73):

$$\left[\begin{array}{cc|cc|cc} 709{,}1814 & 57{,}4099 & 2242{,}3639 & -72{,}0438 & 122{,}5388 & 24{,}6043 \\ 57{,}4099 & 506{,}5581 & 229{,}6397 & 675{,}4108 & -8{,}2014 & 217{,}0963 \\ \hline 2242{,}3639 & 229{,}6397 & 12273{,}1796 & 194{,}5183 & 371{,}2830 & 98{,}4170 \\ -72{,}0438 & 675{,}4108 & 194{,}5183 & 1705{,}8948 & 10{,}2920 & 289{,}4618 \\ \hline 122{,}5388 & -8{,}2014 & 371{,}2830 & 10{,}2920 & 132{,}8308 & 2{,}7338 \\ 24{,}6043 & 217{,}0963 & 98{,}4170 & 289{,}4618 & 2{,}7338 & 184{,}9339 \end{array}\right]\left\{\begin{array}{c} a_{1,u} \\ a_{1,v} \\ \hline a_{2,u} \\ a_{2,v} \\ \hline a_{3,u} \\ a_{3,v} \end{array}\right\} = \left\{\begin{array}{r} -222{,}8856 \\ 0{,}0000 \\ \hline -891{,}5423 \\ 0{,}0000 \\ \hline -31{,}8408 \\ 0{,}0000 \end{array}\right\}, \quad (4.78)$$

з розв'язком:

$$\{\mathbf{a}\} = \{-0{,}2397 \quad 0{,}1137 \mid -0{,}0327 \quad -0{,}0510 \mid 0{,}0838 \quad -0{,}0057\}^{\mathbf{T}}. \qquad (4.79)$$

Знайдене поле переміщень (4.66) можна використати для знаходження поля напружень за формулою (4.61). На *Рис. 4.5* показано як з зростанням кількості базисних функцій отримане апроксимоване рішення на прямих $x = \pm 2$ наближається до точного рішення $t_x = E(1-y^2)/(1+\mu)$, що є природною крайовою умовою для даного рівняння. На *Рис. 4.6* показано апроксимоване поле переміщень пластини під дією заданих навантажень описаним методом зважених нев'язок з параметром $M = 3$.

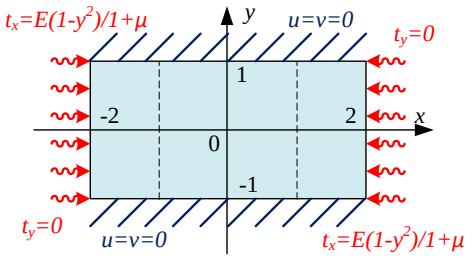


Рис. 4.4 Зображення умов двовимірної задачі лінійної теорії пружності

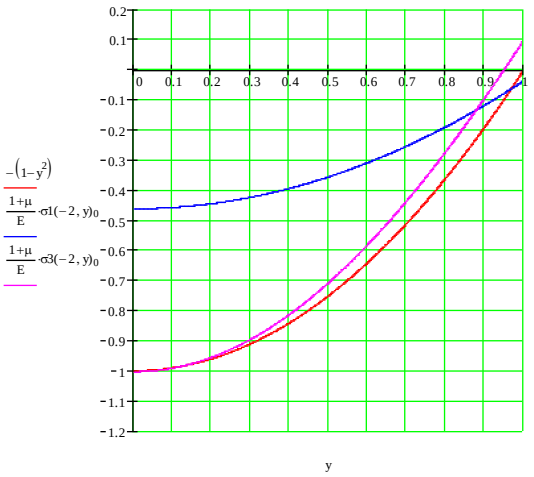


Рис. 4.5 Порівняння точного і наближеного значення напружень пластини на прямих $x = \pm 2$

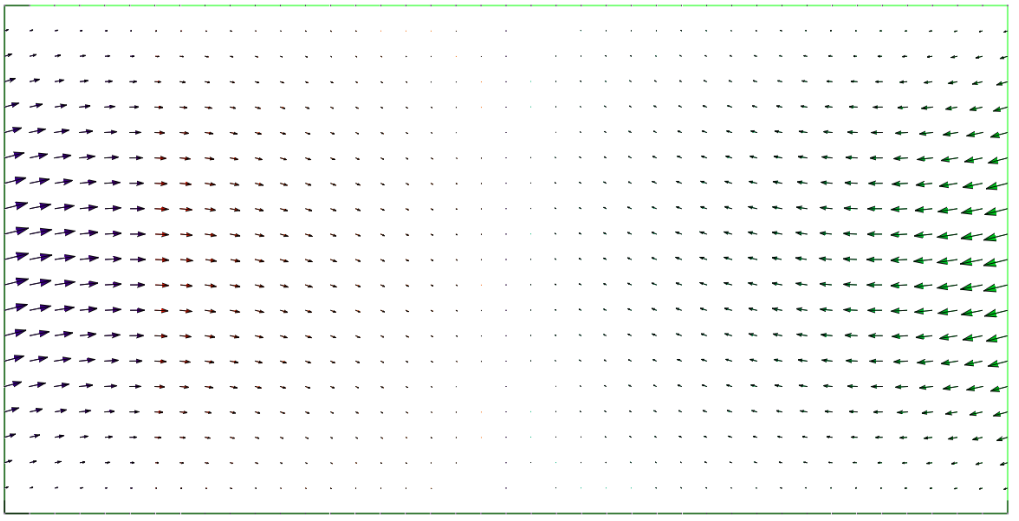


Рис. 4.6 Апроксимоване поле переміщень пластини під дією навантажень

Описаний підхід дуже легко трансформувати на метод скінченних елементів. Оскільки невідомими є кілька величин у вузлі, тобто деякий вектор $\mathbf{u}(\mathbf{r})$, то використовуючи набір функцій форми $N(\mathbf{r})$, розклад наближеного рішення слід розширити для кожної з шуканих величин, аналогічно до (4.51):

$$\begin{aligned} u_{i,1}(\mathbf{r}) &\approx \tilde{u}_{i,1}(\mathbf{r}) = \sum_{j=1}^{M} u_{i,j,1} N_{i,j}(\mathbf{r}), \\ u_{i,2}(\mathbf{r}) &\approx \tilde{u}_{i,2}(\mathbf{r}) = \sum_{j=1}^{M} u_{i,j,2} N_{i,j}(\mathbf{r}), \\ &\dots \\ u_{i,D}(\mathbf{r}) &\approx \tilde{u}_{i,D}(\mathbf{r}) = \sum_{j=1}^{M} u_{i,j,D} N_{i,j}(\mathbf{r}). \end{aligned} \tag{4.80}$$

Щоб отримати вираз у матричній формі, розглянемо двовимірний симплекс елемент, що апроксимує векторне поле, кожна точка якого містить дві шукані змінні, наприклад переміщення по горизонтальній і вертикальній осях. З рівняння (4.80) отримаємо:

$$\begin{aligned} u_{i,1}(\mathbf{r}) &\approx \tilde{u}_{i,1}(\mathbf{r}) = u_{i,1,1} N_{i,1}(\mathbf{r}) + u_{i,2,1} N_{i,2}(\mathbf{r}) + u_{i,3,1} N_{i,3}(\mathbf{r}), \\ u_{i,2}(\mathbf{r}) &\approx \tilde{u}_{i,2}(\mathbf{r}) = u_{i,1,2} N_{i,1}(\mathbf{r}) + u_{i,2,2} N_{i,2}(\mathbf{r}) + u_{i,3,2} N_{i,3}(\mathbf{r}). \end{aligned} \tag{4.81}$$

Об'єднаємо їх як:

$$\mathbf{u}_i(\mathbf{r}) \approx \tilde{\mathbf{u}}_i(\mathbf{r}) = \begin{bmatrix} u_{i,1,1} N_{i,1}(\mathbf{r}) + u_{i,1,2} 0 + u_{i,2,1} N_{i,2}(\mathbf{r}) + u_{i,2,2} 0 + u_{i,3,1} N_{i,3}(\mathbf{r}) + u_{i,3,2} 0 \\ u_{i,1,1} 0 + u_{i,1,2} N_{i,1}(\mathbf{r}) + u_{i,2,1} 0 + u_{i,2,2} N_{i,2}(\mathbf{r}) + u_{i,3,1} 0 + u_{i,3,2} N_{i,3}(\mathbf{r}) \end{bmatrix}, \tag{4.82}$$

або[1]:

$$\mathbf{u}_i(\mathbf{r}) \approx \tilde{\mathbf{u}}_i(\mathbf{r}) = \begin{bmatrix} N_{i,1}(\mathbf{r}) & 0 & N_{i,2}(\mathbf{r}) & 0 & N_{i,3}(\mathbf{r}) & 0 \\ 0 & N_{i,1}(\mathbf{r}) & 0 & N_{i,2}(\mathbf{r}) & 0 & N_{i,3}(\mathbf{r}) \end{bmatrix} \begin{Bmatrix} u_{i,1,1} \\ u_{i,1,2} \\ u_{i,2,1} \\ u_{i,2,2} \\ u_{i,3,1} \\ u_{i,3,2} \end{Bmatrix}. \tag{4.83}$$

Застосовуючи такий підхід для елементів довільної розмірності та довільної кількості вузлів отримаємо загальну матричну форму:

$$\mathbf{u}_i(\mathbf{r}) \approx \tilde{\mathbf{u}}_i(\mathbf{r}) = [[\mathbf{N}_{i,1}(\mathbf{r})] \quad [\mathbf{N}_{i,2}(\mathbf{r})] \quad \dots \quad [\mathbf{N}_{i,D}(\mathbf{r})]]\{\mathbf{u}\}_i = [\mathbf{N}]_i \{\mathbf{u}\}_i, \tag{4.84}$$

де:

[1] Інший можливий варіант записується як:

$$\tilde{\mathbf{u}}_i(\mathbf{r}) = \begin{bmatrix} N_{i,1}(\mathbf{r}) & N_{i,2}(\mathbf{r}) & N_{i,3}(\mathbf{r}) & 0 & 0 & 0 \\ 0 & 0 & 0 & N_{i,1}(\mathbf{r}) & N_{i,2}(\mathbf{r}) & N_{i,3}(\mathbf{r}) \end{bmatrix} \begin{pmatrix} u_{i,1,1} & u_{i,2,1} & u_{i,3,1} & u_{i,1,2} & u_{i,2,2} & u_{i,3,2} \end{pmatrix}^{\mathrm{T}},$$

проте в програмних реалізаціях вигідно тримати в пам'яті підряд обчислені коефіцієнти по кожному з вимірів шуканої векторної величини, тому в літературі по МСЕ переважно використовується саме попередня формула.

$$\{\mathbf{u}\}_i = \left\{[u_{i,1,1}, u_{i,1,2}, \ldots, u_{i,1,D}], [u_{i,2,1}, u_{i,2,2}, \ldots, u_{i,2,D}], \ldots, [u_{i,M,1}, u_{i,M,2}, \ldots, u_{i,M,D}]\right\}^{\mathbf{T}} \quad (4.85)$$

та:

$$[\mathbf{N}_{i,k}(\mathbf{r})] = \begin{bmatrix} N_{i,k}(\mathbf{r}) & 0 & \cdots & 0 \\ 0 & N_{i,k}(\mathbf{r}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & N_{i,k}(\mathbf{r}) \end{bmatrix}, \quad k = 1, 2, \ldots, D. \quad (4.86)$$

Спираючись на те, що $u(\mathbf{r}) = [\mathbf{P}][\mathbf{C}]^{-1}\{\mathbf{u}\}$, для симплекс елементів можна побудувати вираз для функцій форми $N(\mathbf{r})$, де матриця координат симплексу $[\mathbf{C}]$ та матриця поліному $[\mathbf{P}]$ тепер певним чином розширюються, зокрема:

$$[\mathbf{P}] = \begin{bmatrix} 1 & 0 & \cdots & 0 & x_1 & 0 & \cdots & 0 & x_2 & 0 & \cdots & 0 & \cdots & x_N & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & x_1 & \cdots & 0 & 0 & x_2 & \cdots & 0 & \cdots & 0 & x_N & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & x_1 & 0 & 0 & \cdots & x_2 & \cdots & 0 & 0 & \cdots & x_N \end{bmatrix}. \quad (4.87)$$

Подальші судження не є такими ж тривіальними, як для еліптичних рівнянь з шуканим скалярним потенціалом, оскільки матриця градієнтів $[\mathbf{B}]$, тепер є набагато складнішою і залежить від вкладу кожного диференціального оператору з вектору $\mathbf{A}(.)$ системи рівнянь (4.57).

Якщо для задачі можна вивести рівняння в слабкій формі, а для системи рівнянь записати загальний оператор в матричній формі $[\mathcal{L}]$, наприклад як для задачі теорії пружності (4.60), то приймаючи $[\mathbf{B}] = [\mathcal{L}][\mathbf{N}]_i$, локальну матрицю жорсткості та локальний вектор навантаження можна записати у вже звичній формі як:

$$[\mathbf{K}]_i = \int_{\Omega_i} [\mathbf{B}]_i^{\mathbf{T}} [\mathbf{D}]_i [\mathbf{B}]_i \, d\Omega, \quad \{\mathbf{f}\}_i = \int_{\Gamma_i} [\mathbf{N}]_i^{\mathbf{T}} \, f_i \, d\Gamma. \quad (4.88)$$

Якщо ж цього зробити не вдається, то використовують аналог рівняння (4.57):

$$\int_{\Omega_i} [\mathbf{N}]_i^{\mathbf{T}} \mathbf{A}(\tilde{\mathbf{u}}_i) d\Omega = \int_{\Gamma_i} [\mathbf{N}]_i^{\mathbf{T}} \mathbf{B}(\tilde{\mathbf{u}}_i) d\Gamma. \quad (4.89)$$

Для прикладу розглянемо одну з мультфізичних задач, а саме, так звану, зв'язану статичну задачу термопружності [19] в одновимірному випадку [16], [20]. Перш за все запишемо систему диференціальних рівнянь, що визначають фізичний процес – одне рівняння теплопровідності та одне рівняння пружності для плоских напружень:

$$\mathbf{A}(\mathbf{u}) = \begin{bmatrix} dq_x / dx + Q \\ d\sigma_x / dx + X \end{bmatrix}_{\Omega} = 0. \quad (4.90)$$

Зауважимо, що перше рівняння отримується з закону Фур'є: $q = -\lambda \nabla T$, а шуканою величиною є вектор:

$$\mathbf{u}(x, y) = \begin{Bmatrix} T(x, y) \\ u(x, y) \end{Bmatrix}. \quad (4.91)$$

При нагріванні в тілі виникають напруження, спричинені тепловим розширенням матеріалу тіла і навпаки. Розміри L такого розширення при однаковій температурі для різних матеріалів є різними і в загальному випадку описуються лінійним коефіцієнтом теплового розширення α [°C^{-1}] (він може бути анізотропним):

$$\alpha = \frac{1}{L}\frac{\Delta L}{\Delta T} = \frac{1}{L}\frac{dL}{dT}. \tag{4.92}$$

Знаючи лінійний коефіцієнт теплового розширення можна знайти внутрішні сили X, що діють на нескінченно малий об'єм в межах області Ω, а саме:

$$X = \frac{2\alpha E}{1-\mu}\frac{dT}{dx}. \tag{4.93}$$

В свою чергу, під дією напружень матеріал нагрівається – це еквівалентно присутності внутрішнього джерела тепла Q рівному:

$$Q = \frac{2\alpha E}{1-\mu}\frac{du}{dx}. \tag{4.94}$$

Тому, систему диференціальних рівнянь (4.90) можна переписати як:

$$\mathbf{A}(\mathbf{u}) = \begin{bmatrix} dq_x/dx + 2\alpha E/(1-\mu)\cdot du/dx \\ -d\sigma_x/dx + 2\alpha E/(1-\mu)\cdot dT/dx \end{bmatrix}_{\Omega} = 0. \tag{4.95}$$

Як і раніше, доповнимо систему природними крайовими умовами Неймана та головними крайовими умовами Діріхле:

$$\begin{aligned} \mathbf{B}(\mathbf{u}) &= \begin{bmatrix} q_x l_x - q \\ \sigma_x l_x - t_x \end{bmatrix}_{\Gamma_{\text{II}}} = \{\partial \mathbf{u}/\partial \mathbf{n}\} - \{f\}\big|_{\Gamma_{\text{II}}} = 0, \\ \mathbf{B}(\mathbf{u}) &= \begin{bmatrix} T - T_\infty \\ u - u_\infty \end{bmatrix}_{\Gamma_{\text{I}}} = \{\mathbf{u}\} - \{\mathbf{u}_\infty\}\big|_{\Gamma_{\text{I}}} = 0. \end{aligned} \tag{4.96}$$

Систему рівнянь (4.95) можна розписати як:

$$\mathbf{A}(\mathbf{u}) = [\mathcal{L}]^{\mathbf{T}}[\mathbf{D}]\big([\mathcal{L}]\{\mathbf{u}\}\big) + \{\mathbf{X}\} = 0, \tag{4.97}$$

де:

$$[\mathcal{L}] = \begin{bmatrix} d/dx & 0 \\ 0 & d/dx \end{bmatrix}, \quad [\mathbf{D}] = \begin{bmatrix} \lambda & 0 \\ 0 & \dfrac{-E}{1-\mu^2} \end{bmatrix}. \tag{4.98}$$

Перепишемо вектор внутрішніх джерел як:

$$\{\mathbf{X}\} = [\mathcal{L}]^{\mathbf{T}}\big([\mathbf{J}]\{\mathbf{u}\}\big), \tag{4.99}$$

а систему рівнянь (4.74) як:

$$\mathbf{A}(\mathbf{u}) = [\mathcal{L}]^{\mathbf{T}}[\mathbf{D}]\big([\mathcal{L}]\{\mathbf{u}\}\big) + [\mathcal{L}]^{\mathbf{T}}\big([\mathbf{J}]\{\mathbf{u}\}\big) = 0, \tag{4.100}$$

де $[\mathbf{J}]$ – допоміжна матриця виду:

$$[\mathbf{J}] = \frac{2\alpha E}{1-\mu}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{4.101}$$

Побудуємо рівняння методу зважених нев'язок для одновимірного симплекс елементу:

$$\int\limits_{\Omega_i}[\mathbf{N}]_i^{\mathbf{T}}\left([\mathcal{L}]^{\mathbf{T}}[\mathbf{D}]_i\left([\mathcal{L}]\{\tilde{\mathbf{u}}\}_i\right)\right)d\Omega+\int\limits_{\Omega_i}[\mathbf{N}]_i^{\mathbf{T}}\left([\mathcal{L}]^{\mathbf{T}}\left([\mathbf{J}]\{\tilde{\mathbf{u}}\}_i\right)\right)d\Omega-$$
$$-\int\limits_{\Gamma_i}[\mathbf{N}]_i^{\mathbf{T}}\{\partial\mathbf{u}/\partial\mathbf{n}\}d\Gamma+\int\limits_{\Gamma_i}[\mathbf{N}]_i^{\mathbf{T}}\{f\}_i d\Gamma=0. \tag{4.102}$$

Отримане рівняння можна записати у слабкій формі:

$$\left(\int\limits_{\Omega_i}\left([\mathcal{L}][\mathbf{N}]_i\right)^{\mathbf{T}}[\mathbf{D}]_i\left([\mathcal{L}][\mathbf{N}]_i\right)d\Omega-\int\limits_{\Omega_i}[\mathbf{N}]_i^{\mathbf{T}}\left([\mathcal{L}]^{\mathbf{T}}[\mathbf{J}][\mathbf{N}]_i\right)d\Omega\right)\{\mathbf{u}\}_i=\int\limits_{\Gamma_i}[\mathbf{N}]_i^{\mathbf{T}}\{f\}_i d\Gamma. \tag{4.103}$$

Коефіцієнти оберненої матриці координат симплексу $[\mathbf{C}]^{-1}$ рівні:

$$[\mathbf{C}]_i^{-1}=\begin{bmatrix}1 & X_{i,1}\\ 1 & X_{i,2}\end{bmatrix}^{-1}=\begin{bmatrix}X_{i,2}/h_i & -X_{i,1}/h_i\\ -1/h_i & 1/h_i\end{bmatrix}, \tag{4.104}$$

звідки отримаємо:

$$\nabla[\mathbf{N}]_i=\begin{bmatrix}dN_{i,1}/dx & dN_{i,2}/dx\end{bmatrix}=\begin{bmatrix}[\mathbf{C}]_{i,2,1}^{-1} & [\mathbf{C}]_{i,2,2}^{-1}\end{bmatrix}=\begin{bmatrix}-1/h_i & 1/h_i\end{bmatrix}. \tag{4.105}$$

Враховуючи те, що матриця функцій форми симплекс елементу рівна:

$$[\mathbf{N}]_i=\begin{bmatrix}N_{i,1} & 0 & N_{i,2} & 0\\ 0 & N_{i,1} & 0 & N_{i,2}\end{bmatrix}, \tag{4.106}$$

а шуканий вектор вузлових значень:

$$\{\mathbf{u}\}_i=\begin{Bmatrix}T_{i,1} & u_{i,1} & T_{i,2} & u_{i,2}\end{Bmatrix}^{\mathbf{T}} \tag{4.107}$$

вираз $\left([\mathcal{L}][\mathbf{N}]_i\right)^{\mathbf{T}}[\mathbf{D}]_i\left([\mathcal{L}][\mathbf{N}]_i\right)$ можна розписати як:

$$\left([\mathcal{L}][\mathbf{N}]_i\right)^{\mathbf{T}}[\mathbf{D}]_i\left([\mathcal{L}][\mathbf{N}]_i\right)=\frac{1}{h_i^2}\begin{bmatrix}\lambda & 0 & -\lambda & 0\\ 0 & -\dfrac{E}{1-\mu^2} & 0 & \dfrac{E}{1-\mu^2}\\ -\lambda & 0 & \lambda & 0\\ 0 & \dfrac{E}{1-\mu^2} & 0 & -\dfrac{E}{1-\mu^2}\end{bmatrix}, \tag{4.108}$$

і вираз $[\mathbf{N}]_i^{\mathbf{T}}\left([\mathcal{L}]^{\mathbf{T}}[\mathbf{J}][\mathbf{N}]_i\right)$ як:

$$[\mathbf{N}]_i^{\mathbf{T}}\left([\mathcal{L}]^{\mathbf{T}}[\mathbf{J}][\mathbf{N}]_i\right)=\frac{1}{L_i}\begin{bmatrix}0 & -N_{i,1}\dfrac{2\alpha E}{1-\mu} & 0 & N_{i,1}\dfrac{2\alpha E}{1-\mu}\\ -N_{i,1}\dfrac{2\alpha E}{1-\mu} & 0 & N_{i,1}\dfrac{2\alpha E}{1-\mu} & 0\\ 0 & -N_{i,2}\dfrac{2\alpha E}{1-\mu} & 0 & N_{i,2}\dfrac{2\alpha E}{1-\mu}\\ -N_{i,2}\dfrac{2\alpha E}{1-\mu} & 0 & N_{i,2}\dfrac{2\alpha E}{1-\mu} & 0\end{bmatrix}. \tag{4.109}$$

Знайдемо інтеграл від функцій форми:

$$\int_{X_{i,1}}^{X_{i,2}} N_{i,1}dx = \int_{X_{i,1}}^{X_{i,2}} N_{i,2}dx = \frac{h_i}{2} \quad \int_{X_{i,1}}^{X_{i,2}} 1dx = h_i. \tag{4.110}$$

Тепер можна записати вираз для локальних матриці жорсткості та вектору навантажень:

$$[\mathbf{K}]_i = \frac{1}{h_i}\begin{bmatrix} \lambda & -\frac{h_i\alpha E}{1-\mu} & -\lambda & \frac{h_i\alpha E}{1-\mu} \\ -\frac{h_i\alpha E}{1-\mu} & -\frac{E}{1-\mu^2} & \frac{h_i\alpha E}{1-\mu} & \frac{E}{1-\mu^2} \\ -\lambda & -\frac{h_i\alpha E}{1-\mu} & \lambda & \frac{h_i\alpha E}{1-\mu} \\ -\frac{h_i\alpha E}{1-\mu} & \frac{E}{1-\mu^2} & \frac{h_i\alpha E}{1-\mu} & -\frac{E}{1-\mu^2} \end{bmatrix}, \quad \{\mathbf{f}\}_i = \begin{Bmatrix} q_{i,1} \\ t_{x,i,1} \\ q_{i,2} \\ t_{x,i,2} \end{Bmatrix}. \tag{4.111}$$

Нехай алюмінієвий стержень довжиною $0 \le x \le 10$ м, закріплений в точці $x = 0$, там ж підтримується постійна температура 20°C. На протилежному кінці стержень нагрівається потоком тепла з густиною $q = 1000$ Вт/м^2°C. Коефіцієнт теплопровідності $\lambda = 237$ Вт/м°C, модуль пружності матеріалу пластини $E = 70$ ГПа, коефіцієнт Пуассона $\mu = 0{,}34$, коефіцієнт лінійного теплового розширення при температурі 20°C $\alpha = 22{,}2 \times 10^{-6}\,°\text{C}^{-1}$. Необхідно знайти утворене температурне поле та поле переміщень стержня.

Розіб'ємо область регулярною сіткою зі 100 симплекс елементів, для кожного з яких знайдемо локальну матрицю жорсткості, враховуючи що $h_i = 0{,}1$, $\alpha E/(1-\mu) = 0{,}002355$, $E/(1-\mu^2) = 79{,}149706$:

$$[\mathbf{K}]_i = \begin{bmatrix} 2370 & -0{,}001757 & -2370 & 0{,}001757 \\ -0{,}001757 & -791{,}497060 & 0{,}001757 & 791{,}497060 \\ -2370 & -0{,}001757 & 2370 & 0{,}001757 \\ -0{,}001757 & 791{,}497060 & 0{,}001757 & -791{,}497060 \end{bmatrix}. \tag{4.112}$$

Потік тепла q враховується в передостанньому коефіцієнті глобального вектору навантажень, або передостанньому коефіцієнті локального вектору навантажень останнього симплекс елементу:

$$\{\mathbf{f}\}_{100} = \begin{Bmatrix} q_{i,1} \\ t_{x,i,1} \\ q_{i,2} \\ t_{x,i,2} \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 1000 \\ 0 \end{Bmatrix}. \tag{4.113}$$

З умов задачі відомо, що $T(0) = 20$ та $u(0) = 0$, тому перед процесом ансамблювання модифікуємо локальну матрицю жорсткості і вектор навантажень першого симплекс елементу:

$$[\mathbf{K}]_1 = \begin{bmatrix} 2370 & 0 & 0 & 0 \\ 0 & -791{,}497060 & 0 & 0 \\ 0 & 0 & 2370 & 0{,}001757 \\ 0 & 0 & 0{,}001757 & -791{,}497060 \end{bmatrix} \quad \{\mathbf{f}\}_1 = \begin{Bmatrix} 47400 \\ 0 \\ 47400 \\ 0{,}035142 \end{Bmatrix}. \quad (4.114)$$

Тепер можна побудувати глобальну систему рівнянь, фрагмент якої показано нижче:

$$[\mathbf{K}] = \begin{bmatrix} 2370 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & -791{,}497060 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 47400 & 0 & -2370 & 0{,}001757 & \cdots \\ 0 & 0 & 0 & -1582{,}994120 & 0{,}001757 & 791{,}497060 & \cdots \\ 0 & 0 & -2370 & -0{,}001757 & 47400 & 0 & \cdots \\ 0 & 0 & -0{,}001757 & 791{,}497060 & 0 & -1582{,}994120 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad (4.115)$$

$$\{\mathbf{f}\} = \{47400 \quad 0 \quad 47400 \quad 0{,}035142 \quad 0 \quad 0 \quad \ldots\}^{\mathbf{T}}.$$

Після розв'язку системи отримаємо шуканий вектор вузлових значень:

$$\{\mathbf{u}\} = \{20 \quad 0 \quad 20{,}421941 \quad 0{,}000186 \quad 20{,}843882 \quad 0{,}000371 \quad \ldots\}^{\mathbf{T}}. \quad (4.116)$$

На *Рис. 4.7* зображене апроксимоване поле температур, а на *Рис. 4.8* зображено апроксимоване поле переміщень. Перевіримо правильність отриманих результатів: у точці $x = 0$ переміщення рівні нулю за умовою задачі, у точці $x = 10$ ми отримали апроксимований результат рівний $\tilde{u}(10) = 0{,}009367$ м, тобто утворилася різниця довжини стержня $\Delta L = 0{,}009367$ м. Апроксимована температура в точці $x = 10$ рівна $\tilde{T}(10) = 62{,}194092\,^{\circ}\text{C}$, тобто утворилася різниця температур $\Delta T = 42{,}194092\,^{\circ}\text{C}$. Відомо, що $\Delta T = d \cdot q / \lambda = 10 \cdot 1000 / 237 = 42{,}194092\,^{\circ}\text{C}$, саме таке значення ми і отримали. З формули (4.92) відомо, що $\alpha = \Delta L / (L \Delta T) = 0{,}009367 / (10 \cdot 42{,}194092) = 22{,}2 \times 10^{-6}\,^{\circ}\text{C}^{-1}$, саме таке значення визначалося умовами задачі. Отже отримані результати можна вважати правильними.

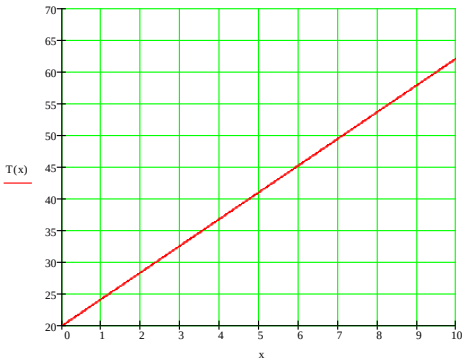


Рис. 4.7 100-елементна апроксимація поля температур задачі термопружності

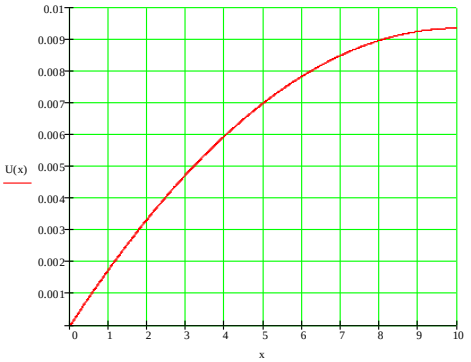


Рис. 4.8 100-елементна апроксимація поля переміщень задачі термопружності

В літературі по теорії термопружності подібні зв'язані задачі прийнято розв'язувати в два етапи – спочатку окремо знайти температурне поле, а потім на основі нього знайти поле переміщень, або навпаки, в залежності від наявних крайових умов, тобто від присутності теплового потоку чи поверхневого навантаження.

Натомість описаний підхід дозволяє враховувати обидві умови одночасно за їх наявності. Більше того, використовуючи такий підхід можна розв'язувати будь-які мультифізичні задачі, за умови наявності визначних диференціальних рівнянь.

Основна проблема при цьому полягає в значній кількості обчислень при виведенні локальних матриць жорсткості та векторів навантаження. Останнє пояснюється наявністю великої кількості, так званих, *степенів свободи* у кожному з вузлів скінченно-елементної сітки. Наприклад, для тривимірної задачі термопружності в плоских напруженнях у кожному вузлі необхідно апроксимувати чотири невідомі – температуру і три переміщення вздовж координатних осей. Використовуючи симплекс елементи, а в тривимірному просторі це тетраедри з чотирма вершинами, отримаємо локальну матрицю жорсткості розмірами шістнадцять на шістнадцять коефіцієнтів. І оскільки для достатньо точної апроксимації необхідно використовувати багато елементів, отримуємо дуже великі глобальні системи рівнянь, процес розв'язку яких неможливо уявити без застосування потужної обчислювальної техніки.

4.3. Рішення нестаціонарних задач

Статичні процеси, що описувалися еліптичними рівняннями та розглядалися до цього моменту, є лише частиною широкого кола фізичних явищ, які піддаються моделюванню за допомогою методів зважених нев'язок. Наступним класом процесів, що будуть розглядатися є динамічні або нестаціонарні процеси. Крайові задачі, що описують такі процеси, передбачають зміну шуканих величин по часовій координаті. У деяких з цих задач розглядається так званий перехідних період, або період релаксації, що описує поведінку об'єкту моделювання між початком фізичного процесу і досягненням його стаціонарного стану. Зустрічаються також задачі, в яких стаціонарний стан взагалі не досяжний і період релаксації займає увесь фізичний процес.

З нестаціонарними задачами дуже часто зустрічаються при дослідженні явищ переносу тепла чи матерії, моделюванні процесів протікання рідин, дослідженні динаміки руху різноманітних конструкцій, моделюванні еволюції складних систем та багато ін. Переважна більшість з таких задач описується параболічними рівняннями, де присутні частинні похідні першого порядку від невідомої функції по часовій координаті. Крім того, сюди відносяться майже всі коливальні системи та хвильові процеси, що описуються гіперболічними рівняннями, де присутні частинні похідні другого порядку від невідомої функції по часовій координаті.

Як правило, крім крайових умов, в задачі задано стан системи в початковий

момент часу, або початкові умови. Необхідно знайти стани системи в деяких наступних моментах часу. Як вже було сказано, формально крайові та початкові умови нічим не відрізняються, оскільки часова координата в принципі володіє такими ж характеристиками, як і просторові. Тому завжди допускається можливість підбирати апроксимаційні функції, що розглядаються по всій просторово-часовій області та будувати наближене рішення так, як це було показано у попередніх прикладах. Проте, затрати зусиль на такий підхід не є виправданими, оскільки [1]:

- у випадку розгляду великих часових інтервалів, задача стає надто громіздкою, при цьому поведінка шуканої функції в цих інтервалах зазвичай не є непередбачуваною;
- отримана система лінійних рівнянь в загальному випадку не є симетричною навіть при використанні методів Бубнова-Гальоркіна;
- геометрична простота часової осі, у порівнянні з просторовими координатами об'єкту моделювання, не стимулює шукати складні апроксимації в просторово-часовій області, крім того в ряді задач наперед невідомо довжину часового інтервалу в якому відбувається фізичний процес;
- результати апроксимації в просторово-часовій області співпадають з ітераційною послідовною апроксимацією спочатку в просторі а потім в часі.

Саме для цього використовується, так звана, процедура часткової дискретизації, відома також під назвами метод Канторовича або метод прямих [1].

Припустимо, що шуканий потенціал u залежить від чотирьох незалежних змінних x, y, z, τ, що б вони не позначали. Спробуємо вибрати у якості коефіцієнтів a_j розкладу наближеного рішення $\tilde{u}$, деякі функції, що залежать від однієї зі змінних, наприклад $a_j(\tau)$:

$$u(x,y,z,\tau) \approx \tilde{u}(x,y,z,\tau) = u_0(x,y,z,\tau) + \sum_{j=1}^{M} a_j(\tau)\varphi_j(x,y,z), \qquad (4.117)$$

де тепер u_0 та φ_j вибрані так, щоб розклад автоматично задовольняв головні крайові умови задачі. Тоді при застосуванні методу зважених нев'язок похідні від a_j по τ зберігаються і отримані рівняння утворюють систему звичайних диференціальних рівнянь з незалежною змінною τ. Як наслідок, стандартна система $[\mathbf{K}]\{\mathbf{a}\} = \{\mathbf{f}\}$ заміняється системою:

$$[\mathbf{K}]\{\mathbf{a}\} + [\mathbf{C}]\frac{d}{d\tau}\{\mathbf{a}\} + \ldots = \{\mathbf{f}\}, \qquad (4.118)$$

порядок якої залежить від порядку найвищої похідної по τ, що входить у вихідне рівняння для потенціалу u. Очевидно, що описаний підхід може бути ефективно застосований до параболічних рівнянь для здійснення подальшої

часткової дискретизації по часовій координаті τ. Знову ж таки, починаючи з рішення задач механіки, матрицю [**C**] прийнято називати *матрицею демпфування* [21] (тобто релаксації, або затухання).

Для багатьох лінійних нестаціонарних фізичних задач диференціальне рівняння можна представити у вигляді:

$$\mathcal{L}u + k - \alpha\frac{du}{d\tau} - \beta\frac{d^2u}{d\tau^2} = 0, \qquad (4.119)$$

де, $\mathcal{L}$ – диференціальний оператор, що включає диференціювання тільки по просторових координатах, а k, α і β – задані функції координат і часу. Зокрема:

- під таке формулювання підпадають такі гіперболічні рівняння, як рівняння опису поперечних коливань натягнутої струни з густиною ρ та натягом T:

  $$\frac{\partial^2u}{\partial x^2} - \frac{\rho}{T}\frac{d^2u}{d\tau^2} = 0, \qquad (4.120)$$

  це рівняння співпадає з (4.119) при $\alpha = k = 0$, $\beta = \rho/T$ і $\mathcal{L}u = \partial^2u/\partial x^2$,
- а також, такі параболічні рівняння, як рівняння опису лінійних нестаціонарних процесів переносу тепла в матеріалі з густиною ρ, питомою теплоємністю c, коефіцієнтом теплопровідності λ і внутрішнім джерелом тепла Q:

  $$\lambda\nabla^2T + Q - \rho c\frac{\partial T}{\partial\tau} = 0, \qquad (4.121)$$

  це рівняння співпадає з (4.119) при $\alpha = \rho c$, $\beta = 0$, $k = Q$ і $\mathcal{L}u = \nabla^2T$.

Застосовуючи метод зважених нев'язок до рівняння (4.119) отримаємо систему звичайних диференціальних рівнянь:

$$[\mathbf{M}]\frac{d^2}{d\tau^2}\{\mathbf{a}\} + [\mathbf{C}]\frac{d}{d\tau}\{\mathbf{a}\} + [\mathbf{K}]\{\mathbf{a}\} = \{\mathbf{f}\}, \qquad (4.122)$$

де компоненти окремих матриць і правої частини визначаються як:

$$[\mathbf{M}]_{i,j} = \int_\Omega \beta\omega_i\varphi_j d\Omega, \quad [\mathbf{C}]_{i,j} = \int_\Omega \alpha\omega_i\varphi_j d\Omega,$$
$$[\mathbf{K}]_{i,j} = -\int_\Omega \omega_i\mathcal{L}\varphi_j d\Omega, \quad \{\mathbf{f}\}_i = \int_\Omega\left(k + \mathcal{L}u_0 - \alpha\frac{du_0}{d\tau} - \beta\frac{d^2u_0}{d\tau^2}\right)d\Omega. \qquad (4.123)$$

Матрицю [**M**] прийнято називати *матрицею маси*. Залишається розв'язати дану систему рівнянь зі вказаними при $\tau = 0$ значеннями {**u**} і (якщо $\beta \neq 0$) $d\{\mathbf{u}\}/d\tau$. Це класична задача теорії звичайних диференціальних рівнянь, і в принципі її можна розв'язати точно. Крім того, при застосуванні методу Бубунова-Гальоркіна, отримані матриці [**M**], [**C**], [**K**] будуть симетричними.

Для прикладу розглянемо одновимірну задачу нестаціонарної

теплопровідності. Нехай золотий стержень займає область $0 \le x \le 0{,}1$ м. На одному з країв підтримується постійна температура $T_0 = T(0,\tau) = 0\,^\circ$C, з іншої сторони подається тепло з швидкістю $q = 1000$ Вт/м^2°C. Коефіцієнт теплопровідності $\lambda = 320$ Вт/м°C, питома теплоємність матеріалу $c = 129$ Дж/кг°C, густина $\rho = 19300$ кг/м^2. В початковий момент часу розподіл температури в стержні рівний $T_0 = T(x,0) = 0\,^\circ$C. Необхідно знайти розподіл температурного поля в період $0 \le \tau \le 180$ с.

$$c\rho\frac{\partial T(x,\tau)}{\partial \tau} = \lambda\frac{\partial^2 T(x,\tau)}{\partial x^2}, \quad \frac{\partial T(0.1,\tau)}{\partial \mathbf{n}} = 1000, \tag{4.124}$$
$$T(0,\tau) = 0, \quad T(x,0) = 0, \quad 0 \le x \le 0{,}1, \quad 0 \le \tau \le 180.$$

Запишемо розклад наближеного розв'язку задачі:

$$T(x,\tau) \approx \tilde{T}(x,\tau) = T_0(x) + \sum_{j=1}^{M} a_j(\tau)\varphi_j(x). \tag{4.125}$$

Виберемо $T_0(x) = T_0$, $\varphi_j(x) = \omega_j(x) = x^j$, очевидно що система обраних базисних функцій автоматично задовольняє початкові та головні крайові умови.

Обчислюючи суму скалярних добутків пробних рішень та вагових функцій отримаємо:

$$\int_0^{0.1}\omega_i\left(\lambda\frac{\partial^2\tilde{T}}{\partial x^2} - c\rho\frac{\partial\tilde{T}}{\partial\tau}\right)dx - \left[\omega_i\left(\frac{\partial T}{\partial \mathbf{n}} - q\right)\right]_{x=0}^{x=0.1} = 0, \quad i = 1,2,\ldots M. \tag{4.126}$$

Еквівалентна слабка форма рівняння, отримана з допомогою інтегрування за частинами запишеться як:

$$\int_0^{0.1}\left(\omega_i c\rho\frac{\partial\tilde{T}}{\partial\tau}\right)dx + \int_0^{0.1}\left(\frac{\partial\omega_i}{\partial x}\lambda\frac{\partial\tilde{T}}{\partial x}\right)dx - \left[\omega_i q\right]_{x=0}^{x=0.1} = 0, \quad i = 1,2,\ldots M. \tag{4.127}$$

Підставляючи сюди апроксимацію (4.125) отримаємо:

$$\left(\int_0^{0.1}\omega_i c\rho\varphi_j dx\right)\frac{\partial a_j}{\partial\tau} + \left(\int_0^{0.1}\frac{\partial\omega_i}{\partial x}\lambda\frac{\partial\varphi_j}{\partial x}dx\right)a_j = \left[\omega_i q\right]_{x=0}^{x=0.1} \quad i = 1,2,\ldots M, \tag{4.128}$$

або скорочено:

$$[\mathbf{C}]\frac{d}{d\tau}\{\mathbf{a}\} + [\mathbf{K}]\{\mathbf{a}\} = \{\mathbf{f}\}, \tag{4.129}$$

де:

$$[\mathbf{C}]_{i,j} = \int_0^{0.1}\omega_i c\rho\varphi_j dx \quad [\mathbf{K}]_{i,j} = \int_0^{0.1}\frac{\partial\omega_i}{\partial x}\lambda\frac{\partial\varphi_j}{\partial x}dx \quad \{\mathbf{f}\} = \left[\omega_i q\right]_{x=0}^{x=0.1}. \tag{4.130}$$

Враховуючи лінійність системи звичайних диференціальних рівнянь (4.129), можна було б побудувати її аналітичне рішення. Але, щоб показати процедуру, яка характерна і для більш складних задач, проведемо чисельне інтегрування системи. Для цього використаємо скінченно-різницеву апроксимацію по часовій координаті (іншим поширеним методом такої

апроксимації є використання тих ж методів зважених нев'язок, що були описані раніше [1], [21], [22]).

Розіб'ємо часовий інтервал на множину моментів $\tau_0, \tau_1, \ldots, \tau_n, \ldots$ з кроком $\Delta\tau$. Тоді розклад в ряд Тейлора шуканих коефіцієнтів $\{\mathbf{a}\}$ можна записати як:

$$\mathbf{a}(\tau_{n+1}) = \mathbf{a}(\tau_n + \Delta\tau) = \mathbf{a}(\tau_n) + \Delta\tau \left.\frac{d\mathbf{a}(\tau)}{d\tau}\right|_{\tau_n} + \frac{\Delta\tau^2}{2} \left.\frac{d^2\mathbf{a}(\tau)}{d\tau^2}\right|_{\tau_n} \ldots \quad (4.131)$$

Вибираючи достатньо малий крок $\Delta\tau >> \Delta\tau^2$, знайдемо вираз для першої похідної, відкинувши всі доданки, після першого:

$$\left.\frac{d\mathbf{a}(\tau)}{d\tau}\right|_{\tau_n} \approx \frac{\mathbf{a}(\tau_{n+1}) - \mathbf{a}(\tau_n)}{\Delta\tau}. \quad (4.132)$$

Отриманий вираз також називають скінченно-різницевою апроксимацією *вперед*. Очевидно, що похибка такої апроксимації задається виразом:

$$\varepsilon = -\frac{\Delta\tau}{2} \left.\frac{d^2\mathbf{a}(\tau)}{d\tau^2}\right|_{\tau_n+\theta}, \quad (4.133)$$

де, θ – деяке число, $0 \le \theta \le 1$, обумовлене наявністю залишкових членів розкладу в ряд Тейлора. Значення похибки ε пропорційне до кроку $\Delta\tau$, така апроксимація має *перший порядок* точності і позначається як $O(\Delta\tau)$. З отриманого виразу неможливо знайти точне значення похибки, оскільки число θ є невідомим, однак відомо, що:

$$|\varepsilon| \le \left(\frac{\Delta\tau}{2}\right) \max_{[\tau_n, \tau_{n+1}]} \left|\frac{d^2\mathbf{a}(\tau)}{d\tau^2}\right|. \quad (4.134)$$

Аналогічним чином, можна отримати формулу для скінченно-різницевої апроксимації *назад*:

$$\left.\frac{d\mathbf{a}(\tau)}{d\tau}\right|_{\tau_n} \approx \frac{\mathbf{a}(\tau_n) - \mathbf{a}(\tau_{n-1})}{\Delta\tau}, \quad (4.135)$$

похибка такої апроксимації також має порядок $O(\Delta\tau)$.

Якщо не обмежуватись першим доданком розкладу в ряд Тейлора і записати:

$$\begin{aligned}
&\mathbf{a}(\tau_{n+1}) = \mathbf{a}(\tau_n + \Delta\tau) = \mathbf{a}(\tau_n) + \Delta\tau \left.\frac{d\mathbf{a}(\tau)}{d\tau}\right|_{\tau_n} + \\
&+ \frac{\Delta\tau^2}{2} \left.\frac{d^2\mathbf{a}(\tau)}{d\tau^2}\right|_{\tau_n} + \frac{\Delta\tau^3}{6} \left.\frac{d^3\mathbf{a}(\tau)}{d\tau^3}\right|_{\tau_n+\theta_1}, \quad 0 \le \theta_1 \le 1, \\
&\mathbf{a}(\tau_{n-1}) = \mathbf{a}(\tau_n - \Delta\tau) = \mathbf{a}(\tau_n) - \Delta\tau \left.\frac{d\mathbf{a}(\tau)}{d\tau}\right|_{\tau_n} + \\
&+ \frac{\Delta\tau^2}{2} \left.\frac{d^2\mathbf{a}(\tau)}{d\tau^2}\right|_{\tau_n} - \frac{\Delta\tau^3}{6} \left.\frac{d^3\mathbf{a}(\tau)}{d\tau^3}\right|_{\tau_n+\theta_2}, \quad 0 \le \theta_2 \le 1,
\end{aligned} \quad (4.136)$$

а потім від першого виразу відняти другий:

$$\mathbf{a}(\tau_{n+1})-\mathbf{a}(\tau_{n-1})=2\Delta\tau\left.\frac{d\mathbf{a}(\tau)}{d\tau}\right|_{\tau_n}+\frac{\Delta\tau^3}{6}\left(\left.\frac{d^3\mathbf{a}(\tau)}{d\tau^3}\right|_{\tau_n+\theta_1}+\left.\frac{\Delta\tau^3}{6}\frac{d^3\mathbf{a}(\tau)}{d\tau^3}\right|_{\tau_n+\theta_2}\right),\quad(4.137)$$

то більш точну апроксимацію похідної можна отримати як, скінченно-різницеву апроксимації з *центральною різницею*:

$$\left.\frac{d\mathbf{a}(\tau)}{d\tau}\right|_{\tau_n}\approx\frac{\mathbf{a}(\tau_{n+1})-\mathbf{a}(\tau_{n-1})}{2\Delta\tau}.\quad(4.138)$$

Похибка апроксимації виражається як:

$$|\varepsilon|\le\left(\frac{\Delta\tau^2}{6}\right)\max_{[\tau_{n-1},\tau_{n+1}]}\left|\frac{d^3\mathbf{a}(\tau)}{d\tau^3}\right|,\quad(4.139)$$

і має порядок $O(\Delta\tau^2)$.

Вирази для більш точних апроксимацій та вирази для похідних вищих порядків, за необхідності, можна знайти аналогічним чином.

Повернемося до розгляду системи рівнянь (4.129). Щоб мати можливість об'єднати різні види скінченно-різницевих апроксимацій, введемо додатковий параметр $0\le\gamma\le1$, який крім того можна виразити і при застосуванні методів зважених нев'язок до окремої апроксимації по часовій координаті, наприклад параметр γ може позначати точку коллокації у відповідному методі. Апроксимуючи розв'язок системи (4.129) за допомогою методу скінченних різниць, для деякого моменту часу отримаємо:

$$[\mathbf{C}]\frac{d}{d\tau}\{\mathbf{a}\}\Big|_{\tau=\tau_n+\gamma\Delta\tau}+[\mathbf{K}]\{\mathbf{a}\}\big|_{\tau=\tau_n+\gamma\Delta\tau}=\{\mathbf{f}\}_{\tau=\tau_n+\gamma\Delta\tau}.\quad(4.140)$$

Використовуючи розклад в ряд Тейлора виду (4.131), отримаємо:

$$\begin{gathered}\mathbf{a}(\tau_n)=\mathbf{a}(\tau_n+\gamma\Delta\tau)+\gamma\Delta\tau\left.\frac{d\mathbf{a}(\tau)}{d\tau}\right|_{\tau_n+\gamma\Delta\tau}+\\+\frac{\gamma^2\Delta\tau^2}{2}\left.\frac{d^2\mathbf{a}(\tau)}{d\tau^2}\right|_{\tau_n+\gamma\Delta\tau}+\frac{\gamma^3\Delta\tau^3}{6}\left.\frac{d^3\mathbf{a}(\tau)}{d\tau^3}\right|_{\tau_n+\theta_1}\\\mathbf{a}(\tau_{n+1})=\mathbf{a}(\tau_n+\gamma\Delta\tau)+(1-\gamma)\Delta\tau\left.\frac{d\mathbf{a}(\tau)}{d\tau}\right|_{\tau_n+\gamma\Delta\tau}+\\+\frac{(1-\gamma)^2\Delta\tau^2}{2}\left.\frac{d^2\mathbf{a}(\tau)}{d\tau^2}\right|_{\tau_n+\gamma\Delta\tau}+\frac{(1-\gamma)^3\Delta\tau^3}{6}\left.\frac{d^3\mathbf{a}(\tau)}{d\tau^3}\right|_{\tau_n+\theta_2}\\\left.\frac{d\mathbf{a}(\tau)}{d\tau}\right|_{\tau_n+\gamma\Delta\tau}\approx\frac{\mathbf{a}(\tau_{n+1})-\mathbf{a}(\tau_n)}{\Delta\tau}.\end{gathered}\quad(4.141)$$

з похибкою ε виду:

$$\varepsilon=\left((1-\gamma)^2-\gamma^2\right)\frac{\Delta\tau}{2}\left.\frac{d^2\mathbf{a}(\tau)}{d\tau^2}\right|_{\tau_n+\gamma\Delta\tau}+$$

$$+\frac{\Delta\tau^2}{6}\left((1-\gamma)^3\frac{d^3\mathbf{a}(\tau)}{d\tau^3}\bigg|_{\tau_n+\theta_2}+\gamma^3\frac{d^3\mathbf{a}(\tau)}{d\tau^3}\bigg|_{\tau_n+\theta_1}\right). \qquad (4.142)$$

Слід зазначити, що ε має порядок $O(\Delta\tau^2)$ при $\gamma=1/2$ і $O(\Delta\tau)$ в усіх інших випадках. Помноживши перше рівняння з (4.141) на $(1-\gamma)$, а друге рівняння на γ і додавши результати отримаємо:

$$\mathbf{a}(\tau_n+\gamma\Delta\tau)\approx(1-\gamma)\mathbf{a}(\tau_n)+\gamma\mathbf{a}(\tau_{n+1}), \qquad (4.143)$$

де похибка має порядок $O(\Delta\tau^2)$.

Застосувавши до вектору навантажень $\{\mathbf{f}\}$ процедуру, аналогічну до (4.143), і підставивши (4.141) та (4.143) в (4.140), отримаємо:

$$\left(\frac{[\mathbf{C}]}{\Delta\tau}+\gamma[\mathbf{K}]\right)\{\mathbf{a}^{\tau_{n+1}}\}+\left(-\frac{[\mathbf{C}]}{\Delta\tau}+(1-\gamma)[\mathbf{K}]\right)\{\mathbf{a}^{\tau_n}\}=(1-\gamma)\{\mathbf{f}^{\tau_n}\}+\gamma\{\mathbf{f}^{\tau_{n+1}}\}. \quad (4.144)$$

Зауважимо, що отримана система рівнянь точно співпадає з системою рівнянь, що отримується при застосуванні методів зважених нев'язок до окремої апроксимації по часовій координаті [1].

Таким чином, за допомогою підбору параметру γ отримаємо одну з відомих скінченно-різницевих схем для рівняння типу (4.129):

- **Схема Ейлера (схема з різницею вперед)**

Схема отримується при $\gamma=0$, що дає:

$$\left(\frac{[\mathbf{C}]}{\Delta\tau}\right)\{\mathbf{a}^{\tau_{n+1}}\}+\left(-\frac{[\mathbf{C}]}{\Delta\tau}+[\mathbf{K}]\right)\{\mathbf{a}^{\tau_n}\}=\{\mathbf{f}^{\tau_n}\}. \qquad (4.145)$$

Якщо матриця демпфування $[\mathbf{C}]$ є діагональною, і як наслідок, елементи оберненої матриці $[\mathbf{C}]^{-1}$ можуть бути отримані безпосередньо, то така схема називається *явною*, оскільки $\{\mathbf{a}^{\tau_{n+1}}\}$ виражається явно через $\{\mathbf{a}^{\tau_n}\}$ через співвідношення:

$$\{\mathbf{a}^{\tau_{n+1}}\}=\Delta\tau[\mathbf{C}]^{-1}\left[\left(\frac{[\mathbf{C}]}{\Delta\tau}-[\mathbf{K}]\right)\{\mathbf{a}^{\tau_n}\}+\{\mathbf{f}^{\tau_n}\}\right]. \qquad (4.146)$$

Дана схема відноситься до класу умовно стійких схем, це означає, що крок $\Delta\tau$ повинен бути достатньо малим і не перевищувати деяку величину, щоб метод був стійким, а отримане апроксимоване рішення з кожною ітерацією не губилося в похибках заокруглення чи в похибках апроксимації.

- **Схема Кранка-Ніколсона (схема з центральною різницею)**

Схема отримується при $\gamma=1/2$, що дає:

$$\left(\frac{[\mathbf{C}]}{\Delta\tau}+\frac{1}{2}[\mathbf{K}]\right)\{\mathbf{a}^{\tau_{n+1}}\}+\left(-\frac{[\mathbf{C}]}{\Delta\tau}+\frac{1}{2}[\mathbf{K}]\right)\{\mathbf{a}^{\tau_n}\}=\frac{1}{2}\{\mathbf{f}^{\tau_n}\}+\frac{1}{2}\{\mathbf{f}^{\tau_{n+1}}\}. \quad (4.147)$$

Дана схема є *неявною*, оскільки щоб знайти $\{\mathbf{a}^{\tau_{n+1}}\}$, необхідно рішити систему

рівнянь, матриця якої не є діагональною. Похибка апроксимації в схемі має другий порядок точності. Схема є безумовно стійкою [21], це означає, що стійкість обчислень не залежить від кроку $\Delta\tau$. Втім, останнє твердження ніяк не впливає на похибки заокруглень при обчисленнях.

- **Схема з різницею назад**

Схема є неявною і отримується при $\gamma = 1$, що дає:

$$\left(\frac{[\mathbf{C}]}{\Delta\tau} + [\mathbf{K}]\right)\{\mathbf{a}^{\tau_{n+1}}\} + \left(-\frac{[\mathbf{C}]}{\Delta\tau}\right)\{\mathbf{a}^{\tau_n}\} = \frac{1}{2}\{\mathbf{f}^{\tau_{n+1}}\}. \tag{4.148}$$

- **Багатошарові схеми**

Аналогічно до скінченно-різницевих схем вищих порядків, які виводяться врахуванням більшої кількості доданків в розкладі в ряд Тейлора, за допомогою застосування методів зважених нев'язок для окремої апроксимації по часовій координаті, при необхідності можна вивести так звані багатошарові схеми бажаного порядку точності [1], [21]. Наприклад, наближене рішення для $\{\mathbf{a}^{\tau_n}\}$ будується аналогічно до звичайного розкладу:

$$\mathbf{a}(\tau) \approx \tilde{\mathbf{a}}(\tau) = \sum_{n=1}^{\infty} \mathbf{a}_{\tau_n} \varphi_{\tau_n}(\tau). \tag{4.149}$$

Систему базисних функцій підберемо так, щоб на кожному часовому кроці:

$$\{\tilde{\mathbf{a}}\} \approx [\mathbf{\Phi}^{\tau_n}]\{\mathbf{a}^{\tau_n}\} + [\mathbf{\Phi}^{\tau_{n+1}}]\{\mathbf{a}^{\tau_{n+1}}\}, \tag{4.150}$$

де:

$$\begin{gathered}[\mathbf{\Phi}^{\tau_n}] = \frac{\tau_n - \tau}{\Delta\tau} \quad \frac{d[\mathbf{\Phi}^{\tau_n}]}{d\tau} = -\frac{1}{\Delta\tau} \\ [\mathbf{\Phi}^{\tau_{n+1}}] = \frac{\tau - \tau_n}{\Delta\tau} \quad \frac{d[\mathbf{\Phi}^{\tau_{n+1}}]}{d\tau} = \frac{1}{\Delta\tau}, \\ \Delta\tau = \tau_{n+1} - \tau_n.\end{gathered} \tag{4.151}$$

Рівняння методів зважених нев'язок запишеться як:

$$\int_{\tau_n}^{\tau_{n+1}} [\mathbf{W}^{\tau_n}]\left([\mathbf{C}]\frac{d}{d\tau}\{\tilde{\mathbf{a}}\} + [\mathbf{K}]\{\tilde{\mathbf{a}}\} - \{\mathbf{f}\}\right)d\tau = 0, \quad n = 1,2,\ldots \tag{4.152}$$

Приймемо $[\mathbf{W}^{\tau_{n-1}}] = [\mathbf{\Phi}^{\tau_{n-1}}]$, $[\mathbf{W}^{\tau_n}] = [\mathbf{\Phi}^{\tau_n}]$ та $[\mathbf{W}^{\tau_{n+1}}] = [\mathbf{\Phi}^{\tau_{n+1}}]$, виразимо $\{\mathbf{f}\}$ аналогічно до (4.150), звідки отримаємо:

$$\begin{gathered}\int_{\tau_{n-1}}^{\tau_n} [\mathbf{W}^{\tau_{n-1}}]\left([\mathbf{C}]\left(\{\mathbf{a}^{\tau_{n-1}}\}\frac{d[\mathbf{W}^{\tau_{n-1}}]}{d\tau} + \{\mathbf{a}^{\tau_n}\}\frac{d[\mathbf{W}^{\tau_n}]}{d\tau}\right) + [\mathbf{K}]\left(\{\mathbf{a}^{\tau_{n-1}}\}[\mathbf{W}^{\tau_{n-1}}] + \{\mathbf{a}^{\tau_n}\}[\mathbf{W}^{\tau_n}]\right)\right)d\tau + \\ + \int_{\tau_n}^{\tau_{n+1}} [\mathbf{W}^{\tau_n}]\left([\mathbf{C}]\left(\{\mathbf{a}^{\tau_n}\}\frac{d[\mathbf{W}^{\tau_n}]}{d\tau} + \{\mathbf{a}^{\tau_{n+1}}\}\frac{d[\mathbf{W}^{\tau_{n+1}}]}{d\tau}\right) + [\mathbf{K}]\left(\{\mathbf{a}^{\tau_n}\}[\mathbf{W}^{\tau_n}] + \{\mathbf{a}^{\tau_{n+1}}\}[\mathbf{W}^{\tau_{n+1}}]\right)\right)d\tau = \\ = \int_{\tau_{n-1}}^{\tau_n} [\mathbf{W}^{\tau_{n-1}}]\left(\{\mathbf{f}^{\tau_{n-1}}\}[\mathbf{W}^{\tau_{n-1}}] + \{\mathbf{f}^{\tau_n}\}[\mathbf{W}^{\tau_n}]\right)d\tau + \int_{\tau_n}^{\tau_{n+1}} [\mathbf{W}^{\tau_n}]\left(\{\mathbf{f}^{\tau_n}\}[\mathbf{W}^{\tau_n}] + \{\mathbf{f}^{\tau_{n+1}}\}[\mathbf{W}^{\tau_{n+1}}]\right)d\tau.\end{gathered} \tag{4.153}$$

Проводячи інтегрування, отримаємо трьохшарову схему:

$$\left(\frac{[\mathbf{C}]}{2\Delta\tau}+\frac{[\mathbf{K}]}{6}\right)\{\mathbf{a}^{\tau_{n+1}}\}+\left(\frac{2}{3}[\mathbf{K}]\right)\{\mathbf{a}^{\tau_n}\}+\left(-\frac{[\mathbf{C}]}{2\Delta\tau}+\frac{[\mathbf{K}]}{6}\right)\{\mathbf{a}^{\tau_{n-1}}\}= \\ =\frac{1}{6}\{\mathbf{f}^{\tau_{n+1}}\}+\frac{2}{3}\{\mathbf{f}^{\tau_n}\}+\frac{1}{6}\{\mathbf{f}^{\tau_{n-1}}\}. \tag{4.154}$$

Зауважимо, щоб знайти за даною схемою $\{\mathbf{a}^{\tau_{n+1}}\}$ необхідно знати крім $\{\mathbf{a}^{\tau_n}\}$, ще і $\{\mathbf{a}^{\tau_{n-1}}\}$.

За необхідності, аналогічним чином можна вивести формули і для більшої кількості часових шарів, наприклад з врахуванням $\{\mathbf{a}^{\tau_{n+2}}\}$ і $\{\mathbf{a}^{\tau_{n-2}}\}$.

Якщо характеристики об'єкту моделювання не змінюються в часі та не залежать від шуканого потенціалу, тобто задача є лінійною, то на кожній ітерації матриці $[\mathbf{C}]$ і $[\mathbf{K}]$ є незмінними, що колосально зменшує кількість обчислень. В іншому випадку, матриці необхідно обчислювати заново на кожній ітерації.

Підсумовуючи вищесказане, для розв'язку системи (4.129) використаємо схему Кранка-Ніколсона (4.147), при кількості пробних функцій $M=3$, та кроком дискретизації $\Delta\tau=0,25\,\text{с}$, що для заданого часового інтервалу еквівалентно 720 точкам на часовій осі. Обчисливши значення симетричних матриць демпфування, жорсткості та вектору навантаження (4.130) отримаємо:

$$[\mathbf{C}]=\begin{bmatrix} 829,9 & 62,2425 & 4,9794 \\ 62,2425 & 4,9794 & 0,41495 \\ 4,9794 & 0,41495 & 0,035567 \end{bmatrix}, \\ [\mathbf{K}]=\begin{bmatrix} 32 & 3,2 & 0,32 \\ 3,2 & 0,426667 & 0,048 \\ 0,32 & 0,048 & 0,00576 \end{bmatrix}, \quad \{\mathbf{f}\}=\begin{Bmatrix} 100 \\ 10 \\ 1 \end{Bmatrix}. \tag{4.155}$$

Задача описується лінійним рівнянням, тому обчислені матриці та вектор навантаження не змінюється з ітераціями в схемі Кранка-Ніколсона, звідки отримаємо:

$$[\mathbf{A}]=2\left(\frac{[\mathbf{C}]}{\Delta\tau}+\frac{1}{2}[\mathbf{K}]\right)=\begin{bmatrix} 6671,2 & 501,14 & 40,1552 \\ 501,14 & 40,261867 & 3,3676 \\ 40,1552 & 3,3676 & 0,290297 \end{bmatrix}, \\ [\mathbf{P}]=-2\left(-\frac{[\mathbf{C}]}{\Delta\tau}+\frac{1}{2}[\mathbf{K}]\right)=\begin{bmatrix} 6671,2 & 494,74 & 39,5152 \\ 494,74 & 39,408533 & 3,2716 \\ 39,5152 & 3,2716 & 0,278777 \end{bmatrix}. \tag{4.156}$$

На кожній ітерації рішення шукається як:

$$\{\mathbf{a}^{\tau_{n+1}}\}=[\mathbf{A}]^{-1}\left([\mathbf{P}]\{\mathbf{a}^{\tau_n}\}-2\{\mathbf{f}\}\right). \tag{4.157}$$

Знайдемо $[\mathbf{A}]^{-1}$:

$$[\mathbf{A}]^{-1} = \begin{bmatrix} 0{,}013565 & -0{,}400607 & 2{,}770873 \\ -0{,}400607 & 12{,}667001 & -91{,}530095 \\ 2{,}770873 & -91{,}530095 & 681{,}962575 \end{bmatrix}. \qquad (4.158)$$

Оскільки обрана система базисних функцій автоматично задовольняє початкові та головні крайові умови, для початкового наближення можна обрати $\{\mathbf{a}^{\tau_0}\} = \{\mathbf{0}\}$. Наведемо кілька векторів отриманих коефіцієнтів за (4.157):

$$\{\mathbf{a}^{\tau_1}\} = \begin{Bmatrix} 0{,}242617 \\ -9{,}841637 \\ 87.497855 \end{Bmatrix}, \quad \{\mathbf{a}^{\tau_{100}}\} = \begin{Bmatrix} 1{,}290766 \\ 1{,}849637 \\ 50{,}240202 \end{Bmatrix},$$
$$\{\mathbf{a}^{\tau_{300}}\} = \begin{Bmatrix} 2{,}749174 \\ 0{,}384169 \\ 10{,}260217 \end{Bmatrix}, \quad \{\mathbf{a}^{\tau_{720}}\} = \begin{Bmatrix} 3{,}111553 \\ 0{,}013745 \\ 0{,}367106 \end{Bmatrix}. \qquad (4.159)$$

Обчисливши значення $\{\mathbf{a}^{\tau_n}\}$, за допомогою (4.125) можна знайти апроксимоване значення температури у відповідному часовому інтервалі. Отримане рішення задачі наведено на *Рис. 4.9*.

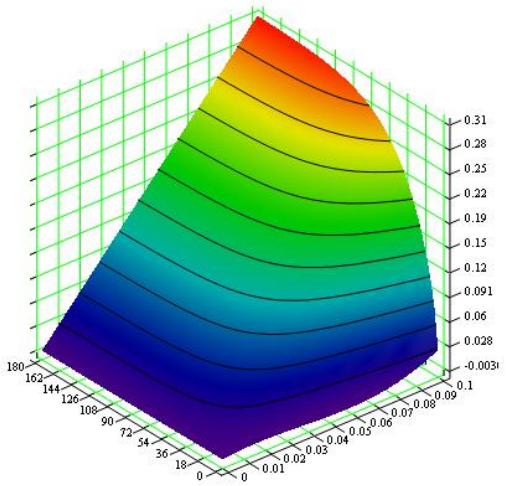


Рис. 4.9 Апроксимоване рішення нестаціонарної задачі теплопровідності

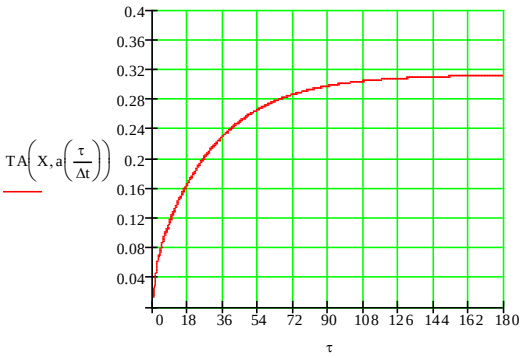


Рис. 4.10 Апроксимоване значення температури на стороні з заданою природною крайовою умовою

При постійному джерелі тепла $q = 1000$ Вт/м^2°С на стороні $x = 0{,}1$ м, через певний період процес стане стаціонарним (цікавий читач може переконатися, що це станеться через шість хвилин від початку нагрівання). Оскільки нам відомо коефіцієнт теплопровідності $\lambda = 320$ Вт/м°С та розміри стержня, можна знайти очікувану різницю температур на кінцях, тобто значення до якого повинен з часом прямувати розв'язок на стороні, де вказана природна крайова

умова. Різниця температур шукається як $\Delta T = d \cdot q/\lambda$, де $d = 0.1$ – довжина стержня, тобто $\Delta T = 0{,}3125$. На *Рис. 4.10* показано апроксимовану зміну температури з часом, як і очікувалося, значення температури прямує до ΔT.

Застосуємо метод часткової дискретизації для розв'язку загального рівняння другого порядку (4.122). Знову ж таки, можна використати вже описані процедури скінченно-різницевої апроксимації по часовій координаті, але щоб продемонструвати більш загальний підхід знаходження значень $\{\mathbf{a}\}$ знову застосуємо для цього метод зважених нев'язок, як це робилося на прикладі багатошарових схем.

Наближене рішення для $\{\mathbf{a}\}$ шукається за розкладом (4.149), однак тепер базисні функції $\varphi_{\tau_n}(\tau)$ повинні мати степінь не нижчу від другої, так як в рівняння входять другі похідні по часу. Для цього, візьмемо подвійний часовий крок з точками $\tau_{2n}, \tau_{2n+1}, \tau_{2n+2}$, звідки побудуємо систему базисних функцій для $\{\mathbf{a}\}$ (квадратичний поліном):

$$\{\tilde{\mathbf{a}}\} \approx [\mathbf{\Phi}^{\tau_{2n}}]\{\mathbf{a}^{\tau_{2n}}\} + [\mathbf{\Phi}^{\tau_{2n+1}}]\{\mathbf{a}^{\tau_{2n+1}}\} + [\mathbf{\Phi}^{\tau_{2n+2}}]\{\mathbf{a}^{\tau_{2n+2}}\}, \qquad (4.160)$$

де:

$$\begin{aligned}
&[\mathbf{\Phi}^{\tau_{2n}}] = -\frac{\xi(1-\xi)}{2}, \quad \frac{d[\mathbf{\Phi}^{\tau_{2n}}]}{d\tau} = \frac{-1/2+\xi}{\Delta\tau}, \quad \frac{d^2[\mathbf{\Phi}^{\tau_{2n}}]}{d\tau^2} = \frac{1}{\Delta\tau^2},\\
&[\mathbf{\Phi}^{\tau_{2n+1}}] = 1-\xi^2, \quad \frac{d[\mathbf{\Phi}^{\tau_{2n+1}}]}{d\tau} = -\frac{2\xi}{\Delta\tau}, \quad \frac{d[\mathbf{\Phi}^{\tau_{2n+1}}]}{d\tau} = -\frac{2}{\Delta\tau^2},\\
&[\mathbf{\Phi}^{\tau_{2n+2}}] = \frac{\xi(1+\xi)}{2}, \quad \frac{d[\mathbf{\Phi}^{\tau_{2n+2}}]}{d\tau} = \frac{1/2+\xi}{\Delta\tau}, \quad \frac{d^2[\mathbf{\Phi}^{\tau_{2n+2}}]}{d\tau^2} = \frac{1}{\Delta\tau^2},\\
&\xi = \frac{\tau-\tau_{2n+1}}{\Delta\tau}, \qquad \Delta\tau = \tau_{2n+2}-\tau_{2n+1}, \qquad \Delta\tau = \tau_{2n+1}-\tau_{2n}.
\end{aligned} \qquad (4.161)$$

Застосування методу зважених нев'язок дає систему рівнянь:

$$\int_{\tau_{2n}}^{\tau_{2n+2}} [\mathbf{W}^{\tau_n}]\left([\mathbf{M}]\frac{d^2}{d\tau^2}\{\tilde{\mathbf{a}}\} + [\mathbf{C}]\frac{d}{d\tau}\{\tilde{\mathbf{a}}\} + [\mathbf{K}]\{\tilde{\mathbf{a}}\} - \{\mathbf{f}\}\right)d\tau = 0, \quad n = 1,2,\ldots \qquad (4.162)$$

Знову ж таки виразимо $\{\mathbf{f}\}$ аналогічно до (4.160) і отримаємо систему:

$$\begin{aligned}
&\left([\mathbf{M}] + \gamma\Delta\tau[\mathbf{C}] + \beta\Delta\tau^2[\mathbf{K}]\right)\{\mathbf{a}^{\tau_{2n+2}}\} +\\
&+\left(-2[\mathbf{M}] + (1-2\gamma)\Delta\tau[\mathbf{C}] + (1/2-2\beta+\gamma)\Delta\tau^2[\mathbf{K}]\right)\{\mathbf{a}^{\tau_{2n+1}}\} +\\
&+\left([\mathbf{M}] - (1-\gamma)\Delta\tau[\mathbf{C}] + (1/2+\beta-\gamma)\Delta\tau^2[\mathbf{K}]\right)\{\mathbf{a}^{\tau_{2n}}\} = \Delta\tau^2\{\mathbf{f}^{\tau_n}\},
\end{aligned} \qquad (4.163)$$

де:

$$\gamma = \int_{-1}^{1} [\mathbf{W}^{\tau_n}]\left(\xi + 1/2\right)d\tau,$$

$$\beta = \int_{-1}^{1} [\mathbf{W}^{\tau_n}]1/2\,\xi\left(\xi+1\right)d\tau \Big/ \int_{-1}^{1} [\mathbf{W}^{\tau_n}]d\tau,$$

$$\{\mathbf{f}^{\tau_n}\} = \int_{-1}^{1} [\mathbf{W}^{\tau_n}]\mathbf{f}\left(\tau_{2n+1} + \xi\Delta\tau\right)d\tau \Big/ \int_{-1}^{1} [\mathbf{W}^{\tau_n}]d\tau = \\ = \beta\{\mathbf{f}^{\tau_{2n+2}}\} + (1/2 + 2\beta - \gamma)\{\mathbf{f}^{\tau_{2n+1}}\} + (1/2 + \beta - \gamma)\{\mathbf{f}^{\tau_{2n}}\}. \quad (4.164)$$

Рівняння (4.163) відповідає загальному алгоритму, вперше отриманому Ньюмарком в 1959 році[1], і є одним з найкращих відомих рекурентних співвідношень для рівнянь другого порядку [1]. Рекомендується в загальному випадку брати значення $\gamma = 1/2$ (як і в схемі Кранка-Ніколсона), що відповідає симетричним ваговим функціям для кожної з точок $\tau_{2n}, \tau_{2n+1}, \tau_{2n+2}$. Якщо $\beta = 0$, а матриці [**C**] і [**K**] є діагональними, то для знаходження $\{\mathbf{a}^{\tau_{2n+2}}\}$ не потрібне ніяке обернення матриць і схема стає явною. Але, в такому випадку (так само, як і для рівнянь першого порядку), можна показати що стійкість буде умовною і крок по часу $\Delta\tau$ має бути належним чином обмежений.

Для початку обчислень за схемою (4.163) необхідно знати крім значення $\{\mathbf{a}^{\tau_0}\}$, ще і $\{\mathbf{a}^{\tau_1}\}$. Щоб отримати останнє, можна застосувати деякі інші стартові схеми, що дозволяють знайти $\{\mathbf{a}^{\tau_1}\}$, за відомим значенням $\{\mathbf{a}^{\tau_0}\}$.

4.4. Рішення нелінійних задач

Природні явища, що описуються лінійними диференціальними рівняннями і відповідними крайовими умовами є лише частковим випадком, або грубим наближенням на фізичному рівні, широкого кола явищ, які зустрічаються у всесвіті. Для математичного опису задач, що неможливо описати в рамках лінійних теорій, використовуються або їх нелінійні розширення, або відносно нові теорії, на кшталт нелінійної динаміки чи теорії хаосу[2]. Прикладом таких задач є моделювання атмосферних процесів, процесів життєдіяльності людини, зокрема мозкової активності та серцебиття, моделювання турбулентності та багато інших.

До цього моменту нами обговорювалося застосування методу скінченних елементів та методів зважених нев'язок до лінійних задач, проте з однаковим успіхом ці методи можуть бути застосовані і для рішення нелінійних задач [1]. В таких випадках, застосування процедури методу зважених нев'язок приводить не до стандартної системи рівнянь, а до системи нелінійних рівнянь, що можна записати як:

$$[\mathbf{K}(\mathbf{a})]\{\mathbf{a}\} = \{\mathbf{f}\}. \quad (4.165)$$

Для рішення даної системи рівнянь зазвичай використовується відповідна ітераційна процедура [23], [24].

Для прикладу розглянемо задачу нелінійної стаціонарної теплопровідності,

---

[1] Newmark N. – A method for computation of structural dynamics // Proc. Amer. Soc. Civ. Eng., 85(EM3):67-94, 1959.

[2] Див. наприклад:
Mandelbrot B. – The Fractal Geometry of Nature // New-York W. H. Freeman and Co., 1982;
Crownover R. – Introduction to Fractals and Chaos // Boston: Jones and Bartlett Publishers Inc., 1995;

тобто задачу поширення тепла в тілі, теплопровідність якого залежить від температури. Диференціальне рівняння, що описує такий процес, має вигляд:

$$\lambda(T(\mathbf{r}))\nabla^2 T(\mathbf{r}) + Q(\mathbf{r}) = 0, \quad \mathbf{r} \in \Omega. \tag{4.166}$$

З крайовими умовами Діріхле та Неймана:

$$T(\mathbf{r})\big|_{\Gamma_T} = T_\infty, \quad \lambda(T(\mathbf{r}))\frac{\partial T(\mathbf{r})}{\partial \mathbf{n}}\bigg|_{\Gamma_q} = q. \tag{4.167}$$

Наближене рішення шукається по розкладу:

$$T(\mathbf{r}) \approx \tilde{T}(\mathbf{r}) = T_0(\mathbf{r}) + \sum_{j=1}^{M} a_j \varphi_j(\mathbf{r}), \tag{4.168}$$

де $u_0(\mathbf{r})$ та $\varphi_j(\mathbf{r})$, як і звичайно вибираються так, щоб автоматично забезпечити виконання крайової умови Діріхле на Γ_T. Використовуючи процедуру пониження порядку, на основі теореми Стокса, а також вибираючи вагові функції рівні базисним, всередині області Ω, і протилежні по знаку на границі Γ_q, приходимо до системи рівнянь методів зважених нев'язок:

$$\begin{aligned} &\int_\Omega \left[\frac{\partial \omega_i}{\partial x}\lambda(\tilde{T})\frac{\partial \tilde{T}}{\partial x} + \frac{\partial \omega_i}{\partial y}\lambda(\tilde{T})\frac{\partial \tilde{T}}{\partial y} + \frac{\partial \omega_i}{\partial z}\lambda(\tilde{T})\frac{\partial \tilde{T}}{\partial z} \right] dxdydz = \\ &= \int_\Omega \omega_i Q dxdydz - \int_{\Gamma_q} \omega_i q d\Gamma \quad i = 1,2,\ldots M. \end{aligned} \tag{4.169}$$

Для рішення цієї системи використаємо метод простої ітерації [1], [22]. Виберемо деяке початкове наближення:

$$\{\mathbf{a}\} = \{\mathbf{a}^0\} = \left\{a_1^0, a_2^0, \ldots, a_M^0\right\}^{\mathbf{T}} \tag{4.170}$$

і відповідне йому початкове рішення $\tilde{T}^0$, а потім отримаємо покращене рішення $\{\mathbf{a}^1\}$ з лінійного рівняння:

$$[\mathbf{K}(\mathbf{a}^0)]\{\mathbf{a}^1\} = \{\mathbf{f}^0\}, \tag{4.171}$$

де:

$$\begin{aligned} [\mathbf{K}(\mathbf{a}^0)]_{i,j} &= \int_\Omega \left[\frac{\partial \omega_i}{\partial x}\lambda(\tilde{T}^0)\frac{\partial \varphi_j}{\partial x} + \frac{\partial \omega_i}{\partial y}\lambda(\tilde{T}^0)\frac{\partial \varphi_j}{\partial y} + \frac{\partial \omega_i}{\partial z}\lambda(\tilde{T}^0)\frac{\partial \varphi_j}{\partial z} \right] dxdydz, \\ \{\mathbf{f}^0\}_i &= \int_\Omega \omega_i Q dxdydz - \int_{\Gamma_q} \omega_i q d\Gamma - \\ &- \int_\Omega \left[\frac{\partial \omega_i}{\partial x}\lambda(\tilde{T}^0)\frac{\partial T_0}{\partial x} + \frac{\partial \omega_i}{\partial y}\lambda(\tilde{T}^0)\frac{\partial T_0}{\partial y} + \frac{\partial \omega_i}{\partial z}\lambda(\tilde{T}^0)\frac{\partial T_0}{\partial z} \right] dxdydz. \end{aligned} \tag{4.172}$$

Після чого, продовжимо обчислення по загальній ітераційній формулі:

$$[\mathbf{K}(\mathbf{a}^{n-1})]\{\mathbf{a}^n\} = \{\mathbf{f}^{n-1}\}, \tag{4.173}$$

застосовуючи її доти, доки процес обчислень не зійдеться в межах обраної точності.

Щоб показати процес обчислень, розглянемо одновимірну задачу

теплопровідності, з джерелом тепла в області, де теплопровідність залежить від температури. Нехай матеріал займає область $0 \le x \le 1$. На краях підтримується постійна температура $T(0) = T(1) = 0$ °C. Коефіцієнт теплопровідності матеріалу залежить від температури $\lambda(T) = 1 + 0{,}1T$ Вт/м°C. В кожній точці тіла міститься внутрішнє джерело тепла $Q(x) = 10x$. Необхідно знайти розподіл температурного поля.

Спочатку запишемо диференціальне рівняння, що описує даний процес:

$$d\left(\lambda(T(x))\,dT(x)/dx\right)/dx = -10x, \quad T(0) = T(1) = 0, \quad 0 \le x \le 1. \quad (4.174)$$

Після цього, виберемо систему базисних функцій, що автоматично задовольняє крайові умови:

$$T_0(x) = 0, \quad \varphi_j = x^j(1-x), \quad \tilde{T}(x) = \sum_{j=1}^{M} a_j x^j (1-x). \quad (4.175)$$

Застосовуючи метод зважених нев'язок, отримаємо систему рівнянь:

$$\int_0^1 \omega_i \left[\frac{d}{dx}\left(\lambda(\tilde{T}(x)) \frac{d}{dx}\tilde{T}(x) \right) + 10x \right] dx = 0, \quad i = 1, 2, \ldots M. \quad (4.176)$$

Для апроксимації виберемо метод поточкових коллокацій при $M = 2$. Нехай $x_1 = 1/3$ та $x_2 = 2/3$. Отримаємо систему рівнянь виду (4.165):

$$[\mathbf{K}(\mathbf{a})] = \begin{bmatrix} -\dfrac{d}{dx}\left(\lambda(\tilde{T}(x)) \dfrac{d}{dx}\varphi_1(x) \right)\Big|_{x=x_1} & -\dfrac{d}{dx}\left(\lambda(\tilde{T}(x)) \dfrac{d}{dx}\varphi_2(x) \right)\Big|_{x=x_1} \\ -\dfrac{d}{dx}\left(\lambda(\tilde{T}(x)) \dfrac{d}{dx}\varphi_1(x) \right)\Big|_{x=x_2} & -\dfrac{d}{dx}\left(\lambda(\tilde{T}(x)) \dfrac{d}{dx}\varphi_2(x) \right)\Big|_{x=x_2} \end{bmatrix}, \quad (4.177)$$

$$\{\mathbf{f}\} = \begin{Bmatrix} Q(x)\big|_{x=x_1} \\ Q(x)\big|_{x=x_2} \end{Bmatrix}.$$

Виберемо початкове наближення $a_1 = a_2 = 0$, тобто $\{\mathbf{a}^0\} = \{\mathbf{0}\}$, в наслідок чого отримаємо:

$$[\mathbf{K}(\mathbf{a}^0)] = \begin{bmatrix} 2{,}000000 & 0{,}000000 \\ 2{,}000000 & 2{,}000000 \end{bmatrix}, \quad \{\mathbf{f}^0\} = \begin{Bmatrix} 3{,}333333 \\ 6{,}666667 \end{Bmatrix}. \quad (4.178)$$

Рішенням є:

$$\{\mathbf{a}^1\} = \begin{Bmatrix} 1{,}666667 \\ 1{,}666667 \end{Bmatrix}. \quad (4.179)$$

Ці значення використаємо для побудови наступної матриці:

$$[\mathbf{K}(\mathbf{a}^1)] = \begin{bmatrix} 2{,}061728 & -0{,}037037 \\ 2{,}104938 & 2{,}123457 \end{bmatrix}, \quad \{\mathbf{f}^1\} = \begin{Bmatrix} 3{,}333333 \\ 6{,}666667 \end{Bmatrix}, \quad \{\mathbf{a}^2\} = \begin{Bmatrix} 1{,}643892 \\ 1{,}509979 \end{Bmatrix}. \quad (4.180)$$

Продовжимо цей процес, поки не отримаємо результат, який не змінюється до четвертого знаку після коми:

$$[\mathbf{K}(\mathbf{a}^2)]=\begin{bmatrix}2{,}060389 & -0{,}035043\\ 2{,}099537 & 2{,}117802\end{bmatrix},\quad \{\mathbf{a}^3\}=\begin{Bmatrix}1{,}643643\\ 1{,}518451\end{Bmatrix},\quad \left|\{\mathbf{a}^3\}-\{\mathbf{a}^2\}\right|=\begin{Bmatrix}0{,}000249\\ 0{,}008471\end{Bmatrix},$$
$$[\mathbf{K}(\mathbf{a}^3)]=\begin{bmatrix}2{,}060412 & -0{,}035134\\ 2{,}099779 & 2{,}118042\end{bmatrix},\quad \{\mathbf{a}^4\}=\begin{Bmatrix}1{,}643685\\ 1{,}518048\end{Bmatrix},\quad \left|\{\mathbf{a}^4\}-\{\mathbf{a}^3\}\right|=\begin{Bmatrix}0{,}000042\\ 0{,}000402\end{Bmatrix}, \qquad (4.181)$$
$$[\mathbf{K}(\mathbf{a}^4)]=\begin{bmatrix}2{,}060412 & -0{,}035130\\ 2{,}099769 & 2{,}118032\end{bmatrix},\quad \{\mathbf{a}^5\}=\begin{Bmatrix}1{,}643683\\ 1{,}518066\end{Bmatrix},\quad \left|\{\mathbf{a}^5\}-\{\mathbf{a}^4\}\right|=\begin{Bmatrix}0{,}000003\\ 0{,}000018\end{Bmatrix}.$$

На *Рис. 4.11* показано отримане апроксимоване рішення.

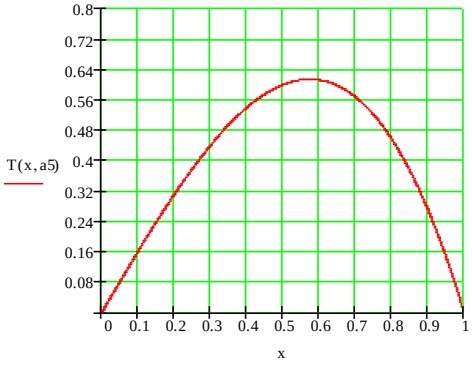


Рис. 4.11 Апроксимоване рішення задачі нелінійної стаціонарної теплопровідності з джерелом тепла в області

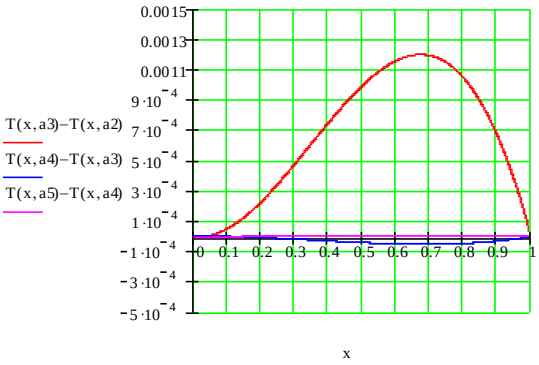


Рис. 4.12 Нев'язка між ітераціями отриманих апроксимованих значень температури

Основною областю застосування чисельних методів є рішення нелінійних нестаціонарних задач, коли компоненти матриць з системи рівнянь (4.118) залежать від шуканих коефіцієнтів $\{\mathbf{a}\}$. У випадку рівнянь першого і другого порядку можна використати попередньо описані процедури скінченно-різницевої апроксимації, чи апроксимації методами зважених нев'язок. Для сумісності отриманих рівнянь необхідне повне чисельне інтегрування рівняння методу зважених нев'язок і в загальному випадку необхідні ітерації на кожному кроці по часу.

Для нелінійних рівнянь не буде справедливий і аналіз стійкості обчислень, однак деяку інформацію про імовірні характеристики чисельних схем можна отримати з допомогою так званої локальної *лінеаризації* рівнянь на одному кроці по часу. Тоді, наприклад для рівнянь першого порядку можна було б розглянути стійкість на інтервалі $[\tau_n,\tau_{n+1}]$ двохшарової схеми (4.144) стосовно вже лінійного рівняння типу:

$$[\mathbf{C}(\mathbf{a}^{\tau_n})]\frac{d}{d\tau}\{\mathbf{a}^{\tau_n}\}+[\mathbf{K}(\mathbf{a}^{\tau_n})]\{\mathbf{a}^{\tau_n}\}=\{\mathbf{f}(\mathbf{a}^{\tau_n})\}. \qquad (4.182)$$

Розгляд задач нелінійної динаміки в загальному випадку є надто складним, щоб розглядати його у даній роботі. Для додаткової інформації слід звернутися до відповідної літератури. Більш детально про застосування методів зважених нев'язок до нелінійних задач можна дізнатися наприклад в [23], [25].

4.5. Список використаної літератури до розділу 4

[1] Zienkiewicz O., Morgan K. – Finite elements and approx. // New-York: Wiley, 1983.

[2] Norrie D., Vries G. – An Introduction to Finite Element Analysis // New-York: Academic press, 1978.

[3] Zienkiewicz O. – The Finite Element Method in Engineering Science / Метод конечных элементов в технике / пер. с англ. под ред Пбедри Б. // Москва: Мир, 1975.

[4] Szabó B., Babuška I. – Introduction to Finite Element Analysis. Formulation, Verification and Validation // New-York: Wiley, 2011.

[5] Bathe, K. – Finite Element Procedures // NJ Englewood Cliffs: Prentice-Hall, 1996.

[6] Olson G. – Dynamical Analogies. 2nd ed. // New-York: Van Nostr. Comp., Inc., 1958.

[7] Jaworski N., Farmaga I., Matviykiv O., Lobur M., Spiewak P., Ciupinski L., Kurzydlowski K. – Thermal Analysis Methods for Design of Composite Materials with Complex Structure // ECS Transactions, 59(1):513-523, 2014.

[8] Веников В. – Теория подобия и моделирования // Москва: Высшая школа, 1976.

[9] Гухман А. – Введение в теорию подобия. 2-е изд. // Москва: Выс. шк., 296 с., 1973.

[10] Нейман Л., Демирчян К. – Теоретические основы электротехники. В 2-х т. Учебник для вузов. Том 1. 3-е изд., перераб. и доп. // Ленинград: Энергоиздат, 1981.

[11] Нейман Л., Демирчян К. – Теоретические основы электротехники. В 2-х т. Учебник для вузов. Том 2. // Ленинград: Энергоиздат. Ленингр. отд-ние, 1967.

[12] Silvester P., Ferrari R. – Метод конечных элементов для радиоинженеров и инженеров-электриков / пер. с англ. // Москва: Мир, 1986.

[13] Дульнев Г. – Тепло- и массообмен в радиоэлектронной аппаратуре: Учебник для вузов по спец. "Конструир. и произв. радиоаппаратуры" // Москва: Выс. Шк., 1984.

[14] Яворський Н., Фармага І., Марікуца У. – Розроблення дискретної моделі знаходження ефективних теплофізичних характеристик композитних матеріалів зі складною структурою. // Вісник НУ "ЛП" "Комп'ютерні науки та інформаційні технології", 744:152-158., 2012.

[15] [Electronic resource] Westendorp G. – Electric circuit diagram equivalents of fields // http://westy31.home.xs4all.nl/Electric.html.

[16] Лурье А. – Теория упругости // Москва: Наука, 1970.

[17] Timoshenko S., Goodier J. – Теория упругости / пер. с англ. // Москва: Наука, 1979.

[18] Клованич С. – Метод конечных элементов в нелинейных задачах инженерной механики // Запорожье: Світ геотехніки, 2009.

[19] Jaworski N., Farmaga I., Karvatskiy R. – Finding the Composite Materials Linear Temperature Expansion Coefficient Based on Thermoelasticity Problem Numerical Simulation // Proc. of CADMD'2014, pp. 77-83 – October 10-11, Lviv, Ukraine, 2014.

[20] Коваленко А. – Основы термоупругости // Київ: Наукова думка, 1970.

[21] Segerlind L. – Applied Finite Element Analysis / Применение метода конечных элементов / пер. с англ. Шестаков А., под. ред. Победри Б. // Москва: Мир, 1979.

[22] Демидович Б., Марон И. – Основы вычислительной математики. 3-е изд., испр. // Москва: Наука, 1966.

[23] Thomee V. – Galerkin Finite Element Methods for Parabolic Problems. 2-nd ed. // New-York: Springer, 2006.

[24] Saad Y. – Iterative Methods for Sparse Linear Systems. 2-nd ed. // Philadelphia: Society for Industrial and Applied Mathematics, 2003.

[25] Knabner P., Angerman L. – Numerical Methods for Elliptic and Parabolic Partial Differential Equations // New-York: Springer, 2003.

5. Особливості апроксимації методом скінченних елементів

5.1. Одновимірні комплекс елементи та інтерполяція вищих порядків

До цього моменту ми розглядали метод скінченних елементів з використанням лінійних кусково-визначених функцій, для апроксимації диференціальних рівнянь другого порядку. У багатовимірних просторах такі функції будувалися на основі простих топологічних структур – симплексів, а відповідні скінченні елементи називалися симплекс елементами. По аналогії з методами зважених нев'язок, які можна розглядати як апроксимацію єдиним суперелементом по всій області, у якості базисних функцій, тобто функцій форми скінченних елементів, можуть виступати і поліноми вищих порядків. Більше того, для гарантування умови повноти, для апроксимації певного класу задач, використання елементів вищого порядку є єдиним можливим варіантом.

Нагадаємо, що для апроксимації задач, які визначаються диференціальними рівняннями в слабкій формі є допустимим використання повних поліномів порядку не нижчого від p, де $2p$ – порядок диференціального рівняння, яке можна записати у слабкій формі. Тому в усіх попередніх прикладах, де задачі описувалися рівняннями другого порядку, використання симплекс елементів було допустимим.

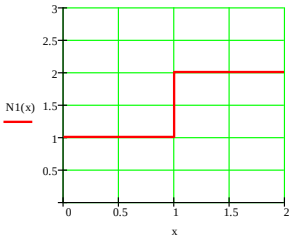
Проводячи аналогію з методами скінченних різниць, використання функцій форми високих порядків, тобто порядків вищих за допустимі, еквівалентно збільшенню порядку точності скінченно-елементної моделі, і як наслідок – швидшій збіжності отриманого апроксимованого рішення до точного рішення.

Іншою, не менш важливою умовою збіжності скінченно-елементної моделі є, так звана, *узгодженість* функцій форми сусідніх елементів. Щоб зрозуміти цю умову розглянемо поведінку трьох пар інтерполяційних функцій та їх похідних поблизу міжелементної границі *Рис. 5.1*. Перша пара функцій має розрив першого роду на границі двох елементів, друга пара має розрив першого роду в похідній, третя пара має розрив першого роду в другій похідній. Очевидно, що в точках розриву відповідно перша друга і третя похідна цих пар функцій не будуть обмеженими, тобто міститимуть розриви другого роду.

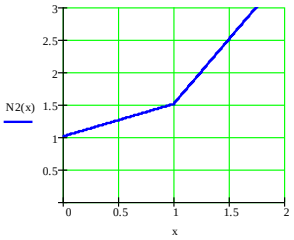
Якщо тепер обчислювати інтеграли з рівняння методів зважених нев'язок, то бажано виключити такі безмежні значення похідних, оскільки вони приводять до невизначеності в інтегралах. Якщо диференціальний оператор $\mathcal{L}$ деякої крайової задачі містить похідні порядку $2p$, і в слабкій формі відповідні диференціальні оператори більш низького порядку містять похідні порядку не вищого p, то для усунення подібних невизначеностей необхідно гарантувати кускову диференційованість похідних порядку $p-1$. Математично це означає, що ми вимагаємо, щоб функції форми належали до класу *гладкості* $C^{p-1}(\Omega)$ [1].

[1] Не вдаючись в деталі, функція належить класу гладкості $C^r(\Omega)$, якщо вона і її похідні до порядку r включно є неперервними і похідні порядку r (а при $r=0$ сама функція) кусково-неперервно диференційовані.

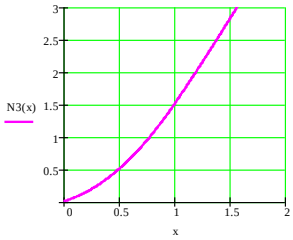
$\varphi_1(x)=1,$
$\varphi_2(x)=2,$



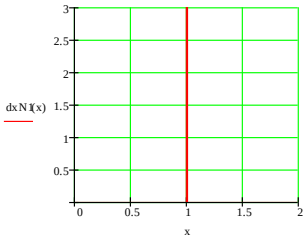
$\varphi_1(x)=0,5x+1,$
$\varphi_2(x)=2x+0,5,$



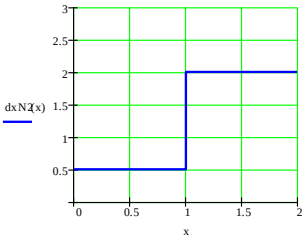
$\varphi_1(x)=x^2+0,5x,$
$\varphi_2(x)=0,25x^2+2x-0,75,$



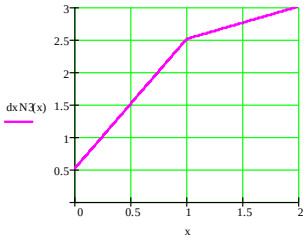
$\left.\frac{d\varphi}{dx}\right|_{x=1}\to\infty,$



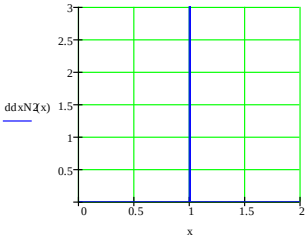
$\frac{d\varphi_1(x)}{dx}=0,5,$
$\frac{d\varphi_2(x)}{dx}=2,$



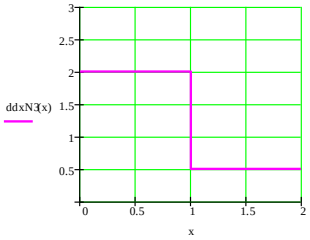
$\frac{d\varphi_1(x)}{dx}=2x+0,5,$
$\frac{d\varphi_2(x)}{dx}=0,5x+2,$



$\left.\frac{d^2\varphi}{dx^2}\right|_{x=1}\to\infty,$



$\frac{d^2\varphi_1(x)}{dx^2}=2,$
$\frac{d^2\varphi_2(x)}{dx^2}=0,5,$



$\left.\frac{d^3\varphi}{dx^3}\right|_{x=1}\to\infty,$

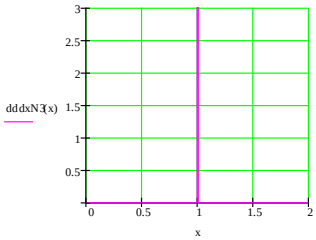


Рис. 5.1 Поведінка трьох типів базисних функцій та їх похідних поблизу границі між сусідніми скінченними елементами в точці $x=1$

Історично, ця умова прийшла з задач механіки де при формулюванні визначального рівняння у термінах деформацій, а шуканого рішення – в термінах переміщень, було звичним описувати поле переміщень як *спільне*, якщо переміщення змінювалися неперервно по області, і в такому випадку деформації були кусково-неперервними. Пізніше визначення було перенесене в область скінченних елементів для того, щоб описувати представлення пробних функцій в неперервній області. А більш загальний термін, під назвою *узгодженість*, вперше був запропонований в 1965 році[1]. Таким чином, функції форми скінченних елементів є узгодженими, якщо самі функції і їх похідні до порядку $p-1$ включно є неперервними при переході через границю між елементами [1], [16].

Оскільки для рівнянь другого порядку, що розглядалися в усіх попередніх прикладах, можливо побудувати слабку форму де містяться частинні похідні максимум першого порядку і допустимим є використання лінійних функцій форм, стає очевидним, що для побудови стійких та збіжних скінченно-елементних схем вимагається $C^0(\Omega)$ гладкість апроксимованого рішення, тобто допускаються розриви першого роду в похідних при переході між елементами.

Для того, щоб зрозуміти, як будувати функції форми високих порядків, що відповідають критеріям повноти і узгодженості для рішення класу гладкості $C^0(\Omega)$, розглянемо спочатку одновимірний елемент (*Рис. 5.2*).

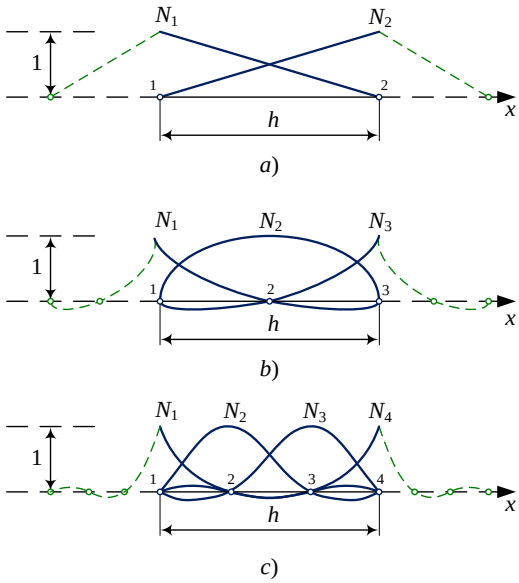


Рис. 5.2 Одновимірні елементи та відповідні стандартні базисні функції: a) лінійна; b) квадратична; c) кубічна

[1] Bizely G., Cheung Y., Irons B., Zienkiewicz O. – Triangular elements in plate bending – conforming and non-conforming solutions // Proc. Conf. Matrix Methods Struct. Mech., Wright-Patterson AFB, Ohio, October 26-28, AFFDL-TR-66-80, pp. 547-576, November 1965.

Симплекс елементом буде звичайний відрізок (*Рис. 5.2.a*), обмежений двома вузлами, на якому визначено лінійні інтерполяційні функції. Значення коефіцієнту u_j для кожної з базисних функцій є значення шуканого потенціалу у відповідному вузлі, тобто кожна N_j рівна одиниці в j-му вузлі та рівна нулю в усіх інших вузлах.

Наступним стандартним одновимірним елементом є квадратичний елемент, що складається вже з трьох вузлів (*Рис. 5.2.b*). Цей елемент є *комплекс* елементом, оскільки кількість його вузлів більша за одиницю від розмірності задачі. Для квадратичного елементу положення внутрішнього вузла не є принциповим і одразу помітно, що відповідна для цього вузла базисна функція є внутрішньою по відношенню до елементу та не поширюється на сусідні елементи.

Збільшивши степінь інтерполяційних функцій ще на один порядок, та відповідно добавивши ще один внутрішній вузол, отримаємо стандартний одновимірний кубічний елемент (*Рис. 5.2.c*) .

У загальному випадку, апроксимація $\tilde{u}$ на елементі з $p+1$ вузлами, що не обов'язково розміщені рівномірно, буде зводитися до поліному степені p. Вираз для кожної з функцій форми такого комплекс елементу можна записати як:

$$N_{i,j}(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \ldots + \alpha_p x^p. \tag{5.1}$$

Оскільки кожна з функцій форми повинна відповідати інтерполяційним умовам:

$$\varphi_j(\mathbf{r}) = \begin{cases} 1, & \mathbf{r} = \mathbf{r}_j, \\ 0, & \mathbf{r} = \mathbf{r}_i, \end{cases} \quad i \neq j, \tag{5.2}$$

та:

$$\sum_{j=1}^{M} \varphi_j(\mathbf{r}) = 1, \quad \mathbf{r} \in \Omega_i, \tag{5.3}$$

то невідомі коефіцієнти α є рішеннями системи рівнянь:

$$\begin{aligned}
x = X_1, \quad & N_{i,j}(x) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_1^2 + \alpha_3 X_1^3 + \ldots + \alpha_p X_1^p = 0, \\
x = X_2, \quad & N_{i,j}(x) = \alpha_0 + \alpha_1 X_2 + \alpha_2 X_2^2 + \alpha_3 X_2^3 + \ldots + \alpha_p X_2^p = 0, \\
& \vdots \\
x = X_j, \quad & N_{i,j}(x) = \alpha_0 + \alpha_1 X_j + \alpha_2 X_j^2 + \alpha_3 X_j^3 + \ldots + \alpha_p X_j^p = 1, \\
& \vdots \\
x = X_{p+1}, \quad & N_{i,j}(x) = \alpha_0 + \alpha_1 X_{p+1} + \alpha_2 X_{p+1}^2 + \alpha_3 X_{p+1}^3 + \ldots + \alpha_p X_{p+1}^p = 0.
\end{aligned} \tag{5.4}$$

Або у вже звичній матричній формі:

$$[\mathbf{N}] = [\mathbf{P}][\mathbf{C}]^{-1}, \tag{5.5}$$

де:

$$[\mathbf{P}] = \begin{bmatrix} 1 & x & x^2 & \ldots & x^p \end{bmatrix}, \tag{5.6}$$

та:

$$[\mathbf{C}]=\begin{bmatrix}1 & X_1 & X_1^2 & \cdots & X_1^p\\ 1 & X_2 & X_2^2 & \cdots & X_2^p\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & X_{p+1} & X_{p+1}^2 & \cdots & X_{p+1}^p\end{bmatrix}. \tag{5.7}$$

Розв'язок цієї системи приведе до виразу:

$$N_{i,j}(x)=\Lambda_j^p(x)=\prod_{k=1,\ j\neq k}^{p+1}\frac{x-X_k}{X_j-X_k}= \\ =\frac{(x-X_1)(x-X_2)\ldots(x-X_{j-1})(x-X_{j+1})\ldots(x-X_{p+1})}{(X_j-X_1)(X_j-X_2)\ldots(X_j-X_{j-1})(X_j-X_{j+1})\ldots(X_j-X_{p+1})}, \tag{5.8}$$

що також відомий як фундаментальний поліном Лагранжа степені p. Тому такі комплекс елементи в літературі часто називають *Лагранжевими елементами* [1], [16].

Для прикладу знову розв'яжемо рівняння:

$$d^2y(x)/dx^2-y(x)=0,\quad y(0)=0,\quad y(1)=1,\quad 0\le x\le 1, \tag{5.9}$$

використовуючи два квадратичні елементи. Розмістимо на відрізку $0\le x\le 1$ п'ять вузлів, наприклад $X_1=0$, $X_2=¼$, $X_3=½$, $X_4=¾$ та $X_5=1$ (вузли спеціально розміщені рівномірно для спрощення обчислень, в загальному випадку це не обов'язково), об'єднаємо їх в два квадратичні елементи $\Omega_1=[X_1,X_2,X_3]$ та $\Omega_2=[X_3,X_4,X_5]$. Знайдемо вирази функцій форм цих елементів:

$$[\mathbf{N}]_1=[1\quad x\quad x^2]\begin{bmatrix}1 & X_{i,1} & X_{i,1}^2\\ 1 & X_{i,2} & X_{i,2}^2\\ 1 & X_{i,3} & X_{i,3}^2\end{bmatrix}^{-1}, \tag{5.10}$$

де:

$$\begin{bmatrix}1 & X_{i,1} & X_{i,1}^2\\ 1 & X_{i,2} & X_{i,2}^2\\ 1 & X_{i,3} & X_{i,3}^2\end{bmatrix}^{-1}=\begin{bmatrix}\dfrac{X_{i,2}X_{i,3}}{(X_{i,1}-X_{i,2})(X_{i,1}-X_{i,3})} & \dfrac{-X_{i,1}X_{i,3}}{(X_{i,1}-X_{i,2})(X_{i,2}-X_{i,3})} & \dfrac{X_{i,1}X_{i,2}}{(X_{i,2}-X_{i,3})(X_{i,1}-X_{i,3})}\\ \dfrac{-X_2-X_3}{(X_{i,1}-X_{i,2})(X_{i,1}-X_{i,3})} & \dfrac{X_1+X_3}{(X_{i,1}-X_{i,2})(X_{i,2}-X_{i,3})} & \dfrac{-X_1+X_2}{(X_{i,2}-X_{i,3})(X_{i,1}-X_{i,3})}\\ \dfrac{1}{(X_{i,1}-X_{i,2})(X_{i,1}-X_{i,3})} & \dfrac{-1}{(X_{i,1}-X_{i,2})(X_{i,2}-X_{i,3})} & \dfrac{1}{(X_{i,2}-X_{i,3})(X_{i,1}-X_{i,3})}\end{bmatrix}, \tag{5.11}$$

звідки:

$$[\mathbf{N}]_i=\left[\frac{(x-X_{i,2})(x-X_{i,3})}{(X_{i,1}-X_{i,2})(X_{i,1}-X_{i,3})}\quad\frac{(x-X_{i,1})(x-X_{i,3})}{(X_{i,2}-X_{i,1})(X_{i,2}-X_{i,3})}\quad\frac{(x-X_{i,1})(x-X_{i,2})}{(X_{i,3}-X_{i,1})(X_{i,3}-X_{i,2})}\right]. \tag{5.12}$$

Знайдемо матриці похідних для відповідних функцій форм:

$$\frac{d[\mathbf{N}]_i}{dx}=\left[\frac{2x-X_{i,2}-X_{i,3}}{(X_{i,1}-X_{i,2})(X_{i,1}-X_{i,3})}\quad\frac{2x-X_{i,1}-X_{i,3}}{(X_{i,2}-X_{i,1})(X_{i,2}-X_{i,3})}\quad\frac{2x-X_{i,1}-X_{i,2}}{(X_{i,3}-X_{i,1})(X_{i,3}-X_{i,2})}\right]. \tag{5.13}$$

Для спрощення, позначимо довжину елементу як h_i. Оскільки внутрішній вузол

розміщено посередині, то відстані між сусідніми вузлами рівні $h_i/2$. Підставивши знайдені вирази у рівняння методів зважених нев'язок знайдемо локальні матриці жорсткості елементів:

$$[\mathbf{K}]_i = \int_{X_{i,1}}^{X_{i,3}} \frac{d[\mathbf{N}]_i^{\mathbf{T}}}{dx}\frac{d[\mathbf{N}]_i}{dx}dx + \int_{X_{i,1}}^{X_{i,3}} [\mathbf{N}]_i^{\mathbf{T}}[\mathbf{N}]_i dx =$$

$$= \frac{1}{-3h_i}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix} + \frac{h_i}{15}\begin{bmatrix} 2 & 1 & -1/2 \\ 1 & 8 & 1 \\ -1/2 & 1 & 2 \end{bmatrix} = \tag{5.14}$$

$$= \frac{1}{15h_i}\begin{bmatrix} 35+2h_i^2 & h_i^2-40 & (10-h_i^2)/2 \\ h_i^2-40 & 80+8h_i^2 & h_i^2-40 \\ (10-h_i^2)/2 & h_i^2-40 & 35+2h_i^2 \end{bmatrix}.$$

Зберемо глобальну систему рівнянь:

$$\begin{bmatrix} 4{,}733333 & -5{,}300000 & 0{,}650000 & 0 & 0 \\ -5{,}300000 & 10{,}933333 & -5{,}300000 & 0 & 0 \\ 0{,}650000 & -5{,}300000 & 9{,}466667 & -5{,}300000 & 0{,}650000 \\ 0 & 0 & -5{,}300000 & 10{,}933333 & -5{,}300000 \\ 0 & 0 & 0{,}650000 & -5{,}300000 & 4{,}733333 \end{bmatrix}\begin{Bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{Bmatrix}. \tag{5.15}$$

Враховуючи початкові умови, тобто відомі y_1 та y_5, систему слід модифікувати:

$$\begin{bmatrix} 4{,}733333 & 0 & 0 & 0 & 0 \\ 0 & 10{,}933333 & -5{,}300000 & 0 & 0 \\ 0 & -5{,}300000 & 9{,}466667 & -5{,}300000 & 0 \\ 0 & 0 & -5{,}300000 & 10{,}933333 & 0 \\ 0 & 0 & 0 & 0 & 4{,}733333 \end{bmatrix}\begin{Bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ -0{,}650000 \\ 5{,}300000 \\ 4{,}733333 \end{Bmatrix}. \tag{5.16}$$

Розв'язком цієї системи є вектор:

$$\{\mathbf{y}\} = \{0{,}000000 \quad 0{,}214948 \quad 0{,}443414 \quad 0{,}699704 \quad 1{,}000000\}^{\mathbf{T}}. \tag{5.17}$$

На *Рис. 5.3* показано графік точного рішення та його отриманої апроксимації. На *Рис. 5.4* показано похибку між точним та отриманим наближеним рішенням.

Як і слід було очікувати, похибка є мінімальною у вузлах дискретизації. Крім того, порівнюючи отриману похибку квадратичної апроксимації, з похибкою лінійної апроксимації, отриманої у попередніх розділах, можна переконатися у швидшій збіжності першої, навіть при використанні меншої кількості елементів.

Дуже часто [1], [3], [4], при роботі з елементами вищих порядків використовують їх барицентричні координати:

$$N_1 = L_1 = \frac{X_2 - x}{X_2 - X_1}, \quad N_2 = L_2 = \frac{x - X_1}{X_2 - X_1}, \tag{5.18}$$

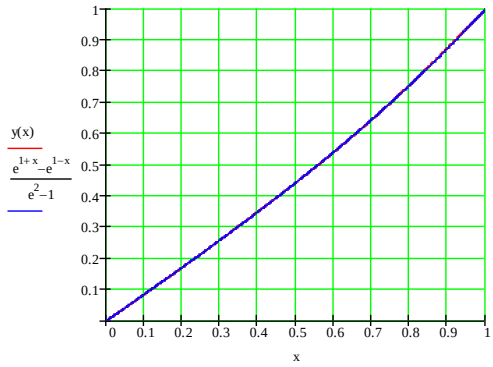


Рис. 5.3 Точне та наближене рішення рівняння $d^2y(x)/dx^2 - y(x) = 0$, отримане квадратичною скінченно-елементною апроксимацією (майже співпадають)

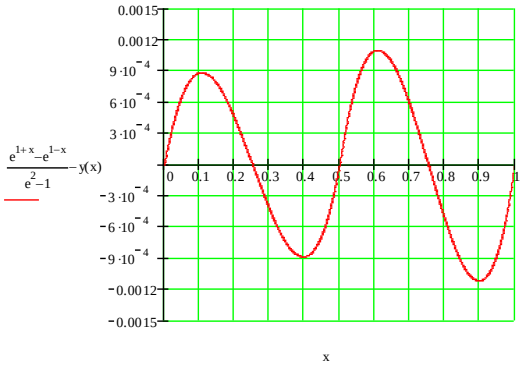


Рис. 5.4 Похибка між точним та отриманим квадратично-наближеним рішенням рівняння $d^2y(x)/dx^2 - y(x) = 0$

звідки знаходять нормовану локальну координату елементу ξ, визначену на відрізку:

$$-1 \le \xi \le 1, \tag{5.19}$$

куди відображений елемент. Позначивши координату середини елементу як $X_c = (X_1 + X_2)/2$, а довжину елементу як $h = X_2 - X_1$ отримаємо:

$$\begin{aligned} \xi = L_2 - L_1 = \frac{x - X_1}{h} - \frac{X_2 - x}{h} = \frac{x - X_1 - X_2 + x}{h} = \\ = \frac{2x - X_1 - X_2}{h} = \frac{2(x - X_c)}{h} = \frac{2(x - X_1)}{h} - 1. \end{aligned} \tag{5.20}$$

Якщо внутрішні вузли комплекс елементів розміщені рівномірно, то використовуючи таку нормовану локальну координату легко отримати зручні формули для функцій форми. Наприклад для лінійного елементу:

$$N_1 = (1 - \xi)/2, \quad N_2 = (1 + \xi)/2; \tag{5.21}$$

для квадратичного елементу:

$$N_1 = -\xi(1 - \xi)/2, \quad N_2 = (1 + \xi)(1 - \xi), \quad N_3 = \xi(1 + \xi)/2; \tag{5.22}$$

для кубічного елементу:

$$\begin{aligned} N_1 &= -(9/16)(\xi + 1)(\xi + 1/3)(\xi - 1/3), \\ N_2 &= (27/16)(\xi + 1)(\xi - 1)(\xi - 1/3), \\ N_3 &= -(27/16)(\xi + 1)(\xi + 1/3)(\xi + 1/3), \\ N_4 &= (9/16)(\xi + 1/3)(\xi - 1/3)(\xi + 1). \end{aligned} \tag{5.23}$$

Щоб знайти похідні від функцій форми, виражених в нормованих локальних координатах знову використаємо матрицю Якобі:

$$\frac{dN_j}{d\xi} = [\mathbf{Jac}_\xi x]\frac{dN_j}{dx} = \frac{dx}{d\xi}\frac{dN_j}{dx}. \tag{5.24}$$

Обернувши останнє рівняння отримаємо:

$$\frac{dN_j}{dx} = [\mathbf{Jac}_{\xi} x]^{-1} \frac{dN_j}{d\xi} = \frac{1}{dx/d\xi} \frac{dN_j}{d\xi}, \tag{5.25}$$

де матрицю Якобі можна знайти як:

$$[\mathbf{Jac}_{\xi} x] = \frac{dN_1}{d\xi} X_1 + \frac{dN_2}{d\xi} X_2 + \ldots + \frac{dN_{p+1}}{d\xi} X_{p+1}. \tag{5.26}$$

Обчислюючи останній вираз отримаємо значення, що відповідає фізичному змісту Якобіана – відношенню об'ємів тіл при деформації. Тобто в даному випадку:

$$[\mathbf{Jac}_{\xi} x] = \frac{2}{h} \quad [\mathbf{Jac}_{\xi} x]^{-1} = \frac{h}{2}, \tag{5.27}$$

звідки:

$$\frac{dN_j}{dx} = \frac{d\xi}{dx} \frac{dN_j}{d\xi} = \frac{2}{h} \frac{dN_j}{d\xi} \quad \Rightarrow \quad dx = \frac{h}{2} d\xi. \tag{5.28}$$

На основі отриманих виразів можна набагато легше знайти матриці жорсткості та вектори навантажень. Наприклад для лінійного елементу матриця жорсткості буде рівна:

$$[\mathbf{K}]_i = \int_{X_{i,1}}^{X_{i,2}} \frac{d[\mathbf{N}]_i^{\mathbf{T}}}{dx} \frac{d[\mathbf{N}]_i}{dx} dx + \int_{X_{i,1}}^{X_{i,2}} [\mathbf{N}]_i^{\mathbf{T}} [\mathbf{N}]_i \, dx = \frac{1}{h_i} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \frac{h}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}; \tag{5.29}$$

для квадратичного елементу (співпадає з (5.14)):

$$\begin{aligned} [\mathbf{K}]_i &= \int_{X_{i,1}}^{X_{i,3}} \frac{d[\mathbf{N}]_i^{\mathbf{T}}}{dx} \frac{d[\mathbf{N}]_i}{dx} dx + \int_{X_{i,1}}^{X_{i,3}} [\mathbf{N}]_i^{\mathbf{T}} [\mathbf{N}]_i \, dx = \\ &= \frac{1}{-3h_i} \begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix} + \frac{h_i}{15} \begin{bmatrix} 2 & 1 & -1/2 \\ 1 & 8 & 1 \\ -1/2 & 1 & 2 \end{bmatrix}; \end{aligned} \tag{5.30}$$

для кубічного елементу:

$$\begin{aligned} [\mathbf{K}]_i &= \int_{X_{i,1}}^{X_{i,4}} \frac{d[\mathbf{N}]_i^{\mathbf{T}}}{dx} \frac{d[\mathbf{N}]_i}{dx} dx + \int_{X_{i,1}}^{X_{i,4}} [\mathbf{N}]_i^{\mathbf{T}} [\mathbf{N}]_i \, dx = \\ &= \frac{1}{h_i} \begin{bmatrix} 37/10 & -189/40 & 27/20 & -13/40 \\ -189/40 & 54/5 & -297/40 & 27/20 \\ 27/20 & -297/40 & 54/5 & -189/40 \\ -13/40 & 27/20 & -189/40 & 37/10 \end{bmatrix} + h_i \begin{bmatrix} 8/105 & 33/560 & -3/140 & 19/1680 \\ 33/560 & 27/70 & -27/560 & -3/140 \\ -3/140 & -27/560 & 27/70 & 33/560 \\ 19/1680 & -3/140 & 33/560 & 8/105 \end{bmatrix}; \end{aligned} \tag{5.31}$$

і так далі для елементів вищих порядків.

Порівнюючи отримані локальні матриці жорсткості з матрицями жорсткості, що розглядалися на початку при апроксимації класичними методами зважених нев'язок, можна помітити, що у першому випадку, на відміну від другого, отримані результати ніяк між собою не пов'язані, і передбачити коефіцієнти матриці, при включенні ще одного порядку інтерполяції стає неможливо. Це пов'язано з тим, що всі функції форми

потрібно обчислювати заново. Натомість при використанні класичного розкладу наближеного рішення методами зважених нев'язок, включаючи новий доданок з вищим порядком, є можливість використовувати вже обчислені коефіцієнти матриць:

$$M=1,\quad [A_{1,1}]\{a_1\}=\{f_1\},$$

$$M=2,\quad \begin{bmatrix} A_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix}\begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix}=\begin{Bmatrix} f_1 \\ f_2 \end{Bmatrix}, \tag{5.32}$$

$$M=3,\quad \begin{bmatrix} A_{1,1} & B_{1,2} & C_{1,3} \\ B_{2,1} & B_{2,2} & C_{2,3} \\ C_{3,1} & C_{3,2} & C_{3,3} \end{bmatrix}\begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix}=\begin{Bmatrix} f_1 \\ f_2 \\ f_3 \end{Bmatrix}.$$

Очевидно, що на кожному кроці уточнення апроксимації, отримані на попередньому кроці матриці та вектори зустрічаються знову і немає необхідності їх обчислювати заново.

Така можливість дозволяє значно зменшити час обчислень, наприклад при уточненні результатів розв'язку задачі. Крім того, при використанні строго ортогональних функцій форм, наприклад як у спектральних методах зважених нев'язок, можна отримати діагональні чи майже діагональні матриці:

$$M=1,\quad [K_{1,1}]\{a_1\}=\{f_1\},$$

$$M=2,\quad \begin{bmatrix} K_{1,1} & 0 \\ 0 & K_{2,2} \end{bmatrix}\begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix}=\begin{Bmatrix} f_1 \\ f_2 \end{Bmatrix}, \tag{5.33}$$

$$M=3,\quad \begin{bmatrix} K_{1,1} & 0 & 0 \\ 0 & K_{2,2} & 0 \\ 0 & 0 & K_{3,3} \end{bmatrix}\begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix}=\begin{Bmatrix} f_1 \\ f_2 \\ f_3 \end{Bmatrix},$$

що веде до лінійної чи майже лінійної складності обчислень систем рівнянь:

$$a_j = K_{j,j}^{-1} f_j. \tag{5.34}$$

Щоб мати можливість працювати за подібною схемою, необхідно відмовитися від відображення конкретного фізичного змісту внутрішньо-вузловими коефіцієнтами u_j розкладу наближеного рішення, тобто знову використовувати замість них абстрактні коефіцієнти a_j [1]. Таким чином інтерполяція вищих порядків буде будуватися шляхом адитивного уточнення інтерполяції нижчих порядків. В літературі по методах скінченних елементів, отримані таким способом базисні функції часто називають *ієрархічними поліномами* [1].

Оскільки ми відмовляємося від визначення внутрішніх вузлів, то для

[1] Це не стосується вузлових коефіцієнтів для вузлів, що розміщенні на границі елементу, оскільки вони повинні забезпечувати міжелементну неперервність та узгодженість апроксимованого рішення.

апроксимації вищих порядків знову можна використовувати симплекс елементи, де тепер кількість вузлів елементу буде меншою за кількість базисних інтерполяційних функцій. Описаний випадок є частиною біль загальної класифікації скінченних елементів, які тепер можна розрізняти як [1], [16], [3], [5]:

- *субпараметричні елементи* – кількість вузлів є меншою за кількість інтерполяційних функцій, сюди входять ієрархічні поліноми на симплекс елементах;
- *ізопараметричні елементи* – кількість вузлів співпадає з кількістю інтерполяційних функцій, сюди входять вже описані симплекс та комплекс елементи;
- *суперпараметричні елементи* – кількість вузлів є більшою за кількість інтерполяційних функцій.

Щоб створити набір ієрархічних базисних функцій для одновимірного елементу, ми повинні використати стандартні лінійні функції N_1 та N_2 для граничних вузлів (*Рис. 5.5.a*), оскільки вони відповідають за виконання міжелементної неперервності та узгодженості апроксимованого рішення.

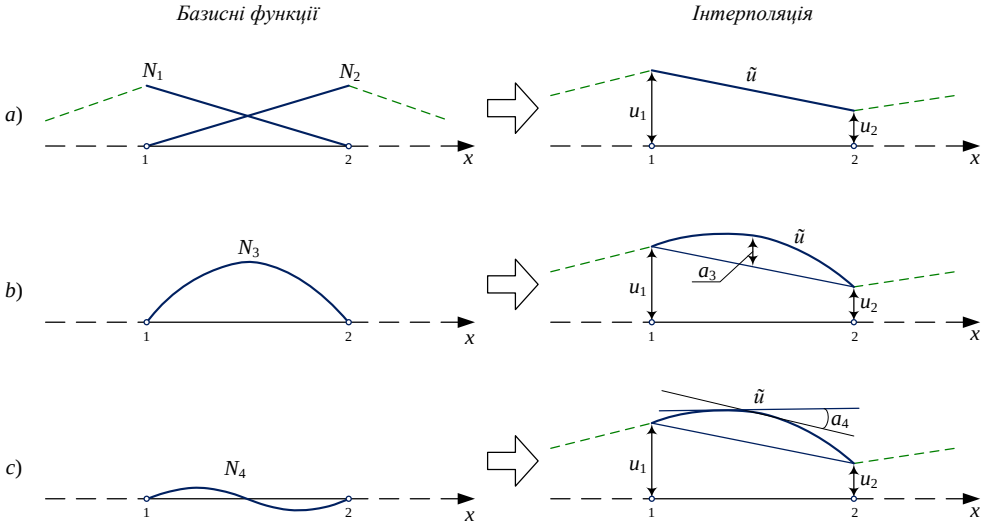


Рис. 5.5 Одновимірний елемент і відповідні ієрархічні базисні функції та інтерполяції: лінійна (a), квадратична (b) і кубічна (c)

Щоб ввести квадратичну інтерполяцію, додамо до розкладу наближеного рішення поліном другого порядку від нормованої локальної координати елементу:

$$N_3 = \alpha_0 + \alpha_1 \xi + \alpha_2 \xi^2, \tag{5.35}$$

з коефіцієнтами, вибраними так, щоб $N_3 = 0$ при $\xi = \pm 1$. Таким чином необхідна гладкість апроксимації між елементами буде збережена. Отримана квадратична інтерполяція (*Рис. 5.5.b*) запишеться як:

$$\tilde{u} = u_1 N_1 + u_2 N_2 + a_3 N_3, \tag{5.36}$$

де (див. (5.21) та (5.22)):

$$N_1 = (1-\xi)/2, \quad N_2 = (1+\xi)/2, \quad N_3 = (1+\xi)(1-\xi). \tag{5.37}$$

Коефіцієнт a_3 тепер не показує значення шуканого потенціалу у вузлі, натомість він описує величину відхилення лінійної інтерполяції $\tilde{u}$ в центрі елементу, оскільки в цій точці N_3 приймає значення одиниці.

Аналогічно для кубічного елементу, до квадратичного представлення (5.36) необхідно додати $a_4 N_4$, де N_4 – кубічний поліном виду:

$$N_4 = \alpha_0 + \alpha_1 \xi + \alpha_2 \xi^2 + \alpha_3 \xi^3, \tag{5.38}$$

що приймає нульове значення при $\xi = \pm 1$. З безмежної кількості таких поліномів, виберемо той, який показано на (*Рис. 5.5.c*). Він приймає нульове значення в точці $\xi = 0$, при чому в цій ж точці $dN_4/d\xi = 1$. Тому:

$$N_4 = \xi(1-\xi^2). \tag{5.39}$$

Тепер коефіцієнт a_4 описує відхилення куту нахилу в центрі елементу, від куту нахилу попередньої інтерполяції.

Аналогічним чином можна вивести формулу для поліному четвертого порядку:

$$N_5 = \xi^2(1-\xi^2), \tag{5.40}$$

однак, зміст коефіцієнту a_5 не є очевидним, та й в загальному випадку в цьому немає необхідності.

Як вже було сказано, існує безліч поліномів заданого порядку $p \geq 2$, що відповідають критеріям (3.70) та (3.71), тому описана система ієрархічних базисних функцій не є єдиною. Наприклад, інша зручна система ієрархічних функцій визначається як [1]:

$$N_{p+1} = \begin{cases} (\xi^p - 1)/p!, & p \text{ парне}, \\ (\xi^p - \xi)/p!, & p \text{ непарне}, \end{cases} \tag{5.41}$$

де $p \geq 2$ – степінь поліному (N_1 та N_2 не змінюються). Це дає систему базисних функцій:

$$\begin{aligned} &N_1 = (1-\xi)/2, && N_2 = (1+\xi)/2, \\ &N_3 = (\xi^2 - 1)/2, && N_4 = (\xi^3 - \xi)/6, \\ &N_5 = (\xi^4 - 1)/24, && N_6 = (\xi^5 - \xi)/120. \end{aligned} \tag{5.42}$$

Неважко визначити, що всі похідні від N_{p+1} другого і більш високих порядків приймають нульове значення при $\xi = 0$, за винятком $d^p N_{p+1}/d\xi^p$, що рівна в цій точці одиниці. Як наслідок, при використанні базисних функцій виду (5.41), коефіцієнти що входять в інтерполяцію можна співставити зі значеннями відповідних похідних:

$$a_{p+1} = d^p \tilde{u}/d\xi^p \Big|_{\xi=0}, \quad p \geq 2. \tag{5.43}$$

Таке співставлення надає їм фізичний зміст, але розуміється, що це не є обов'язковим.

Порівняємо результати – для елементу четвертого порядку, на основі (5.42) отримаємо:

$$[\mathbf{K}]_i = \frac{2}{h_i}\int_{-1}^{1}\frac{d[\mathbf{N}]_i^{\mathrm{T}}}{d\xi}\frac{d[\mathbf{N}]_i}{d\xi}d\xi + \frac{h_i}{2}\int_{-1}^{1}[\mathbf{N}]_i^{\mathrm{T}}[\mathbf{N}]_i d\xi = \tag{5.44}$$

$$= \frac{2}{h_i}\begin{bmatrix} 1/2 & -1/2 & 0 & 0 & 0 \\ -1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 2/3 & 0 & 1/15 \\ 0 & 0 & 0 & 2/45 & 0 \\ 0 & 0 & 1/15 & 0 & 1/126 \end{bmatrix} +$$

$$+\frac{h_i}{2}\begin{bmatrix} 2/3 & 1/3 & -1/3 & 1/45 & -1/30 \\ 1/3 & 2/3 & -1/3 & -1/45 & -1/30 \\ -1/3 & -1/3 & 4/15 & 0 & 8/315 \\ 1/45 & -1/45 & 0 & 4/945 & 0 \\ -1/30 & -1/30 & 8/315 & 0 & 1/405 \end{bmatrix}, \quad \{\mathbf{u}\}_i = \begin{Bmatrix} u_{i,1} \\ u_{i,2} \\ a_{i,3} \\ a_{i,4} \\ a_{i,5} \end{Bmatrix}.$$

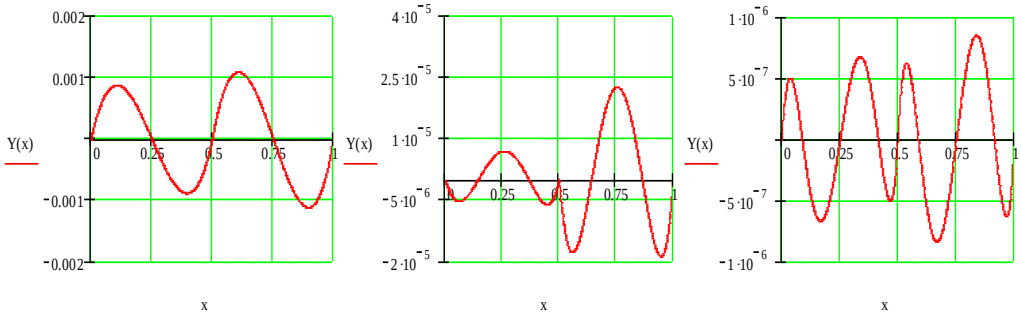
З цих матриць можна виділити підматриці для інтерполяції нижчого порядку шляхом відкидання останніх рядків і стовбців, або навпаки – інтерполяцію вищого порядку шляхом додавання нових рядків і стовбців. Всі вже обчислені коефіцієнти не змінюються.

Щоб виразити отримані $N_j(\xi)$ як $N_j(x)$ слід просто замінити локальну координату за формулою (5.20). При ансамблюванні глобальної матриці, та побудові результатів апроксимації, також необхідно враховувати, що всі коефіцієнти a_j є визначені тільки для конкретного елементу і не поширюються на сусідні елементи. Зв'язок між сусідніми елементами будується завдяки граничним вузлам, тобто в наших термінах завдяки $u_1 N_1$ та $u_2 N_2$, тому ненульові коефіцієнти результуючої глобальної матриці жорсткості будуть в інших рядках і стовбцях, ніж це було при ансамблюванні з використанням ізопараметричних скінченних елементів.

На *Рис. 5.6* зображено похибки апроксимації рішень, отриманих з допомогою двоелементної інтерполяції ієрархічними базисними функціями (5.42), тобто з використанням матриць (5.44).

Спробуємо тепер побудувати сімейство ортогональних базисних функцій, подібних до тих, що використовувалися в спектральних методах зважених нев'язок, що дасть змогу отримати діагональні чи майже діагональні матриці жорсткості, і як наслідок – значно спростити процес обчислень вузлових значень.

Нагадаємо, що дві функції є ортогональними, якщо їх скалярний добуток рівний нулю, в термінах методів зважених нев'язок, це означало, що скалярний добуток пробних і повірочних функцій був повинен давати ненульові значення тільки тоді, коли індекси цих функцій співпадали.



a) *b)* *c)*

Рис. 5.6 Похибка між точним та наближеним рішенням рівняння $d^2y(x)/dx^2 - y(x) = 0$*, отриманим двоелементною апроксимацією ієрархічними базисними функціями: (a) квадратична, (b) кубічна та (c) четвертого порядку*

Переносячи цю ідею в формулювання методу скінченних елементів, ми тепер вимагаємо, щоб для певним чином вибраних базисних функцій, інтеграли для еліптичних рівнянь в слабкій формі типу:

$$\int_{\Omega} \frac{dN_k(x)}{dx}\frac{dN_j(x)}{dx}dx = \frac{2}{h}\int_{-1}^{1} \frac{dN_k(\xi)}{d\xi}\frac{dN_j(\xi)}{d\xi}d\xi, \qquad (5.45)$$

давали нульові значення при $k \neq j$ і ненульові при $k = j$. Приклади вибору таких систем базисних функцій були наведені в попередніх розділах. Одним з них є множина поліномів Лежандра $P_p(\xi)$ для відрізку $-1 \le \xi \le 1$ [1]. Поліном Лежандра степені p визначається як:

$$P_p(\xi) = \frac{1}{2^p\, p!}\frac{d^p}{d\xi^p}\left((\xi^2-1)^p\right). \qquad (5.46)$$

Беручи невизначений інтеграл від цього полінома, попередньо помноживши його на $2p$, отримаємо формулу для знаходження функцій форми:

$$N_{p+1}(\xi) = \int \frac{1}{(p-1)!}\frac{1}{2^{p-1}}\frac{d^p}{d\xi^p}\left((\xi^2-1)^p\right)d\xi - \left[\int \frac{1}{(p-1)!}\frac{1}{2^{p-1}}\frac{d^p}{d\xi^p}\left((\xi^2-1)^p\right)d\xi\right]_{\xi=\pm 1}. \qquad (5.47)$$

При $p = 2, 3, 4, 5, 6, 7$, знаходимо (N_1 та N_2 не змінюються):

$$N_3 = \xi^2 - 1, \qquad N_4 = 2(\xi^3 - \xi),$$
$$N_5 = (15\xi^4 - 18\xi^2 + 3)/4, \qquad N_6 = 7\xi^5 - 10\xi^3 + 3\xi, \qquad (5.48)$$
$$N_7 = (105\xi^6 - 175\xi^4 + 75\xi^2 - 5)/8, \quad N_8 = (99\xi^7 - 189\xi^5 + 105\xi^3 - 15\xi)/4.$$

На основі цих формул, для рівняння (5.45) отримаємо:

$$\frac{2}{h_i}\int_{-1}^{1} \frac{d[\mathbf{N}]_i^{\mathrm{T}}}{d\xi}\frac{d[\mathbf{N}]_i}{d\xi}d\xi =$$

$$= \frac{2}{h_i} \begin{bmatrix} 1/2 & -1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 32/5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 72/7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 128/9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 200/11 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 288/13 \end{bmatrix}. \quad (5.49)$$

Зауважимо, що отримана діагональна (за винятком коефіцієнтів для стандартних Лагранжевих функцій форм) матриця стосується тільки еліптичних рівнянь в слабкій формі типу (5.45), тобто рівнянь де фігурують тільки перші похідні. Другий доданок рівняння, де фігурують самі функції форми скінченних елементів, дає вже не діагональну, проте стрічкову матрицю:

$$\frac{h_i}{2} \int_{-1}^{1} [\mathbf{N}]_i^{\mathrm{T}} [\mathbf{N}]_i \, d\xi =$$

$$= \frac{h_i}{2} \begin{bmatrix} 2/3 & 1/3 & -2/3 & 4/15 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & -2/3 & -4/15 & 0 & 0 & 0 & 0 \\ -2/3 & -2/3 & 16/15 & 0 & -8/35 & 0 & 0 & 0 \\ 4/15 & -4/15 & 0 & 64/105 & 0 & -64/315 & 0 & 0 \\ 0 & 0 & -8/35 & 0 & 16/35 & 0 & -40/231 & 0 \\ 0 & 0 & 0 & -64/315 & 0 & 256/639 & 0 & -64/429 \\ 0 & 0 & 0 & 0 & -40/231 & 0 & 400/1287 & 0 \\ 0 & 0 & 0 & 0 & 0 & -64/429 & 0 & 192/715 \end{bmatrix}. \quad (5.50)$$

Ми не будемо розглядати інші способи побудови ортогональної системи базисних функцій. Більше того, в загальному випадку процес отримання ортогонального базису є не таким очевидним та виходить за рамки нашого розгляду. Цікавому читачу рекомендуємо, за необхідності, ознайомитися з процесом ортогоналізації Грама-Шмідта [6], [7], [8], [9].

5.2. Багатовимірні комплекс і мультиплекс елементи

Тепер розширимо отримані нами результати на багатовимірні задачі. Перш за все, розглянемо білінійну та біквадратичну інтерполяції, як найпростіші інтерполяції, що можуть використовуватися для комплекс елементів у двовимірному просторі.

Також нагадаємо, що ми розглядаємо побудову апроксимацій що належать $C^0(\Omega)$ класу гладкості, тобто апроксимації для задач, що описуються диференціальними рівняннями з частинними похідними максимум другого порядку, для яких можливо побудувати слабку форму.

Як вже відомо, в одновимірному випадку лінійна інтерполяція отримувалася шляхом застосування симплекс елементів – відрізків з двома вузлами. У двовимірному випадку симплекс елементами є трикутники, але в ряді задач більш зручно будувати дискретизацію на чотирикутниках, що в

описаних термінах є комплекс елементами. Крім того, у деяких роботах [3], [10], прямокутні чотирикутники, через простоту виводу їх функцій форми, відносять до окремого підкласу *мультиплекс елементів*. Мультиплекс елементом вважається комплекс елемент, у якого сторони є паралельними координатним осям.

Рішення на чотирикутному елементі відповідає білінійній інтерполяції і описується рівнянням:

$$\tilde{u}_i(x_1, x_2) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + a_3 x_1 x_2. \qquad (5.51)$$

У звичній матричній формі, відповідні функції форми можна виразити як:

$$[\mathbf{N}] = [\mathbf{P}][\mathbf{C}]^{-1} = [1 \quad x_1 \quad x_2 \quad x_1 x_2] \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & X_{1,1}X_{1,2} \\ 1 & X_{2,1} & X_{2,2} & X_{2,1}X_{2,2} \\ 1 & X_{3,1} & X_{3,2} & X_{3,1}X_{3,2} \\ 1 & X_{4,1} & X_{4,2} & X_{4,1}X_{4,2} \end{bmatrix}^{-1}. \qquad (5.52)$$

Така система може бути використана і для комплекс і для мультиплекс елементів, тобто для довільних чотирикутників. Розв'язавши її, отримаємо вираз для функцій форми, що відповідає добутку Лагранжевих поліномів відповідної степені. Так, якщо кожен з вузлів чотирикутника позначити двома індексами $r = 1, 2$ та $s = 1, 2$, по одному на кожну координату, то відповідну вузлу (r, s) базисну функцію елементу можна записати як:

$$N_{rs}(x_1, x_2) = \Lambda_r^p(x_1)\Lambda_s^p(x_2), \qquad (5.53)$$

де Λ_r^p та Λ_s^p – фундаментальні поліноми Лагранжа степені p (в даному випадку $p = 1$), що визначаються рівнянням (5.8).

Отримане співвідношення є загальним і за допомогою нього можна отримати вирази для функцій форм чотирикутних елементів довільного порядку інтерполяції (*Рис. 5.7*). Очевидно, що при подальшому ансамблюванні глобальної системи, вузли на границях сусідніх елементів повинні співпадати.

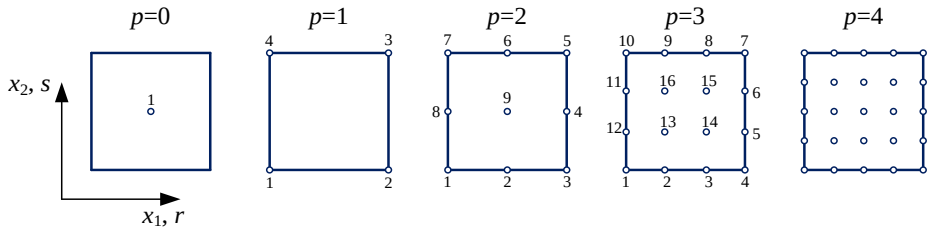


Рис. 5.7 Приклади чотирикутних елементів різних порядків інтерполяції

Як і в одновимірному випадку, для побудови виразів базисних функцій елементів зручно використовувати локальні нормовані координати по одній на кожну глобальну координату. Так для мультиплекс елементів, тобто прямокутників, отримаємо:

$$\begin{aligned} \xi_1 &= 2(x_1 - X_{c,1})/h_1, \quad \partial\xi_1 = 2\partial x_1/h_1, \quad -1 \le \xi_1 \le 1, \\ \xi_2 &= 2(x_2 - X_{c,2})/h_2, \quad \partial\xi_2 = 2\partial x_2/h_2, \quad -1 \le \xi_2 \le 1, \end{aligned} \qquad (5.54)$$

де точка $(X_{c,1}, X_{c,2})$ є центром елементу в системі координат (x_1, x_2) (*Рис. 5.8.a*).

Для довільних чотирикутників, тобто комплекс елементів, співвідношення є складнішими і в загальному випадку будуються на основі деякого відображення з глобальної системи координат в локальну (*Рис. 5.8.b*). Ці співвідношення можна знайти як:

$$\begin{aligned} x_1 &= \frac{1}{4}(1-\xi_1)(1-\xi_2)X_{1,1} + \frac{1}{4}(1+\xi_1)(1-\xi_2)X_{2,1} + \\ &+ \frac{1}{4}(1+\xi_1)(1+\xi_2)X_{3,1} + \frac{1}{4}(1-\xi_1)(1+\xi_2)X_{4,1}, \\ x_2 &= \frac{1}{4}(1-\xi_1)(1-\xi_2)X_{1,2} + \frac{1}{4}(1+\xi_1)(1-\xi_2)X_{2,2} + \\ &+ \frac{1}{4}(1+\xi_1)(1+\xi_2)X_{3,2} + \frac{1}{4}(1-\xi_1)(1+\xi_2)X_{4,2}. \end{aligned} \tag{5.55}$$

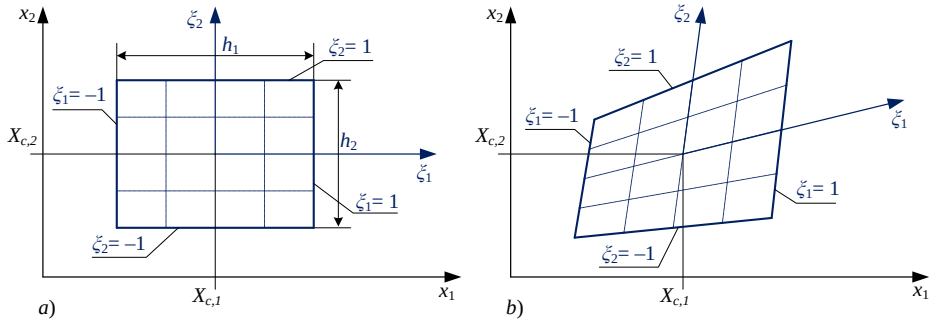


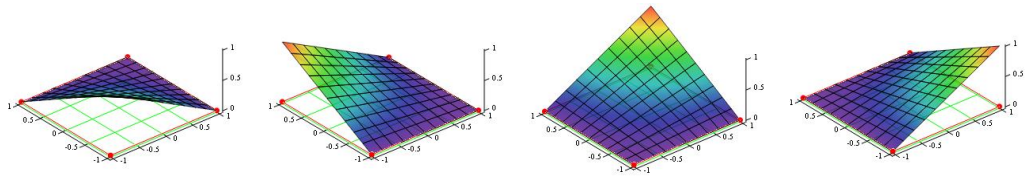
Рис. 5.8 Нормовані координати (ξ_1, ξ_2) для мультиплекс (a) та комплекс (b) елементів в площині (x_1, x_2)

Згідно останніх формул, для білінійного (і комплекс і мультиплекс) елементу отримаємо:

$$\begin{aligned} N_{1,1} = N_1 = (1-\xi_1)(1-\xi_2)/4, \quad N_{1,2} = N_4 = (1-\xi_1)(1+\xi_2)/4, \\ N_{2,1} = N_2 = (1+\xi_1)(1-\xi_2)/4, \quad N_{2,2} = N_3 = (1+\xi_1)(1+\xi_2)/4. \end{aligned} \tag{5.56}$$

Графіки цих функцій зображено на *Рис. 5.9*. Позначивши локальні нормовані координати j-го вузла як $(\Theta_{j,1}, \Theta_{j,2})$ останній вираз можна записати як:

$$N_j(\xi_1, \xi_2) = (1+\xi_1\Theta_{j,1})(1+\xi_2\Theta_{j,2})/4. \tag{5.57}$$



$N_{11}(\xi_1, \xi_2)$ $N_{12}(\xi_1, \xi_2)$ $N_{22}(\xi_1, \xi_2)$ $N_{21}(\xi_1, \xi_2)$

Рис. 5.9 Зображення білінійних базисних функцій

Аналогічно виводяться формули для біквадратичного елементу (*Рис. 5.10*):

$$N_j(\xi_1,\xi_2)=N_{rs}(\xi_1,\xi_2)=\Lambda_r^2(x_1)\Lambda_s^2(x_2)=\prod_{k=1,\,k\neq j}^{3}\frac{\xi_1-\Theta_{k,1}}{\Theta_{j,1}-\Theta_{k,1}}\frac{\xi_2-\Theta_{k,2}}{\Theta_{j,2}-\Theta_{k,2}}, \quad (5.58)$$

тобто:

$$\begin{array}{ll}
\text{Кутові вузли} & N_j(\xi_1,\xi_2)=\dfrac{1}{4}\xi_1\Theta_{j,1}(1+\xi_1\Theta_{j,1})\xi_2\Theta_{j,2}(1+\xi_2\Theta_{j,2}), \quad j=1,3,5,7, \\
\text{Вузли посередині сторони} & \begin{cases} N_j(\xi_1,\xi_2)=\dfrac{1}{2}(1-\xi_1^2)\xi_2\Theta_{j,2}(1+\xi_2\Theta_{j,2}), & j=2,6, \\ N_j(\xi_1,\xi_2)=\dfrac{1}{2}\xi_1\Theta_{j,1}(1+\xi_1\Theta_{j,1})(1-\xi_2^2), & j=4,8, \end{cases} \\
\text{Внутрішній вузол} & N_j(\xi_1,\xi_2)=(1-\xi_1^2)(1-\xi_2^2), \quad j=9,
\end{array} \quad (5.59)$$

і так далі, для елементів вищих порядків.

Знайдемо перші похідні наближеного розкладу на чотирикутних елементах високого порядку $\partial\tilde{u}/\partial x_1$ та $\partial\tilde{u}/\partial x_2$:

$$\frac{\partial\tilde{u}}{\partial x_1}=\sum_{j=1}^{M}\left(u_j\frac{\partial N_j}{\partial x_1}\right), \quad \frac{\partial\tilde{u}}{\partial x_2}=\sum_{j=1}^{M}\left(u_j\frac{\partial N_j}{\partial x_2}\right), \quad (5.60)$$

де:

$$\begin{Bmatrix}\dfrac{\partial N_j}{\partial x_1}\\ \dfrac{\partial N_j}{\partial x_2}\end{Bmatrix}=[\mathbf{Jac}_{\xi}x]^{-1}\begin{Bmatrix}\dfrac{\partial N_j}{\partial \xi_1}\\ \dfrac{\partial N_j}{\partial \xi_2}\end{Bmatrix}, \quad \begin{Bmatrix}\dfrac{\partial N_j}{\partial \xi_1}\\ \dfrac{\partial N_j}{\partial \xi_2}\end{Bmatrix}=[\mathbf{Jac}_{\xi}x]\begin{Bmatrix}\dfrac{\partial N_j}{\partial x_1}\\ \dfrac{\partial N_j}{\partial x_2}\end{Bmatrix}. \quad (5.61)$$

Матриця Якобі в даному випадку, на основі (5.26) та (5.55) визначається як:

$$[\mathbf{Jac}_{\xi}x]=\begin{bmatrix}\dfrac{\partial x_1}{\partial \xi_1} & \dfrac{\partial x_2}{\partial \xi_1}\\ \dfrac{\partial x_1}{\partial \xi_2} & \dfrac{\partial x_2}{\partial \xi_2}\end{bmatrix}=\frac{1}{4}\begin{bmatrix}-(1-\xi_2) & (1-\xi_2) & (1+\xi_2) & -(1+\xi_2)\\ -(1-\xi_1) & -(1+\xi_1) & (1+\xi_1) & (1-\xi_1)\end{bmatrix}\begin{bmatrix}X_{1,1} & X_{1,2}\\ X_{2,1} & X_{2,2}\\ X_{3,1} & X_{3,2}\\ X_{4,1} & X_{4,2}\end{bmatrix}, \quad (5.62)$$

звідки видно, що для застосування комплекс елементів обчислення є на порядок складнішими, ніж це було для симплекс елементів, оскільки похідні від функцій форми вже не є константами, а деякими функціями від координат. Для їх обчислення дуже часто використовують методи чисельного інтегрування, які будуть розглянуті пізніше.

Використовуючи білінійну інтерполяцію на мультиплекс елементах все ж можна вивести прості аналітичні вирази, особливо для однорідних еліптичних рівнянь, що розглядаються в ізотропному середовищі. Наприклад для задачі стаціонарної теплопровідності з попередніх розділів, на основі системи (5.52) матриця жорсткості прийме вигляд:

$$[\mathbf{K}]_i=\int_{\Omega_i}[\mathbf{B}]_i^{\mathrm{T}}[\mathbf{D}]_i[\mathbf{B}]_i\,d\Omega_i=$$

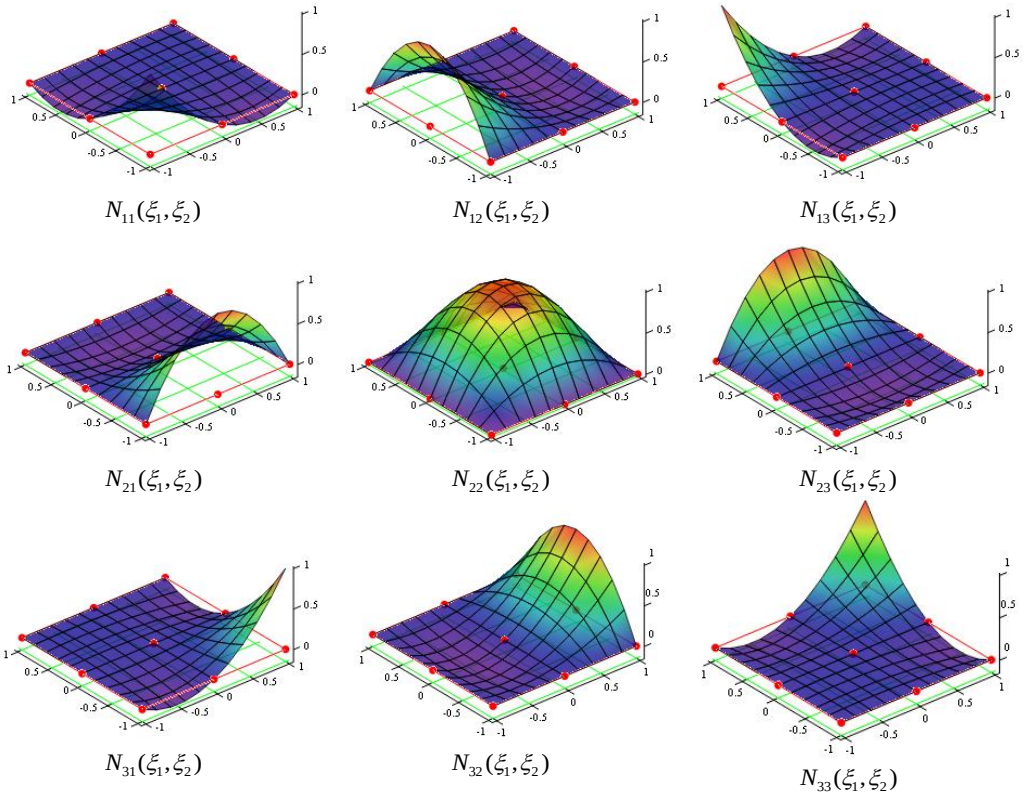


Рис. 5.10 Зображення біквадратичних базисних функцій (квадратичних функцій форми чотирикутника)

$$= \frac{\lambda}{6h_1h_2}\begin{bmatrix} 2h_1^2+2h_2^2 & h_1^2-2h_2^2 & -h_1^2-h_2^2 & -2h_1^2+h_2^2 \\ h_1^2-2h_2^2 & 2h_1^2+2h_2^2 & -2h_1^2+h_2^2 & -h_1^2-h_2^2 \\ -h_1^2-h_2^2 & -2h_1^2+h_2^2 & 2h_1^2+2h_2^2 & h_1^2-2h_2^2 \\ -2h_1^2+h_2^2 & -h_1^2-h_2^2 & h_1^2-2h_2^2 & 2h_1^2+2h_2^2 \end{bmatrix}. \qquad (5.63)$$

Якщо тепер розбити область на 100 квадратів, аналогічно до того, як це робилося у попередніх розділах, отримаємо нову глобальну матрицю жорсткості для того ж глобального вектору навантажень. Різниця між рішенням останньої системи і рішенням, отриманим при використанні симплекс елементів показано на *Рис. 5.11*.

Окремі "піки" нев'язок, що розміщені всередині елементів, пов'язані з тим, що при лінійній інтерполяції, похідна від функцій форми симплекс елементів, а отже і від наближеного рішення, є сталою і обов'язково має розрив першого роду в міжелементних зонах. А оскільки трикутники будувалися по одній з діагоналей квадрату, отримуємо нев'язки біля границь де задані природні крайові умови, при чому напрям такого "зміщення" нев'язок залежить від обраної діагоналі.

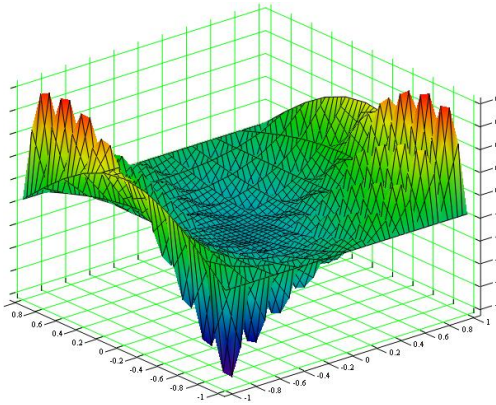


Рис. 5.11 Різниця між рішеннями, отриманими з допомогою білінійної апроксимації, та апроксимації на симплекс елементах, при однаковій кількості вузлів

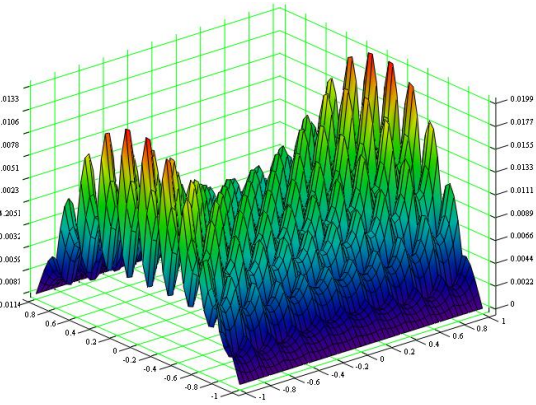


Рис. 5.12 Різниця між рішенням, отриманим з допомогою методу Бубнова-Гальоркіна, при $M=5$, та рішенням білінійної скінченно-елементної апроксимації

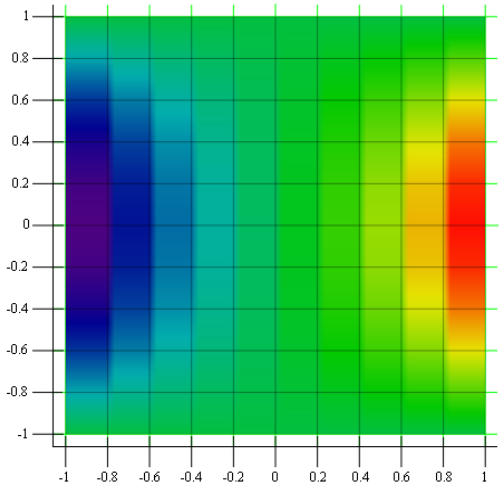


Рис. 5.13 Значення похідної $\partial\tilde{T}/\partial x$ від отриманого апроксимованого рішення

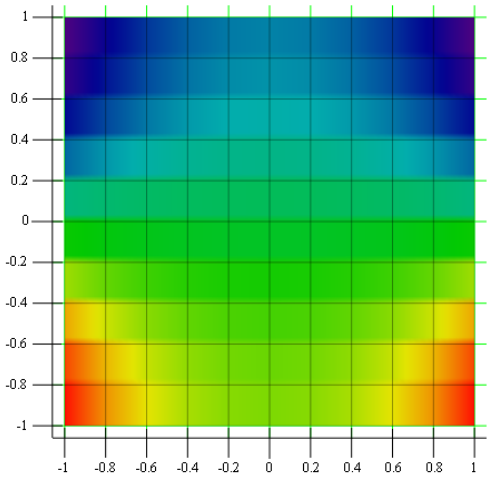


Рис. 5.14 Значення похідної $\partial\tilde{T}/\partial y$ від отриманого апроксимованого рішення

Натомість білінійна інтерполяція, за рахунок члену x_1x_2, що є елементом не повного квадратичного розкладу, компенсує вказаний недолік апроксимації на симплекс елементах і отримане наближене рішення вже не залежить від вибору діагоналі. Тим не менше, похідні є розривними в міжелементних зонах (*Рис. 5.13* та *Рис. 5.14*), оскільки рішення належить $C^0(\Omega)$ класу гладкості. Також за рахунок члену x_1x_2 похідні в межах елементу вже не є константами, а змінюються вздовж ортогональних координат: $\partial\tilde{T}/\partial x_1$ вздовж x_2; $\partial\tilde{T}/\partial x_2$ вздовж x_1.

У деяких дослідженнях, для спрощення процесу обчислень внутрішні вузли подібних елементів інтерполяції вищих порядків просто не враховуються. Це

допустимо, оскільки всі члени добутків Лагранжевих поліномів з формули (5.53) перевищують число членів, необхідних для побудови полінома деякої степені p. Тому зайві члени можна відкинути без порушення необхідних умов до гладкості апроксимованого рішення. Такі елементи були отримані випадково [4] і дістали загальну назву *серендипових*[1]. Вперше серендипові елементи з'явилися в 1968 році[2].

Формули функцій форм для білінійних серендипових елементів співпадають з білінійною інтерполяцією (5.56). Для біквадратичних серендипових елементів (*Рис. 5.15*, *Рис. 5.16*) отримаємо:

$$\begin{aligned}
&\text{Кутові вузли} && N_j(\xi_1,\xi_2)=\frac{1}{4}(1+\xi_1\Theta_{j,1})(1+\xi_2\Theta_{j,2})(\xi_1\Theta_{j,1}+\xi_2\Theta_{j,2}-1), && j=1,3,5,7,\\
&\text{Вузли посередині сторони} && \begin{cases} N_j(\xi_1,\xi_2)=\frac{1}{2}(1-\xi_1^2)(1+\xi_2\Theta_{j,2}), & j=2,6,\\ N_j(\xi_1,\xi_2)=\frac{1}{2}(1+\xi_1\Theta_{j,1})(1-\xi_2^2), & j=4,8. \end{cases}
\end{aligned} \qquad (5.64)$$

Для бікубічних:

$$\begin{aligned}
&\text{Кутові вузли} && N_j(\xi_1,\xi_2)=\frac{1}{32}(1+\xi_1\Theta_{j,1})(1+\xi_2\Theta_{j,2})(-10+9(\xi_1^2+\xi_2^2)), && j=1,4,7,10,\\
&\text{Вузли посередині сторони} && \begin{cases} N_j(\xi_1,\xi_2)=\frac{9}{32}(1+\xi_1\Theta_{j,1})(1-\xi_1^2)(1+9\xi_2\Theta_{j,2}), & j=2,3,8,,9,\\ N_j(\xi_1,\xi_2)=\frac{9}{32}(1+9\xi_1\Theta_{j,1})(1-\xi_2^2)(1+\xi_2\Theta_{j,2}), & j=5,6,11,12. \end{cases}
\end{aligned} \qquad (5.65)$$

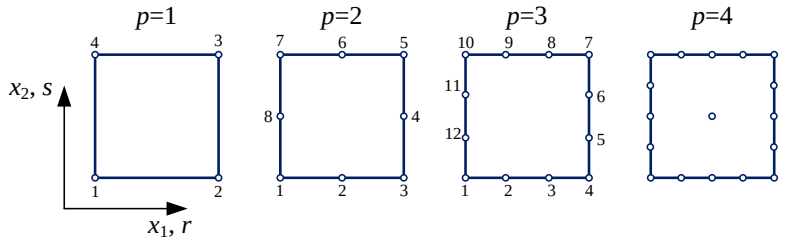


Рис. 5.15 Сімейство двовимірних серендипових елементів

На жаль, кількість необхідних для виконання умов гладкості апроксимованого рішення компонент, що можуть бути отримані тільки з використанням граничних вузлів, є недостатньою для порядків інтерполяції $p \geq 4$. Як наслідок, для отримання інтерполяцій таких високих порядків необхідно знову вводити внутрішні вузли [1].

Графічно, функції форми комплекс (і мультиплекс) елементів дуже зручно зображати за допомогою трикутника Паскаля, звідки одразу можна побудувати загальні матричні вирази для функцій форм типу (5.52), що підходять для всіх нами описаних сімейств елементів вищих порядків (*Рис. 5.17*).

[1] Від англійського слова "*serendipity*", що прийшло від стародавньої назви Цейлону (Serendip) і означало "подарунок несподіваних і цінних відкриттів чи знахідок" – в честь Перської казки "Три принци з Серендипу".

[2] Ergatoudis J., Irons B., Zienkiewicz O. // Int. J. Solids Structures, 4:31-42, 1968.

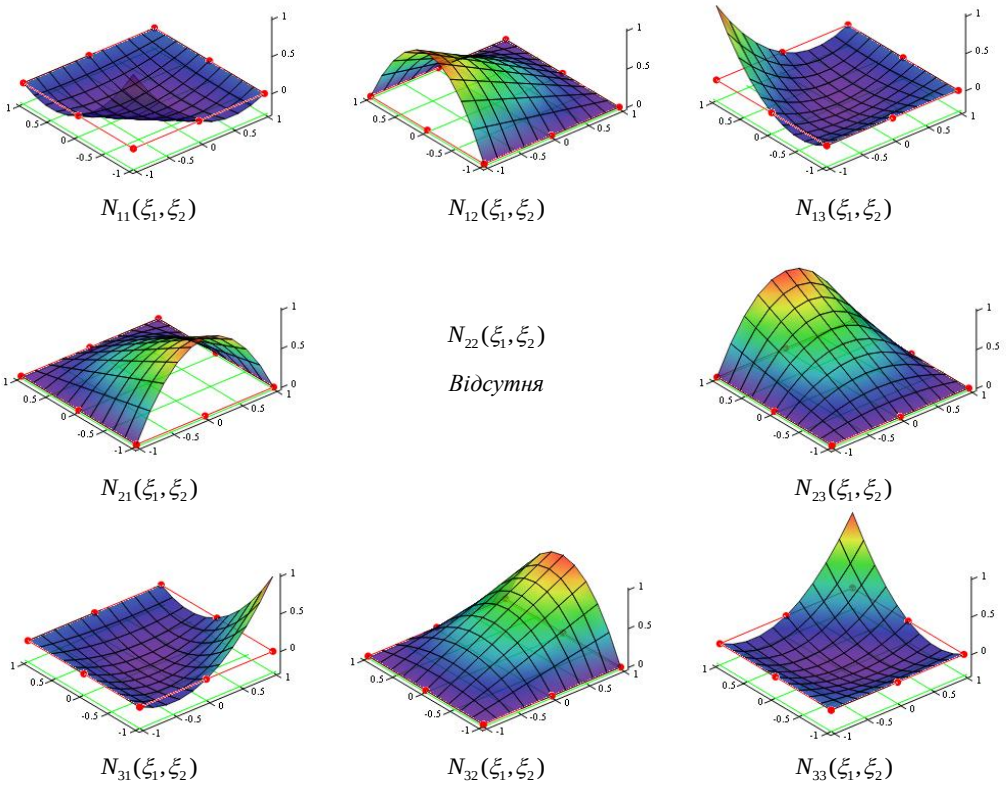


Рис. 5.16 Зображення квадратичних серендипових функцій форми чотирикутника

Кожен степеневий рівень трикутника Паскаля містить повний (двовимірний) поліном відповідного порядку, тому одразу стає очевидним, що наприклад двовимірна Лагранжева інтерполяціє на чотирикутнику містить повну систему членів порядку p в поліноміальному розкладі та окремі члени порядку $2p$. Так білінійна інтерполяція це добуток двох лінійних інтерполяцій по кожній з координат, тобто система є повною відносно першого порядку (містить члени $1, x_1, x_2$) і неповною відносно другого порядку (містить добуток $x_1 x_2$, але не містить членів x_1^2 та x_2^2).

Відомо [10], що швидкість збіжності скінченно-елементної моделі визначається найвищим порядком повного поліному, тому на практиці дуже не ефективно використовувати Лагранжеві чотирикутні елементи через їх надлишковість, і саме через це замість них використовують сімейство серендипових елементів.

У деяких багатовимірних задачах, для спрощення обчислень, можна також використовувати комплекс елементи з непропорційним порядком інтерполяції по різних координатах (*Рис. 5.17.d, Рис. 5.18*). Наприклад лінійною інтерполяцією по одній координаті та квадратичною чи кубічною по іншій [10].

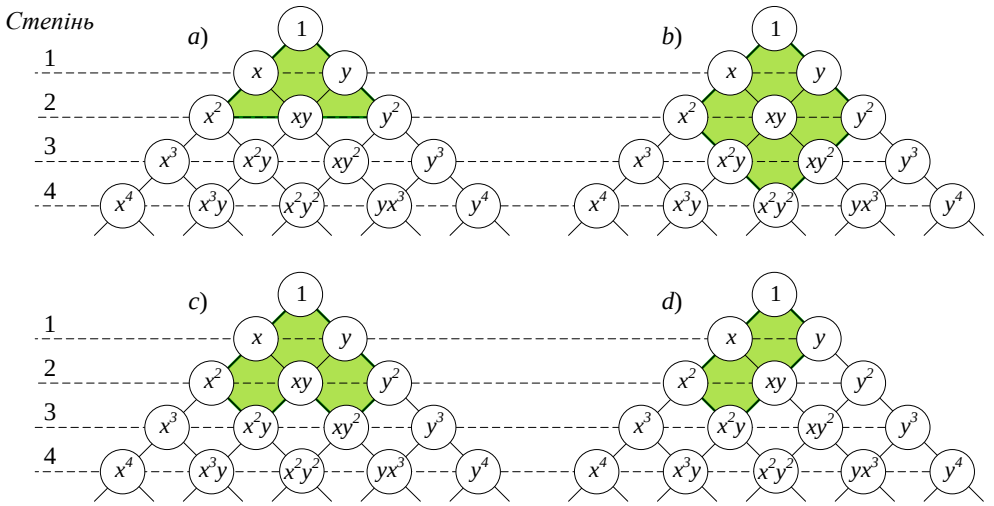


Рис. 5.17 Трикутники Паскаля, де зафарбовані елементи утворюють: (a) повний квадратичний поліном, (b) квадратичний Лагранжевий поліном, (c) квадратичний серендиповий поліном, (d) непропорційний поліном – лінійний по y *та квадратичний по* x

Додаючи проміжні вузли до симплекс елементів, аналогічно до описаних чотирикутних комплекс елементів побудуємо елементи вищих порядків на трикутниках. Як видно з *Рис. 5.17.a*, такі елементи одразу містять повний поліном необхідного порядку без зайвих членів. Функції форми для трикутних елементів вищих порядків можна вивести, знову ж таки, використовуючи матричні формули типу (5.52) з коефіцієнтами, що беруться з трикутника Паскаля.

Наприклад для квадратичного елементу отримаємо:

$$[\mathbf{N}]=[\mathbf{P}][\mathbf{C}]^{-1}=[1 \quad x_1 \quad x_2 \quad x_1^2 \quad x_1x_2 \quad x_2^2]\cdot \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & X_{1,1}^2 & X_{1,1}X_{1,2} & X_{1,2}^2 \\ 1 & X_{2,1} & X_{2,2} & X_{2,1}^2 & X_{2,1}X_{2,2} & X_{2,2}^2 \\ 1 & X_{3,1} & X_{3,2} & X_{3,1}^2 & X_{3,1}X_{3,2} & X_{3,2}^2 \\ 1 & X_{4,1} & X_{4,2} & X_{4,1}^2 & X_{4,1}X_{4,2} & X_{4,2}^2 \\ 1 & X_{5,1} & X_{5,2} & X_{5,1}^2 & X_{5,1}X_{5,2} & X_{5,2}^2 \\ 1 & X_{6,1} & X_{6,2} & X_{6,1}^2 & X_{6,1}X_{6,2} & X_{6,2}^2 \end{bmatrix}^{-1}. \tag{5.66}$$

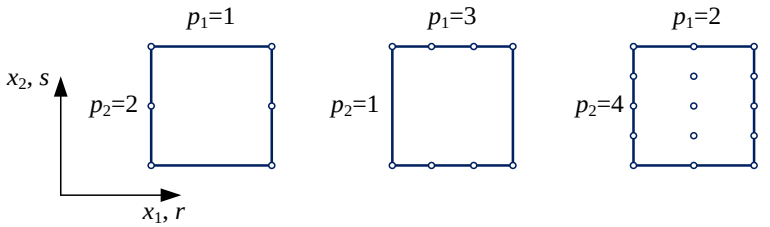


Рис. 5.18 Приклади елементів з непропорційним розміщенням вузлів

Розв'язавши систему, прийдемо до добутку Лагранжевих поліномів другого степеня по кожній з барицентричних координат трикутника (*Рис. 5.20*), або в загальному випадку, для трикутного елементу довільного порядку (*Рис. 5.19*):

$$N_j(x_1,x_2)=N_{abc}(L_1,L_2,L_3)=\Lambda_a^p(L_1)\Lambda_b^p(L_2)\Lambda_c^p(L_3), \qquad (5.67)$$

де a,b,c – локальні індекси вершин відносно кожної барицентричної координати. Зазначимо, що $0\le a,b,c\le p$ та в кожному вузлі $a+b+c=p$.

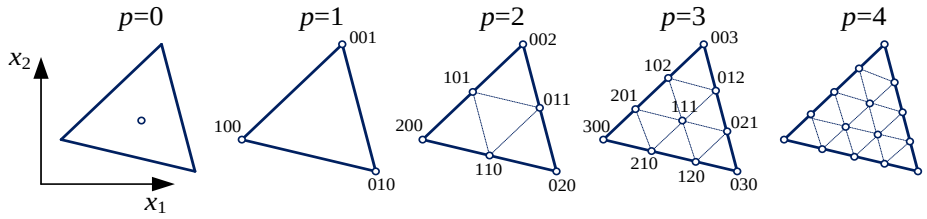


Рис. 5.19 Приклади трикутних елементів різних порядків інтерполяції

Оскільки неможливо безпосередньо використати формулу (5.8) через специфічну індексацію вузлів, то кожен з Лагранжевих поліномів формули (5.67), а також формул для симплексів будь-якої розмірності, через барицентричні координати зручно представляти як:

$$\Lambda_j^p(L_i)\begin{cases}=\prod_{k=1}^{j}\dfrac{pL_i-k+1}{k}, & j\ge 1,\\ =1, & j=0.\end{cases} \qquad (5.68)$$

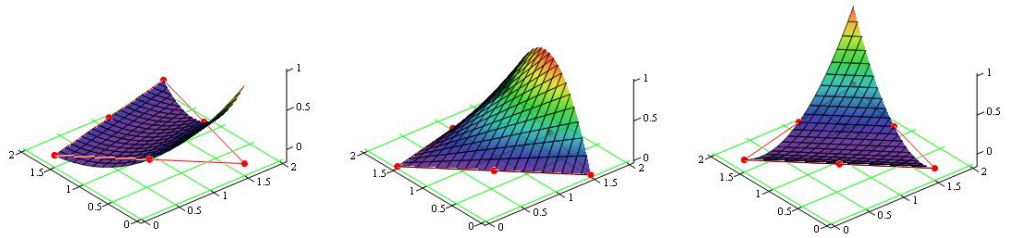
Наприклад для квадратичного елементу (*Рис. 5.20*):

$$N_{200}=\Lambda_2^2(L_1)\Lambda_0^2(L_2)\Lambda_0^2(L_3)=\Lambda_2^2(L_1)\cdot 1\cdot 1=\frac{2L_1-1+1}{1}\frac{2L_1-2+1}{2}=$$
$$=L_1(2L_1-1),\quad N_{011}=4L_2L_3, \qquad (5.69)$$
$$N_{020}=L_2(2L_2-1),\quad N_{110}=4L_1L_2,$$
$$N_{002}=L_3(2L_3-1),\quad N_{101}=4L_1L_3.$$

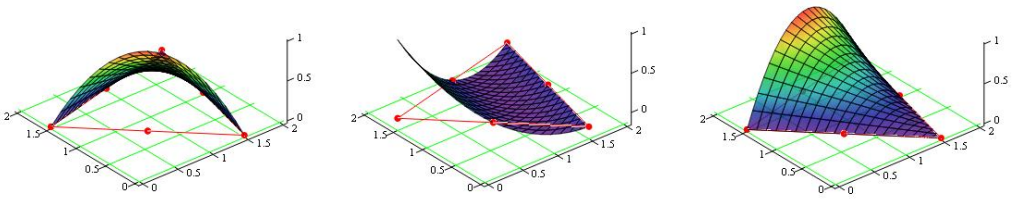
Для кубічного елементу:

$$N_{300}=\frac{1}{2}L_1(3L_1-1)(3L_1-2),\quad N_{210}=\frac{9}{2}L_1L_2(3L_1-1),\quad N_{120}=\frac{9}{2}L_1L_2(3L_2-1),$$
$$N_{030}=\frac{1}{2}L_2(3L_2-1)(3L_2-2),\quad N_{210}=\frac{9}{2}L_2L_3(3L_2-1),\quad N_{120}=\frac{9}{2}L_2L_3(3L_3-1), \qquad (5.70)$$
$$N_{003}=\frac{1}{2}L_3(3L_3-1)(3L_3-2),\quad N_{102}=\frac{9}{2}L_3L_1(3L_3-1),\quad N_{201}=\frac{9}{2}L_1L_3(3L_1-1),$$
$$N_{111}=27L_1L_2L_3.$$

Будуючи функції форми на основі барицентричних координат, дуже зручно використовувати аналог трикутника Паскаля для тривимірного простору (*Рис. 5.21.а*), де змінними вже є не глобальні x_1 та x_2, а барицентричні координати L_1, L_2 та L_3. При такому підході повний поліном степені p, і як наслідок, наближене рішення $\tilde{u}$ можна записати як [10], [11]:



$N_{200}(L_1,L_2,L_3)$ $N_{110}(L_1,L_2,L_3)$ $N_{020}(L_1,L_2,L_3)$



$N_{101}(L_1,L_2,L_3)$ $N_{002}(L_1,L_2,L_3)$ $N_{011}(L_1,L_2,L_3)$

Рис. 5.20 Зображення квадратичних функцій форми трикутника (трикутник з вершинами $(1,5;0),(2;2),(0;1,5)$*)*

$$\tilde{u}=\sum_{j=1}^{M}\alpha_j L_1^a L_2^b L_3^c,\quad a+b+c=p. \tag{5.71}$$

Наприклад для кубічного елементу:

$$\begin{aligned}\tilde{u}=&\alpha_1 L_1^3+\alpha_2 L_2^3+\alpha_3 L_3^3+\alpha_4 L_1^2 L_2+\alpha_5 L_1 L_2^2+\alpha_6 L_1 L_3^2+\\&+\alpha_7 L_2^2 L_3+\alpha_8 L_1^2 L_3+\alpha_9 L_2 L_3^2+\alpha_{10} L_1 L_2 L_3.\end{aligned} \tag{5.72}$$

Використовуючи останні співвідношення при обчисленні Лагранжевих трикутних елементів, стає можливо виписати прості аналітичні формули для інтегралів по функціях форми цих елементів довільного порядку. Вперше такі формули були виведені в 1973 році в роботі [23]. Наприклад, враховуючи що:

$$\left|[\mathbf{Jac}_{\mathbf{L}}\mathbf{r}]\right|=2\Omega, \tag{5.73}$$

на основі (5.67) та (5.71) отримаємо:

$$\begin{aligned}\int_{\Omega}N_j(x_1,x_2)dx_1dx_2&=\int_{\Omega}N_{abc}(L_1,L_2,L_3)\left|[\mathbf{Jac}_{\mathbf{L}}\mathbf{r}]\right|dL_1dL_2dL_3=\\&=2\Omega\int_{\Omega}L_1^aL_2^bL_3^cdL_1dL_2dL_3.\end{aligned} \tag{5.74}$$

Заміняючи $L_3=1-L_1-L_2$, (5.74) можна переписати як:

$$2\Omega\int_{\Omega}L_1^aL_2^bL_3^cdL_1dL_2dL_3=2\Omega\int_0^1\int_0^{1-L_1}L_1^aL_2^b(1-L_1-L_2)^cdL_2dL_1. \tag{5.75}$$

Інтегруючи по частинам, отримаємо:

$$2\Omega\int_0^1\int_0^{1-L_1}L_1^aL_2^b(1-L_1-L_2)^cdL_2dL_1=\frac{c}{b+1}2\Omega\int_0^1\int_0^{1-L_1}L_1^aL_2^{b+1}(1-L_1-L_2)^{c-1}dL_2dL_1. \tag{5.76}$$

Продовживши цей процес отримаємо:

$$2\Omega\int_0^1\int_0^{1-L_1} L_1^a L_2^b (1-L_1-L_2)^c\, dL_2 dL_1 = \frac{b!c!}{(b+c)!} 2\Omega\int_0^1\int_0^{1-L_1} L_1^a L_2^{b+c} (1-L_1-L_2)^0\, dL_2 dL_1, \quad (5.77)$$

$$2\Omega\int_0^1\int_0^{1-L_1} L_1^a L_2^b (1-L_1-L_2)^c\, dL_2 dL_1 = \frac{a!b!c!}{(a+b+c)!} 2\Omega\int_0^1\int_0^{1-L_1} L_1^0 L_2^{a+b+c} (1-L_1-L_2)^0\, dL_2 dL_1. \quad (5.78)$$

Що в результаті приводить до:

$$\int_\Omega N_j(x_1,x_2)dx_1dx_2 = 2\Omega\int_0^1\int_0^{1-L_1} L_1^a L_2^b (1-L_1-L_2)^c\, dL_2 dL_1 = \frac{a!b!c!}{(a+b+c+2)!} 2\Omega. \quad (5.79)$$

Зауважимо тепер, що при використанні симплекс елементів, де функції форми відповідають барицентричним координатам, можна використовувати останню формулу для обчислення інтегралів добутку типу:

$$\int_\Omega N_j N_k d\Omega = \int_\Omega L_1^1 L_2^1 L_3^0 d\Omega = \frac{1!1!0!}{(1+1+0+2)!} 2\Omega = \frac{\Omega}{12}. \quad (5.80)$$

Крім того, виведена формула дає ті ж результати, що й формули обчислення вектору навантажень $\{\mathbf{f}\}$ з попередніх розділів. Інтегрування здійснюється по грані елементу, що є симплексом в просторі з розмірністю $N-1$, в двовимірному випадку границею є відрізок:

$$\int_\Gamma N_j d\Gamma = \int_\Gamma L_1^1 L_2^0 d\Gamma = \frac{1!0!}{(1+0+1)!}\Gamma = \frac{\Gamma}{2}, \quad (5.81)$$

в тривимірному випадку границею є трикутник:

$$\int_\Omega N_j d\Omega = \int_\Omega L_1^1 L_2^0 L_3^0 d\Omega = \frac{1!0!0!}{(1+0+0+2)!} 2\Omega = \frac{\Omega}{3}. \quad (5.82)$$

Знайдемо перші похідні наближеного розкладу на трикутних елементах високого порядку $\partial\tilde{u}/\partial x_1$ та $\partial\tilde{u}/\partial x_2$. Для цього знову використаємо матрицю Якобі при $L_3 = 1-L_1-L_2$:

$$\begin{Bmatrix} \frac{\partial N_{abc}}{\partial L_1} \\ \frac{\partial N_{abc}}{\partial L_2} \end{Bmatrix} = [\mathbf{Jac_L r}] \begin{Bmatrix} \frac{\partial N_{abc}}{\partial x_1} \\ \frac{\partial N_{abc}}{\partial x_2} \end{Bmatrix}, \quad \begin{Bmatrix} \frac{\partial N_{abc}}{\partial x_1} \\ \frac{\partial N_{abc}}{\partial x_2} \end{Bmatrix} = [\mathbf{Jac_L r}]^{-1} \begin{Bmatrix} \frac{\partial N_{abc}}{\partial L_1} \\ \frac{\partial N_{abc}}{\partial L_2} \end{Bmatrix}, \quad (5.83)$$

який рівний:

$$[\mathbf{Jac_L r}] = \begin{bmatrix} \frac{\partial x_1}{\partial L_1} & \frac{\partial x_2}{\partial L_1} \\ \frac{\partial x_1}{\partial L_2} & \frac{\partial x_2}{\partial L_2} \end{bmatrix} = \begin{bmatrix} X_{1,1}-X_{3,1} & X_{1,2}-X_{3,2} \\ X_{2,1}-X_{3,1} & X_{2,2}-X_{3,2} \end{bmatrix}. \quad (5.84)$$

Зауваживши, що:

$$\frac{\partial N_{abc}(L_1,L_2)}{\partial L_1} = \frac{\partial N_{abc}(L_1,L_2,L_3)}{\partial L_1}\frac{\partial L_1}{\partial L_1} +$$

$$+\frac{\partial N_{abc}(L_1,L_2,L_3)}{\partial L_2}\frac{\partial L_2}{\partial L_1}+\frac{\partial N_{abc}(L_1,L_2,L_3)}{\partial L_3}\frac{\partial L_3}{\partial L_1},$$
$$\frac{\partial L_1}{\partial L_1}=1,\quad \frac{\partial L_2}{\partial L_1}=0,\quad \frac{\partial L_3}{\partial L_1}=\frac{\partial(1-L_1-L_2)}{\partial L_1}=-1, \tag{5.85}$$

отримаємо:

$$\frac{\partial N_{abc}(L_1,L_2)}{\partial L_1}=\frac{\partial N_{abc}(L_1,L_2,L_3)}{\partial L_1}-\frac{\partial N_{abc}(L_1,L_2,L_3)}{\partial L_3},$$
$$\frac{\partial N_{abc}(L_1,L_2)}{\partial L_2}=\frac{\partial N_{abc}(L_1,L_2,L_3)}{\partial L_2}-\frac{\partial N_{abc}(L_1,L_2,L_3)}{\partial L_3}. \tag{5.86}$$

Враховуючи те, що трикутні елементи представляють повні поліноми необхідної степені, а також можливість застосування простих аналітичних формул при інтегруванні та наявність алгоритмів автоматичної дискретизації ними об'єктів великої складності, саме ці елементи найчастіше використовують при рішенні практичних задач [10].

Всі описані формули можуть бути розширені для застосування у просторах з довільною кількістю вимірів. Для цього достатньо ввести додаткові множники по кожній з координатних осей у формулах (5.53) або (5.67). Графічно це відповідає використанню аналогів трикутника Паскаля для відповідної розмірності (*Рис. 5.21*), звідки безпосередньо можна взяти коефіцієнти для функцій форм у матричному вигляді типу (5.52) чи (5.66).

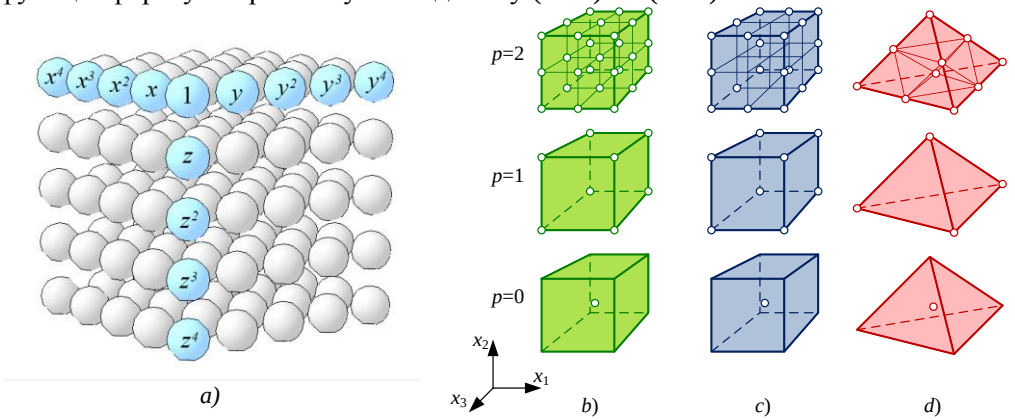


Рис. 5.21 Тривимірні Лагранжеві елементи і члени, що дають вклад у формування базисних функцій: a) аналог трикутника Паскаля для трьох вимірів; b) сімейство шестигранних Лагранжевих елементів; c) сімейство шестигранних серендипових елементів; d) сімейство тетраедральних Лагранжевих елементів

На багатовимірні випадки можна розширити і аналітичні формули інтегрування по функціях форми симплексу (5.79). Кожен вимір додає нову барицентричну координату, тому у загальному випадку можна записати:

$$\int_\Omega N_j(x_1,x_2)d\Omega = N!\Omega\int_0^1\int_0^{1-L_1}\dots\int_0^{1-L_M} L_1^{a_1}L_2^{a_2}\dots L_M^{a_m}dL_M\dots dL_2dL_1 =$$

$$= \frac{a_1!a_2!\ldots a_M!N!}{(a_1+a_1+\ldots+N)!}\Omega, \qquad (5.87)$$

де як і раніше N – кількість вимірів, $M = N+1$ – кількість вузлів симплексу, Ω – об'єм симплексу. Так для трьох вимірів отримаємо:

$$\int_V L_1^a L_2^b L_3^c L_4^d dV = \frac{a!b!c!d!}{(a+b+c+d+3)!}6V. \qquad (5.88)$$

Як і для одновимірних елементів вищих порядків, так і для багатовимірних елементів можна вивести ієрархічні базисні функції [1]. Цей варіант особливо ефективний, коли додаткові ієрархічні поліноми зв'язують значення шуканої величини на границях елементу.

Так, як вже відомі набори ієрархічних поліномів для одновимірного випадку, побудова субпараметричних багатовимірних елементів не є складною задачею, оскільки:

- функції для кутових вузлів співпадають зі стандартними лінійними функціями;
- добуток ієрархічних функцій, що були визначені для одновимірного випадку, в кутових вузлах завжди рівний нулю.

Тому, беручи за основу одновимірні Лагранжеві поліноми низьких порядків, наприклад лінійні, та записуючи їх добуток з ієрархічними функціями з одновимірного випадку на відповідній границі елементу, отримаємо систему ієрархічних базисних функцій для багатовимірних елементів. Наприклад для білінійних функцій форми (5.56) кожна локальна координата ξ_1 і ξ_2 відповідає координатній осі, а отже, парі протилежних границь елементу. Щоб ввести систему ієрархічних функцій потрібно на границях помножити відповідні лінійні Лагранжеві поліноми $\Lambda_r^1(\xi_1)$ та $\Lambda_s^1(\xi_2)$, на одновимірні ієрархічні функції необхідного порядку (5.42) $N_{p+1}(\xi_2)$ та $N_{p+1}(\xi_1)$ відповідно (*Рис. 5.22*).

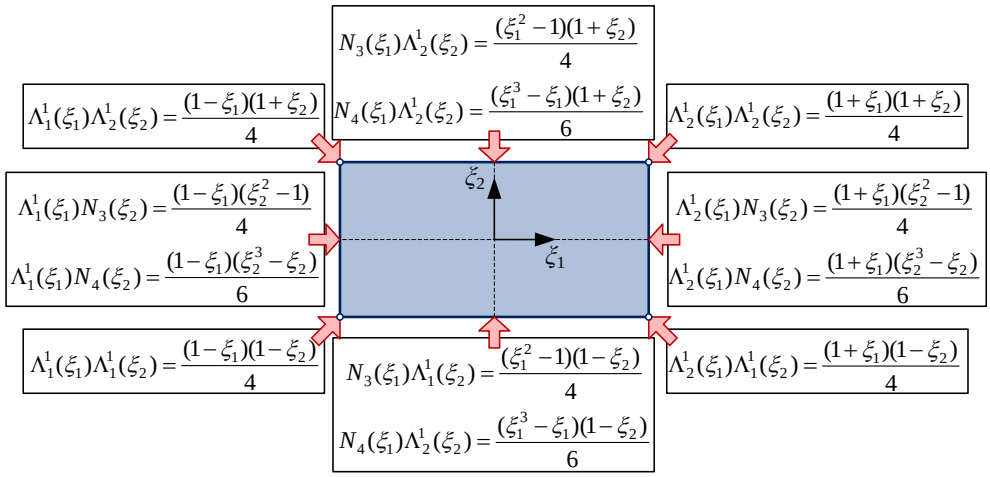


Рис. 5.22 Квадратичні та кубічні ієрархічні базисні функції для чотирикутника

Слід зауважити, що як і у випадку серендипових елементів, для отримання інтерполяційних функцій порядку $p \geq 4$ необхідно включати додаткові внутрішні вузли, щоб поліноми зберігали повноту, так і для ієрархічних базисних функцій на чотирикутниках чи шестигранниках, при інтерполяції порядку $p \geq 4$ необхідно вводити внутрішні базисні функції, що не асоціюються з жодною з границь елементу і рівні на них нулю (див трикутник Паскаля *Рис. 5.17*). Наприклад, для чотирикутника внутрішньою функцією, що відповідає члену $\xi_1^2 \xi_2^2$ може бути:

$$N_{inner} = (\xi_1^2 - 1)(\xi_2^2 - 1)/4. \qquad (5.89)$$

Очевидно, що шукані ієрархічні змінні a_j для кожної границі елементу при ансамблюванні повинні бути ототожнені з відповідними ієрархічними змінними на границях сусідніх елементів, як це відбувається для вузлових значень при використанні класичних комплекс елементів.

Розглянемо тепер ієрархічні базисні функції для симплексів. Враховуючи (5.20), для грані трикутника, наприклад утвореної вершинами $(100,010)$, локальна нормована координата це $\xi = L_2 - L_1$. В той же час, барицентрична координата L_3 на цій грані рівна нулю. Наведені судження справедливі і для інших граней трикутника, чи будь-якого симплексу у просторі з довільною розмірністю, за умови перенумерації вузлів. Тому формули для одновимірних ієрархічних базисних функцій (5.41) можна узагальнити через барицентричні координати, звідки для кожної грані можна будувати набір ієрархічних базисних функцій (*Рис. 5.23*):

$$N_{p+1}^{(100,010)} = \begin{cases} \left((L_2 - L_1)^p - (L_1 + L_2)^p\right)/p!, & p \text{ парне}, \\ \left((L_2 - L_1)^p - (L_2 - L_1)(L_1 + L_2)^{p-1}\right)/p!, & p \text{ непарне}. \end{cases} \qquad (5.90)$$

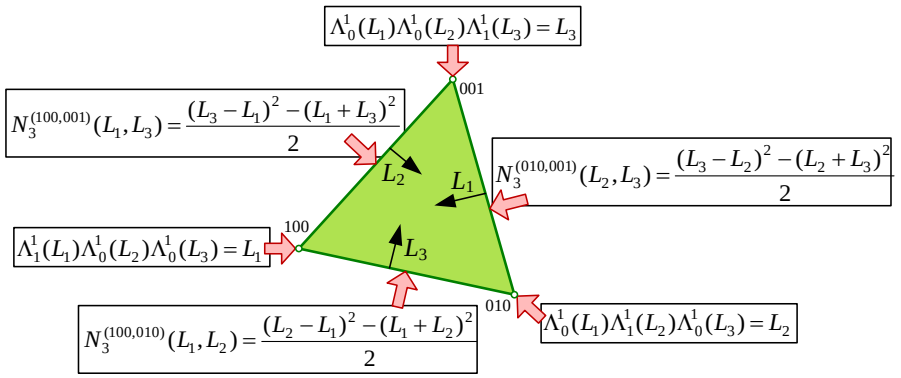


Рис. 5.23 Квадратичні ієрархічні базисні функції для трикутника

Щоб отримати повний набір ієрархічних інтерполяційних функцій порядку $p \geq 3$ для трикутника, слід подібно до чотирикутних елементів, вводити внутрішні базисні функції, що рівні нулю на всіх границях елементу (див трикутник Паскаля *Рис. 5.17*). Наприклад для кубічної інтерполяції можна

використати функцію $L_1L_2L_3$ (див. (5.72)), для інтерполяції четвертого порядку функції $L_1^2L_2L_3$, $L_1L_2^2L_3$ та $L_1L_2L_3^2$, і так далі.

Очевидно, що як і для одновимірного випадку, запропонована система ієрархічних базисних функцій не є єдиною. За необхідності можна вивести альтернативні системи на основі сімейства ортогональних чи майже ортогональних поліномів, що використовуються в спектральних методах зважених нев'язок, наприклад використовуючи поліноми Лежандра з формули (5.46).

З іншої сторони, всі функції форми, що розглядалися, будувалися тільки у вигляді поліномів. Поліноміальна апроксимація набула популярності в зв'язку зі простотою обчислень, і в деякому сенсі є оптимальною [5]. Але, в загальному випадку немає необхідності обмежуватися тільки нею. Наприклад, у якості внутрішніх базисних функцій можна використати аналітичні функції типу:

$$\cos\left(\frac{\pi}{2}\xi_1\right)\cos\left(\frac{\pi}{2}\xi_2\right), \qquad (5.91)$$

що рівні нулю на границях елементів. Чи будь-які інші функції, що забезпечують збіжність скінченно-елементної моделі.

5.3. Чисельне інтегрування при побудові матриць елементів

При обчисленні матриць елементів для елементів високих степенів інтерполяції зростає складність підінтегральних виразів, що робить алгебраїчні викладки дуже громіздкими. Якщо крім того використовується відображення області елементу, яке міняє її форму (наприклад для чотирикутних комплекс елементів *Рис. 5.8.b*), то для обчислення похідних, що входять в ці вирази, необхідно знайти обернену матрицю Якобі (див. (5.61),(5.62)). При цьому, інтеграли стають на стільки складними, що знайти їх точне аналітичне рішення майже неможливо. У таких випадках застосовують процедури чисельного інтегрування, при яких інтеграл рівняння методу зважених нев'язок заміняється на деяку просту в обчисленні суму [1].

Історично чисельне інтегрування вперше використовувалося при розв'язку задач механіки за довго до винайдення методу скінченних елементів. З цієї області прийшли і назви для чисельного знаходження інтегралів. Так чисельне інтегрування по одній змінній називається механічною *квадратурою*[1], а чисельне знаходження подвійного інтегралу – механічною *кубатурою* [13].

Відомо багато методів чисельного інтегрування (див. наприклад [13], [14]), детальний їх аналіз виходить за рамки нашого розгляду, тут будуть описані методи чисельного інтегрування, що застосовуються безпосередньо при побудові скінченно-елементних моделей. Чисельне інтегрування почало застосовуватися в методі скінченних елементів у середині 1960-их років

[1] Від латинського "*quadratura*" – надання квадратної форми, під чим розумілося знаходження площі складної фігури шляхом розбиття її на маленькі квадрати. З винайденням інтегрального числення термін квадратура став синонімом інтегралу.

(вперше в 1966 році[1,2]). Квадратурні та кубатурні формули були адаптовані під потреби МСЕ з робіт по прикладній математиці[3,4,5] і опубліковані в таких тепер відомих роботах як [5] та [15].

Класичний спосіб обчислення квадратур полягає в заміні даної складної чи невідомої підінтегральної функції $G(\xi)$, що в нашому випадку, в силу використання локальних нормованих координат, визначена на відрізку $-1 \le \xi \le 1$, на деяку просту інтерполяційну чи апроксимаційну функцію. Остання функція повинна бути такою, щоб інтеграл обчислювався безпосередньо. Зазвичай у якості інтерполяційних чи апроксимаційних функцій беруться поліноми.

Щоб знайти значення квадратур для виразів, які утворюються при виводі формул для скінченних елементів, використовуються зважені значення цих підінтегральних виразів у спеціально вибраних внутрішніх вузлах, при чому ці вузли зазвичай не співпадають з вузлами комплексів. В одновимірному випадку на проміжку $-1 \le \xi \le 1$ завжди можна визначити набір спеціально вибраних, не обов'язково рівновіддалених вузлів $\xi_1, \xi_2, \ldots, \xi_p$ і знайти деякий поліном $F_g(\xi)$ степені $g \ge p$, що співпадає з невідомою $G(\xi)$ в кожному з цих вузлів. Тоді, інтеграл можна наближено обчислити на основі цього поліному:

$$\int_{-1}^{1} G(\xi)d\xi = \int_{-1}^{1} F_g(\xi)d\xi + \theta(G) = \alpha_1 G(\xi_1) + \alpha_2 G(\xi_2) + \ldots + \\ + \alpha_p G(\xi_p) + \theta(G) = \sum_{i=1}^{p} \alpha_i G(\xi_i) + \theta(G), \tag{5.92}$$

де $\theta(G)$ – залишковий член, що виражає похибку квадратурної формули. Рішення буде точним тоді і тільки тоді, коли початкова підінтегральна функція $G(\xi)$ сама є поліномом степені $\le g$. В іншому випадку завжди існує похибка, утворена не врахованим залишковим членом $\theta(G)$. Як і раніше, похибка між $G(\xi)$ та $F_g(\xi)$ буде зменшуватися з наближенням до визначених вузлів. Саме тому набір вузлів підбирається спеціальним чином так, щоб отримати максимальну точність апроксимації.

---

[1] Irons B. – Numerical integration applied to finite element method // Conf. on Use of digital computers in Srtructural Engineering, Univ. of Newcastle, July 1966.

[2] Felippa C. – Refined finite element analysis of linear and nonlinear two-dimensional structures // Ph.D. Dissertation , Department of Civil Engineering, University of California at Berkeley, Berkeley, CA, 1966.

[3] Hammer P., Marlowe O., Stroud A. – Numerical Integration Over Simpexes and Cones // Math. Tables Aids Comp., 10:130-137, 1956.

[4] Hammer P., Stroud A. – Numerical evaluation of multiple integrals // Math. Tables Aids Comput., 12:272–280, 1958.

[5] Abramowitz M., Stegun L., (eds.) – Handbook of Mathematical Functions with Formulas // Graphs and Mathematical Tables, Applied Mathematics Series 55, Natl. Bur. Standards, U.S. Department of Commerce, Washington, D.C., 1964.

Спробуємо вивести такі квадратурні формули, щоб апроксимація давала точне значення інтегралу кожного разу, коли $G(\xi)$ є поліномом степені не вище $g \geq p$. Необхідно так підібрати вузли $\xi_1, \xi_2, \ldots, \xi_p$ і коефіцієнти $\alpha_1, \alpha_2, \ldots, \alpha_p$, щоб квадратурна формула (5.92) була точною для всіх поліномів $G(\xi)$ найвищої можливої степені g, тобто $\theta(G) = 0$. Ми маємо $2p$ невідомих (α_i та ξ_i). Поліном степені $2p-1$ визначається $2p$ коефіцієнтами, тому найвища можлива степінь g рівна:

$$g = 2p - 1. \tag{5.93}$$

Так як p – ціле число, то g завжди буде непарним числом, наприклад для одного вузла найвища можлива степінь g, при якій $\theta(G) = 0$ – рівна одиниці, при двох вузлах – трьом, при трьох – п'яти, при чотирьох – семи, і так далі.

Для справедливості виразу (5.92) необхідно і достатньо щоб він був вірним при:

$$G(\xi) = 1, \xi, \xi^2, \ldots, \xi^{2p-1}. \tag{5.94}$$

Справді, припускаючи що:

$$\int_{-1}^{1} \xi^k d\xi = \sum_{i=1}^{p} \alpha_i \xi_i^k \quad k = 0, 1, 2, \ldots, 2p-1, \tag{5.95}$$

та:

$$G(\xi) = \sum_{k=0}^{2p-1} C_k \xi^k, \tag{5.96}$$

отримаємо:

$$\int_{-1}^{1} G(\xi) d\xi = \sum_{k=0}^{2p-1} C_k \int_{-1}^{1} \xi^k d\xi = \sum_{k=0}^{2p-1} C_k \sum_{i=1}^{p} \alpha_i \xi_i^k = \sum_{i=1}^{p} \alpha_i \sum_{k=0}^{2p-1} C_k \xi_i^k = \sum_{i=1}^{p} \alpha_i G(\xi_i). \tag{5.97}$$

Враховуючи що:

$$\int_{-1}^{1} \xi^k d\xi = \frac{1-(-1)^{k+1}}{k+1} = \begin{cases} 2/(k+1), & k \text{ парне}, \\ 0, & k \text{ непарне}, \end{cases} \tag{5.98}$$

отримаємо:

$$\begin{aligned}
\alpha_1 + \alpha_2 + \alpha_3 + \ldots + \alpha_p &= 2, \\
\alpha_1\xi_1 + \alpha_2\xi_2 + \alpha_3\xi_3 + \ldots + \alpha_p\xi_p &= 0, \\
\alpha_1\xi_1^2 + \alpha_2\xi_2^2 + \alpha_3\xi_3^2 + \ldots + \alpha_p\xi_p^2 &= 2/3, \\
&\ldots \\
\alpha_1\xi_1^k + \alpha_2\xi_2^k + \alpha_3\xi_3^k + \ldots + \alpha_p\xi_p^k &= \left(1-(-1)^{p+1}\right)/\left(p+1\right), \\
&\ldots \\
\alpha_1\xi_1^{2p-2} + \alpha_2\xi_2^{2p-2} + \alpha_3\xi_3^{2p-2} + \ldots + \alpha_p\xi_p^{2p-2} &= 2/\left(2p-1\right), \\
\alpha_1\xi_1^{2p-1} + \alpha_2\xi_2^{2p-1} + \alpha_3\xi_3^{2p-1} + \ldots + \alpha_p\xi_p^{2p-1} &= 0.
\end{aligned} \tag{5.99}$$

Щоб розв'язати останню систему потрібно мати набір вузлів $\xi_1, \xi_2, \ldots, \xi_p$, вибраних так, щоб отримати найвищу точність квадратурної формули (5.92). В даному випадку використовують спеціальний математичний прийом: Розглянемо ортогональний поліном Лежандра $P_p(\xi)$ з формули (5.46) (*Рис. 5.24*).

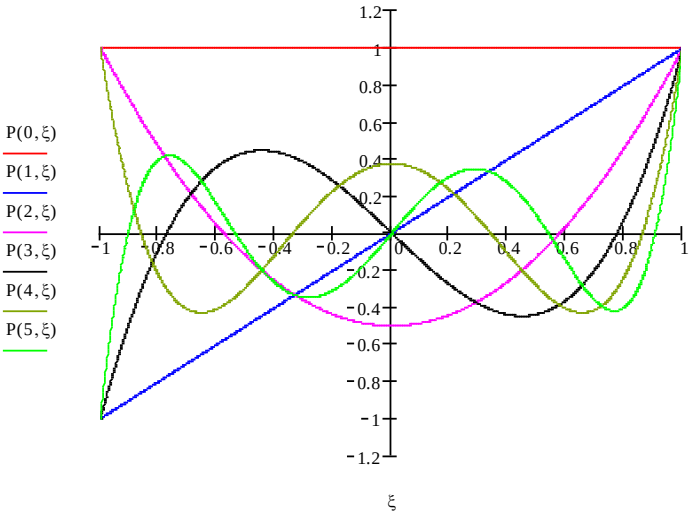


Рис. 5.24 Графік поліномів Лежандра від нульової до п'ятої степені на відрізку $-1 \le \xi \le 1$

Для нього можна виділити наступні основні характеристики:

- $P_p(1) = 1$, $P_p(-1) = (-1)^p$ для $p = 0, 1, 2, \ldots$;
- $\int_{-1}^{1} P_p(\xi) Q_k(\xi) d\xi = 0$ при $k < p$, де $Q_k(\xi)$ – будь-який поліном степені k, меншої p;
- поліном Лежандра $P_p(\xi)$ має p різних дійсних коренів на інтервалі $-1 \le \xi \le 1$ (див. *Таблиця 5.1*).

Виберемо у якості інтерполяційної функції поліном виду:

$$F_g(\xi) = \xi^k P_p(\xi), \quad k = 0, 1, 2, \ldots, p-1. \tag{5.100}$$

Так як степінь цього поліному не перевищує $2p-1$, то на основі системи (5.99) для нього повинна бути справедлива формула (5.92) та:

$$\int_{-1}^{1} \xi^k P_p(\xi) d\xi = \sum_{i=1}^{p} \alpha_i \xi_i^k P_p(\xi_i). \tag{5.101}$$

З іншої сторони, в силу ортогональності поліномів Лежандра:

$$\int_{-1}^{1} \xi^k P_p(\xi) d\xi = 0, \quad k < p, \tag{5.102}$$

звідки:

Таблиця 5.1

Корені поліномів Лежандра від нульової до п'ятої степені та відповідні їм вагові коефіцієнти квадратури Гауса-Лежандра

Поліном	Корені	Вагові коефіцієнти
$P_1(\xi)=\xi$	$\xi_1=0$	$\alpha_1=2$
$P_2(\xi)=\frac{1}{2}\left(3\xi^2-1\right)$	$\xi_1=-\sqrt{\frac{1}{3}},\ \xi_2=\sqrt{\frac{1}{3}}$	$\alpha_1=1,\ \alpha_2=1$
$P_3(\xi)=\frac{1}{2}\left(5\xi^3-3\xi\right)$	$\xi_1=-\sqrt{\frac{3}{5}},\ \xi_2=0,\ \xi_3=\sqrt{\frac{3}{5}}$	$\alpha_1=\frac{5}{9},\ \alpha_2=\frac{8}{9},\ \alpha_3=\frac{5}{9}$
$P_4(\xi)=\frac{1}{8}\left(35\xi^4-30\xi^2+3\right)$	$\xi_1=-\sqrt{\frac{15+2\sqrt{30}}{35}},\ \xi_2=-\sqrt{\frac{15-2\sqrt{30}}{35}},$ $\xi_3=\sqrt{\frac{15-2\sqrt{30}}{35}},\ \xi_4=\sqrt{\frac{15+2\sqrt{30}}{35}}$	$\alpha_1=\frac{18-\sqrt{30}}{36},\ \alpha_2=\frac{18+\sqrt{30}}{36},$ $\alpha_3=\frac{18+\sqrt{30}}{36},\ \alpha_4=\frac{18-\sqrt{30}}{36}$
$P_5(\xi)=\frac{1}{8}\left(63\xi^5-70\xi^3+16\xi\right)$	$\xi_1=-\sqrt{\frac{35+2\sqrt{70}}{63}},$ $\xi_2=-\sqrt{\frac{35-2\sqrt{70}}{63}},$ $\xi_3=0,$ $\xi_4=\sqrt{\frac{35-2\sqrt{70}}{63}},$ $\xi_5=\sqrt{\frac{35+2\sqrt{70}}{63}}$	$\alpha_1=\frac{322-13\sqrt{70}}{900},$ $\alpha_2=\frac{322+13\sqrt{70}}{900},$ $\alpha_3=\frac{128}{225},$ $\alpha_4=\frac{322+13\sqrt{70}}{900},$ $\alpha_5=\frac{322-13\sqrt{70}}{900}$

$$\int_{-1}^{1}\xi^k P_p(\xi)d\xi=\sum_{i=1}^{p}\alpha_i\xi_i^k P_p(\xi_i)=0. \qquad (5.103)$$

Останнє рівняння завжди буде вірним при будь-яких значеннях $\alpha_1,\alpha_2,\ldots,\alpha_p$ якщо:

$$P_p(\xi_i)=0,\quad i=1,2,\ldots,p, \qquad (5.104)$$

тобто, для досягнення максимальної точності квадратурної формули (5.92) у якості вузлів $\xi_1,\xi_2,\ldots,\xi_p$ достатньо взяти корені відповідного поліному Лежандра. Формула (5.92), де $\xi_1,\xi_2,\ldots,\xi_p$ корені поліномів Лежандра, а коефіцієнти $\alpha_1,\alpha_2,\ldots,\alpha_p$ визначаються з системи (5.99) називається *квадратурною формулою Гауса-Лежандра* [1], [3], [13], [14], [15], [4].

Корені поліному Лежандра можна знайти ітеративно за методом Ньютона:

$$\xi_i^{(k+1)}=\xi_i^{(k)}-\frac{P_p(\xi_i^{(k)})}{dP_p(\xi_i^{(k)})/d\xi}\quad \xi_i^{(0)}=\cos\left(\frac{\pi(4i-1)}{4p+2}\right),\quad i=1,2,\ldots,p. \qquad (5.105)$$

Похідну поліному можна знайти за допомогою безпосереднього диференціювання, або застосувавши співвідношення:

$$\frac{dP_p(\xi)}{d\xi}=\frac{p}{1-\xi^2}\left(P_{p-1}(\xi)-\xi P_p(\xi)\right). \tag{5.106}$$

Залишковий член квадратури Гауса-Лежандра для початкової функції $G(\xi)$ рівний:

$$\theta(G)=\frac{2^{2p+1}(p!)^4}{(2p+1)((2p)!)^3}\frac{d^{2p}G(\xi)}{d\xi^{2p}}. \tag{5.107}$$

Очевидно, коли $G(\xi)$ є поліномом степені $g=2p-1$, то похідна $d^{2p}G(\xi)/d\xi^{2p}=0$, і як наслідок, залишковий член $\theta(G)$ також рівний нулю. В *Таблиця 5.1* наведено корені для поліномів Лежандра перших п'яти степенів та відповідні їм вагові коефіцієнти в квадратурі Гауса-Лежандра.

Наприклад, необхідно знайти інтеграл від поліному п'ятої степені $G(\xi)=5\xi^5+2\xi^4+\xi^3+6\xi^2-4\xi+4$ на відрізку $-1\le\xi\le1$. Точне рішення буде рівне:

$$\begin{aligned}&\int_{-1}^{1}G(\xi)d\xi=\int_{-1}^{1}\left(5\xi^5+2\xi^4+\xi^3+6\xi^2-4\xi+4\right)d\xi=\\&=\left[\frac{5}{6}\xi^6+\frac{2}{5}\xi^5+\frac{1}{4}\xi^4+2\xi^3-2\xi^2+4\xi\right]_{\xi=1}-\\&-\left[\frac{5}{6}\xi^6+\frac{2}{5}\xi^5+\frac{1}{4}\xi^4+2\xi^3-2\xi^2+4\xi\right]_{\xi=-1}=\frac{64}{5}.\end{aligned} \tag{5.108}$$

Щоб отримати точну апроксимацію достатньо використати поліном Лежандра третьої степені. Обчисливши його корені та відповідні квадратурні коефіцієнти, або взявши вже обчислені з *Таблиця 5.1*, отримаємо:

$$\begin{aligned}&\xi_1=-\sqrt{\frac{3}{5}},\quad \xi_2=0,\quad \xi_3=\sqrt{\frac{3}{5}},\quad \alpha_1=\frac{5}{9},\quad \alpha_2=\frac{8}{9},\quad \alpha_3=\frac{5}{9},\\&G(\xi_1)=\frac{8\sqrt{15}}{25}+\frac{208}{25},\quad G(\xi_2)=4,\quad G(\xi_3)=\frac{-8\sqrt{15}}{25}+\frac{208}{25},\\&\int_{-1}^{1}G(\xi)d\xi=\alpha_1G(\xi_1)+\alpha_2G(\xi_2)+\alpha_3G(\xi_3)=\\&=\frac{5}{9}\left(\frac{8\sqrt{15}}{25}+\frac{208}{25}\right)+\frac{8}{9}4+\frac{5}{9}\left(\frac{-8\sqrt{15}}{25}+\frac{208}{25}\right)=\frac{208}{45}+\frac{32}{9}+\frac{208}{45}=\frac{64}{5}.\end{aligned} \tag{5.109}$$

Точне рішення можна отримати і для всіх поліномів степені нижчої п'ятої, наприклад для кубічного поліному:

$$Q(\xi)=5\xi^3+3\xi^2-2\xi+2,\quad \int_{-1}^{1}Q(\xi)d\xi=6. \tag{5.110}$$

При тому ж поліномі Лежандра:

$$Q(\xi_1)=\frac{-\sqrt{15}}{5}+\frac{19}{5},\quad Q(\xi_2)=2,\quad Q(\xi_3)=\frac{\sqrt{15}}{5}+\frac{19}{5},$$

$$\int_{-1}^{1} G(\xi)d\xi = \frac{5}{9}\left(\frac{-\sqrt{15}}{5}+\frac{19}{5}\right)+\frac{8}{9}2+\frac{5}{9}\left(\frac{\sqrt{15}}{5}+\frac{19}{5}\right)= \\ =\frac{19}{9}+\frac{16}{9}+\frac{19}{9}=\frac{54}{9}=6. \tag{5.111}$$

Крім використання квадратурної формули Гауса-Лежандра, дуже поширеним і в дечому простішим способом чисельного інтегрування є використання квадратурних формул *Ньютона-Котеса* [1], [13], [14], [15], [4], [5]. Вони відрізняються тим, що вузли $\xi_1, \xi_2, \ldots, \xi_p$ розміщуються рівномірно з деяким кроком, а максимальна степінь поліному, що може бути інтерпольований точно, рівна кількості вузлів, тобто $g = p$. Це є очевидним недоліком у порівнянні з попередньо описаними квадратурами. На базі квадратур Ньютона-Котеса виводяться такі знайомі методи чисельного інтегрування, як метод трапецій чи метод Сімпсона (метод парабол). Оскільки квадратури Гауса-Лежандра дають вищу точність апроксимації навіть при меншій кількості вузлів, у скінченно-елементних моделях використовуються саме вони, на відміну від зазначених квадратур Ньютона-Котеса [1], [3], [15], і саме тому ми не наводимо тут останні.

Для чисельного інтегрування можна використати і інші підходи. Наприклад, в деяких випадках може бути корисним апріорне фіксування деяких вузлів ξ_i з подальшим знаходженням наступних. У такому випадку, при заданій кількості вузлів степінь поліному, що може бути апроксимований точно, була б не вищою, ніж для відповідної квадратури Гауса-Лежандра і не нижчою, ніж для відповідної квадратури Ньютона-Котеса, тобто $2p-1 \geq g \geq p$. Зокрема, іноді корисно фіксувати вузли в граничних точках області, коли $\xi_p = -\xi_1 = 1$, але зберегти вільність вибору внутрішніх вузлів. Такий підхід називають *квадратурами Гауса-Лобатто* [1], [14]. Ця квадратура є точною для поліномів степені $g = 2p-3$. Квадратурна формула записується як:

$$\int_{-1}^{1} G(\xi)d\xi = \frac{2p}{p(p-1)}G(1)+\sum_{i=2}^{p-1}\alpha_i G(\xi_i)+\frac{2p}{p(p-1)}G(-1)+\theta(G). \tag{5.112}$$

Вільними вузлами ξ_i є $(i-1)$ корені похідних Лежандревих поліномів $dP_p(\xi_i^{(k)})/d\xi$ при $i = 2, 3, \ldots, p-1$. Відповідні вагові коефіцієнти визначаються як:

$$\alpha_i = \frac{2}{p\left(p-1\right)\left(P_{p-1}(\xi_i)\right)^2}. \tag{5.113}$$

Залишковий член рівний:

$$\theta(G) = \frac{-p(p-1)^3 2^{2p-1}((p-2)!)^4}{(2p-1)((2p-2)!)^3}\frac{d^{2p-2}G(\xi)}{d\xi^{2p-2}}. \tag{5.114}$$

Очевидно, коли $G(\xi)$ є поліномом степені $g = 2p-3$, то похідна

$d^{2p-2}G(\xi)/d\xi^{2p-2}=0$, і як наслідок, залишковий член $\theta(G)$ також рівний нулю.

У *Таблиця 5.2* наведено обчислені корені та відповідні їм вагові коефіцієнти квадратури Гауса-Лобатто для трьох, чотирьох та п'яти вузлів. Зауважимо, що при трьох вузлах отримана квадратура точно співпадає з квадратурою Ньютона-Котеса для трьох рівновіддалених вузлів, що як ми знаємо, дає точний результат для поліномів третьої степені. Ця трьох вузлова квадратура також відома як формула Сімпсона або метод парабол.

Таблиця 5.2

Корені та відповідні їм вагові коефіцієнти квадратури Гауса-Лобатто для трьох, чотирьох та п'яти вузлів

Кількість вузлів	Корені	Вагові коефіцієнти
3	$\xi_1=-1,\ \xi_2=0,\ \xi_3=1$	$\alpha_1=1/3,\ \alpha_2=4/3,\ \alpha_3=1/3$
4	$\xi_1=-1,\ \xi_2=-\sqrt{1/5},\ \xi_3=\sqrt{1/5},$ $\xi_4=1$	$\alpha_1=1/6,\ \alpha_2=5/6,\ \alpha_3=5/6,\ \alpha_4=1/6$
5	$\xi_1=-1,\ \xi_2=-\sqrt{3/7},\ \xi_3=0,$ $\xi_4=\sqrt{3/7},\ \xi_5=1$	$\alpha_1=1/10,\ \alpha_2=49/90,\ \alpha_3=32/45,$ $\alpha_4=49/90,\ \alpha_5=1/10$

Цікавий читач може знайти в літературі й інші методи чисельного інтегрування, наприклад квадратури Гауса-Чебишова, які будуються на основі тригонометричних функцій, чи квадратури Гауса-Радо, де фіксованим є тільки перший вузол. Але як вже зазначалося, при побудові скінченно-елементних моделей найчастіше використовуються квадратури Гауса-Лежандра, оскільки в них необхідно проводити найменшу кількість обчислень для точного знаходження інтегралів від поліномів $g=2p-1$, де p – кількість вузлів інтегрування.

Розглянемо тепер кубатурні формули, тобто формули чисельного інтегрування подвійних інтегралів, які виникають у двовимірних задачах. Знову ж таки, зупинимося на формулах Гуса-Лежандра, як таких, що дають найточніші результати при мінімальній кількості обчислень.

У двовимірному випадку складна підінтегральна функція $G(\xi_1,\xi_2)$ залежить від двох локальних нормованих координат, які визначені на квадраті $-1\le\xi_1,\xi_2\le 1$. Найпростішим способом виведення кубатурної формули буде застосування поетапного чисельного інтегрування окремо по кожній з координат. Тобто спочатку знайти (передбачається, що кількість вузлів інтегрування по кожній з координат буде однаковою):

$$\int_{-1}^{1} G(\xi_1,\xi_2)d\xi_1 \approx \sum_{i=1}^{p}\alpha_i G(\xi_{i,1},\xi_2), \qquad (5.115)$$

а потім:

$$\int_{-1}^{1}\left(\sum_{i=1}^{p}\alpha_i G(\xi_{i,1},\xi_2)+\theta_1(G)\right)d\xi_2 \approx \sum_{j=1}^{p}\alpha_j\left(\sum_{i=1}^{p}\alpha_i G(\xi_{i,1},\xi_{j,2})\right), \qquad (5.116)$$

і в результаті отримати:

$$\iint_{\Omega} G(x_1,x_2)dx_1dx_2 = \int_{-1}^{1}\int_{-1}^{1} G(\xi_1,\xi_2)\left|[\mathbf{Jac_L r}]\right| d\xi_1 d\xi_2 \approx$$
$$\approx \sum_{i=1}^{p}\sum_{j=1}^{p}\left|[\mathbf{Jac_L r}]\right|\alpha_i\alpha_j G(\xi_{i,1},\xi_{j,2}), \tag{5.117}$$

де $(\xi_{i,1},\xi_{j,2})$ – координати вузлів, точне положення яких визначається типом формули інтегрування. Використовуючи квадратури Гауса-Лежандра, це будуть корені поліномів Лежандра.

Якщо формули інтегрування окремо по ξ_1 та ξ_2 точні для поліному степені g, то кубатурна формула (5.117) буде давати точне значення для всіх виразів виду $\xi_1^{p_1}\xi_2^{p_2}$, де $p_1, p_2 \le g$. Стандартні квадратурні правила Гауса-Лежандра такого типу зображені на *Рис. 5.25*.

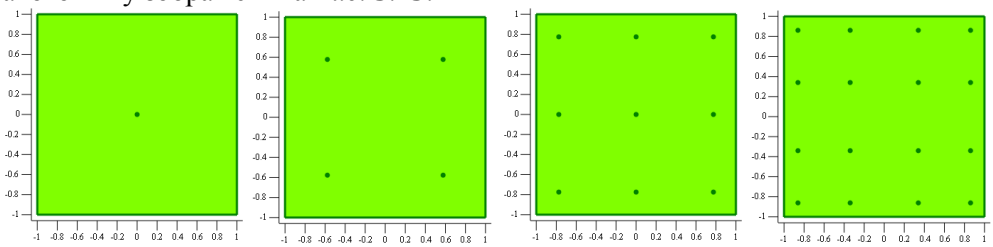


Рис. 5.25 Розміщення вузлів кубатур Гауса-Лежандра на чотирикутниках

Очевидне узагальнення квадратур на тривимірні випадки для обчислення інтегралів по кубу $-1 \le \xi_1,\xi_2,\xi_3 \le 1$, приводить до співвідношення виду:

$$\iiint_{\Omega} G(x_1,x_2,x_3)dx_1dx_2dx_3 = \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} G(\xi_1,\xi_2,\xi_3)\left|[\mathbf{Jac_L r}]\right| d\xi_1 d\xi_2 d\xi_3 \approx$$
$$\approx \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\left|[\mathbf{Jac_L r}]\right|\alpha_i\alpha_j\alpha_k G(\xi_{i,1},\xi_{j,2},\xi_{k,3}). \tag{5.118}$$

Цей процес можна продовжити для довільної кількості вимірів. Отримані формули називаються *мультиплікативними* [1] і широко застосовуються при побудові двовимірних і тривимірних скінченно-елементних моделей.

Описаний процес чисельного інтегрування буде точним і для членів, що виникають додатково до повних поліномів степені p від незалежних змінних ξ_1,ξ_2,ξ_3 (див. трикутних Паскаля *Рис. 5.17*). Як наслідок, можна отримати формули чисельного інтегрування, що будуть точними для повних поліномів заданої степені, але потребують меншої кількості вузлів, ніж описані мультиплікативні формули (вперше такі формули були запропоновані в 1971 році[1]). Їх аналіз виходить за рамки нашого розгляду, тому цікавий читач може звернутися наприклад до [17] чи [18].

[1] Irons B. – Quadrature rules for brick based finite elements // Int. Journ. Num. Meth. Eng., 3:293-294, 1971.

Розглянемо кубатурні формули для симплексів, тобто для трикутників чи тетраедрів. Як і раніше, складна підінтегральна функція на симплексах виражається в його барицентричних координатах. Наприклад для трикутника, знову здійснивши проекцію на універсальний симплекс елемент з вершинами $(0,0)$, $(0,1)$, $(1,0)$, межі інтегрування стануть змінними (див. наприклад (5.79) або (5.87)):

$$\iint\limits_{\Omega} G(x_1,x_2)dx_1dx_2 = \int\limits_0^1 \int\limits_0^{1-L_1} G(L_1,L_2,L_3)\left|[\mathbf{Jac_L r}]\right| dL_2dL_1, \quad L_3 = 1-L_1-L_2. \tag{5.119}$$

Щоб застосувати попередньо описаний мультиплікативний підхід (вперше це було зроблено в 1968 році[1] на основі квадратур Гауса-Радо [15]), здійснимо заміну змінних, яка відповідає проекції трикутника на квадрат з одиничною стороною:

$$\begin{aligned} &L_1 = u_1, \\ &L_2 = (1-u_1)u_2, \end{aligned} \quad \left|[\mathbf{Jac_u L}]\right| = \left\| \begin{bmatrix} \dfrac{\partial L_1}{\partial u_1} & \dfrac{\partial L_1}{\partial u_2} \\ \dfrac{\partial L_2}{\partial u_1} & \dfrac{\partial L_2}{\partial u_2} \end{bmatrix} \right\| = \left\| \begin{bmatrix} 1 & 0 \\ -u_2 & 1-u_1 \end{bmatrix} \right\| = 1-u_1. \tag{5.120}$$

Це дає змогу замінити межі інтегрування:

$$\begin{aligned} \iint\limits_{\Omega} G(x_1,x_2)dx_1dx_2 &= \int\limits_0^1 \int\limits_0^{1-L_1} G(L_1,L_2)\left|[\mathbf{Jac_L r}]\right| dL_2dL_1 = \\ &= \int\limits_0^1 \int\limits_0^1 \Big(G(u_1,(1-u_1)u_2)\cdot(1-u_1)\cdot\left|[\mathbf{Jac_L r}]\right| \Big) du_1du_2. \end{aligned} \tag{5.121}$$

Після цього здійснимо заміну змінних, що відповідає проекції отриманого квадрата в квадрат з межами $-1;1$:

$$\begin{aligned} &u_1 = \frac{1+\xi_1}{2}, \\ &u_2 = \frac{1+\xi_2}{2}, \end{aligned} \quad \left|[\mathbf{Jac}_{\xi}\mathbf{u}]\right| = \left\| \begin{bmatrix} \dfrac{\partial u_1}{\partial \xi_1} & \dfrac{\partial u_1}{\partial \xi_2} \\ \dfrac{\partial u_2}{\partial \xi_1} & \dfrac{\partial u_2}{\partial \xi_2} \end{bmatrix} \right\| = \left\| \begin{bmatrix} \dfrac{1}{2} & 0 \\ 0 & \dfrac{1}{2} \end{bmatrix} \right\| = \frac{1}{4}. \tag{5.122}$$

Звідки отримаємо межі інтегрування:

$$\begin{aligned} &\iint\limits_{\Omega} G(x_1,x_2)dx_1dx_2 = \int\limits_0^1 \int\limits_0^{1-L_1} G(L_1,L_2)\left|[\mathbf{Jac_L r}]\right| dL_2dL_1 = \\ &= \int\limits_0^1 \int\limits_0^1 \Big(G(u_1,(1-u_1)u_2)\cdot(1-u_1)\cdot\left|[\mathbf{Jac_L r}]\right| \Big) du_1du_2 = \\ &= \int\limits_{-1}^1 \int\limits_{-1}^1 \left(G\left(\frac{1+\xi_1}{2},\frac{(1-\xi_1)(1+\xi_2)}{4}\right)\cdot\left(\frac{1-\xi_1}{8}\right)\cdot\left|[\mathbf{Jac_L r}]\right| \right) d\xi_1 d\xi_2, \end{aligned} \tag{5.123}$$

[1] Anderson R., Irons B., Zienkiewicz O. – Vibration and Stability of Plates Using Finite Elements // Int. Jour. Solids Struct., 4:1031-1055, 1968.

придатні для застосування кубатурної формули:

$$\iint\limits_{\Omega} G(x_1,x_2)dx_1dx_2 \approx \sum_{i=1}^{p}\sum_{j=1}^{p}\left(\frac{1-\xi_{i,1}}{8}\right)\left|[\mathbf{Jac_L r}]\right|\alpha_i\alpha_j G\left(\frac{1+\xi_{i,1}}{2},\frac{(1-\xi_{i,1})(1+\xi_{j,2})}{4}\right), \quad (5.124)$$

де координати вузлів $\xi_{i,1},\xi_{j,2}$ та вагові коефіцієнти α_i,α_j, у випадку використання квадратур Гауса-Лежандра, можна знайти з системи (5.99), вони наведені в *Таблиця 5.3*.

Таблиця 5.3

Вузли та відповідні їм вагові коефіцієнти кубатури Гауса-Лежандра на трикутнику

Кількість вузлів	Вузли $\left(L_1=\frac{1+\xi_1}{2}, L_2=\frac{(1-\xi_1)(1+\xi_2)}{4}\right)$	Вагові коефіцієнти $\frac{1-\xi_1}{8}\alpha_i\alpha_j$
1	$(1/2,\ 1/4)$	$25/648$
4	$\left((3-\sqrt{3})/3,\ 1/6\right)$ $\left((3-\sqrt{3})/3,\ (3+\sqrt{3})/3\right)$ $\left((3+\sqrt{3})/3,\ (3-\sqrt{3})/3\right)$ $\left((3+\sqrt{3})/3,\ 1/6\right)$	$(3+\sqrt{3})/24$ $(3+\sqrt{3})/24$ $(3-\sqrt{3})/24$ $(3-\sqrt{3})/24$
9	$\left((5-\sqrt{15})/10,\ 1/10\right)$ $\left((5-\sqrt{15})/10,\ (5+\sqrt{15})/20\right)$ $\left((5-\sqrt{15})/10,\ (4+\sqrt{15})/10\right)$ $\left(1/2,\ (5-\sqrt{15})/20\right)$ $(1/2,\ 1/4)$ $\left(1/2,\ (5+\sqrt{15})/20\right)$ $\left((5+\sqrt{15})/10,\ (4-\sqrt{15})/10\right)$ $\left((5+\sqrt{15})/10,\ (5-\sqrt{15})/20\right)$ $\left((5+\sqrt{15})/10,\ 1/10\right)$	$(25+5\sqrt{15})/648$ $(5+\sqrt{15})/81$ $(25+5\sqrt{15})/648$ $5/81$ $8/81$ $5/81$ $(25-5\sqrt{15})/648$ $(5-\sqrt{15})/81$ $(25-5\sqrt{15})/648$

В отриманій кубатурній формулі розміщення вузлів інтегрування буде не рівномірним і не симетричним (*Рис. 5.26*). Це призводить до різної точності інтегрування по напрямках кожної з барицентричних координат L_1, L_2 та L_3. Цей недолік можна обійти, якщо спробувати вивести симетричні кубатурні формули, тобто інваріантні формули, де при циклічній перенумерації вузлів, що змінює порядок барицентричних координат, результат б не змінювався. Вимога симетрії є очевидною і математично, оскільки згідно теорії, інтеграли повинні залишатися незмінними при застосуванні будь-яких афінних (лінійних обертань, переносів, деформацій, тощо) перетворень області в саму себе.

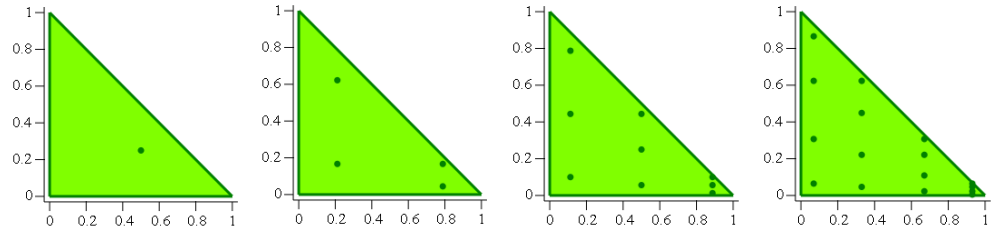


Рис. 5.26 Розміщення вузлів кубатур Гауса-Лежандра на трикутниках

Необхідно вивести кубатурну формулу типу:

$$\iint_{\Omega} G(x_1,x_2)dx_1dx_2 = \int_0^1 \int_0^{1-L_1} G(L_1,L_2,L_3)\left|[\mathbf{Jac_L r}]\right| dL_1dL_2 \approx$$
$$\approx \sum_{i=1}^{p} \left|[\mathbf{Jac_L r}]\right| \alpha_i G(L_{i,1},L_{i,2},L_{i,3}), \quad L_3 = 1-L_1-L_2, \tag{5.125}$$

де p тепер позначає кількість вузлів інтегрування $(L_{i,1},L_{i,2},L_{i,3})$, при чому ця кількість повинна бути мінімально можливою для точного інтегрування поліному $G(L_1,L_2,L_3)$ деякої степені g. Для визначення вузлів інтегрування вже не можна використати корені поліномів Лежандра, а для визначення вагових коефіцієнтів систему рівнянь (5.99).

Нагадаємо, що в загальному випадку, кубатури типу Гауса деякої степені точності g визначаються як кубатури (5.125), що точні для всіх лінійних комбінацій виразів $L_1^iL_2^j$ (L_3 не враховується, оскільки це комбінація з L_1 та L_2) де $0 \le i,j \le g$, і відповідно для всіх поліномів $G(L_1,L_2,L_3)=G(L_1,L_2)$ степені g. Перелік таких виразів можна вивести на основі трикутника Паскаля (*Рис. 5.17*)[1]. Знову використовуючи аналітичні формули для інтегрування в барицентричних координатах (5.79) отримаємо:

---

[1] Оболонкою span{**X**} деякої підмножини **X** множини **V**, $\mathbf{X} \subset \mathbf{V}$ називають перетин всіх підпросторів **V**, що містять **X**. Іншими словами оболонка span{**X**} складається з усіх можливих комбінацій елементів **X**. Кубатури типу Гауса деякої степені точності g визначаються як:

$$\iint_{\Omega} G(\xi_1,\xi_2)d\xi_1d\xi_2 \approx \frac{1}{2}\sum_{i=1}^{p} \alpha_i G(\xi_{i,1},\xi_{j,2}),$$

є точними для всіх поліномів $G(\xi_1,\xi_2)$, що містяться у повному поліноміальному (функціональному) просторі степені g, що є оболонкою підмножини поліномів двовимірного простору типу $\xi_1^i\xi_2^j$ [9], [19]:

$$\forall G(\xi_1,\xi_2) \in \mathbf{P}_g(\xi_1,\xi_2), \quad \mathbf{P}_g(\xi_1,\xi_2) = \text{span}\{\xi^i\xi^j, \quad 0 \le i,j \le g\}.$$

Наприклад:

$$\mathbf{P}_1(\xi_1,\xi_2) = \text{span}\{1 \quad \xi_1 \quad \xi_2\},$$
$$\mathbf{P}_2(\xi_1,\xi_2) = \text{span}\{1 \quad \xi_1 \quad \xi_2 \quad \xi_1^2 \quad \xi_1\xi_2 \quad \xi_2^2\}.$$

Очевидно, що елементи типу $\xi_1^i\xi_2^j$ де $0 \le i,j \le g$ для підмножин також можна безпосередньо вивести на основі трикутника Паскаля.

$$\iint_{\Omega} L_1^i L_2^j dL_1 dL_2 = \frac{i!\,j!}{(i+j+2)!},$$

$$\iint_{\Omega} \{1 \quad L_1 \quad L_2 \quad L_1^2 \quad L_1L_2 \quad L_2^2 \quad L_1^3 \quad L_1^2L_2 \quad L_1L_2^2 \quad L_2^3\} dL_1 dL_2 = \quad (5.126)$$

$$= \left\{\frac{1}{2} \quad \frac{1}{6} \quad \frac{1}{6} \quad \frac{1}{12} \quad \frac{1}{24} \quad \frac{1}{12} \quad \frac{1}{20} \quad \frac{1}{60} \quad \frac{1}{60} \quad \frac{1}{20}\right\}.$$

і так далі, для виразів вищих порядків.

При $g=1$, за визначенням, кубатура повинна бути точною для поліномів $G(L_1,L_2)=\{1 \quad L_1 \quad L_2\}$, звідки $p=g=1$ та:

$$\begin{aligned} G(L_1,L_2)=1 \quad &\Rightarrow \quad \frac{1}{2}=\sum_{i=1}^{1}\alpha_i \quad &=\alpha_1, \\ G(L_1,L_2)=L_1 \quad &\Rightarrow \quad \frac{1}{6}=\sum_{i=1}^{1}\alpha_i L_{i,1} \quad &=\alpha_1 L_{1,1}, \\ G(L_1,L_2)=L_2 \quad &\Rightarrow \quad \frac{1}{6}=\sum_{i=1}^{1}\alpha_i L_{i,2} \quad &=\alpha_1 L_{1,2}. \end{aligned} \qquad (5.127)$$

Легко побачити, що $\alpha_1=1/2$, $L_{1,1}=1/3$, $L_{1,2}=1/3$ та $L_{1,3}=1-L_{1,1}-L_{1,2}=1/3$, тобто вузлом інтегрування є барицентр, а ваговим коефіцієнтом – площа трикутника (*Таблиця 5.4.1*). Так ми отримали Гаусову кубатурну формулу для трикутника, точну для поліномів першого порядку:

$$\int_0^1 \int_0^{1-L_1} G(L_1,L_2,L_3)dL_1dL_2 = \frac{1}{2}G\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right). \qquad (5.128)$$

Перед тим як продовжити, зауважимо, що оскільки i-ий вузол інтегрування визначається барицентричними координатами $(L_{i,1},L_{i,2},L_{i,3})$ зв'язаних між собою відношенням $L_{i,1}+L_{i,2}+L_{i,3}=1$, то всі вузли, отримані перестановкою цих координат, для дотримання симетрії повинні мати однакові вагові коефіцієнти α_i з кубатурної формули (5.125). Наприклад, якщо всі три барицентричні координати вузла є різними – отримаємо шість симетричних вузлів: $(L_{i,1},L_{i,2},L_{i,3})$, $(L_{i,1},L_{i,3},L_{i,2})$, $(L_{i,2},L_{i,1},L_{i,3})$, $(L_{i,2},L_{i,3},L_{i,1})$, $(L_{i,3},L_{i,1},L_{i,2})$, $(L_{i,3},L_{i,2},L_{i,1})$. Отриману множину вузлів називають зіркою[1] і позначають S_{111}. Якщо дві барицентричні координати вузла є рівними, то множина зводиться до трьох різних вузлів, а відповідна зірка позначається як S_{21}. Нарешті, якщо всі барицентричні координати є рівними, тобто вузол є барицентром (центроїдом), то відповідна зірка позначається як S_3. Комбінації таких зірок повністю визначають набори вузлів інтегрування для кубатур типу (5.125): зірки S_3, S_{21}

[1] Назва *simple point star* або просто *star* (зірка) S_k взята з теорії графів, де вона позначає топологію повного дводольного графу $K_{1,k}$: дерева з одним внутрішнім вузлом і k листками.

та S_{111} містять 1, 3 та 6 вузлів відповідно – як наслідок, симетрична кубатурна формула повинна містити $i+3j+6k$ вузлів, де i, j та k є невід'ємними цілими числами, i рівне 0 або 1 [19].

При $g=2$, за визначенням, кубатура повинна бути точною для поліномів $G(L_1,L_2)=\{1 \quad L_1 \quad L_2 \quad L_1^2 \quad L_1L_2 \quad L_2^2\}$. Кількість вузлів інтегрування p вже не може бути рівною одиниці. При двох вузлах інтегрування отримана формула буде не симетричною, тому мінімальна необхідна кількість вузлів рівна трьом[1], що відповідає S_{21} зірці:

$$\begin{aligned}
G(L_1,L_2)=1 \quad &\Rightarrow \quad \frac{1}{2}=\sum_{i=1}^{3}\alpha_i \quad &&=\alpha_1 \quad &&+\alpha_2 \quad &&+\alpha_3,\\
G(L_1,L_2)=L_1 \quad &\Rightarrow \quad \frac{1}{6}=\sum_{i=1}^{3}\alpha_i L_{i,1} \quad &&=\alpha_1 L_{1,1} \quad &&+\alpha_2 L_{2,1} \quad &&+\alpha_3 L_{3,1},\\
G(L_1,L_2)=L_2 \quad &\Rightarrow \quad \frac{1}{6}=\sum_{i=1}^{3}\alpha_i L_{i,2} \quad &&=\alpha_1 L_{1,2} \quad &&+\alpha_2 L_{2,2} \quad &&+\alpha_3 L_{3,2},\\
G(L_1,L_2)=L_1^2 \quad &\Rightarrow \quad \frac{1}{12}=\sum_{i=1}^{3}\alpha_i L_{i,1}^2 \quad &&=\alpha_1 L_{1,1}^2 \quad &&+\alpha_2 L_{2,1}^2 \quad &&+\alpha_3 L_{3,1}^2,\\
G(L_1,L_2)=L_1L_2 \quad &\Rightarrow \quad \frac{1}{24}=\sum_{i=1}^{3}\alpha_i L_{i,1}L_{i,2} \quad &&=\alpha_1 L_{1,1}L_{1,2} \quad &&+\alpha_2 L_{2,1}L_{2,2} \quad &&+\alpha_3 L_{3,1}L_{3,2},\\
G(L_1,L_2)=L_2^2 \quad &\Rightarrow \quad \frac{1}{12}=\sum_{i=1}^{3}\alpha_i L_{i,2}^2 \quad &&=\alpha_1 L_{1,2}^2 \quad &&+\alpha_2 L_{2,2}^2 \quad &&+\alpha_3 L_{3,2}^2.
\end{aligned} \qquad (5.129)$$

Отримано шість рівнянь і дев'ять невідомих, тому рішення не буде єдиним. Щоб отримати афінно-інваріантну формулу, тобто формулу симетричну відносно барицентру, вузли інтегрування можна визначити як $r\mathbf{V}_i+(1-r)\mathbf{C}$, де r – коефіцієнт, що потрібно знайти, $\mathbf{V}_i$ – вершини трикутника в барицентричних координатах, тобто $(1,0,0)$, $(0,1,0)$ та $(0,0,1)$, а $\mathbf{C}=(\mathbf{V}_1+\mathbf{V}_2+\mathbf{V}_3)/3$ – барицентр. Ваговими коефіцієнтами для кожного з трьох симетричних вузлів інтегрування буде третина площі трикутника, а коефіцієнт r рівний $\pm 1/2$. У літературі [1], [16], [3], [15] найчастіше використовують значення $r=-1/2$, при якому (*Таблиця 5.4.2*):

$$(L_1,L_2,L_3)=\left\{\left(\frac{1}{2},\frac{1}{2},0\right) \quad \left(0,\frac{1}{2},\frac{1}{2}\right) \quad \left(\frac{1}{2},0,\frac{1}{2}\right)\right\}, \quad \alpha=\left\{\frac{1}{6} \quad \frac{1}{6} \quad \frac{1}{6}\right\}, \qquad (5.130)$$

і відповідна кубатурна формула для трикутника, що точна для поліномів другого порядку:

$$\int_0^1\int_0^{1-L_1} G(L_1,L_2,L_3)\,dL_1dL_2=\frac{1}{6}\left(G\left(\frac{1}{2},\frac{1}{2},0\right)+G\left(0,\frac{1}{2},\frac{1}{2}\right)+G\left(\frac{1}{2},0,\frac{1}{2}\right)\right). \qquad (5.131)$$

При $g=3$, за визначенням, кубатура повинна бути точною для поліномів

[1] Dunavant D. – High Degree Efficient Symmetrical Gaussian Quadrature Rules for Triangle // Int. Jour. for Numerical Methods in Engineering, 21:1129-1148, 1985.

$G(L_1,L_2)=\{1 \quad L_1 \quad L_2 \quad L_1^2 \quad L_1L_2 \quad L_2^2 \quad L_1^3 \quad L_1^2L_2 \quad L_1L_2^2 \quad L_2^3\}$. Для отримання точного рішення трьох вузлів інтегрування буде недостатньо, тому $p=4$, що відповідає комбінації зірок S_3 та S_{21}. Побудувавши систему рівнянь, аналогічну попереднім, отримаємо десять рівнянь і дванадцять невідомих – рішення знову не буде єдиним. Зірка S_3 відповідає барицентру $\mathbf{C}$, зірка S_{21} знову шукається як $r\mathbf{V}_i+(1-r)\mathbf{C}$, звідки $r=1/2\pm1/10=2/5;3/5$. Наприклад, при $r=3/5$ отримаємо (*Таблиця 5.4.4*):

$$(L_1,L_2,L_3)=\left\{\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right) \quad \left(\frac{11}{15},\frac{2}{15},\frac{2}{15}\right) \quad \left(\frac{2}{15},\frac{11}{15},\frac{2}{15}\right) \quad \left(\frac{2}{15},\frac{2}{15},\frac{11}{15}\right)\right\},$$
$$\alpha=\left\{\frac{-27}{96} \quad \frac{25}{96} \quad \frac{25}{96} \quad \frac{25}{96}\right\}, \tag{5.132}$$

і відповідна кубатурна формула для трикутника, що точна для поліномів третього порядку:

$$\int_0^1\int_0^{1-L_1} G(L_1,L_2,L_3)dL_1dL_2=\frac{-27}{96}G\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)+$$
$$+\frac{25}{96}G\left(\frac{11}{15},\frac{2}{15},\frac{2}{15}\right)+\frac{25}{96}G\left(\frac{2}{15},\frac{11}{15},\frac{2}{15}\right)+\frac{25}{96}G\left(\frac{2}{15},\frac{2}{15},\frac{11}{15}\right). \tag{5.133}$$

При $g=4$, кількість вузлів p рівна семи, що відповідає комбінації зірки S_3 та двох S_{21}. Вузли інтегрування можна знайти як барицентр $\mathbf{C}$, відношення $r\mathbf{V}_i+(1-r)\mathbf{C}$ для першої S_{21} та $s\mathbf{V}_i+(1-s)\mathbf{C}$ для другої S_{21}. Звідки (*Таблиця 5.4.7*) $r=(1+\sqrt{15})/7$ з ваговими коефіцієнтами $\alpha^{(r)}=(155-\sqrt{15})/2400$, $s=(1-\sqrt{15})/7$ з ваговими коефіцієнтами $\alpha^{(s)}=(155+\sqrt{15})/2400$ і ваговий коефіцієнт для барицентру $\mathbf{C}$ рівний $9/80$. Цікаво, що отримана кубатурна формула є точною і для поліномів п'ятого порядку [1], [3], [15].

З іншої сторони, використовуючи семи вузлову формулу, можна отримати апроксимацію, що точна тільки для поліномів третього порядку. Набір вузлів цієї формули складається з барицентру $\mathbf{C}$, вузлів з формули для $g=2$ (5.130) та вузлів трикутника $\mathbf{V}_i$ (*Таблиця 5.4.6*). Таку формулу часто використовують замість (5.132), оскільки в ній відсутні від'ємні вагові коефіцієнти [15]. Вперше описані тут симетричні формули чисельного інтегрування на симплексах були виведені у 1956 році[1].

Аналізуючи зірки, на основі яких виведені вищеописані симетричні формули чисельного інтегрування, можна зауважити, що: зірка $S_3(1/3)$ завжди є барицентром; зірка $S_{21}(a)$ завжди має три різні вузли, які можна визначити як

[1] Hammer P., Marlowe O., Stroud A. – Numerical Integration Over Simpexes and Cones // Math. Tables Aids Comp., 10:130-137, 1956.

$(a,a,1-2a)$, $(1-2a,a,a)$ та $(a,1-2a,a)$; зірка $S_{111}(a,b)$ завжди має шість різних вузлів, що визначаються перестановкою з $(a,b,1-a-b)$.

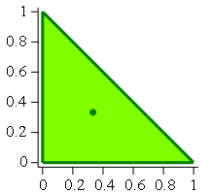
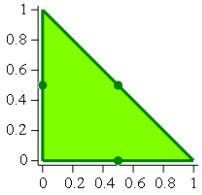
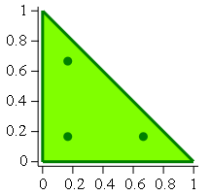
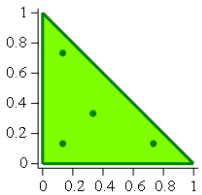
Очевидно, що описані двовимірні формули можна аналогічно розширити і на довільну кількість вимірів. Зокрема мультиплікативну формулу Гауса-Лежандра (5.124) можна розширити на тетраедр ввівши проекцію з заміною координат:

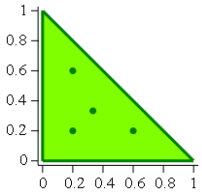
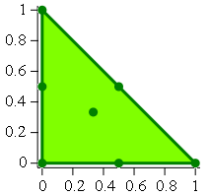
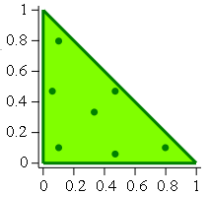
$$L_1 = u_1u_2u_3, \quad L_2 = u_1u_2(1-u_3), \quad L_3 = u_1(1-u_2),$$

$$\left|[\mathbf{Jac_u L}]\right| = \left\|\begin{bmatrix} \frac{\partial L_1}{\partial u_1} & \frac{\partial L_1}{\partial u_2} & \frac{\partial L_1}{\partial u_3} \\ \frac{\partial L_2}{\partial u_1} & \frac{\partial L_2}{\partial u_2} & \frac{\partial L_2}{\partial u_3} \\ \frac{\partial L_3}{\partial u_1} & \frac{\partial L_3}{\partial u_2} & \frac{\partial L_3}{\partial u_3} \end{bmatrix}\right\| = \left\|\begin{bmatrix} u_2u_3 & u_1u_3 & u_1u_2 \\ u_2(1-u_3) & u_1(1-u_3) & -u_1u_2 \\ 1-u_2 & -u_1 & 0 \end{bmatrix}\right\| = -u_1^2u_2, \quad (5.134)$$

Таблиця 5.4

Вузли та відповідні їм вагові коефіцієнти симетричної кубатури Гауса на трикутнику

№	g	p	Розміщення вузлів	Координати вузлів (L_1,L_2,L_3)	Коефіцієнти α
1	1	1		(1/3 1/3 1/3)	1/2
2	2	3		(1/2 1/2 0) (0 1/2 1/2) (1/2 0 1/2)	1/6 1/6 1/6
3	2	3		(1/6 1/6 2/3) (2/3 1/6 1/6) (1/6 2/3 1/6)	1/6 1/6 1/6
4	3	4		(1/3 1/3 1/3) (2/15 11/15 2/15) (2/15 2/15 11/15) (11/15 2/15 2/15)	–27/96 25/96 25/96 25/96

5	3	4		$(1/3 \quad 1/3 \quad 1/3)$ $(1/5 \quad 3/5 \quad 1/5)$ $(1/5 \quad 1/5 \quad 3/5)$ $(3/5 \quad 1/5 \quad 1/5)$	$-27/96$ $25/96$ $25/96$ $25/96$
6	3	7		$(1/3 \quad 1/3 \quad 1/3)$ $(1/2 \quad 1/2 \quad 0)$ $(0 \quad 1/2 \quad 1/2)$ $(1/2 \quad 0 \quad 1/2)$ $(1 \quad 0 \quad 0)$ $(0 \quad 1 \quad 0)$ $(0 \quad 0 \quad 1)$	$27/120$ $8/120$ $8/120$ $8/120$ $3/120$ $3/120$ $3/120$
7	5	7		$(1/3 \quad 1/3 \quad 1/3)$ $\left((9+2\sqrt{15})/21 \quad (6-\sqrt{15})/21 \quad (6-\sqrt{15})/21\right)$ $\left((6-\sqrt{15})/21 \quad (9+2\sqrt{15})/21 \quad (6-\sqrt{15})/21\right)$ $\left((6-\sqrt{15})/21 \quad (6-\sqrt{15})/21 \quad (9+2\sqrt{15})/21\right)$ $\left((9-2\sqrt{15})/21 \quad (6+\sqrt{15})/21 \quad (6+\sqrt{15})/21\right)$ $\left((6+\sqrt{15})/21 \quad (9-2\sqrt{15})/21 \quad (6+\sqrt{15})/21\right)$ $\left((6+\sqrt{15})/21 \quad (6+\sqrt{15})/21 \quad (9-2\sqrt{15})/21\right)$	$9/80$ $(155-\sqrt{15})/2400$ $(155-\sqrt{15})/2400$ $(155-\sqrt{15})/2400$ $(155+\sqrt{15})/2400$ $(155+\sqrt{15})/2400$ $(155+\sqrt{15})/2400$

звідки отримаємо:

$$\iiint\limits_{\Omega} G(x_1,x_2,x_3)dx_1dx_2dx_3 = \int\limits_0^1 \int\limits_0^{1-L_1} \int\limits_0^{1-L_1-L_3} G(L_1,L_2,L_3)\left|[\mathbf{Jac}_{\mathbf{L}}\mathbf{r}]\right| dL_3dL_2dL_1 = \tag{5.135}$$

$$= \int\limits_0^1\int\limits_0^1\int\limits_0^1 \Big(G(u_1u_2u_3,\quad u_1u_2(1-u_3),\quad u_1(1-u_2))\cdot(u_1^2u_2)\cdot\left|[\mathbf{Jac}_{\mathbf{L}}\mathbf{r}]\right|\Big)du_1du_2du_3.$$

Після цього здійснимо заміну змінних, що відповідає проекції отриманого куба в куб з межами $-1;1$:

$$\begin{aligned} u_1 &= (1+\xi_1)/2, \\ u_2 &= (1+\xi_2)/2, \\ u_3 &= (1+\xi_3)/2, \end{aligned} \quad \left|[\mathbf{Jac}_{\xi}\mathbf{u}]\right| = \left\|\begin{bmatrix} \frac{\partial L_1}{\partial u_1} & \frac{\partial L_1}{\partial u_2} & \frac{\partial L_1}{\partial u_3} \\ \frac{\partial L_2}{\partial u_1} & \frac{\partial L_2}{\partial u_2} & \frac{\partial L_2}{\partial u_3} \\ \frac{\partial L_3}{\partial u_1} & \frac{\partial L_3}{\partial u_2} & \frac{\partial L_3}{\partial u_3} \end{bmatrix}\right\| = \left\|\begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}\right\| = \frac{1}{8}. \tag{5.136}$$

отримаємо межі інтегрування:

$$\iiint\limits_{\Omega} G(x_1,x_2,x_3)dx_1dx_2dx_3 =$$

$$= \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}\left(G\left(\frac{(1-\xi_1)(1+\xi_2)(1+\xi_3)}{8}, \frac{(1-\xi_1)(1+\xi_2)(1-\xi_3)}{8}, \frac{(1-\xi_1)(1-\xi_2)}{4} \right) \cdot \right.$$
$$\left. \cdot \left(\frac{1+\xi_1}{2} \right)^2 \left(\frac{1+\xi_2}{2} \right) \frac{1}{8} \left| [\mathbf{Jac}_{\mathrm{L}}\mathbf{r}] \right| \right) d\xi_1 d\xi_2 d\xi_3, \qquad (5.137)$$

придатні для застосування формули:

$$\iiint_{\Omega} G(x_1, x_2, x_3) dx_1 dx_2 dx_3 \approx \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p} \left(\frac{1+\xi_{i,1}}{2} \right)^2 \left(\frac{1+\xi_{j,2}}{2} \right) \frac{1}{8} \left| [\mathbf{Jac}_{\mathrm{L}}\mathbf{r}] \right| \alpha_i \alpha_j \alpha_k \cdot$$
$$\cdot G\left(\frac{(1-\xi_{i,1})(1+\xi_{j,2})(1+\xi_{k,3})}{8}, \frac{(1-\xi_{i,1})(1+\xi_{j,2})(1-\xi_{k,3})}{8}, \frac{(1-\xi_{i,1})(1-\xi_{j,2})}{4} \right). \qquad (5.138)$$

З іншої сторони, використовуючи тривимірні зірки: S_4 – є барицентром $(1/4,1/4,1/4,1/4)$; $S_{31}(a)$ – чотири вузли визначаються перестановкою $(a,a,a,1-3a)$; $S_{22}(a)$ – шість вузлів визначаються перестановкою $(a,a,1/2-a,1/2-a)$; $S_{211}(a,b)$ – дванадцять вузлів визначаються як $(a,a,b,1-2a-b)$; $S_{1111}(a,b,c)$ – перестановка $(a,b,c,1-a-b-c)$ з двадцяти чотирьох вузлів; можна вивести набір симетричних формул чисельного інтегрування для тетраедра.

Так, аналогічно до двовимірного випадку, при необхідному порядку формули $g=1$ кількість вузлів $p=1$, відповідає зірці S_4, тобто барицентру $(1/4,1/4,1/4,1/4)$, а відповідний ваговий коефіцієнт рівний об'єму тетраедра $\alpha=1/6$ (*Таблиця 5.5.1*).

При $g=2$, використовується зірка S_{31} з кількістю вузлів $p=4$. Ці вузли можна визначити як $r\mathbf{V}_i+(1-r)\mathbf{C}$, де r – коефіцієнт, що потрібно знайти, $\mathbf{V}_i$ – вершини тетраедра в барицентричних координатах, тобто $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$ та $(0,0,0,1)$, а $\mathbf{C}=(\mathbf{V}_1+\mathbf{V}_2+\mathbf{V}_3+\mathbf{V}_4)/4$ – барицентр. Ваговими коефіцієнтами для кожного з чотирьох симетричних вузлів інтегрування буде чверть об'єму тетраедра, а коефіцієнт r рівний $\pm 1/\sqrt{5}$. При коефіцієнті $r=-1/\sqrt{5}$ одна з координат завжди буде від'ємною, тобто вузли будуть знаходитися за межами області інтегрування, через це таку формулу не використовують, а вузли шукають при $r=1/\sqrt{5}$ (*Таблиця 5.5.2*).

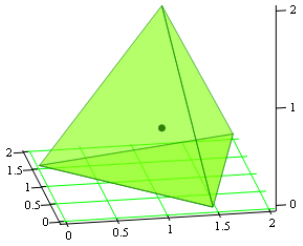
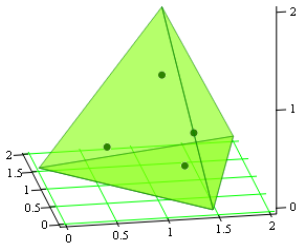
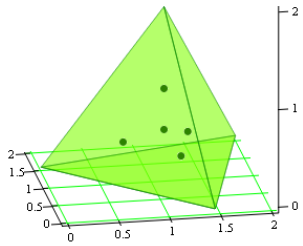
При $g=3$, використовують комбінацію зірок S_4 та S_{31} і загальна кількість вузлів рівна п'яти. У цьому випадку коефіцієнт r рівний $1/3$. Ваговий коефіцієнт для барицентру рівний $-4/30$, а для решти $9/120$ (*Таблиця 5.5.3*).

Симетричні формули чисельного інтегрування вищих порядків, аналогічно до двовимірного випадку, будуються на основі комбінацій зірок S_4, S_{31}, S_{22}, S_{211} та S_{1111}. Формула буде складатися з $i+4j+6k+12u+24v$ вузлів, де i, j,

k, u та v є невід'ємними цілими числами та i рівне 0 або 1. За необхідності можна також вивести формули для більшої кількості вимірів.

Таблиця 5.5

Вузли та відповідні їм вагові коефіцієнти симетричних формул чисельного інтегрування на тетраедрах

№	g	p	Розміщення вузлів	Координати вузлів (L_1, L_2, L_3, L_4)	Коефіцієнти α
1	1	1		$(1/4 \quad 1/4 \quad 1/4 \quad 1/4)$	1/6
2	2	4		$(b \quad a \quad a \quad a)$ $(a \quad b \quad a \quad a)$ $(a \quad a \quad b \quad a)$ $(a \quad a \quad a \quad b)$ $a=\frac{5-\sqrt{5}}{20} \quad b=1-3a=\frac{5+3\sqrt{5}}{20}$	1/24 1/24 1/24 1/24
3	3	5		$(1/4 \quad 1/4 \quad 1/4 \quad 1/4)$ $(1/2 \quad 1/6 \quad 1/6 \quad 1/6)$ $(1/6 \quad 1/2 \quad 1/6 \quad 1/6)$ $(1/6 \quad 1/6 \quad 1/2 \quad 1/6)$ $(1/6 \quad 1/6 \quad 1/6 \quad 1/2)$	–4/30 9/120 9/120 9/120 9/120

У даному підрозділі були наведені лише основні відомості про чисельне інтегрування при побудові матриць скінченних елементів. Розвиток описаних квадратурних і кубатурних формул не завершився і досі є предметом наукових досліджень, особливо для симетричних формул високих порядків. Про це свідчить наявність великої кількості наукових публікацій. Цікавий читач може частково ознайомитися з їх переліком, наприклад в Інтернет ресурсі "Encyclopaedia of Cubature Formulas" (http://nines.cs.kuleuven.be/ecf). Зокрема, подальший розвиток симетричних формул базується на використанні рівнянь моментів і відображеннях в полярні чи сферичні координати[1,2]. Інший

[1] Dunavant D. – High Degree Efficient Symmetrical Gaussian Quadrature Rules for Triangle // Int. Jour. for Numerical Methods in Engineering, 21:1129-1148, 1985.

[2] Heo S., Xu Y. – Constructing symmetric cubature formulae on a triangle // In Advances in Computational Mathematics, eds. Chen Z. et al, Marcel Dekker, New York, pp 203-221, 1999.

ефективний підхід базується на пошуку коренів поліномів методом найменших квадратів[1]. Існують й такі оригінальні підходи, як наприклад підходи на основі аналогії з щільним упакуванням сфер у об'ємі симплексу[2].

Для прикладу обчислимо матрицю жорсткості трикутного елементу другого порядку (5.69) з вершинами $(3/2;0)$, $(2;2)$ та $(0;3/2)$ для задачі стаціонарної теплопровідності з коефіцієнтом теплопровідності $\lambda=1$, визначальне рівняння методу зважених нев'язок якої, у матричній формі записується як:

$$[\mathbf{K}]=\int_{\Omega}[\mathbf{B}]^{\mathbf{T}}[\mathbf{D}][\mathbf{B}]d\Omega. \tag{5.139}$$

Перш за все побудуємо проекцію елементу в універсальний елемент, з заміною глобальних координат в барицентричні:

$$[\mathbf{K}]=\int_{\Omega}[\mathbf{B}]^{\mathbf{T}}[\mathbf{D}][\mathbf{B}]d\Omega=\int_{0}^{1}\int_{0}^{1-L_1}[\mathbf{B}]^{\mathbf{T}}[\mathbf{D}][\mathbf{B}]\left|[\mathbf{Jac_L r}]\right|dL_2dL_1, \tag{5.140}$$

звідки матриця Якобі та Якобіан рівні (5.84):

$$[\mathbf{Jac_L r}]=\begin{bmatrix}\frac{\partial x_1}{\partial L_1} & \frac{\partial x_2}{\partial L_1}\\ \frac{\partial x_1}{\partial L_2} & \frac{\partial x_2}{\partial L_2}\end{bmatrix}=\begin{bmatrix}X_{1,1}-X_{3,1} & X_{1,2}-X_{3,2}\\ X_{2,1}-X_{3,1} & X_{2,2}-X_{3,2}\end{bmatrix}=$$
$$=\begin{bmatrix}\frac{3}{2}-0 & 0-\frac{3}{2}\\ 2-0 & 2-\frac{3}{2}\end{bmatrix}=\begin{bmatrix}3/2 & -3/2\\ 2 & -1/2\end{bmatrix}\quad \left|[\mathbf{Jac_L r}]\right|=\frac{15}{4}. \tag{5.141}$$

Тепер знайдемо матрицю градієнтів $[\mathbf{B}]$. З (5.83)-(5.86) відомо, що:

$$\begin{Bmatrix}\frac{\partial N_{abc}}{\partial x_1}\\ \frac{\partial N_{abc}}{\partial x_2}\end{Bmatrix}=[\mathbf{Jac_L r}]^{-1}\begin{Bmatrix}\frac{\partial N_{abc}}{\partial L_1}\\ \frac{\partial N_{abc}}{\partial L_2}\end{Bmatrix}=[\mathbf{Jac_L r}]^{-1}\begin{Bmatrix}\frac{\partial N_{abc}(L_1,L_2,L_3)}{\partial L_1}-\frac{\partial N_{abc}(L_1,L_2,L_3)}{\partial L_3}\\ \frac{\partial N_{abc}(L_1,L_2,L_3)}{\partial L_2}-\frac{\partial N_{abc}(L_1,L_2,L_3)}{\partial L_3}\end{Bmatrix}. \tag{5.142}$$

Підставляючи в останнє рівняння вирази функцій форми квадратичного трикутного елементу (5.69), отримаємо:

$$[\mathbf{B}]=\begin{bmatrix}3/2 & -3/2\\ 2 & -1/2\end{bmatrix}^{-1}\cdot\begin{bmatrix}\frac{\partial N_{200}}{\partial L_1} & \frac{\partial N_{110}}{\partial L_1} & \frac{\partial N_{020}}{\partial L_1} & \frac{\partial N_{011}}{\partial L_1} & \frac{\partial N_{002}}{\partial L_1} & \frac{\partial N_{101}}{\partial L_1}\\ \frac{\partial N_{200}}{\partial L_2} & \frac{\partial N_{110}}{\partial L_2} & \frac{\partial N_{020}}{\partial L_2} & \frac{\partial N_{011}}{\partial L_2} & \frac{\partial N_{002}}{\partial L_2} & \frac{\partial N_{101}}{\partial L_2}\end{bmatrix}=$$

[1] Zhang L., Cui T., Liu H. – A set of symmetric quadrature rules on triangles and tetrahedral // Jour. of Comp. Math., 27(1):89-96, 2009.

[2] Williams D., Shunn L., Jameson A. – Symmetric quadrature rules for simplexes based on sphere close packed lattice arrangements // Jour. of Comp. and Appl. Math., 266:18-38, 2014.

$$= \begin{bmatrix} 2/15 & 2/5 \\ -8/15 & 2/5 \end{bmatrix} \cdot \begin{bmatrix} 4L_1 - 1 & 4L_2 & 0 & -4L_2 & 1-4L_3 & 4L_3 - 4L_1 \\ 0 & 4L_1 & 4L_2 - 1 & 4L_3 - 4L_2 & 1-4L_3 & -4L_1 \end{bmatrix}. \quad (5.143)$$

Тензор $[\mathbf{D}]$ у даному випадку є одиничною матрицею розміру 2×2.

Оскільки отримана матриця градієнтів $[\mathbf{B}]$ містить поліноми максимум першого порядку, тензор властивостей середовища $[\mathbf{D}]$ є одиничною матрицею, а матриця Якобі перетворення $\left|[\mathbf{Jac_L r}]\right|$ складається тільки з констант, то підінтегральний вираз в (5.140) міститиме максимум поліноми другого порядку, тобто $g = 2$.

Для точного обчислення цього інтегралу використаємо симетричну кубатурну[1] формулу (5.125) з вузлами (5.131) (див. *Таблиця 5.4.2*):

$$[\mathbf{K}] = \int_{\Omega} [\mathbf{B}]^{\mathbf{T}} [\mathbf{D}][\mathbf{B}] d\Omega = \int_0^1 \int_0^{1-L_1} [\mathbf{B}]^{\mathbf{T}} [\mathbf{D}][\mathbf{B}] \left|[\mathbf{Jac_L r}]\right| dL_2 dL_1 =$$

$$= \sum_{i=1}^{3} \left|[\mathbf{Jac_L r}]\right| \alpha_i [\mathbf{B}]^{\mathbf{T}} [\mathbf{D}][\mathbf{B}] = \quad (5.144)$$

$$= \frac{15}{4} \cdot \frac{1}{6} \left(G_{[\mathbf{B}]^{\mathbf{T}}[\mathbf{D}][\mathbf{B}]}\left(\frac{1}{2}, \frac{1}{2}, 0\right) + G_{[\mathbf{B}]^{\mathbf{T}}[\mathbf{D}][\mathbf{B}]}\left(0, \frac{1}{2}, \frac{1}{2}\right) + G_{[\mathbf{B}]^{\mathbf{T}}[\mathbf{D}][\mathbf{B}]}\left(\frac{1}{2}, 0, \frac{1}{2}\right) \right),$$

де під $G_{[\mathbf{B}]^{\mathbf{T}}[\mathbf{D}][\mathbf{B}]}(L_1, L_2, L_3)$ розуміється $[\mathbf{B}(L_1, L_2, L_3)]^{\mathbf{T}} [\mathbf{D}][\mathbf{B}(L_1, L_2, L_3)]$. З останнього виразу отримаємо:

$$[\mathbf{K}] = \frac{15}{4} \cdot \frac{1}{6} \left(\frac{1}{225} \begin{bmatrix} 68 & 64 & -36 & -64 & 32 & -64 \\ 64 & 272 & 72 & -272 & 136 & -272 \\ -36 & 72 & 72 & -72 & 36 & -72 \\ -64 & -272 & -72 & 272 & -136 & 272 \\ 32 & 136 & 36 & -136 & 68 & -136 \\ -64 & -272 & -72 & 272 & -136 & 272 \end{bmatrix} + \frac{1}{225} \begin{bmatrix} 68 & -136 & 36 & 136 & 32 & -136 \\ -136 & 272 & -72 & -272 & -64 & 272 \\ 36 & -72 & 72 & 72 & -36 & -72 \\ 136 & -272 & 72 & 272 & 64 & -272 \\ 32 & -64 & -36 & 64 & 68 & -64 \\ -136 & 272 & -72 & -272 & -64 & 272 \end{bmatrix} + \right. \quad (5.145)$$

$$\left. + \frac{1}{225} \begin{bmatrix} 68 & -72 & 36 & -72 & -32 & 72 \\ -72 & 288 & -144 & 288 & -72 & -288 \\ 36 & -144 & 72 & -144 & 36 & 144 \\ -72 & 288 & -144 & 288 & -72 & -288 \\ -32 & -72 & 36 & -72 & 68 & 72 \\ 72 & -288 & 144 & -288 & 72 & 288 \end{bmatrix} \right) = \frac{1}{90} \begin{bmatrix} 51 & -36 & 9 & 0 & 8 & -32 \\ -36 & 208 & -36 & -64 & 0 & -72 \\ 9 & -36 & 54 & -36 & 9 & 0 \\ 0 & -64 & -36 & 208 & -36 & -72 \\ 8 & 0 & 9 & -36 & 51 & -32 \\ -32 & -72 & 0 & -72 & -32 & 208 \end{bmatrix}.$$

Щоб задача була повною і мала розв'язок, необхідно вказати крайові умови. Тоді, коли визначення крайових умов Діріхле є тривіальною задачею вказування значення шуканого потенціалу у вузлі, визначення крайових умов Неймана потребує знаходження інтегралів від функцій форми:

$$\{\mathbf{f}\} = \int_{\Gamma} [\mathbf{N}]^{\mathbf{T}} f d\Gamma. \quad (5.146)$$

Оскільки функції форми вже не є лінійними, неможливо використати формули

[1] Оскільки елемент є симплексом, для побудови його матриці жорсткості також можна використати аналітичні формули інтегрування в барицентричних координатах (5.79).

для симплекс елементів. Спробуємо здійснити безпосереднє інтегрування функцій форми в барицентричних координатах (5.69) по необхідній стороні трикутника. Нехай цією стороною буде "перша" $\left((3/2;0)\quad(2;2)\right)$, для якої $L_3=0$ та $L_2=1-L_1$, у іншому випадку достатньо просто змінити нумерацію вузлів. Інтеграл записується як:

$$\{\mathbf{f}\}=\int_0^1 \mathrm{diag}\begin{Bmatrix} f_1\\ f_2\\ f_3\\ 0\\ 0\\ 0\end{Bmatrix}\begin{bmatrix} N_{200}=L_1(2L_1-1)\\ N_{110}=4L_1L_2\\ N_{020}=L_2(2L_2-1)\\ N_{011}=4L_2L_3\\ N_{002}=L_3(2L_3-1)\\ N_{101}=4L_1L_3\end{bmatrix}_{\substack{L_2=1-L_1\\ L_3=0}}\left|[\mathbf{Jac_L r}]^{\Gamma}\right|dL_1= \tag{5.147}$$

$$=\int_0^1 \mathrm{diag}\begin{Bmatrix} f_1\\ f_2\\ f_3\\ 0\\ 0\\ 0\end{Bmatrix}\begin{bmatrix} L_1(2L_1-1)\\ 4L_1(1-L_1)\\ (1-L_1)(1-2L_1)\\ 0\\ 0\\ 0\end{bmatrix}\left|[\mathbf{Jac_L r}]^{\Gamma}\right|dL_1=\frac{\Omega^{\Gamma}}{6}\mathrm{diag}\begin{Bmatrix} f_1\\ f_2\\ f_3\\ 0\\ 0\\ 0\end{Bmatrix}\begin{Bmatrix} 1\\ 4\\ 1\\ 0\\ 0\\ 0\end{Bmatrix},$$

де $\Omega^{\Gamma}=\Gamma$ – довжина сторони $\left((3/2;0)\quad(2;2)\right)$ рівна $\sqrt{17}/2$.

З іншої сторони, для визначення інтегралу (5.146), замість безпосереднього інтегрування можна використати аналітичні формули інтегрування в барицентричних координатах (5.79):

$$\begin{aligned}
&\int_{\Gamma} N_{200}d\Gamma=\int_{\Gamma}\left(2L_1L_1-L_1\right)d\Gamma=2\int_{\Gamma}L_1^2L_2^0d\Gamma-\int_{\Gamma}L_1^1L_2^0d\Gamma=\\
&\qquad=2\frac{2!\,0!}{(2+0+1)!}\Gamma-\frac{1!\,0!}{(1+0+1)!}\Gamma=\frac{2}{3}\Gamma-\frac{1}{2}\Gamma=\frac{1}{6}\Gamma,\\
&\int_{\Gamma} N_{110}d\Gamma=\int_{\Gamma}4L_1L_2d\Gamma=4\int_{\Gamma}L_1^1L_2^1d\Gamma=4\frac{1!1!}{(1+1+1)!}\Gamma=\frac{2}{3}\Gamma,\\
&\int_{\Gamma} N_{020}d\Gamma=\int_{\Gamma}\left(2L_2L_2-L_2\right)d\Gamma=2\int_{\Gamma}L_1^0L_2^2d\Gamma-\int_{\Gamma}L_1^0L_2^1d\Gamma=\\
&\qquad=2\frac{0!\,2!}{(0+2+1)!}\Gamma-\frac{0!1!}{(0+1+1)!}\Gamma=\frac{2}{3}\Gamma-\frac{1}{2}\Gamma=\frac{1}{6}\Gamma.
\end{aligned}\tag{5.148}$$

Оскільки тепер відомо, як точно будувати локальні матриці жорсткості та вектори навантажень, а процес ансамблювання нічим не відрізняється від попередніх, то можна побудувати апроксимації високих порядків для довільних еліптичних задач.

Порівняємо результати квадратичної апроксимації на вже знайомому з попередніх розділів прикладі задачі стаціонарної теплопровідності. Для цього знову використаємо дискретизацію з тією ж кількістю вузлів, що й для лінійної

та білінійної апроксимацій. Нагадаємо, що нами використовувалася дискретизація пластини регулярною сіткою 200 симплекс елементів. При тій ж кількості вузлів, кількість квадратичних елементів буде рівна 50, але розміри глобальної матриці жорсткості та вектору навантажень не зміняться (*Рис. 5.27*). Натомість матриця буде містити менше нульових коефіцієнтів, що еквівалентно включенню в розгляд додаткових зв'язків між вузлами за рахунок квадратичних функцій форми, які були відсутні при використанні симплекс елементів.

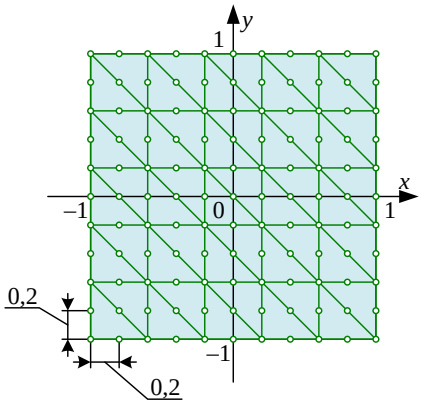


Рис. 5.27 Дискретизація пластини регулярною сіткою з 50 квадратичних трикутних елементів

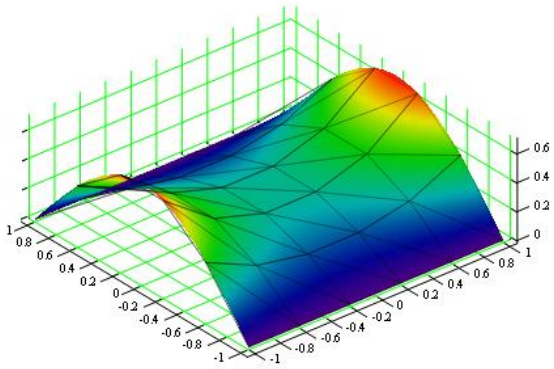


Рис. 5.28 Апроксимоване рішення задачі, при використанні регулярної сітки 50 квадратичних трикутних елементів

З *Рис. 5.29* та *Рис. 5.30* видно, що як і для одновимірного випадку, квадратична апроксимація дає набагато точніші рішення, ніж лінійна апроксимація, навіть при меншій кількості елементів.

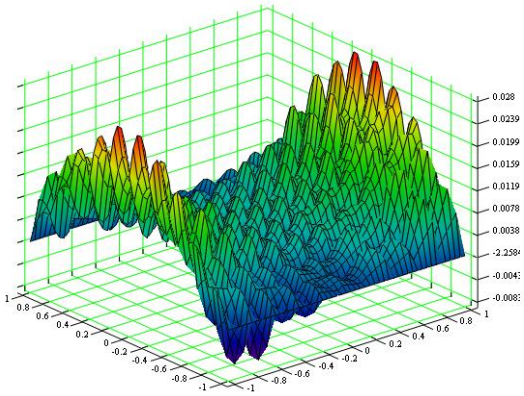


Рис. 5.29 Різниця між рішеннями, отриманими з допомогою квадратичної апроксимації, та апроксимації на симплекс елементах, при однаковій кількості вузлів

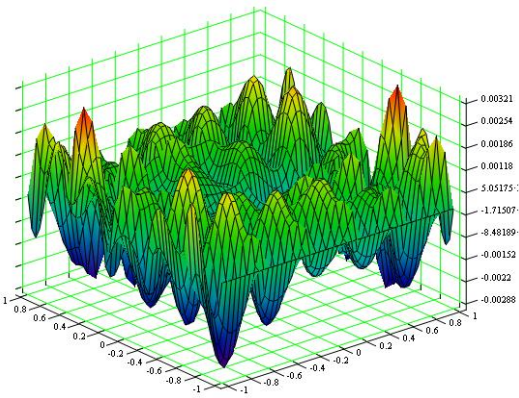


Рис. 5.30 Різниця між апроксимованим рішенням, отриманим з допомогою методу Бубнова-Гальоркіна, при $M = 5$, та рішенням квадратичної апроксимації

З *Рис. 5.31* та *Рис. 5.32* видно, що завдяки використанню повних двовимірних квадратичних поліномів, частинні похідні отриманого рішення можуть лінійно мінятися в межах елементу, на відміну від похідних симплекс

елементів, що були константами, чи на відміну від похідних білінійних елементів, що могли мінятися тільки вздовж однієї з координатних осей. Тим не менше, оскільки рішення належить $C^0(\Omega)$ класу гладкості, частинні похідні квадратичної апроксимації все ж мають розриви в міжелементних зонах.

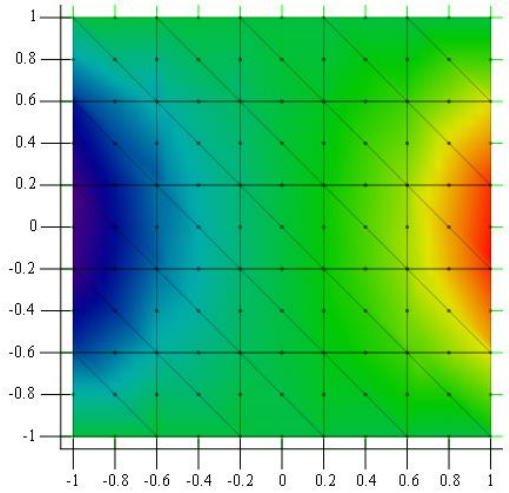


Рис. 5.31 Значення похідної $\partial\tilde{T}/\partial x$ від отриманого апроксимованого рішення

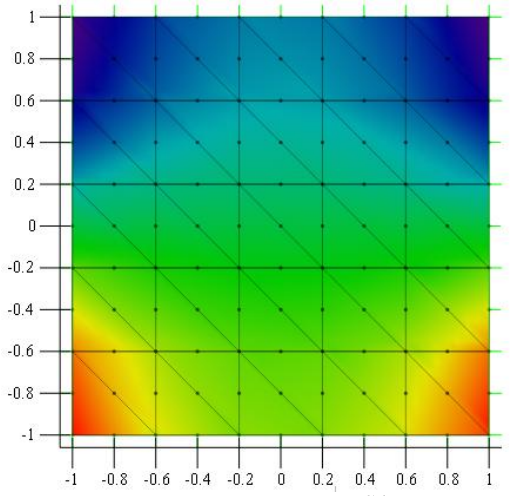


Рис. 5.32 Значення похідної $\partial\tilde{T}/\partial y$ від отриманого апроксимованого рішення

5.4. Криволінійні елементи

Виведені елементи високих порядків забезпечують швидку збіжність апроксимації до точного рішення, і тому з'являється можливість використовувати невелику кількість елементів. На жаль, примітивність форм цих елементів суперечить такій можливості при дискретизації складних об'єктів, і особливо тих, що містять криволінійні границі. Щоб апроксимувати форму криволінійних границь, в інженерних розрахунках дуже часто використовують велику кількість звичайних прямолінійних елементів, навіть у місцях, де градієнти шуканого потенціалу завідомо є незмінними. Це призводить до великої кількості надлишкових обчислень.

Щоб уникнути описаних проблем, необхідна можливість будувати криволінійні елементи. Найпростіше це можна зробити за допомогою відображення звичайних прямолінійних елементів, описаних в своїх локальних нормованих чи барицентричних координатах, в більш складну криволінійну фігуру, що розміщується в глобальній системі координат.

Розглянемо криволінійне відображення з простору локальних нормованих координат (ξ_1, ξ_2) в простір полярних координат (r, θ) і потім в простір глобальних координат (x_1, x_2). Взаємно-однозначне відображення між полярними і декартовими координатами будується за допомогою співвідношень:

$$x_1 = r\cos(\theta), \quad x_2 = r\sin(\theta). \tag{5.149}$$

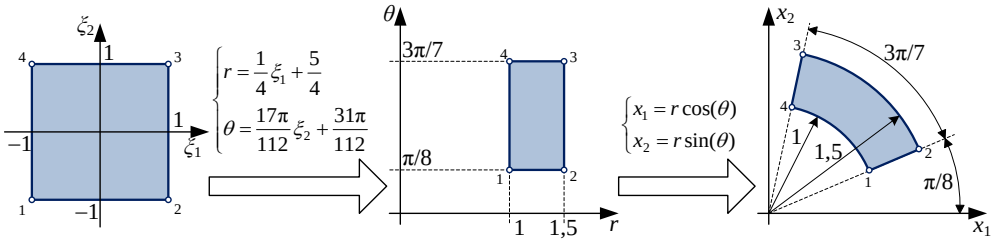


Рис. 5.33 Приклад відображення чотирикутного елементу з локальної системи координат (ξ_1,ξ_2) *в полярну систему координат* (r,θ) *і потім в глобальну систему координат* (x_1,x_2)

Використовуючи таке відображення, можна переводити вже описані прямолінійні елементи в криволінійні, так як це зображено на *Рис. 5.33*.

Щоб побудувати скінченно-елементну модель на основі таких криволінійних елементів, потрібно як і у випадку афінних перетворень з локальних координат в глобальні (див. (5.61)), знайти матрицю Якобі.

Так, для взаємно-однозначного відображення з простору локальних нормованих координат (ξ_1,ξ_2) в простір полярних координат (r,θ) для прямокутного елементу, на основі (5.54) та (5.62), матриця Якобі буде рівна:

$$[\mathbf{Jac}_{(\xi_1,\xi_2)}(r,\theta)]=\begin{bmatrix}\frac{\partial r}{\partial\xi_1} & \frac{\partial\theta}{\partial\xi_1}\\ \frac{\partial r}{\partial\xi_2} & \frac{\partial\theta}{\partial\xi_2}\end{bmatrix}=2\begin{bmatrix}r_2-r_1 & 0\\ 0 & \theta_4-\theta_1\end{bmatrix}=\begin{bmatrix}2h_1 & 0\\ 0 & 2h_2\end{bmatrix}, \qquad (5.150)$$

а з простору полярних координат (r,θ) в глобальні (x_1,x_2), на основі (5.149):

$$[\mathbf{Jac}_{(r,\theta)}(x_1,x_2)]=\begin{bmatrix}\frac{\partial x_1}{\partial r} & \frac{\partial x_2}{\partial r}\\ \frac{\partial x_1}{\partial\theta} & \frac{\partial x_2}{\partial\theta}\end{bmatrix}=$$
$$=\begin{bmatrix}\frac{\partial}{\partial r}r\cos(\theta) & \frac{\partial}{\partial r}r\sin(\theta)\\ \frac{\partial}{\partial\theta}r\cos(\theta) & \frac{\partial}{\partial\theta}r\sin(\theta)\end{bmatrix}=\begin{bmatrix}\cos(\theta) & \sin(\theta)\\ -r\sin(\theta) & r\cos(\theta)\end{bmatrix}, \qquad (5.151)$$

звідки можна побудувати матрицю Якобі для взаємно-однозначного відображення з (ξ_1,ξ_2) в (x_1,x_2) через (r,θ):

$$[\mathbf{Jac}_{\xi}x]=[\mathbf{Jac}_{(\xi_1,\xi_2)}(r,\theta)]\cdot[\mathbf{Jac}_{(r,\theta)}(x_1,x_2)]=$$
$$=\begin{bmatrix}2h_1 & 0\\ 0 & 2h_2\end{bmatrix}\cdot\begin{bmatrix}\cos(\theta) & \sin(\theta)\\ -r\sin(\theta) & r\cos(\theta)\end{bmatrix}=\begin{bmatrix}2h_1\cos(\theta) & 2h_1\sin(\theta)\\ -2h_2 r\sin(\theta) & 2h_2 r\cos(\theta)\end{bmatrix}. \qquad (5.152)$$

Знаючи матрицю Якобі для відображення, можна знайти похідні по глобальних координатах (x_1,x_2) від функцій форми елементів, що визначені в

локальних координатах (ξ_1, ξ_2) за допомогою (5.61). Після чого, стає можливим розв'язати рівняння методу зважених нев'язок. Наприклад, будь-який інтеграл:

$$I = \int_{\Omega_i} \left(\frac{\partial N_g}{\partial x_1} \lambda_{x_1} \frac{\partial N_j}{\partial x_1} + \frac{\partial N_g}{\partial x_2} \lambda_{x_2} \frac{\partial N_j}{\partial x_2} \right) dx_1 dx_2, \qquad (5.153)$$

що, зазвичай отримується при розв'язку однорідних еліптичних рівнянь, можна перетворити на інтеграл по квадрату в локальних нормованих координатах:

$$\begin{aligned} I &= \int_{-1}^{1}\int_{-1}^{1} \left(\frac{\partial N_g}{\partial x_1} \lambda_{x_1} \frac{\partial N_j}{\partial x_1} + \frac{\partial N_g}{\partial x_2} \lambda_{x_2} \frac{\partial N_j}{\partial x_2} \right) \left|[\mathbf{Jac}_{\xi} x]\right| d\xi_1 d\xi_2 = \\ &= \int_{-1}^{1}\int_{-1}^{1} \left([\mathbf{Jac}_{\xi} x]^{-1} [\mathbf{B}] \right)^{\mathbf{T}} [\mathbf{D}] [\mathbf{Jac}_{\xi} x]^{-1} [\mathbf{B}] \left|[\mathbf{Jac}_{\xi} x]\right| d\xi_1 d\xi_2. \end{aligned} \qquad (5.154)$$

Аналогічно шукаються й інші інтеграли (див наприклад, (5.29) та (5.44)).

Недоліком описаних криволінійних елементів є складність їх використання у поєднанні з іншими елементами, оскільки в таких випадках важко побудувати нерозривні міжелементні границі і, таким чином, забезпечити неперервність та допустимість отриманого апроксимованого рішення.

Тому, в більшості випадків, досить зручно використовувати спеціальні параметричні відображення, що базуються на тих ж функціях форми скінченного елементу. Наприклад, якщо базисні функції двовимірного елементу є квадратичними, то три вузли, що описують його границю, в загальному випадку можуть утворити криву другого порядку. Таким чином, шляхом зміни положення цих вузлів можна інтерполювати криволінійні поверхні кривими другого порядку.

Аналогічні дії можна застосувати для елементів довільного порядку і в довільній розмірності. Більше того, за умови використання суперпараметричних елементів, їх форма може описуватися функціями вищого порядку, ніж базисні, та навпаки, у випадку використання субпараметричних елементів.

Вперше ідея використання функцій форми елементів для інтерполяції криволінійних границь була запропонована в 1961 році[1], для відображення прямокутника в довільний чотирикутник, так, як це вже було показано на *Рис. 5.8* та (5.55). Пізніше цю ідею розширили на довільні елементи, при чому це зробили незалежно дві групи дослідників – у США в 1966 році[2,3] та в Великобританії в 1967 році[4].

Аналогічно до (5.55), можна показати [16], [15], що при параметричному відображенні з локальної системи координат $(\xi_1, \xi_2, \ldots)$ в глобальну систему

[1] Taig I. – Structural Analysis by the Matrix Displacement Method // Engl. Electric Aviation Rept. No. SO17, 1961.

[2] Irons B. – Numerical Integration Applied to Finite Element Methods // Conf. Use of Digital Computers in Struct. Eng., Univ. of Newcastle, 1966.

[3] Irons B. – Engineering Application of Numerical Integration in Stiffness Method // JAIAA, 14:2035-2037, 1966.

[4] Coons S. – Curves and Surfaces for Computer Aided Design // Comp. Aided Design Group, Cambridge, 1967.

координат $(x_1, x_2, \ldots)$, залежність між ними, виражається як:

$$x_1 = N_1 X_{1,1} + N_2 X_{2,1} + \ldots + N_M X_{M,1} = \sum_{j=1}^{M} N_j X_{j,1},$$

$$x_2 = N_1 X_{1,2} + N_2 X_{2,2} + \ldots + N_M X_{M,2} = \sum_{j=1}^{M} N_j X_{j,2}, \qquad (5.155)$$

$$\ldots$$

або в матричній формі:

$$\{x_1 \quad x_2 \quad \ldots\} = [\mathbf{N}] \cdot \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots \\ X_{2,1} & X_{2,2} & \cdots \\ \vdots & \vdots & \ddots \\ X_{M,1} & X_{M,2} & \cdots \end{bmatrix}. \qquad (5.156)$$

Тут [**N**] – функції форми скінченного елементу. У випадку, коли інтерполяція буде здійснюватися ними ж, а не з допомогою іншого набору функцій, наприклад при використанні більшої чи меншої кількості вузлів, або ієрархічних функцій, то отриманий елемент і відповідне відображення будуть ізопараметричними.

На *Рис. 5.34* показано відображення квадратичного чотирикутного елементу з дев'ятьма вузлами з локальних нормованих координат (ξ_1, ξ_2) в глобальні (x_1, x_2), де:

$$[X_{j,1} \quad X_{j,2}] = \begin{bmatrix} -0,5 & 1 & 2 & -0,5 & 1,25 & 2,25 & -1 & 1 & 3 \\ 0 & -0,25 & -1 & 1 & 1,25 & 1 & 2 & 2 & 3 \end{bmatrix}^{\mathbf{T}}. \qquad (5.157)$$

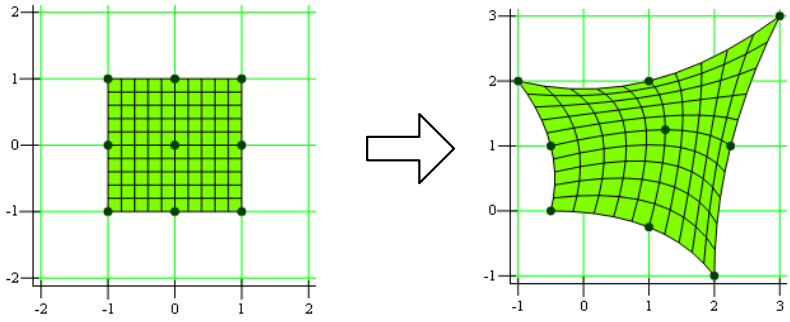


Рис. 5.34 Приклад параметричного відображення чотирикутного квадратичного елементу з локальної системи координат (ξ_1, ξ_2) в глобальну систему координат (x_1, x_2)

Доведено [5], [15], якщо два суміжні криволінійні елементи утворюються з первинних функцій форми, що задовольняють умови неперервності, ці елементи матимуть неперервні границі. У випадку параметричних відображень, з останніх виразів видно, що функції форми визначаються положенням вузлів елементу, і тому апроксимація по спільній границі сусідніх елементів буде неперервною (*Рис. 5.35*).

Очевидно, якщо базисні функції, що використовуються, належать до $C^0(\Omega)$ класу міжелементної гладкості, то і параметричні відображення матимуть ту ж гладкість.

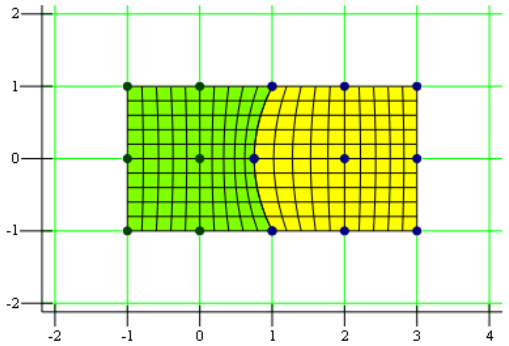


Рис. 5.35 Приклад неперервної границі між двома криволінійними елементами, утвореними параметричним відображенням

При використанні будь-яких відображень, для забезпечення збіжності обчислювального процесу та отримання правильних результатів, в першу чергу необхідно забезпечити їх невиродженість. Відображення є виродженим, коли Якобіан перетворення міняє свій знак на протилежний, або обертається в нуль. Наприклад для чотирикутних елементів, відображення стає виродженим, коли один з внутрішніх кутів стає більшим 180°, або у випадку квадратичних елементів, коли відстань між центральними та кутовими вузлами стає меншою третини сторони, на якій вони розміщені.

Матрицю Якобі, та Якобіан параметричного відображення можна знайти на основі (5.155) як:

$$[\mathbf{Jac}_{\xi} x] = \begin{bmatrix} \sum_{j=1}^{M} \frac{\partial N_j}{\partial \xi_1} X_{j,1} & \sum_{j=1}^{M} \frac{\partial N_j}{\partial \xi_1} X_{j,2} & \cdots \\ \sum_{j=1}^{M} \frac{\partial N_j}{\partial \xi_2} X_{j,1} & \sum_{j=1}^{M} \frac{\partial N_j}{\partial \xi_2} X_{j,2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} = [\mathbf{B}] \cdot \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots \\ X_{2,1} & X_{2,2} & \cdots \\ \vdots & \vdots & \ddots \\ X_{M,1} & X_{M,2} & \cdots \end{bmatrix}, \quad (5.158)$$

звідки на основі (5.154), та застосовуючи методи чисельного інтегрування можна побудувати апроксимацію розв'язку еліптичних задач.

Всі наведені співвідношення також справедливі для трикутних чи тетраедральних елементів, в обчисленнях яких використовуються барицентричні координати. Для прикладу, розглянемо однорідне еліптичне рівняння в фрагменті області, що описується двома квадратичними трикутними криволінійними елементами (*Рис. 5.36*), з координатами:

$$[\mathbf{C}]_1 = [X_{j,1} \quad X_{j,2}]_{\Omega_1} = \begin{bmatrix} 0 & 1 & 0 & 0{,}5 & 0{,}6 & 0 \\ 0 & 0 & 1 & 0 & 0{,}6 & 0{,}5 \end{bmatrix}^{\mathbf{T}},$$

$$[\mathbf{C}]_2 = [X_{j,1} \quad X_{j,2}]_{\Omega_2} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0{,}5 & 0{,}6 \\ 0 & 1 & 1 & 0{,}5 & 1 & 0{,}6 \end{bmatrix}^{\mathbf{T}}. \qquad (5.159)$$

Нехай на стороні $x_1 = 0$ задано крайову умову Діріхле – відомий потенціал $u_\infty = 0$, а на стороні $x_1 = 1$ задано крайову умову Неймана – потік потенціалу нормально до границі $\partial u / \partial \mathbf{n} = 100$. Також приймемо коефіцієнт провідності першого елементу $\lambda_1 = 3$, а другого елементу $\lambda_2 = 1$.

Функції форми квадратичного трикутного елементу в барицентричних координатах задаються співвідношеннями (5.69). Відповідні похідні можна знайти з (5.83)-(5.86) як (5.143). Враховуючи порядок нумерації вузлів елементів (5.159) поміняємо місцями необхідні стовбці та отримаємо:

$$[\mathbf{B}] = \begin{bmatrix} \frac{\partial N_{200}}{\partial L_1} - \frac{\partial N_{200}}{\partial L_3} & \frac{\partial N_{020}}{\partial L_1} - \frac{\partial N_{020}}{\partial L_3} & \frac{\partial N_{002}}{\partial L_1} - \frac{\partial N_{002}}{\partial L_3} & \frac{\partial N_{110}}{\partial L_1} - \frac{\partial N_{110}}{\partial L_3} & \frac{\partial N_{011}}{\partial L_1} - \frac{\partial N_{011}}{\partial L_3} & \frac{\partial N_{101}}{\partial L_1} - \frac{\partial N_{101}}{\partial L_3} \\ \frac{\partial N_{200}}{\partial L_2} - \frac{\partial N_{200}}{\partial L_3} & \frac{\partial N_{020}}{\partial L_2} - \frac{\partial N_{020}}{\partial L_3} & \frac{\partial N_{002}}{\partial L_2} - \frac{\partial N_{002}}{\partial L_3} & \frac{\partial N_{110}}{\partial L_2} - \frac{\partial N_{110}}{\partial L_3} & \frac{\partial N_{011}}{\partial L_2} - \frac{\partial N_{011}}{\partial L_3} & \frac{\partial N_{101}}{\partial L_2} - \frac{\partial N_{101}}{\partial L_3} \end{bmatrix} = \qquad (5.160)$$

$$= \begin{bmatrix} 4L_1 - 1 & 0 & 1 - 4L_3 & 4L_2 & -4L_2 & 4L_3 - 4L_1 \\ 0 & 4L_2 - 1 & 1 - 4L_3 & 4L_1 & 4L_3 - 4L_2 & -4L_1 \end{bmatrix}.$$

Матрицю Якобі відображення, згідно з (5.158) можна знайти як:

$$[\mathbf{Jac}_{\mathbf{L}}\mathbf{r}]_i = [\mathbf{B}][\mathbf{C}]_i =$$

$$= \begin{bmatrix} \sum_{j=1}^{M} \left(\frac{\partial N_j}{\partial L_1} X_{j,1} - \frac{\partial N_j}{\partial L_3} X_{j,1} \right) & \sum_{j=1}^{M} \left(\frac{\partial N_j}{\partial L_1} X_{j,2} - \frac{\partial N_j}{\partial L_3} X_{j,2} \right) \\ \sum_{j=1}^{M} \left(\frac{\partial N_j}{\partial L_2} X_{j,1} - \frac{\partial N_j}{\partial L_3} X_{j,1} \right) & \sum_{j=1}^{M} \left(\frac{\partial N_j}{\partial L_2} X_{j,2} - \frac{\partial N_j}{\partial L_3} X_{j,2} \right) \end{bmatrix}. \qquad (5.161)$$

Оскільки використовується ізопараметричні відображення, для обчислення інтегралів рівняння методу зважених нев'язок, використаємо симетричну кубатурну формулу (5.125) п'ятого порядку з семи вузлами (*Таблиця 5.4.7*). У такому випадку локальні матриці жорсткості, на основі (5.154) можна знайти як:

$$[\mathbf{K}]_i = \int_0^1 \int_0^{1-L_1} \left([\mathbf{Jac}_{\mathbf{L}}\mathbf{r}]_i^{-1} [\mathbf{B}] \right)^{\mathbf{T}} [\mathbf{D}]_i [\mathbf{Jac}_{\mathbf{L}}\mathbf{r}]_i^{-1} [\mathbf{B}] \left| [\mathbf{Jac}_{\mathbf{L}}\mathbf{r}]_i \right| dL_2 dL_1 =$$

$$= \sum_{j=1}^{p} \alpha_j G_{\left([\mathbf{Jac}_{\mathbf{L}}\mathbf{r}]_i^{-1} [\mathbf{B}] \right)^{\mathbf{T}} [\mathbf{D}]_i [\mathbf{Jac}_{\mathbf{L}}\mathbf{r}]_i^{-1} [\mathbf{B}] \left| [\mathbf{Jac}_{\mathbf{L}}\mathbf{r}]_i \right|} (L_{j,1}, L_{j,2}, L_{j,3}), \qquad (5.162)$$

звідки отримаємо:

$$[\mathbf{K}]_1 = \begin{bmatrix} 2{,}636062 & 0{,}459091 & 0{,}459091 & -1{,}676365 & -0{,}201515 & -1{,}676365 \\ 0{,}459091 & 1{,}831113 & 0{,}123853 & -1{,}837190 & -0{,}354196 & -0{,}222670 \\ 0{,}459091 & 0{,}123853 & 1{,}831113 & -0{,}222670 & -0{,}354196 & -1{,}837190 \\ -1{,}676365 & -1{,}837190 & -0{,}222670 & 6{,}774241 & -2{,}583217 & -0{,}454799 \\ -0{,}201515 & -0{,}354196 & -0{,}354196 & -2{,}583217 & 6{,}076340 & -2{,}583217 \\ -1{,}676365 & -0{,}222670 & -1{,}837190 & -0{,}454799 & -2{,}583217 & 6{,}774241 \end{bmatrix}, \qquad (5.163)$$

та:

$$[\mathbf{K}]_2 = \begin{bmatrix} 0{,}482924 & 0{,}221417 & -0{,}077353 & -0{,}964753 & 0{,}089133 & 0{,}248631 \\ 0{,}221417 & 1{,}231627 & 0{,}221417 & -0{,}939002 & -0{,}939002 & 0{,}203543 \\ -0{,}077353 & 0{,}221417 & 0{,}482924 & 0{,}089133 & -0{,}964753 & 0{,}248631 \\ -0{,}964753 & -0{,}939002 & 0{,}089133 & 3{,}791794 & 0{,}350687 & -2{,}327858 \\ 0{,}089133 & -0{,}939002 & -0{,}964753 & 0{,}350687 & 3{,}791794 & -2{,}327858 \\ 0{,}248631 & 0{,}203543 & 0{,}248631 & -2{,}327858 & -2{,}327858 & 3{,}954911 \end{bmatrix}. \quad (5.164)$$

Після ансамблювання системи та врахування крайових умов Діріхле, та крайових умов Неймана за допомогою (5.147)-(5.148), отримаємо систему лінійних рівнянь, розв'язком якого буде вектор вузлових потенціалів (*Рис. 5.37*):

$$\{\mathbf{u}\}_1 = \{0{,}000000 \quad 38{,}956911 \quad 0{,}000000 \quad 18{,}887374 \quad 21{,}824092 \quad 0{,}000000\}^{\mathbf{T}},$$
$$\{\mathbf{u}\}_2 = \{38{,}956911 \quad 61{,}388441 \quad 0{,}000000 \quad 53{,}995640 \quad 22{,}690935 \quad 21{,}824092\}^{\mathbf{T}}. \quad (5.165)$$

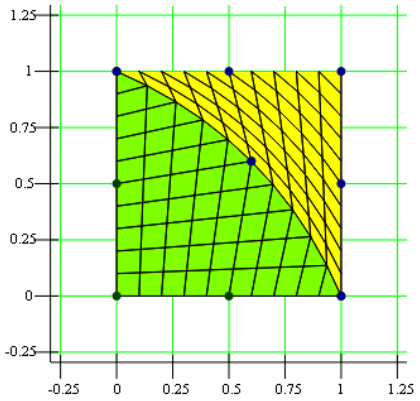


Рис. 5.36 Фрагмент дискретизації області двома квадратичними трикутними криволінійними елементами

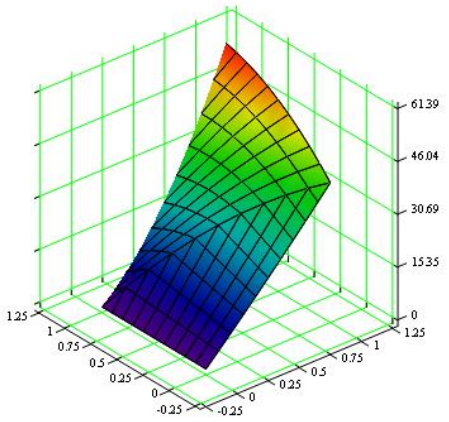


Рис. 5.37 Апроксимоване рішення однорідного еліптичного рівняння на квадратичних трикутних криволінійних елементах

"Класичні" відображення в полярні чи сферичні системи координат, а також параметричні відображення, є далеко не єдиним способом побудови криволінійних елементів. На практиці застосовують і багато інших видів відображень. Одним з них є відображення на основі змішувального процесу (в оригіналі "blending process") [1], [4], вперше запропоноване в 1971 році[1].

Розглянемо цей процес на прикладі чотирикутного елементу першого порядку. Нехай одна зі сторін чотирикутника задається параметричною кривою, координати якої в загальному випадку можна знайти як $x_1 = x_1(t)$ та $x_2 = x_2(t)$ (*Рис. 5.38*).

Щоб здійснити таке відображення та знайти залежність між локальними і глобальними координатами, застосовують змішувальний процес, що складається з таких етапів:

[1] Gordon W. – Blending-Function Methods of Bivariate and Multivariate Interpolation and Approximation // SIAM Journal on Numerical Analysis, 8(1):158-177, March 1971.

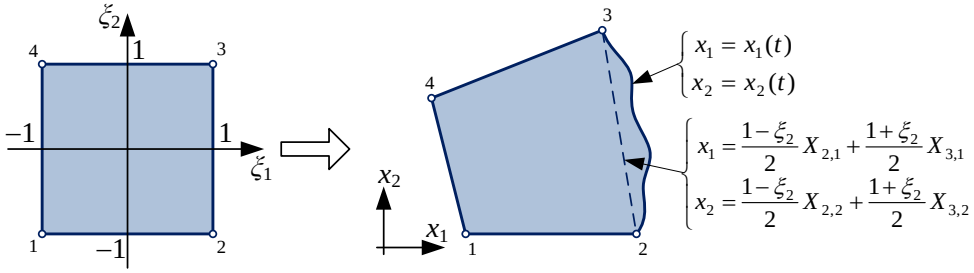


Рис. 5.38 Приклад відображення чотирикутного елементу з однією параметрично заданою криволінійною стороною на основі змішувального процесу

- параметрично задану криволінійну поверхню, нормують так, щоб параметр змінювався в межах локальних нормованих координат, в даному випадку $-1 \le t \le 1$ – тепер параметр відповідає локальній нормованій координаті на відповідній стороні елемента, в даному випадку це ξ_2;
- від отриманої функції віднімають функції стандартної інтерполяції для вузлових значень (або ієрархічні, якщо вони використовуються) по відповідній координаті, в даному випадку це $(1 \pm \xi_2)/2$ помножені на $(X_{2,1}, X_{2,2})$ та $(X_{3,1}, X_{3,2})$ відповідно;
- будують необхідну кількість функцій, що здійснюють лінійну інтерполяцію по решті локальних нормованих координат на відповідній стороні, в даному випадку це $(1+\xi_1)/2$ – стандартна одновимірна лінійна функція форми;
- добуток отриманої різниці та побудованих лінійних функцій по решті координат, додають до стандартного добутку функцій форми та координат елементу (див. (5.55),(5.155),(5.156)).

В результаті такого змішування функцій отримаємо залежність між локальними (ξ_1, ξ_2) та глобальними (x_1, x_2) координатами:

$$x_1 = \frac{1}{4}(1-\xi_1)(1-\xi_2)X_{1,1} + \frac{1}{4}(1+\xi_1)(1-\xi_2)X_{2,1} +$$
$$+\frac{1}{4}(1+\xi_1)(1+\xi_2)X_{3,1} + \frac{1}{4}(1-\xi_1)(1+\xi_2)X_{4,1} +$$
$$+\left(x_1(t=\xi_2) - \frac{1-\xi_2}{2}X_{2,1} - \frac{1+\xi_2}{2}X_{3,1} \right)\frac{1+\xi_1}{2},$$
$$x_2 = \frac{1}{4}(1-\xi_1)(1-\xi_2)X_{1,2} + \frac{1}{4}(1+\xi_1)(1-\xi_2)X_{2,2} +$$
$$+\frac{1}{4}(1+\xi_1)(1+\xi_2)X_{3,2} + \frac{1}{4}(1-\xi_1)(1+\xi_2)X_{4,2} +$$

$$+\left(x_2(t=\xi_2)-\frac{1-\xi_2}{2}X_{2,2}-\frac{1+\xi_2}{2}X_{3,2}\right)\frac{1+\xi_1}{2}. \qquad (5.166)$$

З останнього відношення видно, що перші чотири доданки це стандартні доданки лінійного відображення в довільний чотирикутник, а останній доданок, отриманий на основі змішувального процесу, перетворює сторону чотирикутника в криволінійну.

Відкривши дужки, цей вираз можна переписати як:

$$\begin{aligned} x_1 &= \frac{1}{4}(1-\xi_1)(1-\xi_2)X_{1,1}+\frac{1}{4}(1-\xi_1)(1+\xi_2)X_{4,1}+\frac{1+\xi_1}{2}x_1(t=\xi_2),\\ x_2 &= \frac{1}{4}(1-\xi_1)(1-\xi_2)X_{1,2}+\frac{1}{4}(1-\xi_1)(1+\xi_2)X_{4,2}+\frac{1+\xi_1}{2}x_2(t=\xi_2). \end{aligned} \qquad (5.167)$$

Наприклад, побудуємо криволінійний чотирикутник, що описує фрагмент кола. Для цього знову розглянемо залежність між полярними і декартовими координатами (5.149), але на відміну від попереднього прикладу (*Рис. 5.33*), застосуємо змішувальний процес, завдяки якому лише одна сторона буде криволінійною, а решта прямими.

Нехай, необхідно побудувати елемент, одна зі сторін якого точно описує дугу радіусом $r=2$ та кутом $0\le\theta\le\pi/2$. Також приймемо координати вузлів:

$$[\mathbf{C}]=[X_{j,1}\quad X_{j,2}]_\Omega=\begin{bmatrix}1&2&0&0\\0&0&1&2\end{bmatrix}^{\mathbf{T}}. \qquad (5.168)$$

Записуючи параметричні рівняння так, щоб параметр t мінявся в межах $-1\le t\le 1$, отримаємо:

$$x_1(t)=2\cos\left(\frac{\pi+t/2}{2}\right),\quad x_2(t)=2\sin\left(\frac{\pi+t/2}{2}\right). \qquad (5.169)$$

Тепер залишилося підставити останні співвідношення в (5.167), після чого отримаємо залежність, що описує необхідне відображення, показане на *Рис. 5.39*.

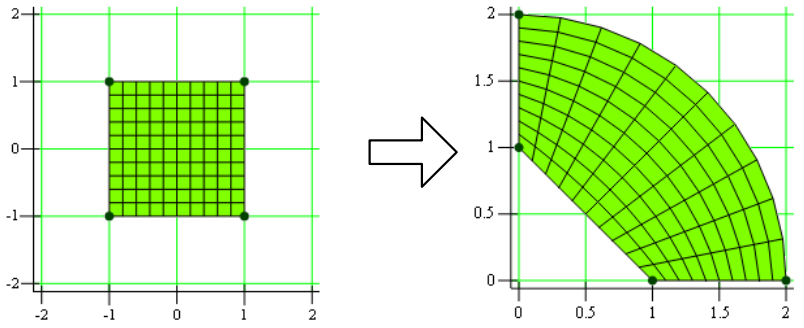


Рис. 5.39 Приклад відображення чотирикутного елементу з стороною, що точно описує дугу, побудованого на основі змішувального процесу

Застосовуючи описану техніку побудови криволінійних елементів, з'являється можливість точно описувати довільні криві. Це особливо актуально

при розв'язку задач динаміки нев'язких рідин, де штучна лінійна чи ізопараметрична апроксимації можуть повністю змінити поведінку рішення біля границь [20]. Проте, слід пам'ятати, що використовуючи таку можливість, в жертву необхідно привести простоту обчислень, оскільки апроксимація на складних кривих поверхнях потребує обчислення такого ж складного Якобіана при інтегруванні рівнянь методу зважених нев'язок.

Розміщення вузлів при формуванні сторін елементів, і відповідне визначення вектору навантажень $\{\mathbf{f}\}$, не є складною задачею у випадку використання прямолінійних елементів. З (5.147) видно, що для цього необхідно обчислити довжину сторони елементу. Для криволінійних елементів ситуація є аналогічною. Довжину (об'єм) криволінійної поверхні в $\Re^N$, заданої параметрично набором функцій $x_1(t), x_2(t), \ldots, x_N(t)$, можна визначити як [21]:

$$\Omega^{\Gamma} = \int_a^b \sqrt{\sum_{i=1}^{N} \left(\frac{dx_i(t)}{dt} \right)^2}\, dt. \qquad (5.170)$$

Або, коли крайові умови задані складною функцією $f(t)$, а не константою f, можна безпосередньо обчислити криволінійний інтеграл першого роду:

$$\{\mathbf{f}\} = \int_a^b [\mathbf{N}]^{\mathbf{T}} \left| [\mathbf{Jac}_t x] \right| f(t) \sqrt{\sum_{i=1}^{N} \left(\frac{dx_i(t)}{dt} \right)^2}\, dt. \qquad (5.171)$$

Спосіб обчислення наведених інтегралів вибирається в залежності від їх складності. Це можуть бути аналітичні вирази, формули чисельного інтегрування, і навіть одновимірна скінченно-елементна апроксимація [3].

Одним з найбільш цікавих і практично корисних видів відображень є таке, при якому безмежна область переводиться в скінченну. Подібні ситуації часто зустрічаються при моделюванні явищ електромагнетизму, чи будь-яких інших явищ, що розглядаються в частині об'єкту моделювання, яка набагато менша за весь об'єкт. У таких випадках використовують спеціальні теоретичні моделі необмежених чи напівобмежених тіл [22], [23], [24].

Існує два основні підходи до чисельного розв'язку цих задач. У першому випадку приймається прагматична точка зору і зовнішня границя фіксується на великій, але скінченній відстані, а область дискретизується тільки до цієї границі. Описана процедура в результаті дає велику кількість вузлів та елементів. Крім того, виникає питання визначення величини цієї "великої" відстані, тому зазвичай для цього необхідно проводити ряд чисельних експериментів. У другому випадку, обчислення проводяться безпосередньо для нескінченної області. Для цього використовують великий набір методів, починаючи від використання аналітичних рішень, що справедливі для нескінченних областей, і завершуючи найпростішими методами, при яких нескінченну область відображають в скінченну, використовуючи спеціальні *нескінченні скінченні елементи* [1] (вперше в 1977 році[1]).

[1] Bettess P. – Infinite elements // International Journal for Numerical Methods in Engineering, 11(1):53-64, 1977.

Спочатку розглянемо одновимірний випадок (*Рис. 5.40*). Нехай елемент починається у вузлі X_1, містить деякий проміжний вузол X_Q, і продовжується до безмежності в X_2. Побудуємо взаємно однозначне відображення такого елементу в локальні нормовані координати $-1 \le \xi \le 1$.

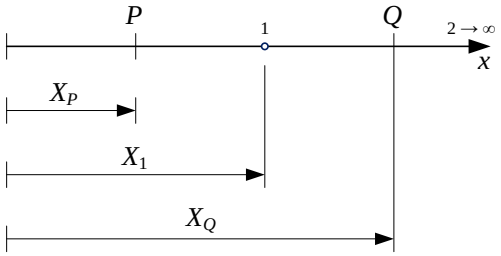


Рис. 5.40 Одновимірний нескінченний скінченний елемент

Глобальну координату можна знайти як:

$$x = N_P(\xi) X_P + N_Q(\xi) X_Q,$$
$$N_P(\xi) = -\frac{\xi}{1-\xi}, \quad N_Q(\xi) = 1 + \frac{\xi}{1-\xi}. \tag{5.172}$$

Ці вирази є аналогічними по формі до параметричного відображення (5.155), але функції форми N спеціально підібрані так, щоб вони приймали безмежні значення у вузлі при $\xi = 1$. Вузол X_P поки що не визначений. Зауважимо, що:

$$\begin{aligned} &\xi = 1: \quad x \equiv \frac{\xi}{1-\xi}(X_Q - X_P) + X_Q = \infty, \quad X_P \neq X_Q, \\ &\xi = 0: \quad x = X_Q, \\ &\xi = -1: \quad x \equiv X_1 = \frac{1}{2} X_P + \frac{1}{2} X_Q. \end{aligned} \tag{5.173}$$

Останні відношення визначають вузол X_P через X_1 та X_Q, і одразу видно, що вузол X_1 лежить посередині відрізку $[X_P, X_Q]$. Тобто відображення (5.172) можна переписати як:

$$x = N_P(\xi)(2X_1 - X_Q) + N_Q(\xi) X_Q = \frac{2(X_Q - X_1)\xi}{1-\xi} + X_Q. \tag{5.174}$$

Для побудови подібних відображень можна використати й багато інших функцій, тому важливо, щоб вони задовольняли умову:

$$N_P(\xi) + N_Q(\xi) = 1. \tag{5.175}$$

Така необхідність випливає з того, що відображення повинно залишатися незмінним при зміщенні початку координат x. Наприклад при:

$$X'_P = X_P + \Delta x, \quad X'_Q = X_Q + \Delta x, \tag{5.176}$$

необхідно, щоб для заданого ξ виконувалась рівність $x' = x + \Delta x$. Можна перевірити, що (5.172) відповідає цій умові.

Для апроксимації потенціалу використаємо ієрархічні базисні функції. Необхідно, щоб при $\xi = 1$, тобто при $x = \infty$, пробне рішення було $\tilde{u}(x) = 0$. Ця умова буде автоматично виконуватися, прийнявши вузлове значення $u_2 = 0$, звідки можна побудувати пробне рішення у вигляді поліному:

$$\begin{aligned}\tilde{u}(\xi) &= u_1 N_1(\xi) + \sum_{j=3}^{p+1} a_j N_j(\xi) + 0 \cdot N_2(\xi) = \\ &= \alpha_0 + \alpha_1 \xi + \alpha_2 \xi^2 + \ldots \alpha_p \xi^p.\end{aligned} \tag{5.177}$$

Тепер відображення можна побудувати, виразивши ξ через x:

$$\xi = 1 - \frac{X_Q - X_P}{x - X_P}. \tag{5.178}$$

Підставляючи це відношення в (5.177) отримаємо пробне рішення в глобальних координатах:

$$\tilde{u}(x) = \beta_0 + \frac{\beta_1}{r} + \frac{\beta_2}{r^2} + \ldots + \frac{\beta_p}{r^p}, \quad r = x - X_P, \tag{5.179}$$

де кількість членів залежить від порядку p інтерполяційного поліному.

Останній вираз відображає типову поведінку точного рішення на достатньо великій відстані та може бути використаний для опису функції "затухання" з будь-яким порядком точності. Очевидно, що оскільки вибір вузла X_Q (або X_1) є довільним, то, щоб отримати таким чином правильне скінченно-елементне рішення, необхідно знати, як веде себе рішення на достатньо великих відстанях і де приблизно починається затухання.

Наприклад, розглянемо рівняння [1]:

$$\begin{cases} d^2u(x)/dx^2 = 2/x^3, \\ u(2) = 1/2, \quad u(x \to \infty) \to 0, \\ 2 \le x < \infty. \end{cases} \tag{5.180}$$

Щоб мати можливість порівняти результати, знайдемо аналітичне рішення:

$$\begin{aligned} &\frac{du(x)}{dx} = \int \frac{2}{x^3} dx \qquad \Rightarrow \quad \frac{du(x)}{dx} = \frac{-1}{x} + C_1, \\ &u(x) = \int \left(\frac{-1}{x} + C_1 \right) dx \quad \Rightarrow \quad u(x) = \frac{1}{x} + C_1 x + C_2, \\ &u(2) = \frac{1}{2} = \frac{1}{2} + C_1 2 + C_2 \quad \Rightarrow \quad C_2 = -2C_1, \\ &u(x \to \infty) \to 0 = \frac{1}{x \to \infty} + C_1 (x \to \infty) - 2C_1 \quad \Rightarrow \quad C_1 = 0, C_2 = 0, \\ &u(x) = \frac{1}{x}. \end{aligned} \tag{5.181}$$

Апроксимуємо рішення єдиним квадратичним елементом, побудованим з допомогою ієрархічних базисних функцій. Для цього приймемо початок

елементу $X_1 = 2$ та довільно виберемо X_Q, нехай $X_Q = 3$. Тоді, згідно (5.173) $X_P = 1$. На основі (5.172) або (5.174) побудуємо відображення з локальних координат в глобальні:

$$x = N_P(\xi) X_P + N_Q(\xi) X_Q = \frac{3-\xi}{1-\xi}. \tag{5.182}$$

І, відповідно (5.178), обернене відображення з глобальних координат в локальні:

$$\xi = 1 - \frac{X_Q - X_P}{x - X_P} = \frac{x-3}{x-1}. \tag{5.183}$$

Оскільки ми наперед приймаємо рівним нулю пробне рішення в безмежності, використовуючи квадратичний елемент можна записати (5.177):

$$\tilde{u}(\xi) = u_1 N_1(\xi) + a_3 N_3(\xi) + 0 \cdot N_2(\xi), \tag{5.184}$$

де $N_1(\xi) = (1-\xi)/2$ та $N_3(\xi) = \xi^2 - 1$.

Рівняння методу зважених нев'язок у слабкій формі для даної задачі можна записати як:

$$\int_2^\infty \frac{\partial [\mathbf{N}]^{\mathrm{T}}}{\partial x} \frac{\partial [\mathbf{N}]}{\partial x} dx \{\mathbf{u}\} = -\int_2^\infty [\mathbf{N}]^{\mathrm{T}} \frac{2}{x^3} dx. \tag{5.185}$$

Оскільки коефіцієнт біля $N_2(\xi)$ (при $\xi = 1$, тобто у безмежності) є рівним нулю, з останніх виразів отримаємо матричне рівняння 2×2:

$$\begin{bmatrix} \int_{-1}^{1} \frac{\partial N_1}{\partial \xi} \left(\frac{\partial x(\xi)}{\partial \xi} \right)^{-1} \left(\frac{\partial x(\xi)}{\partial \xi} \right)^{-1} \frac{\partial N_1}{\partial \xi} \frac{\partial x(\xi)}{\partial \xi} d\xi & \int_{-1}^{1} \frac{\partial N_1}{\partial \xi} \left(\frac{\partial x(\xi)}{\partial \xi} \right)^{-1} \left(\frac{\partial x(\xi)}{\partial \xi} \right)^{-1} \frac{\partial N_3}{\partial \xi} \frac{\partial x(\xi)}{\partial \xi} d\xi \\ \int_{-1}^{1} \frac{\partial N_3}{\partial \xi} \left(\frac{\partial x(\xi)}{\partial \xi} \right)^{-1} \left(\frac{\partial x(\xi)}{\partial \xi} \right)^{-1} \frac{\partial N_1}{\partial \xi} \frac{\partial x(\xi)}{\partial \xi} d\xi & \int_{-1}^{1} \frac{\partial N_3}{\partial \xi} \left(\frac{\partial x(\xi)}{\partial \xi} \right)^{-1} \left(\frac{\partial x(\xi)}{\partial \xi} \right)^{-1} \frac{\partial N_3}{\partial \xi} \frac{\partial x(\xi)}{\partial \xi} d\xi \end{bmatrix} \cdot \begin{Bmatrix} u_1 \\ a_3 \end{Bmatrix} = \begin{Bmatrix} -\int_{-1}^{1} N_1 \frac{2}{x(\xi)^3} \frac{\partial x(\xi)}{\partial \xi} d\xi \\ -\int_{-1}^{1} N_3 \frac{2}{x(\xi)^3} \frac{\partial x(\xi)}{\partial \xi} d\xi \end{Bmatrix}, \tag{5.186}$$

де $\partial x(\xi)/\xi = [\mathbf{Jac}_\xi x]$ шукається на основі відображення (5.182) і є рівним $2/(\xi-1)^2$. Щоб не обчислювати всі інтеграли, підставимо в рівняння головні крайові умови, тобто $u(2) = u_1 = 1/2$, отримаємо:

$$\begin{bmatrix} 1 & 0 \\ 0 & \int_{-1}^{1} \frac{\partial N_3}{\partial \xi} \left(\frac{\partial x(\xi)}{\partial \xi} \right)^{-1} \left(\frac{\partial x(\xi)}{\partial \xi} \right)^{-1} \frac{\partial N_3}{\partial \xi} \frac{\partial x(\xi)}{\partial \xi} d\xi \end{bmatrix} \cdot \begin{Bmatrix} u_1 \\ a_3 \end{Bmatrix} = \begin{Bmatrix} \frac{1}{2} \\ -\int_{-1}^{1} N_3 \frac{2}{x(\xi)^3} \frac{\partial x(\xi)}{\partial \xi} d\xi - \int_{-1}^{1} \frac{\partial N_3}{\partial \xi} \left(\frac{\partial x(\xi)}{\partial \xi} \right)^{-1} \left(\frac{\partial x(\xi)}{\partial \xi} \right)^{-1} \frac{\partial N_1}{\partial \xi} \frac{\partial x(\xi)}{\partial \xi} d\xi \end{Bmatrix}, \tag{5.187}$$

або:

$$a_3 = \frac{-\int_{-1}^{1} N_3 \frac{2}{x(\xi)^3}\frac{\partial x(\xi)}{\partial \xi}d\xi - u_1\int_{-1}^{1}\frac{\partial N_3}{\partial \xi}\left(\frac{\partial x(\xi)}{\partial \xi}\right)^{-1}\left(\frac{\partial x(\xi)}{\partial \xi}\right)^{-1}\frac{\partial N_1}{\partial \xi}\frac{\partial x(\xi)}{\partial \xi}d\xi}{\int_{-1}^{1}\frac{\partial N_3}{\partial \xi}\left(\frac{\partial x(\xi)}{\partial \xi}\right)^{-1}\left(\frac{\partial x(\xi)}{\partial \xi}\right)^{-1}\frac{\partial N_3}{\partial \xi}\frac{\partial x(\xi)}{\partial \xi}d\xi} =$$

$$= \frac{-\int_{-1}^{1}(\xi^2-1)\frac{2(\xi-1)^3}{(\xi-3)^3}\frac{2}{(\xi-1)^2}d\xi - \frac{1}{2}\int_{-1}^{1}2\xi\frac{(\xi-1)^2}{2}\frac{(\xi-1)^2}{2}\frac{-1}{2}\frac{2}{(\xi-1)^2}d\xi}{\int_{-1}^{1}2\xi\frac{(\xi-1)^2}{2}\frac{(\xi-1)^2}{2}2\xi\frac{2}{(\xi-1)^2}d\xi} = \quad (5.188)$$

$$= \frac{-\int_{-1}^{1}\frac{4(\xi-1)(\xi^2-1)}{(\xi-3)^3}d\xi + \int_{-1}^{1}\frac{\xi(\xi-1)^2}{4}d\xi}{\int_{-1}^{1}2\xi^2(\xi-1)^2 d\xi} \approx -0,071542.$$

Звідки апроксимація квадратичним елементом, на основі (5.183) та (5.184), будується як (*Рис. 5.41*):

$$\tilde{u}(x) = u_1 N_1(x(\xi)) + a_3 N_3(x(\xi)) = 0,5\frac{1}{x-1} + 0,071542\frac{4(x-2)}{(x-1)^2}. \quad (5.189)$$

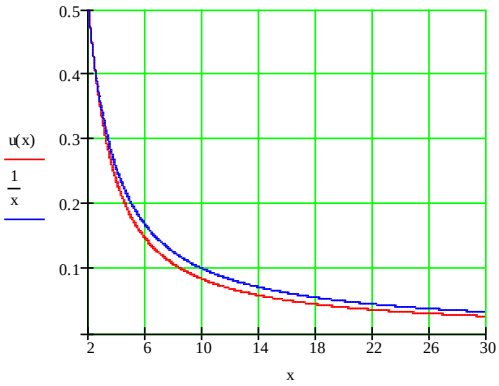


Рис. 5.41 Точне та наближене рішення рівняння $d^2u(x)/dx^2 = 2/x^3$, отримане квадратичним нескінченним елементом

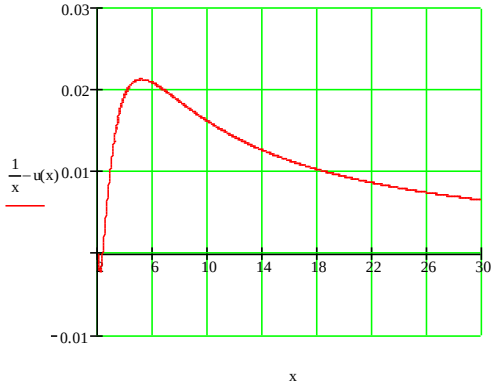


Рис. 5.42 Похибка між точним і отриманим наближеним рішенням рівняння $d^2u(x)/dx^2 = 2/x^3$

Якщо одновимірне відображення нескінченного елементу знайдене, то розширити його на дво- чи тривимірні простори не є складною задачею (*Рис. 5.43*). Спочатку розглянемо відображення прямої, що проходить через вузол 1 та Q, яка утворює сторону нескінченного елементу. Тут можна застосувати одновимірне відображення:

$$x_1 = N_P(\xi_1)(2X_{1,1} - X_{Q,1}) + N_Q(\xi_1)X_{Q,1} \equiv N_P(\xi_1)X_{P,1} + N_Q(\xi_1)X_{Q,1},$$
$$x_2 = N_P(\xi_2)(2X_{1,2} - X_{Q,2}) + N_Q(\xi_2)X_{Q,2} \equiv N_P(\xi_2)X_{P,2} + N_Q(\xi_2)X_{Q,2}, \quad (5.190)$$

де координати $(X_{P,1}, X_{P,2})$ вузла P визначаються як і раніше.

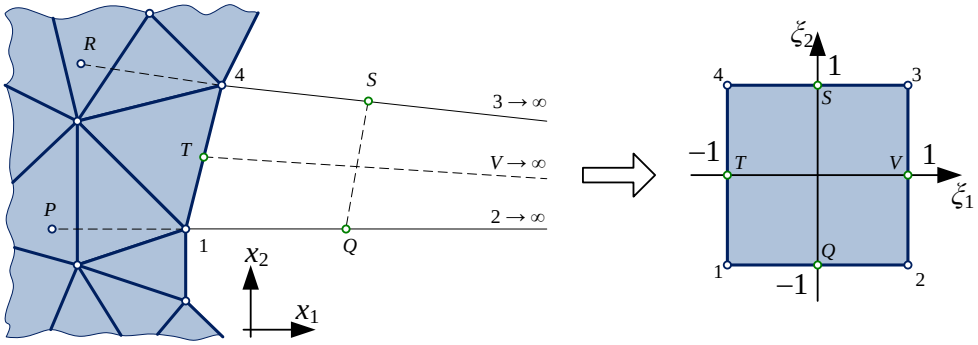


Рис. 5.43 Двовимірний нескінченний елемент

Якщо тепер положення вузла S визначене, при відповідному початку затухання R, то можна записати повне відображення нескінченного елементу:

$$\begin{aligned} x_1 &= N_1(\xi_2)\Big(N_P(\xi_1)(2X_{1,1} - X_{Q,1}) + N_Q(\xi_1)X_{Q,1}\Big) + \\ &+ N_4(\xi_2)\Big(N_P(\xi_1)(2X_{4,1} - X_{S,1}) + N_Q(\xi_1)X_{S,1}\Big), \\ x_2 &= N_1(\xi_2)\Big(N_R(\xi_1)(2X_{1,1} - X_{S,1}) + N_S(\xi_1)X_{S,1}\Big) + \\ &+ N_4(\xi_2)\Big(N_R(\xi_1)(2X_{4,1} - X_{Q,1}) + N_S(\xi_1)X_{Q,1}\Big), \end{aligned} \tag{5.191}$$

де N_1 та N_4 – стандартні лінійні одновимірні базисні функції, що задаються виразами:

$$N_1(\xi_2) = \frac{1-\xi_2}{2}, \quad N_4(\xi_2) = \frac{1+\xi_2}{2}. \tag{5.192}$$

Для інтерполяції знову можна використати ієрархічні базисні функції, при чому не важко помітити, що вздовж прямих $(1,2)$ та $(4,3)$ (і як наслідок вздовж всіх прямих $\xi_2 = \text{const}$) отримуються вирази типу (5.179), де r – відстань від відповідним чином вибраного полюса. Якщо такий полюс фіксується поблизу центру області, то він фактично визначає апроксимацію, тотожну на великих відстанях до точного рішення. При використанні таких елементів може бути отримана найкраща апроксимація [1].

Тепер, коли показано, як при побудові скінченно-елементних моделей використовувати елементи довільного порядку інтерполяції, а також елементи довільної форми, виникає питання, які саме елементи практично використовувати в обчисленнях? Очевидно, що при використанні фіксованих елементів, з послідовним нарощуванням їх порядку p, зростатиме швидкість збіжності чисельного методу, яку в літературі так і позначають "p-збіжність" [1]. У поєднанні з технікою відображень та методами чисельного інтегрування, з'являється можливість розглядати задачі на єдиному суперелементі високого порядку, що глобально описує одразу весь об'єкт моделювання, аналогічно до того, як це робилося класичними методами зважених нев'язок. Такий підхід дійсно застосовується на практиці (на основі змішувального процесу, вперше в

1973 році[1]). Однак він має серйозний недолік, що полягає у великій складності та кількості обчислень. Сюди також можна приписати вже наведені недоліки чисельної реалізації методів зважених нев'язок.

З іншої сторони, при дискретизації області великою кількістю елементів низького порядку, з деяким розміром h, значення потенціалу в сусідніх вузлах перестануть суттєво відрізнятися, що також приведе до зростання швидкості збіжності чисельного методу, яку в літературі так і позначають "*h*-збіжність" [1]. Такий підхід є найбільш простим у реалізації і тому користується великою популярністю в прикладних дослідженнях. Крім того, він дозволяє будувати скінченно-елементні моделі, де міжелементним залежностям приписують безпосередній фізичний зміст, що в ряді випадків є не менш важливим.

Не зовсім зрозуміло, яка збіжність буде швидшою. Практичні результати показують, що швидкість *p*-збіжності завжди є більшою [1], проте формально це не доведено, і не виключено, що таке твердження взагалі може бути доведено в загальному випадку. Тому, при побудові моделей зазвичай йдуть на компроміс, при якому використовують достатню кількість елементів максимум другого чи третього порядку, і таким чином, беруть переваги обох наведених способів, нівелюючи їх недоліки. В 1990-их рр.[2] така компромісна техніка зародила нову модифікації методу скінченних елементів під назвою "*hp*-FEM".

5.5. Список використаної літератури до розділу 5

[1] Zienkiewicz O., Morgan K. – Finite elements and approximation // New-York: Wiley, 1983.
[2] Norrie D., Vries G. – An Introduction to Finite Element Analysis // New-York: Academic press, 1978.
[3] Segerlind L. – Applied Finite Element Analysis / Применение метода конечных элементов / пер. с англ. Шестаков А., под. ред. Победри Б. // Москва: Мир, 1979.
[4] Fletcher C. – Computational Galerkin Methods / Численные методы на основе метода Галёркина / пер. с англ. Соколовская Л., под ред. Шидловский В. // Москва: Мир, 1988.
[5] Strang G., Fix G. – An Analysis of the Finite Element Method / Теория метода конечных элементов / пер с англ. Агошков В., Василенко В., Шайдурова В., под ред. Марчук Г. // Москва: Мир, 1977.
[6] Гантмахер Ф. – Теория матриц. 2-е изд., доп. // Москва: Наука, 1966.
[7] Винберг Э. – Курс Алгебры. 2-е изд., испр. и доп. // Москва: Факториал Пресс, 2001.
[8] Гельфанд И. – Лекции по линейной алгебре. 4-е изд., доп. // Москва: Наука, 1971.
[9] Strang G. – Linear Algebra and its Applications / Линейная алгебра и ее приминения / пер. с англ. // Москва: Мир, 1980.
[10] Gallagher R. – Finite Element Analysis. Fundamentals / Метод конечных элементов. Основы / пер. с англ. Картвелишвили В., под ред. Баничук Н. // Москва: Мир, 1984.

[1] Gordon W., Hall C. – Transfinite Element Methods: Blending-Function Interpolation over Arbitrary Curved Element Domains // Numer. Math. 21:109-129, 1973.

[2] Babuska I. , Guo B. – The h, p and h-p version of the finite element method: basis theory and applications // Advances in Engineering Software, Volume 15, Issue 3-4, 1992.

[11] Silvester P., Ferrari R. – Finite Elements for Electrical Engineers / Метод конечных элементов для радиоинженеров и инженеров-электриков / пер. с англ. Хотяинцева С., под ред. Дубровка Ф. // Москва: Мир, 1986.
[12] Eisenberg M., Malvern L. – On finite element integration in natural coordinates // International Journal for Numerical Methods in Engineering Volume 7, Issue 4, 574-575 pp., 1973.
[13] Демидович Б., Марон И. – Основы вычислительной математики. 3-е изд., испр. // Москва: Наука, 1966.
[14] Бахвалов Н., Жидков Н., Кобельков Г. – Численные методы // Москва: Бином. Лаборатория знаний, 2003.
[15] Zienkiewicz O. – The Finite Element Method in Engineering Science / Метод конечных элементов в технике / пер. с англ. под ред Пбедри Б. // Москва: Мир, 1975.
[16] Szabó B., Babuška I. – Introduction to Finite Element Analysis. Formulation, Verification and Validation // New-York: Wiley, 2011.
[17] Bathe, K. – Finite Element Procedures // NJ Englewood Cliffs: Prentice-Hall, 1996.
[18] Мысовских И. – Интерполяционные кубатурные формулы // Москва: Наука, 1981.
[19] Felippa C. – A Compendium of FEM Integration Rules for Finite Element Work // Eng. Computation, v. 21, pp. 867–890, 2004.
[20] Mitchell A. – The Finite Element Method in Partial Differential Equations / Метод конечных элементов для уравнений с частными производными / пер. с англ. Кондрашов В., Курякин А., под ред. Яненко Н. // Москва: Мир, 1981.
[21] Banach S. – Rachunek Rozniczkowy i Calkowy / Дифференциальное и интегральное исчисление. 2-е изд., испр. и доп. / пер. с польск. Зуховицкий С. // Москва: Наука, 1966.
[22] Лыков А. – Теория теплопроводности // Москва: Высшая школа, 1967.
[23] Лурье А. – Теория упругости // Москва: Наука, 1970.
[24] Нейман Л., Демирчян К. – Теоретические основы электротехники. В 2-х т. Учебник для вузов. Том 2. // Ленинград: Энергоиздат. Ленингр. отд-ние, 1967.

6. Декомпозиція обчислень на компонентному рівні проектування МЕМС

6.1. Доменна декомпозиція та розпаралелювання обчислень

Розвиток методу скінченних елементів не зупинився, про що свідчить значна кількість найсвіжіших наукових публікацій. Навпаки, завдяки взаємодії з іншими методами в галузі комп'ютерних наук, він поступово стає універсальним для будь-яких задач та засобів їх рішення. Так з розвитком технологій паралельних і розподілених обчислень, з'явилася можливість застосовувати різні модифікації методу скінченних елементів для моделювання надвеликих (за мірками перших десятиліть XXI століття) задач з допомогою високопродуктивних обчислень на суперкомп'ютерах або кластерних системах. Відповідно до цього, на зламі тисячоліть отримали розвиток модифікації під загальною назвою чисельних методів доменної декомпозиції, що спрямовані на декомпозицію задач і подальше їх паралельне рішення.

У попередніх розділах показано, що математичне, і як наслідок, програмне забезпечення для моделювання мікроелектромеханічних систем будується на основі чисельних методів. Використання цих методів є трудомісткою обчислювальною задачею, для вирішення якої доцільно використовувати паралельні розподілені обчислення [1], [2]. Тому, в даному розділі буде коротко розглянуто основні сучасні (на момент написання цієї роботи) модифікації методу скінченних елементів у контексті розпаралелювання чи моделювання надскладних або надвеликих об’єктів.

У галузі комп'ютерних наук, дві події називаються одночасними, коли вони відбуваються протягом одного і того ж часового інтервалу [1]. Якщо кілька задач виконуються протягом одного і того ж часового інтервалу, то говорять, що вони виконуються паралельно. Розрізняють фізично одночасне та конкурентне паралельне виконання задач [2]. У другому випадку, програми виконуються одночасно протягом одного і того ж часового інтервалу (паралельно), але послідовно в межах цього інтервалу.

Мета будь-яких комп'ютерних технологій паралелізму – забезпечити умови, що дозволяють обчислювальним пристроям здійснювати великі об'єми роботи за одні і ті ж часові інтервали. Розрізняють дві основні комп'ютерні технології паралелізму (парадигми) – методи паралельного програмування, що забезпечують паралельне виконання задач в межах фізично чи віртуально єдиного обчислювального пристрою; та методи розподіленого програмування, що забезпечують паралельне виконання задач з допомогою кількох фізично чи віртуально розподілених обчислювальних пристроїв [1]. На практиці обидві технології використовуються взаємно.

З позиції технічного забезпечення, у загальному випадку, можна виділити два основні напрямки паралельних обчислень: високопродуктивні обчислення з використанням суперкомп'ютерів (HPC) та розподілені обчислення, в тому числі з використанням кластерів (Distributed computing) [1]. Перевагою використання розподілених обчислювальних систем над суперкомп'ютерами є їх дешевизна за рахунок використання гетерогенних (з неоднорідною архітектурою та системним програмним забезпеченням) обчислювальних

пристроїв та можливість необмеженого нарощування продуктивності за рахунок масштабування; недоліком – низька пропускна здатність каналів зв'язку. Вибір розподілених обчислень, також дає можливість організації так званих волонтерських обчислень, при яких ресурси окремої машини використовуються тільки у вільний від її основної роботи час.

При вирішенні складних трудомістких обчислювальних задач, використовують два основні підходи до їх спрощення та подальшого паралельного рішення – паралелізм даних та паралелізм задач (чи підзадач) [3]. У першому випадку основна ідея полягає в тому, що одні і ті ж обчислювальні операції паралельно застосовуються до різних, відносно незалежних частин вхідних даних; у другому випадку, основна ідея полягає в тому, що початкова задача розбивається на кілька умовно незалежних підзадач, які виконуються паралельно.

Загальний алгоритм розпаралелювання, що може бути застосований з використанням обох описаних підходів, відображений в методології Фостера [3], [6], і передбачає послідовне виконання таких кроків як: декомпозиція (partitioning), планування комунікацій (communication), укрупнення (agglomeration) та планування обчислень (mapping). На основі цієї методології створено ряд парадигм паралельного програмування [7], на яких базується набір спеціальних шаблонів [8], що дозволяють здійснювати розпаралелювання. Так наприклад для генерації скінченно-елементної сітки можна застосувати парадигму "розділяй і володарюй" (Divide-and-Conquer), а для рішення СЛАР – "Конвеєрування" (Pipelining and Systolic).

Розділяють два типи декомпозиції – функціональну та доменну (декомпозицію даних). Такий поділ відображає підходи паралелізму задач та паралелізму даних відповідно. В залежності від конкретної задачі, можна використовувати одразу кілька видів декомпозиції. Наприклад, у межах парадигми "розділяй і володарюй" успішно розроблено ряд методів доменної декомпозиції, зокрема для чисельного розв'язку задач математичної фізики [9], [10]. Саме їх і постараємося розглянути в цьому розділі.

Не кожна задача може піддаватися ефективній декомпозиції, тобто бути розбитою на відносно незалежні підзадачі чи дані [3]. Прискорення S, що отримується при використанні паралельного алгоритму, у порівнянні з послідовним варіантом виконання обчислень, визначається як відношення часу затраченого на рішення задач одним процесором, до часу, затраченого на виконання цієї ж задачі деякою заданою кількістю процесорів [11]. Формально, отримане прискорення описується як:

$$S_K(n) = \frac{\tau_1(n)}{\tau_K(n)}, \qquad (6.1)$$

де K – кількість процесорів, τ – час виконання , n – деякий параметр обчислювальної складності задачі, наприклад величина вхідних даних.

Час виконання алгоритму $\tau_K(n)$ можна розділити на час виконання операцій, що можуть бути виконані паралельно $\tau_P(n)$, та час операцій, що

можуть бути виконані тільки послідовно $\tau_S(n)$. Тоді, останній вираз можна переписати як:

$$S_K(n) = \frac{\tau_S(n) + \tau_P(n)}{\tau_S(n) + \frac{\tau_P(n)}{K}}. \qquad (6.2)$$

Ефективність використання паралельним алгоритмом процесорів при вирішенні задачі визначається відношенням:

$$E_K(n) = \frac{\tau_1(n)}{K \cdot \tau_K(n)} = \frac{S_K(n)}{K}. \qquad (6.3)$$

Величина ефективності показує середній період часу вирішення алгоритму, протягом якого процесори реально використовуються для вирішення задачі, тобто не простоюють.

Оцінки максимально досяжних значень прискорення та ефективності паралельних алгоритмів рішення конкретних задач даються законом Амдала[1], що описує залежність прискорення до кількості процесорів і формально виражається як (6.2). Практично, закон Амдала дає оцінку можливості ефективного нарощування кількості процесорів [3], [11].

6.2. Основи методу скінченних елементів розривів і з'єднань

Застосування методів декомпозиції для спрощення (розпаралелювання) рішення диференціальних рівнянь з частинними похідними почали застосовувати задовго до створення перших обчислювальних машин. Першим з таких методів був метод альтернуючий метод Шварца 1870 року[2]. Його основна ідея полягала в пошуку рішення складної задачі на основі розбиття її області на деякі підобласті, що при тому могли перекривати одна одну [9].

Сучасні методи декомпозиції розв'язку ДРЧП є доменними і будуються на основі парадигми "розділяй і володарюй" [7], де вхідна задача для деякої великої області, розв'язується шляхом розбиття її на підзадачі для множини доменів (підобластей), що утворюють область. Завдяки цьому, підзадачі мають простіші кількісні (обсяг обчислень, обчислювальна складність) і якісні (форми підобластей, їх однотипність) показники, та можуть бути розв'язані паралельно з певним прискоренням [9].

Оскільки методи доменної декомпозиції з самого початку розраховані на паралелізм, то можуть бути максимально ефективно реалізовані технологіями паралельних або розподілених обчислень [12], [13].

Розглядаючи метод скінченних елементів в контексті доменної декомпозиції, можна зауважити, що кожен елемент фактично є доменом. Але, в конкретних реалізаціях, завдяки етапу укрупнення (agglomeration), для

[1] Amdahl G. – Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities // AFIPS Conference Proceedings, 30:483–485, New-York, 1967.

[2] Schwartz, H. – Über einen Grenzübergang durch alternierendes Verfahren // Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich 15:272–286, 1870.

досягнення максимальної швидкодії, у якості доменів вибираються сукупності елементів. При цьому, оцінка складності та відповідного прискорення методу є пропорційна відношенню розміру домену до розміру елементів, з яких він складається [13], [14], [15], [16].

Перші роботи по об'єднанню методу скінченних елементів і методів доменної декомпозиції з'явилися на початку 1990-их рр. Новий метод отримав назву методу скінченних елементів розривів і з'єднань, МСЕРЗ (Finite Element Tearing and Interconnecting Method, FETI)[1]. В подальшому з'явилися модифікації для прикладного використання і специфічних задач [17], основними з яких є FETI-DP (Dual-Primal) [16] і TFETI (Total, у деяких джерелах можна зустріти All-Floating FETI) [18], а також методи з використанням так званих "мортарних" функцій, що не вимагають відповідності дискретизацій доменів на границях [19].

В основі методу FETI лежить не нова ідея використання одного з методів пошуку екстремумів – методу множників Лагранжа [20], [21]. Розглянемо систему лінійних алгебраїчних рівнянь, що отримується методом скінченних елементів при розв'язку еліптичних задач:

$$[\mathbf{K}]\{\mathbf{u}\} = \{\mathbf{f}\}, \tag{6.4}$$

як похідну від деякого функціоналу, що є квадратичною формою для вектору шуканих значень $\{\mathbf{u}\}$:

$$\begin{gathered} \mathcal{F}(\mathbf{u}) = \frac{1}{2}\{\mathbf{u}\}^{\mathbf{T}}[\mathbf{K}]\{\mathbf{u}\} - \{\mathbf{u}\}^{\mathbf{T}}\{\mathbf{f}\} \to \text{extr}, \\ \frac{\partial \mathcal{F}(\mathbf{u})}{\partial \{\mathbf{u}\}^{\mathbf{T}}} = [\mathbf{K}]\{\mathbf{u}\} - \{\mathbf{f}\} = 0. \end{gathered} \tag{6.5}$$

Нехай домени не перетинаються, а границя між ними $\Gamma_{i,i'}$ утворена сторонами скінченних елементів, тобто вузли на границях співпадають. Введемо для кожного домену булеву матрицю граничних вузлових коефіцієнтів $[\mathbf{B}]$, яку в літературі часто називають *оператором стрибку* (jump operator) з простору рішень в простір множників Лагранжа, таких, що при додаванні по всій області:

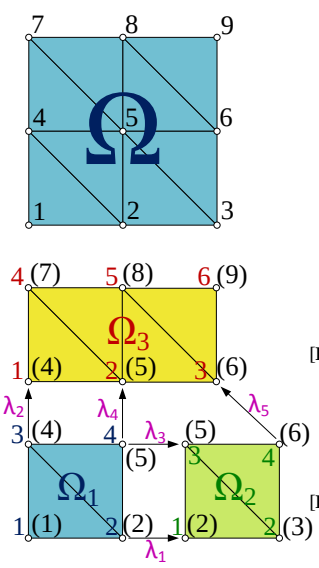
$$\sum_{i=1}^{D} [\mathbf{B}]_i \{\mathbf{u}\}_i^{\Gamma_{i,i'}} = 0. \tag{6.6}$$

Тобто, для кожного вузла, що належить границі між доменами записується рівність:

$$B^{(i)}u + B^{(i')}u = 0 \quad \Leftrightarrow \quad B^{(i)} = -B^{(i')}. \tag{6.7}$$

У класичному методі FETI в якості коефіцієнтів $[\mathbf{B}]$ вибирають значення 1 та -1 відповідно. А для граничних вузлів, що не належать границі, коефіцієнти у відповідному рядку матриці прирівнюють до нуля (*Рис. 6.1*).

[1] Farhat C, Roux F. – A method of finite element tearing and interconnecting and its parallel solution algorithm // Int. J. Numerical Methods in Engineering. 32:1205–1227, 1991.



$$[\mathbf{K}]=\begin{bmatrix} K_{1,1} & K_{1,2} & 0 & K_{1,4} & 0 & 0 & 0 & 0 & 0 \\ K_{2,1} & K_{2,2} & K_{2,3} & K_{2,4} & K_{2,5} & 0 & 0 & 0 & 0 \\ 0 & K_{3,2} & K_{3,3} & 0 & K_{3,5} & K_{3,6} & 0 & 0 & 0 \\ K_{1,1} & K_{4,2} & 0 & K_{4,4} & K_{4,5} & 0 & K_{4,7} & 0 & 0 \\ 0 & K_{5,2} & K_{5,3} & K_{5,4} & K_{5,5} & K_{5,6} & K_{5,7} & K_{5,8} & 0 \\ 0 & 0 & K_{6,3} & 0 & K_{6,5} & K_{6,6} & 0 & K_{6,8} & K_{6,9} \\ 0 & 0 & 0 & K_{7,4} & K_{7,5} & 0 & K_{7,7} & K_{7,8} & 0 \\ 0 & 0 & 0 & 0 & K_{8,5} & K_{8,6} & K_{8,7} & K_{8,8} & K_{8,9} \\ 0 & 0 & 0 & 0 & 0 & K_{9,6} & 0 & K_{9,8} & K_{9,9} \end{bmatrix} \quad \{\mathbf{u}\}=\begin{Bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \end{Bmatrix} \quad \{\mathbf{F}\}=\begin{Bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_7 \\ f_8 \\ f_9 \end{Bmatrix}$$

$$[\mathbf{K}]^{\Omega_1}=\begin{bmatrix} K_{1,1} & K_{1,2} & K_{1,4} & 0 \\ K_{2,1} & K_{2,2} & K_{2,4} & K_{2,5} \\ K_{4,1} & K_{4,2} & K_{4,4} & K_{4,5} \\ 0 & K_{5,2} & K_{5,4} & K_{5,5} \end{bmatrix} \quad \{\mathbf{u}\}^{\Omega_1}=\begin{Bmatrix} u_1 \\ u_2 \\ u_4 \\ u_5 \end{Bmatrix} \quad \{\mathbf{F}\}^{\Omega_1}=\begin{Bmatrix} f_1 \\ f_2 \\ f_4 \\ f_5 \end{Bmatrix} \quad [\mathbf{B}]^{\Omega_1}=\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$[\mathbf{K}]^{\Omega_2}=\begin{bmatrix} K_{2,2} & K_{2,3} & K_{2,5} & 0 \\ K_{3,2} & K_{3,3} & K_{3,5} & K_{3,6} \\ K_{5,2} & K_{5,3} & K_{5,5} & K_{5,6} \\ 0 & K_{6,3} & K_{6,5} & K_{6,6} \end{bmatrix} \quad \{\mathbf{u}\}^{\Omega_2}=\begin{Bmatrix} u_2 \\ u_3 \\ u_5 \\ u_6 \end{Bmatrix} \quad \{\mathbf{F}\}^{\Omega_2}=\begin{Bmatrix} f_2 \\ f_3 \\ f_5 \\ f_6 \end{Bmatrix} \quad [\mathbf{B}]^{\Omega_2}=\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$[\mathbf{K}]^{\Omega_3}=\begin{bmatrix} K_{4,4} & K_{4,5} & 0 & K_{4,7} & 0 & 0 \\ K_{5,4} & K_{5,5} & K_{5,6} & K_{5,7} & K_{5,8} & 0 \\ 0 & K_{6,5} & K_{6,6} & 0 & K_{6,8} & K_{6,9} \\ K_{7,4} & K_{7,5} & 0 & K_{7,7} & K_{7,8} & 0 \\ 0 & K_{8,5} & K_{8,6} & K_{8,7} & K_{8,8} & K_{8,9} \\ 0 & 0 & K_{9,6} & 0 & K_{9,8} & K_{9,9} \end{bmatrix} \quad \{\mathbf{u}\}^{\Omega_3}=\begin{Bmatrix} u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \end{Bmatrix} \quad \{\mathbf{F}\}^{\Omega_3}=\begin{Bmatrix} f_4 \\ f_5 \\ f_6 \\ f_7 \\ f_8 \\ f_9 \end{Bmatrix} \quad [\mathbf{B}]^{\Omega_3}=\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \end{bmatrix}$$

Рис. 6.1 Приклад без-надлишкової декомпозиції дискретизації на домени класичним методом FETI. Кількість рядків матриць $[\mathbf{B}]$*, рівна кількості множників Лагранжа* $\{\boldsymbol{\lambda}\}$*, кількість стовпців матриць* $[\mathbf{B}]$*, рівна кількості вузлів домену. Стрілками показано коефіцієнти, так що у вузлі звідки виходить стрілка, коефіцієнт рівний* 1*, а у вузлі куди стрілка напрямлена, коефіцієнт рівний* -1

У методі FETI-DP в матрицю $[\mathbf{B}]$ не включають коефіцієнти для вузлів, що лежать на границі трьох і більше доменів одночасно (так звані "кутові" або "перехресні" вузли). Натомість, значення шуканої величини в цих вузлах шукається окремою спеціальною процедурою.

У методі TFETI в матрицю $[\mathbf{B}]$ включають також коефіцієнти для вузлів, що лежать на границі початкової області. Це дозволяє відокремити вхідні граничні умови, і таким чином, для кожного домену можна паралельно використовувати однакові процедури обчислень.

У методах з використанням мортарних функцій матриця $[\mathbf{B}]$ в кожному з доменів не містить прості вузлові коефіцієнти, а містить деякі функції, що ставлять у відповідність одна одній дискретизації сусідніх доменів по границі, яка розглядається. Завдяки цьому, не вимагається відповідність дискретизацій по границі – сусідні елементи можуть перекриватися, а їх вузли можуть попадати на сторони чи всередину інших елементів без будь-якого узгодження.

Введемо в квадратичну форму (6.5) множники Лагранжа $\{\boldsymbol{\lambda}\}$, зміст яких, в

даному контексті, полягає у відображенні допоміжних крайових умов $\ell_{i,i'}(u_i(\mathbf{r}))\big|_{\Gamma_{i,i'}} = f_{i,i'}(\mathbf{r})$. Тоді, матрицю $[\mathbf{B}]$, для кожного з доменів, можна побудувати за схемою:

$$[\mathbf{B}]^{\Omega_i} = \begin{array}{c|cccc} b=\{-1;0;1\} & u_1^{\Omega_i} & u_2^{\Omega_i} & \cdots & u_M^{\Omega_i} \\ \hline \lambda_1 & b & b & \cdots & b \\ \lambda_2 & b & b & \cdots & b \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_H & b & b & \cdots & b \end{array}. \qquad (6.8)$$

а загальну систему рівнянь записати у вигляді:

$$\begin{gathered} \mathcal{F}(\mathbf{u},\boldsymbol{\lambda}) = \frac{1}{2}\{\mathbf{u}\}^{\mathrm{T}}[\mathbf{K}]\{\mathbf{u}\} - \{\mathbf{u}\}^{\mathrm{T}}\{\mathbf{f}\} + \{\mathbf{u}\}^{\mathrm{T}}[\mathbf{B}]^{\mathrm{T}}\{\boldsymbol{\lambda}\} \to \mathrm{extr}, \\ \{\mathbf{u}\}^{\mathrm{T}}[\mathbf{B}]^{\mathrm{T}}\{\boldsymbol{\lambda}\} = \{\boldsymbol{\lambda}\}^{\mathrm{T}}[\mathbf{B}]\{\mathbf{u}\}, \end{gathered} \qquad (6.9)$$

$$\begin{cases} \dfrac{\partial \mathcal{F}(\mathbf{u},\boldsymbol{\lambda})}{\partial \{\mathbf{u}\}^{\mathrm{T}}} = [\mathbf{K}]\{\mathbf{u}\} - \{\mathbf{f}\} + [\mathbf{B}]^{\mathrm{T}}\{\boldsymbol{\lambda}\} = \{\mathbf{0}\}, \\ \dfrac{\partial \mathcal{F}(\mathbf{u},\boldsymbol{\lambda})}{\partial \{\boldsymbol{\lambda}\}^{\mathrm{T}}} = [\mathbf{B}]\{\mathbf{u}\} = \{\mathbf{0}\}, \end{cases} \Leftrightarrow \begin{bmatrix} [\mathbf{K}] & [\mathbf{B}]^{\mathrm{T}} \\ [\mathbf{B}] & [\mathbf{0}] \end{bmatrix} \begin{Bmatrix} \{\mathbf{u}\} \\ \{\boldsymbol{\lambda}\} \end{Bmatrix} = \begin{Bmatrix} \{\mathbf{f}\} \\ \{\mathbf{0}\} \end{Bmatrix}.$$

Враховуючи всі домени, вона буде мати вигляд:

$$\begin{bmatrix} [\mathbf{K}]_1 & [\mathbf{0}] & \cdots & [\mathbf{0}] & [\mathbf{B}]_1^{\mathrm{T}} \\ [\mathbf{0}] & [\mathbf{K}]_2 & \cdots & [\mathbf{0}] & [\mathbf{B}]_2^{\mathrm{T}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ [\mathbf{0}] & [\mathbf{0}] & \cdots & [\mathbf{K}]_D & [\mathbf{B}]_D^{\mathrm{T}} \\ [\mathbf{B}]_1 & [\mathbf{B}]_2 & \cdots & [\mathbf{B}]_D & [\mathbf{0}] \end{bmatrix} \begin{Bmatrix} \{\mathbf{u}\}_1 \\ \{\mathbf{u}\}_2 \\ \vdots \\ \{\mathbf{u}\}_D \\ \{\boldsymbol{\lambda}\} \end{Bmatrix} = \begin{Bmatrix} \{\mathbf{f}\}_1 \\ \{\mathbf{f}\}_2 \\ \vdots \\ \{\mathbf{f}\}_D \\ \{\mathbf{0}\} \end{Bmatrix}. \qquad (6.10)$$

У класичному методі FETI, кількість множників Лагранжа відповідає кількості вузлових зв'язків між доменами всієї декомпозиції та кількості рядків для матриці $[\mathbf{B}]$. При цьому, для "перехресних" вузлів, розрізняють надлишкову декомпозицію – коли множники Лагранжа вводяться, так щоб зв'язати домени кожен-з-кожним; без-надлишкову декомпозицію – так щоб домени зв'язувалися хоча б з одним з сусідніх (*Рис. 6.1*); та ортогональну декомпозицію, при якій матриця $[\mathbf{B}]$ вже не буде булевою [22]. Очевидно, що надлишкова декомпозиція збільшує кількість множників Лагранжа, і як наслідок, загальний обсяг обчислень. Але, при подальшому використанні ітераційних методів розв'язку СЛАР, ці додаткові множники приводять до швидшої збіжності, завдяки посиленню зв'язків між доменами.

6.3. Наближене рішення несумісних систем

Рішення будь-якої системи лінійних алгебраїчних рівнянь типу (6.4) існує та є єдиним тоді і тільки тоді, коли кількість рівнянь рівна кількості невідомих і всі рівняння є лінійно незалежними. У формах матричних рівнянь, це умови існування відмінного від нуля визначника матриці. Тобто рішення існує і єдине

тоді і тільки тоді, коли матриця є не виродженою. У такому випадку завжди можна знайти обернену матрицю (обернений оператор відображення з простору розв'язків):

$$\begin{gathered}[\mathbf{A}]\{\mathbf{x}\}=\{\mathbf{b}\},\\ \{\mathbf{x}\}=[\mathbf{A}]^{-1}\{\mathbf{b}\} \quad \Leftrightarrow \quad \left|[\mathbf{A}]\right|\neq 0,\\ [\mathbf{A}][\mathbf{A}]^{-1}=[\mathbf{A}]^{-1}[\mathbf{A}]=[\mathbf{E}],\end{gathered} \tag{6.11}$$

де $[\mathbf{E}]$ – одинична матриця.

Матриця жорсткості з системи лінійних рівнянь (6.4) відповідає цьому критерію у випадку коректності постановки задачі (теореми про існування та єдиність рішення за Адамаром), тобто у випадках, коли коректно задані крайові умови задачі. При декомпозиції області на домени, для деяких з них (а у випадку використання TFETI – для всіх) локальна постановка задачі стає некоректною, оскільки втрачається зв'язок з границями, де визначені ці крайові умови. Відповідні домени називають *плаваючими* (floating). Для них застосовують не просто обернені матриці (6.11), а певне сімейство їх узагальнень.

Щоб знайти сімейство узагальнено обернених матриць, які підходять для FETI, спочатку розглянемо узагальнене обернення Мура-Пенроуза, яке прийнято називати *псевдооберненою матрицею* (вперше введено Муром в 1920 р[1], а пізніше, узагальнено Пенроузом в 1955 р[2]) [23], [24], [25], [26], [27], [28].

Псевдооберненою до будь-якої матриці $[\mathbf{A}]$, називають матрицю $[\mathbf{A}]^{\dagger}$, що відповідає чотирьом рівнянням Пенроуза:

$$\begin{gathered}[\mathbf{A}][\mathbf{A}]^{\dagger}[\mathbf{A}]=[\mathbf{A}],\\ [\mathbf{A}]^{\dagger}[\mathbf{A}][\mathbf{A}]^{\dagger}=[\mathbf{A}]^{\dagger},\\ \left([\mathbf{A}][\mathbf{A}]^{\dagger}\right)^{*}=[\mathbf{A}][\mathbf{A}]^{\dagger},\\ \left([\mathbf{A}]^{\dagger}[\mathbf{A}]\right)^{*}=[\mathbf{A}]^{\dagger}[\mathbf{A}],\end{gathered} \tag{6.12}$$

де, оператор * – означає комплексне (Ермітове) спряження матриці. У випадку, коли елементи матриці є дійсними числами, то $[\mathbf{A}]^{*}=[\mathbf{A}]^{\mathbf{T}}$. У випадку, коли матриця $[\mathbf{A}]$ є не виродженою, то $[\mathbf{A}]^{\dagger}=[\mathbf{A}]^{-1}$. Доведено [23], [25], що псевдообернена матриця завжди існує і є єдиною. Зауважте, що на відміну від обернених матриць, добуток $[\mathbf{A}][\mathbf{A}]^{\dagger}$ або $[\mathbf{A}]^{\dagger}[\mathbf{A}]$ не обов'язково рівний одиничній матриці $[\mathbf{E}]$.

Якщо система лінійних рівнянь типу (6.4) не має єдиного розв'язку, тобто не існує оберненої матриці (6.11), то використовуючи псевдообернену матрицю

[1] Moore E. – On the reciprocal of the general algebraic matrix // Bulletin of the American Mathematical Society, No. 26(9):394–395, 1920.

[2] Penrose R. – A generalized inverse for matrices // Proceedings of the Cambridge Philosophical Society, 51:406–413, 1955.

$[\mathbf{A}]^{\dagger}$, методом найменших квадратів, завжди можна знайти один і тільки один *оптимальний по довжині* наближений розв'язок, який мінімізує квадрат нев'язки, тобто квадратичне відхилення $\left\|[\mathbf{A}]\{\mathbf{x}\}-\{\mathbf{b}\}\right\|^2_{\mathcal{L}_2(\Omega)}$.

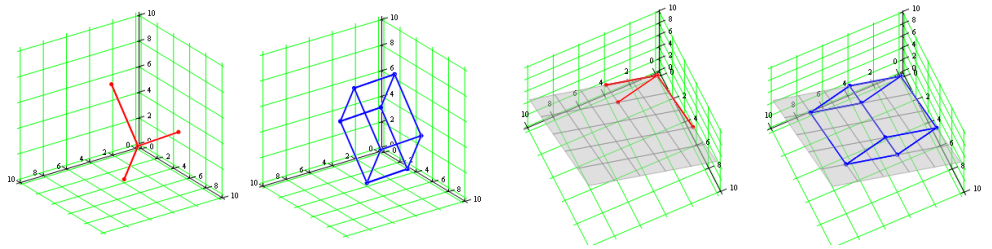
Спробуємо пояснити попереднє твердження наступними прикладами. Річ у тому, що будь-яку матрицю $[\mathbf{A}]$ можна розглядати як оператор відображення, що діє в певному функціональному просторі. Допустимо, матриця $[\mathbf{A}]$ є виродженою – її визначник рівний нулю. Що це означає в контексті оператору відображення?

Оскільки кожен з рядків чи стовпців матриці можна представити як вектор у деякому багатовимірному просторі, визначник матриці, або узагальнений векторний добуток, описує об'єм (гіпер)паралелепіпеду, утвореного зі стовпців або рядків цієї матриці. Виродженість матриці, це відсутність лінійної незалежності між рядками або стовпцями. Один чи кілька з них є лінійними комбінаціями решти. Об'єм такого (гіпер)паралелепіпеду буде рівним нулю – в даному просторі він умовно займатиме тільки площу, що лежить в деякій гіперплощині (*Рис. 6.2*).

Матриця та її визначник:

$$[\mathbf{A}]=\begin{bmatrix}3&4&0\\1&4&5\\6&0&3\end{bmatrix},\quad |[\mathbf{A}]|=144 \qquad\qquad [\mathbf{A}]=\begin{bmatrix}4&0&4\\0&6&2\\2&3&3\end{bmatrix},\quad |[\mathbf{A}]|=0$$

Стовпці матриці, та відповідний їм паралелепіпед:



Рядки матриці, та відповідний їм паралелепіпед:

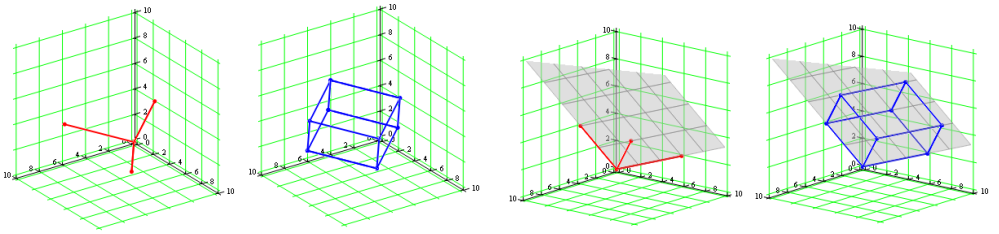


Рис. 6.2 Геометричний зміст визначника не виродженої та виродженої матриць. В обох випадках визначник рівний об'єму утвореного паралелепіпеду. Оскільки для виродженої матриці вектори лежать в одній площині, об'єм паралелепіпеду рівний нулю

При застосуванні матриці $[\mathbf{A}]$ у контексті оператора відображення, праві частини рівняння $[\mathbf{A}]\{\mathbf{x}\}=\{\mathbf{b}\}$, тобто всі можливі вектори $\{\mathbf{b}\}$, можуть належати тільки простору, що породжують стовпці матриці. Іншими словами, вектори $\{\mathbf{b}\}$ утворюються лінійними комбінаціями стовпців $[\mathbf{A}]$. Утворений простір в літературі називають *образом* оператора і позначають $\mathrm{im}([\mathbf{A}])$.

У випадку, коли матриця $[\mathbf{A}]$ є виродженою, простір її стовпців утворює деяку гіперплощину, а не заповнює увесь багатовимірний простір можливих правих частин. І, якщо вектор $\{\mathbf{b}\}$ не належить цій гіперплощині – неможливо підібрати таке рішення $\{\mathbf{x}\}$, яке б відображалося цією матрицею в $\{\mathbf{b}\}$. Таку систему прийнято називати *несумісною*. Найпростіший приклад: рівняння $ax=b$; якщо $a=0$ та $b\neq 0$, то не існує такого x, що б задовольняв систему – отже, вона несумісна.

Що робити у випадку $\{\mathbf{b}\}=\{\mathbf{0}\}$? Подібний випадок завжди допускає рішення $\{\mathbf{x}\}=\{\mathbf{0}\}$ при будь-якій матриці (чи операторі) $[\mathbf{A}]$, але, чи є єдиним це рішення? Іншими словами, необхідно перевірити, чи розв'язок системи $[\mathbf{A}]\{\mathbf{x}\}=\{\mathbf{0}\}$, що називається *однорідною* (до $[\mathbf{A}]\{\mathbf{x}\}=\{\mathbf{b}\}$), є єдиним?

Якщо матриця є не виродженою, то кожен з компонентів вектору рішення $\{\mathbf{x}\}$ відповідає одному з стовпців матриці. Можна умовно назвати всі компоненти вектору $\{\mathbf{x}\}$ *базисними змінними*. Якщо матриця вироджена – завжди залишається один, чи кілька вільних компонент, тобто *вільних змінних*, що можуть приймати довільні значення. Рішення однорідної системи вибиратиметься з простору, що утворений всіма можливими значеннями цих вільних змінних. Цей простір називають *нуль-простором*, або *ядром* оператора та позначають в літературі як $\ker([\mathbf{A}])$.

Наприклад, розглянемо однорідну систему:

$$\begin{bmatrix}4 & 0 & 4\\ 0 & 6 & 2\\ 2 & 3 & 3\end{bmatrix}\cdot\begin{Bmatrix}x_1\\ x_2\\ x_3\end{Bmatrix}=\begin{Bmatrix}0\\ 0\\ 0\end{Bmatrix}, \qquad (6.13)$$

(простір стовпців і рядків матриці цієї системи зображено на *Рис. 6.2*). За допомогою елементарних перетворень зведемо матрицю системи до верхньої трикутної матриці $[\mathbf{U}]$: перший рядок залишимо без змін; другий рядок залишимо без змін; третій рядок запишемо як суму половини першого і половини другого. Використані коефіцієнти утворюють нижню трикутну матрицю $[\mathbf{L}]$. У результаті отримаємо:

$$[\mathbf{L}]=\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0{,}5 & 0{,}5 & 1\end{bmatrix},\quad [\mathbf{U}]=\begin{bmatrix}4 & 0 & 4\\ 0 & 6 & 2\\ 0 & 0 & 0\end{bmatrix},\quad [\mathbf{L}]\cdot[\mathbf{U}]=\begin{bmatrix}4 & 0 & 4\\ 0 & 6 & 2\\ 2 & 3 & 3\end{bmatrix}=[\mathbf{A}]. \qquad (6.14)$$

Рішення $\{\mathbf{x}\}$ отриманої еквівалентної системи $[\mathbf{U}]\{\mathbf{x}\}=\{\mathbf{0}\}$ рівне:

$$\begin{bmatrix}4&0&4\\0&6&2\\0&0&0\end{bmatrix}\cdot\begin{Bmatrix}x_1\\x_2\\x_3\end{Bmatrix}=\begin{Bmatrix}0\\0\\0\end{Bmatrix}\;\Rightarrow\;\begin{cases}4x_1&+4x_3&=0\\6x_2&+2x_3&=0\end{cases}\;\Rightarrow\;\{\mathbf{x}\}=\begin{Bmatrix}-x_3\\-x_3/3\\x_3\end{Bmatrix}.\quad(6.15)$$

Змінні x_1 та x_2 є базисними. Змінна x_3 є вільною і може приймати будь-які значення. Сукупність всіх $\{\mathbf{x}\}$ при $-\infty<x_3<\infty$ утворюють нуль-простір матриці $[\mathbf{A}]$. У загальному випадку, розмірність нуль-простору рівна кількості вільних змінних. В конкретному випадку він є одновимірним і утворює пряму в тривимірному просторі. Особливістю будь-якого нуль-простору є те, що він ортогональний до простору рядків. Тобто всі вектори нуль-простору є ортогональними до всіх векторів простору рядків (*Рис. 6.3*).

$$[\mathbf{A}]=\begin{bmatrix}4&0&4\\0&6&2\\2&3&3\end{bmatrix},\quad\left|[\mathbf{A}]\right|=0$$

$$\ker\left([\mathbf{A}]\right)=\begin{Bmatrix}-x_3\\-x_3/3\\x_3\end{Bmatrix},\quad-\infty<x_3<\infty$$

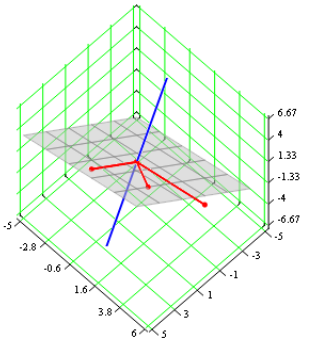


Рис. 6.3 Ортогональність нуль-простору виродженої матриці до простору її рядків

Повернемося знову до початкової системи $[\mathbf{A}]\{\mathbf{x}\}=\{\mathbf{b}\}$, при $\{\mathbf{b}\}\neq\{\mathbf{0}\}$. Повторюючи попередні кроки, отримаємо систему $[\mathbf{U}]\{\mathbf{x}\}=[\mathbf{L}]^{-1}\{\mathbf{b}\}$:

$$\begin{bmatrix}4&0&4\\0&6&2\\0&0&0\end{bmatrix}\cdot\begin{Bmatrix}x_1\\x_2\\x_3\end{Bmatrix}=\begin{Bmatrix}b_1\\b_2\\-0,5b_1-0,5b_2+b_3\end{Bmatrix}.\qquad(6.16)$$

Як і в найпростішому випадку, що був наведений раніше, отримана система може бути сумісною тоді і тільки тоді, коли $-0,5b_1-0,5b_2+b_3=0$. Це обмеження є нічим іншим, ніж рівнянням для простору стовпців матриці $[\mathbf{A}]$. Твердження легко перевірити, переписавши вираз як $z=(x+y)/2$, і побудувавши його графік – отримана площина буде співпадати з площиною, в якій лежать стовпці матриці $[\mathbf{A}]$ на *Рис. 6.2*.

Допустимо, наведене обмеження справджується для деякого заданого вектору $\{\mathbf{b}\}\neq\{\mathbf{0}\}$. Спробуємо знайти всі рішення для $[\mathbf{A}]\{\mathbf{x}\}=\{\mathbf{b}\}$, тобто *загальне рішення*:

$$\begin{cases}4x_1&+4x_3&=b_1\\6x_2&+2x_3&=b_2\end{cases}\;\Rightarrow\;\{\mathbf{x}\}=\begin{Bmatrix}-x_3\\-x_3/3\\x_3\end{Bmatrix}+\begin{Bmatrix}b_1/4\\b_2/6\\(2b_3-b_1-b_2)/2\end{Bmatrix}.\qquad(6.17)$$

Вектор, на який отримане рішення відрізняється від рішення однорідної системи, тобто $\{b_1/4 \quad b_2/6 \quad (2b_3 - b_1 - b_2)/2\}^{\mathbf{T}}$, називається *частковим рішенням* системи $[\mathbf{A}]\{\mathbf{x}\} = \{\mathbf{b}\}$. Загальне рішення системи завжди будується як сума рішення однорідної системи (у даному випадку (6.15)) та часткового рішення.

З геометричної точки зору, часткове рішення, в залежності від заданого вектору $\{\mathbf{b}\}$, утворює вектор, на який зміщується нуль-простір однорідної системи (*Рис. 6.4*). У результаті утворюється множина[1], що "паралельна" нуль-простору. Якщо матриця не вироджена, множина буде містити єдину точку. У конкретному випадку, існує безмежна кількість рішень ($-\infty < x_3 < \infty$).

$$\begin{bmatrix} 4 & 0 & 4 \\ 0 & 6 & 2 \\ 2 & 3 & 3 \end{bmatrix} \cdot \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} -4 \\ -8 \\ -6 \end{Bmatrix}$$

$$\begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} -x_3 - 1 \\ (-x_3 - 4)/3 \\ x_3 \end{Bmatrix}$$

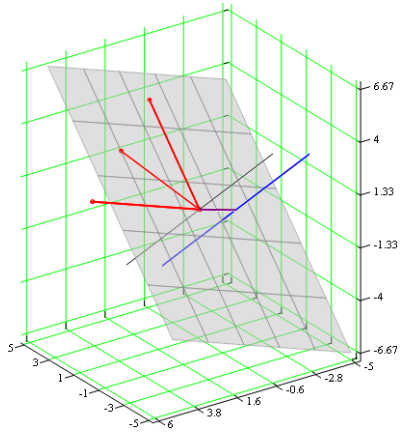


Рис. 6.4 Загальне рішення системи, як сума рішення однорідної системи та часткового рішення. Часткове рішення зміщує нуль-простір та утворює множину, "паралельну" до нього, що і є загальним рішенням системи

Наприклад (*Рис. 6.4*), виберемо $\{\mathbf{b}\} = \{-4 \quad -8 \quad -6\}^{\mathbf{T}}$. Часткове рішення, згідно (6.17), буде рівне $\{-1 \quad -4/3 \quad 0\}^{\mathbf{T}}$. Загальне рішення системи рівне $\{\mathbf{x}\} = \{(-x_3 - 1) \quad (-x_3 - 4)/3 \quad x_3\}^{\mathbf{T}}$:

$$\begin{aligned} x_3 = 2: & \begin{bmatrix} 4 & 0 & 4 \\ 0 & 6 & 2 \\ 2 & 3 & 3 \end{bmatrix} \cdot \begin{Bmatrix} -3 \\ -2 \\ 2 \end{Bmatrix} = \begin{Bmatrix} -4 \\ -8 \\ -6 \end{Bmatrix}; \quad x_3 = 5: \begin{bmatrix} 4 & 0 & 4 \\ 0 & 6 & 2 \\ 2 & 3 & 3 \end{bmatrix} \cdot \begin{Bmatrix} -6 \\ -3 \\ 5 \end{Bmatrix} = \begin{Bmatrix} -4 \\ -8 \\ -6 \end{Bmatrix}; \\ x_3 = 8: & \begin{bmatrix} 4 & 0 & 4 \\ 0 & 6 & 2 \\ 2 & 3 & 3 \end{bmatrix} \cdot \begin{Bmatrix} -9 \\ -4 \\ 8 \end{Bmatrix} = \begin{Bmatrix} -4 \\ -8 \\ -6 \end{Bmatrix}; \quad x_3 = -1: \begin{bmatrix} 4 & 0 & 4 \\ 0 & 6 & 2 \\ 2 & 3 & 3 \end{bmatrix} \cdot \begin{Bmatrix} 0 \\ -1 \\ -1 \end{Bmatrix} = \begin{Bmatrix} -4 \\ -8 \\ -6 \end{Bmatrix}; \quad \dots \end{aligned} \tag{6.18}$$

Що робити у випадку, коли система несумісна, тобто коли вектор $\{\mathbf{b}\}$ не лежить в просторі стовпців матриці? Подібні задачі найчастіше виникають на

[1] Простір обов'язково включає в себе нульовий елемент. Оскільки рішення $\{\mathbf{x}\} = \{\mathbf{0}\}$ вже не задовольняє систему, отриманий набір рішень не може утворити простір. Натомість, він утворює просто деяку множину рішень.

практиці. Наприклад, деяку спрощену модель реального об'єкту зводять до форми оператора чи системи лінійних рівнянь. По цій моделі пробують шукати деяке рішення, яке б описувало стан чи поведінку об'єкту. Вхідні параметри для моделі беруть з реального експерименту. І виявляється, ці вхідні параметри не зовсім підходять під розроблену модель – так і має бути! Модель лише приблизно описує реальний об'єкт.

Звичайно, можна спробувати відкинути деякі несумісні вхідні параметри з отриманого вектору $\{\mathbf{b}\}$, і таким чином спростити собі задачу. Але, подібний спосіб важко обґрунтувати, особливо, якщо всі компоненти вектору $\{\mathbf{b}\}$ отримуються з одного ж, чи рівноцінних реальних експериментів. Очевидним виходом в такій ситуації є вибір деякого допустимого моделлю рішення, що максимально близьке до вхідного вектору $\{\mathbf{b}\}$. Замість того, щоб знаходити точне рішення одних рівнянь, і при цьому, ігнорувати великі похибки в інших, потрібно вибрати рішення так, щоб мінімізувати середню похибку одразу для всіх рівнянь.

Як і раніше, в термінах функціональних просторів, ми будемо мінімізувати нев'язку між точним і отриманим наближеним рішенням, що допустиме, а отже й сумісне з моделлю. Потрібно знайти таке рішення, щоб $[\mathbf{A}]\{\mathbf{x}\}$ було максимально близько до $\{\mathbf{b}\}$. З метою подальшого формулювання методу скінченних елементів розривів і з'єднань, замість попередньо описаних методів подібної апроксимації, використаємо *метод найменших квадратів*. Як вже згадувалося в попередніх розділах, основна ідея цього методу полягає в мінімізації не просто норми нев'язки, а її квадрату.

Маючи $\mathcal{L}_2(\Omega)$ норму, тобто відстань у функціональному просторі:

$$\left\|[\mathbf{A}]\{\mathbf{x}\}-\{\mathbf{b}\}\right\|_{\mathcal{L}_2(\Omega)}=\left(\sum_{i=1}^{M}\left(\left([\mathbf{A}]\{\mathbf{x}\}-\{\mathbf{b}\}\right)_i\right)^2\right)^{\frac{1}{2}}= \tag{6.19}$$

$$=\sqrt{\left(([\mathbf{A}]\{\mathbf{x}\})_1-\{\mathbf{b}\}_1\right)^2+\left(([\mathbf{A}]\{\mathbf{x}\})_2-\{\mathbf{b}\}_2\right)^2+\ldots+\left(([\mathbf{A}]\{\mathbf{x}\})_M-\{\mathbf{b}\}_M\right)^2},$$

та підносячи її до квадрату, отримаємо квадратичну форму (параболу в одновимірному випадку):

$$\left\|[\mathbf{A}]\{\mathbf{x}\}-\{\mathbf{b}\}\right\|^2_{\mathcal{L}_2(\Omega)}=\left([\mathbf{A}]\{\mathbf{x}\}-\{\mathbf{b}\}\right)^{\mathbf{T}}\left([\mathbf{A}]\{\mathbf{x}\}-\{\mathbf{b}\}\right)= \tag{6.20}$$

$$=\{\mathbf{x}\}^{\mathbf{T}}[\mathbf{A}]^{\mathbf{T}}[\mathbf{A}]\{\mathbf{x}\}-2\{\mathbf{x}\}^{\mathbf{T}}[\mathbf{A}]^{\mathbf{T}}\{\mathbf{b}\}+\{\mathbf{b}\}^{\mathbf{T}}\{\mathbf{b}\}.$$

Вона має мінімум в місці, де:

$$\frac{\partial\left(\left\|[\mathbf{A}]\{\mathbf{x}\}-\{\mathbf{b}\}\right\|^2_{\mathcal{L}_2(\Omega)}\right)}{\partial\{\mathbf{x}\}^{\mathbf{T}}}=2\cdot[\mathbf{A}]^{\mathbf{T}}[\mathbf{A}]\{\mathbf{x}\}-2\cdot[\mathbf{A}]^{\mathbf{T}}\{\mathbf{b}\}=0. \tag{6.21}$$

Нам відомо, що ортогональна проекція завжди є єдиною, а норма, тобто відстань, між об'єктом і його ортогональною проекцією є мінімально можливою. Таким чином, задача зводиться до пошуку такого рішення $\{\mathbf{x}\}$, при якому вектор $[\mathbf{A}]\{\mathbf{x}\}$ є ортогональною проекцією вектору $\{\mathbf{b}\}$ – норма, та

відповідний квадрат нев'язки $\|[\mathbf{A}]\{\mathbf{x}\}-\{\mathbf{b}\}\|^2_{L_2(\Omega)}$ для нього є мінімальними. Це означає, що вектор $[\mathbf{A}]\{\mathbf{x}\}-\{\mathbf{b}\}$ повинен бути ортогональним до простору стовпців матриці $[\mathbf{A}]$ (*Рис. 6.5*):

$$\forall\{\mathbf{y}\}\in \mathrm{span}\left(\mathrm{cols}\left([\mathbf{A}]\right)\right):\quad \left([\mathbf{A}]\{\mathbf{y}\}\right)^{\mathbf{T}}\left([\mathbf{A}]\{\mathbf{x}\}-\{\mathbf{b}\}\right)=0. \qquad (6.22)$$

Останній вираз можна переписати як:

$$\forall\{\mathbf{y}\}\in \mathrm{span}\left(\mathrm{cols}\left([\mathbf{A}]\right)\right):\quad \{\mathbf{y}\}^{\mathbf{T}}\left([\mathbf{A}]^{\mathbf{T}}[\mathbf{A}]\{\mathbf{x}\}-[\mathbf{A}]^{\mathbf{T}}\{\mathbf{b}\}\right)=0. \qquad (6.23)$$

Оскільки вектор $\{\mathbf{y}\}$ можна вибрати довільно, вираз в дужках повинен дорівнювати нулю. Звідки приходимо до так званої системи *нормальних рівнянь*, що є фундаментальною для методу найменших квадратів:

$$[\mathbf{A}]^{\mathbf{T}}[\mathbf{A}]\{\mathbf{x}\}=[\mathbf{A}]^{\mathbf{T}}\{\mathbf{b}\}. \qquad (6.24)$$

Якщо система сумісна, то завжди можна знайти:

$$\{\mathbf{x}\}=\left([\mathbf{A}]^{\mathbf{T}}[\mathbf{A}]\right)^{-1}[\mathbf{A}]^{\mathbf{T}}\{\mathbf{b}\}, \qquad (6.25)$$

звідки, ортогональна проекція вектору $\{\mathbf{b}\}$ на простір стовбців матриці $[\mathbf{A}]$ рівна (*Рис. 6.5*):

$$[\mathbf{A}]\{\mathbf{x}\}=[\mathbf{A}]\left([\mathbf{A}]^{\mathbf{T}}[\mathbf{A}]\right)^{-1}[\mathbf{A}]^{\mathbf{T}}\{\mathbf{b}\}. \qquad (6.26)$$

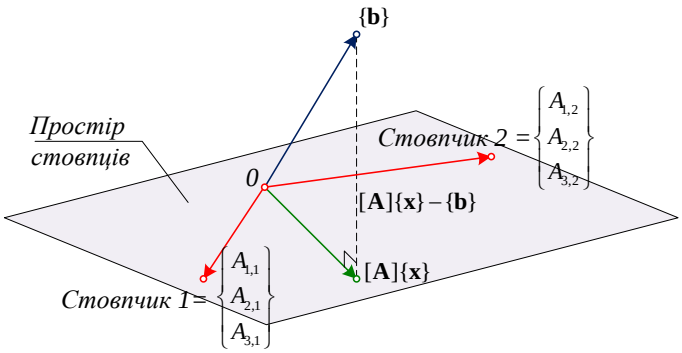


Рис. 6.5 Ортогональна проекція на простір стовпців матриці 3×3 *з двома лінійно незалежними стовпцями*

Наприклад, розглянемо несумісну систему:

$$\begin{bmatrix} 2 & 6 \\ -1 & -3 \end{bmatrix}\cdot\begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix}=\begin{Bmatrix} -2 \\ 6 \end{Bmatrix}. \qquad (6.27)$$

Квадратична форма (6.20) для цієї системи наведена на *Рис. 6.7*. З рисунку видно, що в ній існує безліч точок мінімуму, які формують жолоб. Простір стовпців матриці утворює пряму $b_2=-2b_1$, зображену на *Рис. 6.6*. Очевидно, що вектор $\{\mathbf{b}\}=\{-2 \quad 6\}^{\mathbf{T}}$ не належить цій прямій.

Відкинемо другий стовпець матриці, оскільки він є першим, помноженим на три. Знайдемо ортогональну проекцію $\{\mathbf{b}\}$:

$$[\mathbf{A}]\{\mathbf{x}\}=\begin{bmatrix}2\\-1\end{bmatrix}\left(\begin{bmatrix}2 & -1\end{bmatrix}\begin{bmatrix}2\\-1\end{bmatrix}\right)^{-1}\begin{bmatrix}2 & -1\end{bmatrix}\begin{Bmatrix}-2\\6\end{Bmatrix}=\begin{bmatrix}2\\-1\end{bmatrix}\cdot\frac{1}{5}\cdot(-10)=\begin{Bmatrix}-4\\2\end{Bmatrix}. \quad (6.28)$$

Маючи сумісну проекцію, знайдемо загальний розв'язок системи. Спочатку побудуємо верхню і нижню трикутні матриці:

$$[\mathbf{L}]=\begin{bmatrix}1 & 0\\-0{,}5 & 1\end{bmatrix},\quad [\mathbf{U}]=\begin{bmatrix}2 & 6\\0 & 0\end{bmatrix},\quad [\mathbf{L}]\cdot[\mathbf{U}]=\begin{bmatrix}2 & 6\\-1 & -3\end{bmatrix}=[\mathbf{A}]. \quad (6.29)$$

Далі, знайдемо рішення однорідної системи:

$$\begin{bmatrix}2 & 6\\0 & 0\end{bmatrix}\begin{Bmatrix}x_1\\x_2\end{Bmatrix}=\begin{Bmatrix}0\\0\end{Bmatrix}\ \Rightarrow\ 2x_1+6x_2=0\ \Rightarrow\ \{\mathbf{x}\}=\begin{Bmatrix}-3x_2\\x_2\end{Bmatrix}. \quad (6.30)$$

І загальне рішення, як суму попереднього і часткового:

$$\begin{cases}2x_1+6x_2=b_1\\0=b_2\end{cases}\ \Rightarrow\ \{\mathbf{x}\}=\begin{Bmatrix}-3x_2\\x_2\end{Bmatrix}+\begin{Bmatrix}b_1/2\\b_1/2+b_2\end{Bmatrix},$$
$$[\mathbf{A}]\{\mathbf{x}\}=\begin{Bmatrix}-4\\2\end{Bmatrix}\ \Rightarrow\ \{\mathbf{x}\}=\begin{Bmatrix}-3x_2\\x_2\end{Bmatrix}+\begin{Bmatrix}-2\\0\end{Bmatrix}=\begin{Bmatrix}-3x_2-2\\x_2\end{Bmatrix}. \quad (6.31)$$

Отриманий результат, це рівняння прямої, що проходить по жолобу мінімальних значень квадратичної форми початкової системи (*Рис. 6.7*). Будь-яка точка цієї прямої утворює ортогональну проекцію $[\mathbf{A}]\{\mathbf{x}\}$ і квадрат нев'язки $\left\|[\mathbf{A}]\{\mathbf{x}\}-\{\mathbf{b}\}\right\|^2_{L_2(\Omega)}$ для неї є мінімальним.

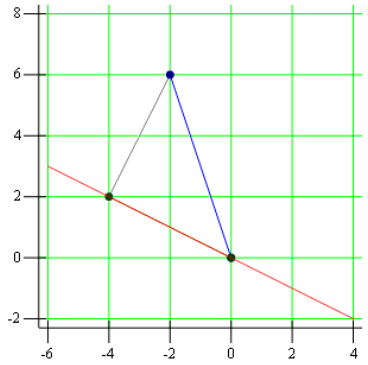


Рис. 6.6 Простір стовбців виродженої матриці системи (6.27), відповідний вектор $\{\mathbf{b}\}$ *та його ортогональна проекція на цей простір*

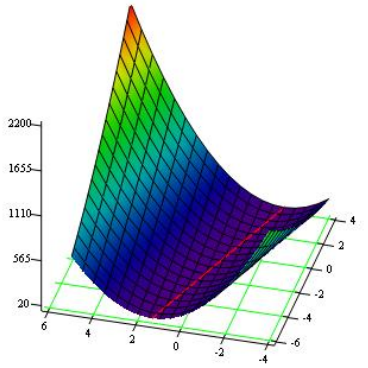


Рис. 6.7 Квадратична форма системи (6.27) та наближене загальне рішення системи, отримане методом найменших квадратів

Чи всі отримані рішення можна вважати однаково корисними? Виявляється, що не завжди. Спробуємо знайти рішення, яке не просто мінімізує квадрат нев'язки (в загальному випадку таких рішень безліч), а рішення яке крім того є мінімальним по своїй довжині. Для його знаходження знову розглянемо простір рядків матриці та її нуль-простір. Нагадаємо, що вони завжди ортогональні (див. наприклад *Рис. 6.4*). Це означає, що будь-який вектор $\{\mathbf{x}\}$ з

безмежної кількості отриманих рішень, може бути розкладений як сума двох ортогональних складових – ортогональної проекції на простір рядків $\{\mathbf{x}_0\}$ і ортогональної проекції на нуль-простір $\{\boldsymbol{\omega}\}$:

$$\{\mathbf{x}\}=\{\mathbf{x}_0\}+\{\boldsymbol{\omega}\},\quad \{\mathbf{x}_0\}\in\operatorname{span}\left(\operatorname{rows}\left([\mathbf{A}]\right)\right),\quad \{\boldsymbol{\omega}\}\in\ker\left([\mathbf{A}]\right). \tag{6.32}$$

Довжина такого рішення, згідно теореми Піфагора, рівна:

$$\left\|\{\mathbf{x}\}\right\|^2_{\mathcal{L}_2(\Omega)}=\left\|\{\mathbf{x}_0\}+\{\boldsymbol{\omega}\}\right\|^2_{\mathcal{L}_2(\Omega)}=\left\|\{\mathbf{x}_0\}\right\|^2_{\mathcal{L}_2(\Omega)}+\left\|\{\boldsymbol{\omega}\}\right\|^2_{\mathcal{L}_2(\Omega)}. \tag{6.33}$$

Зверніть увагу, що $\{\mathbf{x}_0\}$ також мінімізує квадрат нев'язки, оскільки належить до загальних рішень системи. Мінімальним за довжиною, буде рішення, для якого $\left\|\{\boldsymbol{\omega}\}\right\|^2_{\mathcal{L}_2(\Omega)}=0$, тобто $\{\mathbf{x}_0\}$. Воно єдине найближче до початку координат. Щоб його знайти, використовують псевдообернену матрицю:

$$\{\mathbf{x}_0\}=[\mathbf{A}]^{\dagger}\{\mathbf{b}\}. \tag{6.34}$$

Ця псевдообернена матриця фактично здійснює дві операції: будує проекцію вектору $\{\mathbf{b}\}$ на просторі стовпців матриці, тобто будує $[\mathbf{A}]\{\mathbf{x}\}$; та вибирає таке рішення $\{\mathbf{x}_0\}$, яке єдине з усіх можливих належить ще й простору рядків матриці. Знайдене рішення є мінімальним за довжиною. Іншими словами, воно є оптимальним за методом найменших квадратів наближеним рішенням несумісної системи.

6.4. Методи знаходження псевдообернених матриць

Якщо матриця $[\mathbf{A}]$ прямокутна, без лінійно залежних рядків чи стовпців, псевдообернену матрицю $[\mathbf{A}]^{\dagger}$ можна знайти з допомогою границі [27], [29]:

$$[\mathbf{A}]^{\dagger}=\lim_{\delta\to 0}[\mathbf{A}]^{*}\left([\mathbf{A}][\mathbf{A}]^{*}+\delta[\mathbf{E}_{m\times m}]\right)^{-1}=\lim_{\delta\to 0}\left([\mathbf{A}]^{*}[\mathbf{A}]+\delta[\mathbf{E}_{n\times n}]\right)^{-1}[\mathbf{A}]^{*}, \tag{6.35}$$

де, m і n позначають кількість рядків і стовпців матриці $[\mathbf{A}]$ відповідно. Тобто:

$$[\mathbf{A}]^{\dagger}=\begin{cases}[\mathbf{A}]^{*}\left([\mathbf{A}][\mathbf{A}]^{*}\right)^{-1}, & \operatorname{rank}\left([\mathbf{A}]\right)=n,\\ \left([\mathbf{A}]^{*}[\mathbf{A}]\right)^{-1}[\mathbf{A}]^{*}, & \operatorname{rank}\left([\mathbf{A}]\right)=m.\end{cases} \tag{6.36}$$

Хоча б один з цих варіантів обов'язково існує. У першому випадку, отримаємо так звану праву обернену до $[\mathbf{A}]$ матрицю, звідки $[\mathbf{A}][\mathbf{A}]^{\dagger}=[\mathbf{E}]$. В другому випадку, навпаки – отримаємо так звану ліву обернену до $[\mathbf{A}]$ матрицю, звідки $[\mathbf{A}]^{\dagger}[\mathbf{A}]=[\mathbf{E}]$.

У загальному випадку, можливим варіантом знаходження псевдообернених матриць є використання ітераційних методів їх обчислення [25], [26]. Аналогічно до методів рішення систем лінійних рівнянь чи методів оптимізації, вони базуються на методах градієнтного спуску у функціональних просторах. Одним з таких методів, є метод *скалярної корекції* (Scalar correction, SC) детально описаний в [26].

Він базується на градієнтному методі *найскорішого спуску* (Steepest descent, SD) [20], що ітераційно шукає наближене рішення за схемою:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \tag{6.37}$$

де α_k – крок, при якому значення $f(x_k - \alpha_k \nabla f(x_k))$ є мінімальним. Кожна ітерація наближає деяке початкове значення у напрямку, протилежному до зростання функції, і в результаті отримується її локальний мінімум. Метод також може бути застосований для рішення систем з невизначеним оператором чи матрицею, тобто, для пошуку сідельних точок, що робить його в деякій мірі універсальним.

В матричній формі, задача полягає у знаходженні мінімуму квадратичної форми, тому можна використати рівняння типу (6.5). На відміну від класичного методу найскорішого спуску, крім іншого, метод скалярної корекції представляє ітераційну схему, що збігається до наближеного рішення нормального рівняння типу (6.24). Але, шуканим рішенням тепер є не вектор, а матриця (псевдообернена):

$$[\mathbf{A}]^{\mathbf{T}}[\mathbf{A}][\mathbf{X}] = [\mathbf{A}]^{\mathbf{T}}[\mathbf{E}] \quad \Rightarrow \quad [\mathbf{A}]^{\mathbf{T}}\left([\mathbf{A}][\mathbf{X}] - [\mathbf{E}]\right) = 0, \tag{6.38}$$

звідки:

$$\begin{gathered} \mathcal{F}(\mathbf{X}) = \frac{1}{2}[\mathbf{X}]^{\mathbf{T}}[\mathbf{A}]^{\mathbf{T}}[\mathbf{A}][\mathbf{X}] - [\mathbf{X}]^{\mathbf{T}}[\mathbf{A}]^{\mathbf{T}}[\mathbf{E}], \\ \frac{\partial \mathcal{F}(\mathbf{X})}{\partial [\mathbf{X}]^{\mathbf{T}}} = [\mathbf{A}]^{\mathbf{T}}\left([\mathbf{A}][\mathbf{X}] - [\mathbf{E}]\right) = 0. \end{gathered} \tag{6.39}$$

Останнє рівняння є градієнтом та виражає шуканий мінімум квадратичної форми. Тому, можна записати ітераційну схему:

$$[\mathbf{X}_{k+1}] = [\mathbf{X}_k] - \gamma_k [\mathbf{A}]^{\mathbf{T}}\left([\mathbf{A}][\mathbf{X}_k] - [\mathbf{E}]\right), \tag{6.40}$$

де γ_k – крок, при якому значення $\mathcal{F}(\mathbf{X}_{k+1})$ є мінімальним. Оскільки $[\mathbf{A}]^{\mathbf{T}}[\mathbf{A}]$ завжди симетрична, γ_k для методу найскорішого спуску, можна знайти як:

$$\gamma_k = \frac{\left\| [\mathbf{A}]^{\mathbf{T}}\left([\mathbf{A}][\mathbf{X}_k] - [\mathbf{E}]\right) \right\|_F}{\left\| [\mathbf{A}][\mathbf{A}]^{\mathbf{T}}\left([\mathbf{A}][\mathbf{X}_k] - [\mathbf{E}]\right) \right\|_F}, \tag{6.41}$$

де $\| . \|_F$ – норма Фробеніуса:

$$\|[\mathbf{A}]\|_F = \sqrt{\langle [\mathbf{A}], [\mathbf{A}] \rangle_F} = \sqrt{\mathrm{tr}\left([\mathbf{A}]\right)} = \sqrt{\sum_i \sum_j \left|[\mathbf{A}]_{i,j}\right|^2}, \tag{6.42}$$

(модуль означає довжину, у випадку використання комплексних чисел).

Тепер, якщо вибирати початкове наближення $[\mathbf{X}_0]$ так, щоб воно належало простору рядків матриці $[\mathbf{A}]$, наприклад $[\mathbf{X}_0] = [\mathbf{A}]^{\mathbf{T}}$, отримана схема повинна привести до наближеного за методом найменших квадратів рішення нормальної системи, тобто до $[\mathbf{A}]^{\dagger}[\mathbf{E}] = [\mathbf{A}]^{\dagger}$. Питання оптимального вибору початкового наближення та його вплив на збіжність методу досліджується в [25].

Останньою відмінністю методу скалярної корекції, від наведеної схеми, є специфічний вибір кроку γ_k, що забезпечує монотонність зміни градієнту

квадратичної форми, при пошуку рішення. Позначимо:

$$\begin{aligned}[\mathbf{G}_k]&=[\mathbf{A}]^{\mathrm{T}}([\mathbf{A}][\mathbf{X}_k]-[\mathbf{E}]),\\ [\mathbf{S}_k]&=[\mathbf{X}_{k+1}]-[\mathbf{X}_k],\\ [\mathbf{Y}_k]&=[\mathbf{G}_{k+1}]-[\mathbf{G}_k],\\ [\mathbf{R}_k]&=[\mathbf{S}_k]-\gamma_k[\mathbf{Y}_k].\end{aligned} \tag{6.43}$$

На кожному кроці нове значення γ_{k+1} шукається як:

$$\gamma_{k+1}=\begin{cases}\dfrac{\langle[\mathbf{S}_k],[\mathbf{R}_k]\rangle_F}{\langle[\mathbf{Y}_k],[\mathbf{R}_k]\rangle_F}, & \langle[\mathbf{Y}_k],[\mathbf{R}_k]\rangle_F>0,\\ \dfrac{\|[\mathbf{S}_k]\|_F}{\|[\mathbf{Y}_k]\|_F}, & \langle[\mathbf{Y}_k],[\mathbf{R}_k]\rangle_F\le 0.\end{cases} \tag{6.44}$$

Критерієм зупинки роботи ітераційної схеми, можна вибрати вираз типу:

$$\|[\mathbf{S}_k]\|_F=\|[\mathbf{X}_{k+1}]-[\mathbf{X}_k]\|_F\le\varepsilon. \tag{6.45}$$

Підсумовуючи результати, отримаємо загальний алгоритм:

Алгоритм методу скалярної корекції для наближеного пошуку псевдооберненої матриці

Вхідні дані: Матриця $[\mathbf{A}]$;
матриця початкового наближення $[\mathbf{X}_0]$ (можна прийняти $[\mathbf{X}_0]=[\mathbf{A}]^{\mathrm{T}}$);
константа $0<\varepsilon<<1$;
константа $0<\xi_1<<2(1-\varepsilon)\big/\|[\mathbf{A}]\|_F^2$;
максимальна кількість ітерацій N.

1: Прийняти $k=0$;
прийняти $\gamma_k=1$;
обчислити $[\mathbf{G}_k]=[\mathbf{A}]^{\mathrm{T}}([\mathbf{A}][\mathbf{X}_k]-[\mathbf{E}])$;

2: Обчислити $[\mathbf{X}_{k+1}]=[\mathbf{X}_k]-\gamma_k[\mathbf{G}_k]$;
обчислити $[\mathbf{S}_k]=[\mathbf{X}_{k+1}]-[\mathbf{X}_k]$.

3: Якщо $\|[\mathbf{S}_k]\|_F\le\varepsilon$, перейти до кроку 8.

4: Обчислити $[\mathbf{G}_{k+1}]=[\mathbf{A}]^{\mathrm{T}}([\mathbf{A}][\mathbf{X}_{k+1}]-[\mathbf{E}])$;
обчислити $[\mathbf{Y}_k]=[\mathbf{G}_{k+1}]-[\mathbf{G}_k]$;
обчислити $[\mathbf{R}_k]=[\mathbf{S}_k]-\gamma_k[\mathbf{Y}_k]$.

5: Обчислити γ_{k+1} з допомогою (6.44);
обчислити $\xi_2^{(k+1)}=2(1-\varepsilon)\dfrac{\|[\mathbf{G}_{k+1}]\|_F^2}{\|[\mathbf{A}][\mathbf{G}_{k+1}]\|_F^2}$;
якщо $\gamma_{k+1}<\xi_1$ або $\gamma_{k+1}>\xi_2^{(k+1)}$, то $\gamma_{k+1}=\xi_2^{(k+1)}$.

6: Якщо $(k+1)\ge N$, перейти до кроку 8.

7: Прийняти $k=k+1$;
перейти до кроку 2.

8: Повернути $[\mathbf{X}_{k+1}]$.

Вихідні дані: Наближена псевдообернена матриця $[\mathbf{A}]^{\dagger}$.

Класичний метод найскорішого спуску приводить до наближеного рішення за $O(kn^2)$ операцій, де кількість ітерацій k зазвичай менша за n. Оскільки в даному випадку, рішенням є матриця, а не вектор, воно буде отримано за $O(kn^3)$ операцій, при умові, що складність процедури множення двох матриць рівна $O(n^3)$. Однак, останню процедуру можна реалізувати паралельно, та зменшити загальну складність алгоритму. Також дає змогу зменшити складність алгоритму розрідженість вхідних даних.

У якості прикладу знаходження псевдообернені матриці, застосуємо алгоритм скалярної корекції до матриці з системи (6.27):

Вхідні дані: $[\mathbf{A}]=\begin{bmatrix}2 & 6\\ -1 & -3\end{bmatrix}$; $[\mathbf{X}_0]=[\mathbf{A}]^{\mathrm{T}}=\begin{bmatrix}2 & -1\\ 6 & -3\end{bmatrix}$; $\varepsilon=10^{-2}$;

$\xi_1=\varepsilon\cdot\dfrac{2(1-\varepsilon)}{\|[\mathbf{A}]\|_F^2}=3{,}96\cdot10^{-4}$; $N=100$.

1: $k=0$; $\gamma_0=1$; $[\mathbf{G}_0]=[\mathbf{A}]^{\mathrm{T}}([\mathbf{A}][\mathbf{X}_0]-[\mathbf{E}])=\begin{bmatrix}98 & -49\\ 294 & -147\end{bmatrix}$.

Ітерація 0, 2: $[\mathbf{X}_1]=[\mathbf{X}_0]-\gamma_0[\mathbf{G}_0]=\begin{bmatrix}-96 & 48\\ -288 & 144\end{bmatrix}$; $[\mathbf{S}_0]=[\mathbf{X}_1]-[\mathbf{X}_0]=\begin{bmatrix}-98 & 49\\ -294 & 147\end{bmatrix}$.

3: $\|[\mathbf{S}_0]\|_F=346{,}482323>\varepsilon$

4: $[\mathbf{G}_1]=[\mathbf{A}]^{\mathrm{T}}([\mathbf{A}][\mathbf{X}_1]-[\mathbf{E}])=\begin{bmatrix}-4802 & 2401\\ -14406 & 7203\end{bmatrix}$;

$[\mathbf{Y}_0]=[\mathbf{G}_1]-[\mathbf{G}_0]=\begin{bmatrix}-4900 & 2450\\ -14700 & 7350\end{bmatrix}$; $[\mathbf{R}_0]=[\mathbf{S}_0]-\gamma_0[\mathbf{Y}_0]=\begin{bmatrix}4802 & -2401\\ 14406 & -7203\end{bmatrix}$.

5: $\langle[\mathbf{Y}_0],[\mathbf{R}_0]\rangle_F=-2{,}941225\cdot10^{8}>0 \;\Rightarrow\; \gamma_1=\dfrac{\langle[\mathbf{S}_0],[\mathbf{R}_0]\rangle_F}{\langle[\mathbf{Y}_0],[\mathbf{R}_0]\rangle_F}=0{,}02;$

$\xi_2^{(1)}=2(1-\varepsilon)\dfrac{\|[\mathbf{G}_1]\|_F^2}{\|[\mathbf{A}][\mathbf{G}_1]\|_F^2}=0{,}0396;\quad(\gamma_1>\xi_1)\vee(\gamma_1<\xi_2^{(1)})\;\Rightarrow\;\gamma_1=0{,}02.$

6: $(k+1)<N$.

7: $k=1$.

Ітерація 1, 2: $[\mathbf{X}_2]=[\mathbf{X}_1]-\gamma_1[\mathbf{G}_1]=\begin{bmatrix}0{,}04 & -0{,}02\\ 0{,}12 & -0{,}06\end{bmatrix}$; $[\mathbf{S}_1]=[\mathbf{X}_2]-[\mathbf{X}_1]=\begin{bmatrix}96{,}04 & -48{,}02\\ 288{,}12 & -144{,}06\end{bmatrix}$.

3: $\|[\mathbf{S}_1]\|_F=339{,}552676>\varepsilon$

4: $[\mathbf{G}_2]=[\mathbf{A}]^{\mathrm{T}}([\mathbf{A}][\mathbf{X}_2]-[\mathbf{E}])=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$; $[\mathbf{Y}_1]=[\mathbf{G}_2]-[\mathbf{G}_1]=\begin{bmatrix}4802 & -2401\\ 14406 & -7203\end{bmatrix}$;

$[\mathbf{R}_1]=[\mathbf{S}_1]-\gamma_1[\mathbf{Y}_1]=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$.

5: $\langle[\mathbf{Y}_1],[\mathbf{R}_1]\rangle_F=0 \;\Rightarrow\; \gamma_2=\dfrac{\|[\mathbf{S}_1]\|_F}{\|[\mathbf{Y}_1]\|_F}=0{,}02;\quad \xi_2^{(2)}=2(1-\varepsilon)\dfrac{\|[\mathbf{G}_2]\|_F^2}{\|[\mathbf{A}][\mathbf{G}_2]\|_F^2}=0{,}0396;$

$(\gamma_2 > \xi_1) \vee (\gamma_2 < \xi_2^{(2)}) \;\Rightarrow\; \gamma_2 = 0{,}02.$

6: $(k+1) < N.$

7: $k = 2.$

Ітерація 2, 2: $[\mathbf{X}_3] = [\mathbf{X}_2] - \gamma_2[\mathbf{G}_2] = \begin{bmatrix} 0{,}04 & -0{,}02 \\ 0{,}12 & -0{,}06 \end{bmatrix}$; $[\mathbf{S}_2] = [\mathbf{X}_3] - [\mathbf{X}_2] = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$.

3: $\|[\mathbf{S}_1]\|_F = 0 < \varepsilon \;\Rightarrow$ перейти до кроку 8.

8: Повернути $[\mathbf{X}_3]$.

Підставляючи знайдену псевдообернену матрицю в рівняння (6.34), отримаємо оптимальний по довжині наближений за методом найменших квадратів розв'язок несумісної системи (6.27) $\{\mathbf{x}_0\} = [\mathbf{A}]^{\dagger}\{\mathbf{b}\} = \{-0{,}2 \quad -0{,}6\}^{\mathbf{T}}$. З *Рис. 6.8* видно, що це рішення є єдиним, яке:

- найближче до початку координат, тобто має мінімальну довжину;
- належить одночасно множині загальних рішень системи та простору рядків матриці.

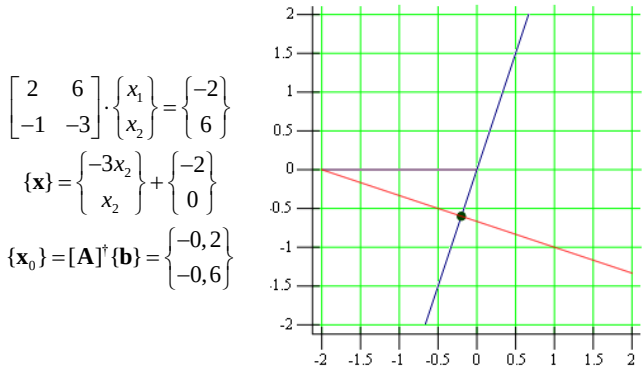


Рис. 6.8 Простір рядків матриці несумісної системи; її часткове рішення, що зміщує нуль-простір матриці; множина загальних рішень несумісної системи; її оптимальний по довжині наближений за методом найменших квадратів розв'язок

Наведений спосіб обчислення псевдообернених матриць не єдиний. Дуже часто, замість нього використовують *сингулярний розклад матриці* (singular value decomposition, SVD) [24], [27], [28], [30]. Будь-яка матриця $[\mathbf{A}_{m\times n}]$ може бути розкладена у вигляді:

$$[\mathbf{A}] = [\mathbf{Q}_1][\mathbf{\Sigma}][\mathbf{Q}_2]^{\mathbf{T}}, \qquad (6.46)$$

де $[\mathbf{Q}_1]$ – *унітарна* матриця розміру $m \times m$; $[\mathbf{Q}_2]$ – *унітарна* матриця розміру $n \times n$; $[\mathbf{\Sigma}]$ – діагональна матриця розміру $m \times n$ (елементами якої є так звані *сингулярні числа*). Під унітарною матрицею (іноді також можна зустріти назву ортогональна, або ортонормована) розуміють матрицю, що має ортонормовані рядки та стовпці – кожен рядок або стовпець ортогональний всім іншим рядкам

або стовпцям відповідно. Іншими словами, добуток унітарної матриці і транспонованої до неї завжди рівний одиничній матриці:

$$[\mathbf{Q}_1]^{\mathbf{T}}[\mathbf{Q}_1]=[\mathbf{Q}_1][\mathbf{Q}_1]^{\mathbf{T}}=[\mathbf{Q}_2]^{\mathbf{T}}[\mathbf{Q}_2]=[\mathbf{Q}_2][\mathbf{Q}_2]^{\mathbf{T}}=[\mathbf{E}]. \qquad (6.47)$$

З виразу (6.46) знаходимо, що:

$$[\mathbf{A}]^{\dagger}=[\mathbf{Q}_2][\mathbf{\Sigma}]^{\dagger}[\mathbf{Q}_1]^{\mathbf{T}}, \qquad (6.48)$$

де, в силу діагональності $[\mathbf{\Sigma}]$, псевдообернену матрицю $[\mathbf{\Sigma}]^{\dagger}$ можна знайти просто обернувши всі ненульові елементи:

$$[\mathbf{\Sigma}_{m\times m}]=\begin{bmatrix}\mu_{1,1} & 0 & \cdots & 0\\ 0 & \mu_{2,2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \mu_{m,m}\end{bmatrix} \Rightarrow [\mathbf{\Sigma}_{m\times m}]^{\dagger}=\begin{bmatrix}1/\mu_{1,1} & 0 & \cdots & 0\\ 0 & 1/\mu_{2,2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1/\mu_{m,m}\end{bmatrix}, \qquad (6.49)$$

$$\mu_{i,i}>\varepsilon, \quad i=1,2,\ldots,m,$$

де ε – довільно вибране мале додатне число.

Існує декілька алгоритмів побудови сингулярного розкладу матриць. Зазвичай вони вже реалізовані в пакетах прикладного програмного забезпечення для задач моделювання, чи у програмних бібліотеках високорівневих алгоритмічних мов. Цікавий читач може ознайомитися з цими алгоритмами зокрема в [28] або [30]. В силу своєї складності та громіздкості, тут вони не наводяться. За необхідності, детальний опис одного з найбільш поширених – алгоритму *Голуба-Кахана-Рейнча*[1,2] (Golub Kahan Reinsch algorithm) можна знайти, наприклад в [30]. Алгоритм дає наближений SVD-розклад. Він приводить до нього за $O(n^3)$ операцій, тобто швидше, ніж метод скалярної корекції.

Ще одним поширеним способом є використання QR-розкладу [24], [25], [28], [30]. Насправді, QR-розклад часто використовується в деяких алгоритмах побудови SVD-розкладу, зокрема в зазначеному алгоритмі Голуба-Кахана-Рейнча. Порівняння основних методів знаходження псевдообернених матриць, а також деякі інші методи, можна знайти зокрема в [31].

Повернемося до нашої початкової мети – знайти сімейство узагальнено обернених матриць, які підходять для методів FETI. Очевидно, що псевдообернена матриця та всі вищенаведені методи її побудови підходять у загальному випадку. Однак, існує частковий випадок, що значно спрощує задачу. Насправді, при симетричності та позитивній визначеності матриць жорсткості, для реалізації FETI методів достатньо знайти ліву обернену матрицю, тобто матрицю, що обов'язково відповідає тільки першому рівнянню Пенроуза (6.12) [32]. Таку матрицю прийнято називати {1}-*оберненням* ({1}-inverse) [25], [26]. Для вироджених матриць {1}-обернення зазвичай не єдине.

[1] Golub G., Kahan W. – Calculating the singular values and pseudo-inverse of a matrix // SIAM J. Num. Anal. Ser. B 2, pp. 205-224, 1965.

[2] Golub G., Reinsch C. – Singular value decomposition and least squares solutions // Numer. Math. 14, pp. 403-420, 1970; HACLA, pp. 134-151, 1970.

Одну з таких матриць можна знайти за допомогою розкладу:

$$[\mathbf{A}]=[\mathbf{P}]\begin{bmatrix}[\mathbf{A}_{11}] & [\mathbf{A}_{12}]\\ [\mathbf{A}_{21}] & [\mathbf{A}_{22}]\end{bmatrix}[\mathbf{P}]^{\mathbf{T}}, \tag{6.50}$$

де $[\mathbf{P}]$ – матриця перестановок; $[\mathbf{A}_{11}]$ – елементи матриці $[\mathbf{A}]$, що лежать на перетині лінійно незалежних рядків і стовпців; $[\mathbf{A}_{22}]$ – елементи матриці $[\mathbf{A}]$, що лежать на перетині лінійно залежних рядків і стовпців; $[\mathbf{A}_{12}]$ та $[\mathbf{A}_{21}]$– елементи матриці $[\mathbf{A}]$, що залишилися. {1}-обернена матриця будується як:

$$[\mathbf{A}]^{\{1\}}=[\mathbf{P}]\begin{bmatrix}[\mathbf{A}_{11}]^{-1} & [\mathbf{0}]\\ [\mathbf{0}] & [\mathbf{0}]\end{bmatrix}[\mathbf{P}]^{\mathbf{T}}. \tag{6.51}$$

Для побудови розкладу (6.50) використовують *розклад Холецького* (Cholesky factorization) [30], [33]:

$$[\mathbf{A}]=[\mathbf{L}][\mathbf{L}]^{\mathbf{T}}, \tag{6.52}$$

де $[\mathbf{L}]$ – нижня трикутна матриця. Щоб знайти лінійно-незалежні рядки, достатньо занулювати рядки $[\mathbf{L}]$ з нульовим діагональним елементом. Опис даного алгоритму можна знайти наприклад в [33].

6.5. Рішення систем методу скінченних елементів розривів і з'єднань

Спробуємо виразити загальне рішення системи через псевдообернену матрицю. Можна зауважити, що вираз $[\mathbf{A}][\mathbf{A}]^{\dagger}$ є оператором ортогональної проекції на простір стовпців, оскільки:

$$[\mathbf{A}][\mathbf{A}]^{\dagger}\{\mathbf{b}\}=[\mathbf{A}]\{\mathbf{x}_0\}. \tag{6.53}$$

Аналогічно до цього, вираз $[\mathbf{A}]^{\dagger}[\mathbf{A}]$ буде оператором ортогональної проекції на простір рядків (див. (6.12)). Щоб знайти проектор на нуль-простір матриці, який є ортогональним до простору рядків, достатньо використати $[\mathbf{E}]-[\mathbf{A}]^{\dagger}[\mathbf{A}]$, де $[\mathbf{E}]$ – одинична матриця [23], [24].

Для довільного вектору $\{\boldsymbol{\alpha}\}$, його ортогональна проекція в нуль-простір будується як $\left([\mathbf{E}]-[\mathbf{A}]^{\dagger}[\mathbf{A}]\right)\{\boldsymbol{\alpha}\}$. Останній вираз є ніщо інше, ніж компонента $\{\boldsymbol{\omega}\}$ з системи (6.32).

Тепер, можна записати загальне рішення несумісної системи, як суму часткового рішення та рішень однорідної системи, з використанням псевдообернені матриці:

$$\begin{aligned}
&[\mathbf{A}]\{\mathbf{x}\}=\{\mathbf{b}\},\quad [\mathbf{A}]\{\mathbf{x}\}=\{\mathbf{0}\},\\
&[\mathbf{A}]\{\mathbf{x}\}=\{\mathbf{b}\}+\{\mathbf{0}\},\\
&[\mathbf{A}]\{\mathbf{x}\}=[\mathbf{A}][\mathbf{A}]^{\dagger}\{\mathbf{b}\}+[\mathbf{A}]\left([\mathbf{E}]-[\mathbf{A}]^{\dagger}[\mathbf{A}]\right)\{\boldsymbol{\alpha}\},\\
&\{\mathbf{x}\}=[\mathbf{A}]^{\dagger}\{\mathbf{b}\}+\left([\mathbf{E}]-[\mathbf{A}]^{\dagger}[\mathbf{A}]\right)\{\boldsymbol{\alpha}\}.
\end{aligned} \tag{6.54}$$

Отримані, перший та другий доданки це $\{\mathbf{x}_0\}$ та $\{\boldsymbol{\omega}\}$ з системи (6.32).

Застосуємо отриманий вираз до системи (6.9), попередньо знайшовши лінійно незалежні стовпці $[\mathbf{E}]-[\mathbf{K}]^{\dagger}[\mathbf{K}]$ та позначивши їх як $[\mathbf{R}]$. Останню матрицю також називають оператором *обмеження* (restriction operator). З першого рівняння отримаємо:

$$\begin{aligned}&[\mathbf{K}]\{\mathbf{u}\}-\{\mathbf{f}\}+[\mathbf{B}]^{\mathbf{T}}\{\boldsymbol{\lambda}\}=\{\mathbf{0}\},\\&\{\mathbf{u}\}=[\mathbf{K}]^{\dagger}\left(\{\mathbf{f}\}-[\mathbf{B}]^{\mathbf{T}}\{\boldsymbol{\lambda}\}\right)+[\mathbf{R}]\{\boldsymbol{\alpha}\}.\end{aligned} \tag{6.55}$$

Перенесемо всі доданки в ліву сторону та розкриємо дужки:

$$\begin{aligned}&[\mathbf{K}]^{\dagger}\left(\{\mathbf{f}\}-[\mathbf{B}]^{\mathbf{T}}\{\boldsymbol{\lambda}\}\right)+[\mathbf{R}]\{\boldsymbol{\alpha}\}-\{\mathbf{u}\}=\{\mathbf{0}\},\\&[\mathbf{K}]^{\dagger}\{\mathbf{f}\}-[\mathbf{K}]^{\dagger}[\mathbf{B}]^{\mathbf{T}}\{\boldsymbol{\lambda}\}+[\mathbf{R}]\{\boldsymbol{\alpha}\}-\{\mathbf{u}\}=\{\mathbf{0}\}.\end{aligned} \tag{6.56}$$

Помножимо обидві частини отриманого виразу на $[\mathbf{B}]$:

$$[\mathbf{B}][\mathbf{K}]^{\dagger}\{\mathbf{f}\}-[\mathbf{B}][\mathbf{K}]^{\dagger}[\mathbf{B}]^{\mathbf{T}}\{\boldsymbol{\lambda}\}+[\mathbf{B}][\mathbf{R}]\{\boldsymbol{\alpha}\}-[\mathbf{B}]\{\mathbf{u}\}=[\mathbf{B}]\{\mathbf{0}\}. \tag{6.57}$$

Беручи до уваги останнє рівняння системи (1.4), тобто $[\mathbf{B}]\{\mathbf{u}\}=\{\mathbf{0}\}$, а також те, що $[\mathbf{B}]\{\mathbf{0}\}=\{\mathbf{0}\}$, отримаємо:

$$[\mathbf{B}][\mathbf{K}]^{\dagger}\{\mathbf{f}\}-[\mathbf{B}][\mathbf{K}]^{\dagger}[\mathbf{B}]^{\mathbf{T}}\{\boldsymbol{\lambda}\}+[\mathbf{B}][\mathbf{R}]\{\boldsymbol{\alpha}\}=\{\mathbf{0}\}. \tag{6.58}$$

Перепишемо це рівняння як:

$$[\mathbf{B}][\mathbf{K}]^{\dagger}[\mathbf{B}]^{\mathbf{T}}\{\boldsymbol{\lambda}\}-[\mathbf{B}][\mathbf{R}]\{\boldsymbol{\alpha}\}=[\mathbf{B}][\mathbf{K}]^{\dagger}\{\mathbf{f}\}. \tag{6.59}$$

Оскільки всі отримані змінні були виражені з першого рівняння (1.4), знову ж таки, воно, і як наслідок, останнє рівняння, є сумісними тоді і тільки тоді, коли вектор $\{\mathbf{f}\}+[\mathbf{B}]^{\mathbf{T}}\{\boldsymbol{\lambda}\}$ належить простору стовпців матриці жорсткості $[\mathbf{K}]$. Раніше було показано, що дане твердження, це те саме, що ортогональність вектору до нуль-простору матриці жорсткості:

$$\left(\{\mathbf{f}\}+[\mathbf{B}]^{\mathbf{T}}\{\boldsymbol{\lambda}\}\right)\perp\ker\left([\mathbf{K}]\right). \tag{6.60}$$

Нагадаємо, що проектор на нуль-простір матриці жорсткості, це $[\mathbf{E}]-[\mathbf{K}]^{\dagger}[\mathbf{K}]$, звідки:

$$\begin{aligned}&[\mathbf{R}]^{\mathbf{T}}\left(\{\mathbf{f}\}-[\mathbf{B}]^{\mathbf{T}}\{\boldsymbol{\lambda}\}\right)=0,\\&-[\mathbf{R}]^{\mathbf{T}}[\mathbf{B}]^{\mathbf{T}}\{\boldsymbol{\lambda}\}=-[\mathbf{R}]^{\mathbf{T}}\{\mathbf{f}\},\\&\left(-[\mathbf{B}][\mathbf{R}]\right)^{\mathbf{T}}\{\boldsymbol{\lambda}\}=-[\mathbf{R}]^{\mathbf{T}}\{\mathbf{f}\}.\end{aligned} \tag{6.61}$$

Введемо нові позначення:

$$\begin{aligned}&[\mathbf{F}]=[\mathbf{B}][\mathbf{K}]^{\dagger}[\mathbf{B}]^{\mathbf{T}},\quad [\mathbf{G}]=-[\mathbf{B}][\mathbf{R}],\\&\{\mathbf{d}\}=[\mathbf{B}][\mathbf{K}]^{\dagger}\{\mathbf{f}\},\qquad \{\mathbf{e}\}=-[\mathbf{R}]^{\mathbf{T}}\{\mathbf{f}\}.\end{aligned} \tag{6.62}$$

На основі (6.59) та (6.61), систему (6.9) можна переписати у вигляді:

$$\begin{cases}[\mathbf{F}]\{\boldsymbol{\lambda}\}+[\mathbf{G}]\{\boldsymbol{\alpha}\}=\{\mathbf{d}\},\\ [\mathbf{G}]^{\mathbf{T}}\{\boldsymbol{\alpha}\}=\{\mathbf{e}\},\end{cases}\Leftrightarrow\begin{bmatrix}[\mathbf{F}] & [\mathbf{G}]\\ [\mathbf{G}]^{\mathbf{T}} & [\mathbf{0}]\end{bmatrix}\begin{Bmatrix}\{\boldsymbol{\lambda}\}\\ \{\boldsymbol{\alpha}\}\end{Bmatrix}=\begin{Bmatrix}\{\mathbf{d}\}\\ \{\mathbf{e}\}\end{Bmatrix}. \tag{6.63}$$

Отриману систему називають *грубою* (coarse problem) або *інтерфейсною* (interface problem) [13], [34].

Обчисливши незалежно для кожного домену псевдообернену матрицю $[\mathbf{K}]_i^{\dagger}$ та відповідно на її основі $[\mathbf{F}]_i$, $[\mathbf{G}]_i$, $\{\mathbf{d}\}_i$ та $\{\mathbf{e}\}_i$, можна зібрати загальну систему рівнянь (6.63) для змінних $\{\boldsymbol{\lambda}\}$ та $\{\boldsymbol{\alpha}\}$:

$$\begin{aligned}
&[\mathbf{F}] = \sum_{i=1}^{D} [\mathbf{B}]_i [\mathbf{K}]_i^{\dagger} [\mathbf{B}]_i^{\mathbf{T}},\\
&[\mathbf{G}] = \left[-[\mathbf{B}]_1 [\mathbf{R}]_1, \;\; -[\mathbf{B}]_2 [\mathbf{R}]_2, \;\; \ldots, \;\; -[\mathbf{B}]_D [\mathbf{R}]_D \right],\\
&\{\mathbf{d}\} = \sum_{i=1}^{D} [\mathbf{B}]_i [\mathbf{K}]_i^{\dagger} \{\mathbf{f}\}_i,\\
&\{\mathbf{e}\} = \left[\left(-[\mathbf{R}]_1^{\mathbf{T}} \{\mathbf{f}\}_1 \right)^{\mathbf{T}}, \;\; \left(-[\mathbf{R}]_2^{\mathbf{T}} \{\mathbf{f}\}_2 \right)^{\mathbf{T}}, \;\; \ldots, \;\; \left(-[\mathbf{R}]_D^{\mathbf{T}} \{\mathbf{f}\}_D \right)^{\mathbf{T}} \right]^{\mathbf{T}}.
\end{aligned} \tag{6.64}$$

Щоб виконати друге рівняння з (6.63), введемо новий проектор $[\mathbf{P}(\mathbf{Q})]$ на нуль-простір матриці $[\mathbf{G}]^{\mathbf{T}}$:

$$[\mathbf{P}(\mathbf{Q})] = [\mathbf{E}] - [\mathbf{Q}][\mathbf{G}] \left([\mathbf{G}]^{\mathbf{T}} [\mathbf{Q}][\mathbf{G}] \right)^{-1} [\mathbf{G}]^{\mathbf{T}}, \tag{6.65}$$

де $[\mathbf{Q}]$ – це так званий передобумовлювач, тобто матриця, введення якої, має за мету пришвидшити ітераційні процеси пошуку рішення СЛАР, або спростити останню, у випадку використання прямих методів рішення. Доведено [37], що $\left([\mathbf{G}]^{\mathbf{T}} [\mathbf{Q}][\mathbf{G}] \right)^{-1}$ завжди існує, за умови, що оператор $[\mathbf{R}]$ побудований з лінійно незалежних стовпців проектору $[\mathbf{E}] - [\mathbf{K}]^{\dagger} [\mathbf{K}]$.

Для довільних систем лінійних рівнянь типу $[\mathbf{A}]\{\mathbf{x}\} = \{\mathbf{b}\}$, передобумовлювач завжди вибирається так, щоб $[\mathbf{Q}][\mathbf{A}] \to [\mathbf{E}]$. Геометрично, це сильно спрощує квадратичну форму матриці. Очевидно, що найкращим передобуомвлювачем є обернена матриця $[\mathbf{A}]^{-1}$ – ітераційний процес пошуку займатиме єдиний крок. У введених термінах це еквівалентно $[\mathbf{Q}] = [\mathbf{F}]^{-1}$.

Оскільки пошук оберненої матриці є дорогою, в сенсі обчислень, операцією, необхідно йти на компроміс. Якщо рішення шукається без передобумовлювача, можна прийняти $[\mathbf{Q}] = [\mathbf{E}]$.

Іншим можливим варіантом, що використовують на практиці, є передобумовлювач Діріхле:

$$[\mathbf{D}] = [\mathbf{B}][\mathbf{S}][\mathbf{B}]^{\mathbf{T}} = \sum_{i=1}^{D} [\mathbf{B}]_i [\mathbf{S}]_i [\mathbf{B}]_i^{\mathbf{T}}, \tag{6.66}$$

де $[\mathbf{S}]_i$ – розклад Шура (Schur complement) [30] для $[\mathbf{K}]_i$, що будується як:

$$[\mathbf{S}] = [\mathbf{K}_{BB}] - [\mathbf{K}_{BI}][\mathbf{K}_{II}]^{-1}[\mathbf{K}_{IB}]. \tag{6.67}$$

Матриці у правій частині останнього виразу беруться з розкладу $[\mathbf{K}]_i$:

$$[\mathbf{K}] = [\mathbf{P}] \begin{bmatrix} [\mathbf{K}_{BB}] & [\mathbf{K}_{BI}] \\ [\mathbf{K}_{IB}] & [\mathbf{K}_{II}] \end{bmatrix} [\mathbf{P}]^{\mathbf{T}}, \tag{6.68}$$

де $[\mathbf{P}]$ – матриця перестановок (не плутати з $[\mathbf{P}(\mathbf{Q})]$); $[\mathbf{K}_{II}]$ – елементи матриці $[\mathbf{K}]$, що відповідають її внутрішнім вузлам (*I*-interior); $[\mathbf{K}_{BB}]$ – елементи матриці $[\mathbf{K}]$, що відповідають її граничним вузлам (*B*-boundary); $[\mathbf{K}_{BI}]$ та $[\mathbf{K}_{IB}]$ – елементи матриці $[\mathbf{K}]$, що утворюються на перетині рядків та стовбців для граничних і внутрішніх вузлів та навпаки.

Оскільки добуток $[\mathbf{P}(\mathbf{Q})]^{\mathbf{T}}[\mathbf{G}]$ завжди рівний нулю, помноживши (6.63) на $[\mathbf{P}(\mathbf{Q})]^{\mathbf{T}}$, отримаємо нову систему рівнянь, з відсутніми змінними $\{\boldsymbol{\alpha}\}$:

$$\begin{cases}[\mathbf{P}(\mathbf{Q})]^{\mathbf{T}}\left([\mathbf{F}]\{\boldsymbol{\lambda}\}-\{\mathbf{d}\}\right)=\{\mathbf{0}\},\\ [\mathbf{G}]^{\mathbf{T}}\{\boldsymbol{\lambda}\}=\{\mathbf{e}\}.\end{cases} \tag{6.69}$$

Будь-яке рішення цієї системи $\{\boldsymbol{\lambda}\}$ відрізняється від іншого тільки на вектор з нуль-простору $[\mathbf{G}]^{\mathbf{T}}$. Таке рішення задовольняє початкову систему (6.9) та вираз (6.55), при умові:

$$\{\boldsymbol{\alpha}\}=-\left([\mathbf{G}]^{\mathbf{T}}[\mathbf{Q}][\mathbf{G}]\right)^{-1}[\mathbf{G}]^{\mathbf{T}}[\mathbf{Q}]\left([\mathbf{F}]\{\boldsymbol{\lambda}\}-\{\mathbf{d}\}\right). \tag{6.70}$$

Розкладемо вектор Лагранжевих множників як:

$$\{\boldsymbol{\lambda}\}=\{\boldsymbol{\lambda}_0\}+[\mathbf{P}(\mathbf{Q})]\{\tilde{\boldsymbol{\lambda}}\}, \tag{6.71}$$

де $\{\boldsymbol{\lambda}_0\}=[\mathbf{Q}][\mathbf{G}]\left([\mathbf{G}]^{\mathbf{T}}[\mathbf{Q}][\mathbf{G}]\right)^{-1}\{\mathbf{e}\}$ – часткове рішення останнього рівняння з (6.69); $\{\tilde{\boldsymbol{\lambda}}\}$ – деяке загальне рішення, для якого $\{\tilde{\boldsymbol{\lambda}}\}\in\ker\left([\mathbf{G}]^{\mathbf{T}}\right)$. Воно може бути знайдене з першого рівняння (6.69):

$$\begin{gathered}[\mathbf{P}(\mathbf{Q})]^{\mathbf{T}}\left([\mathbf{F}]\left(\{\boldsymbol{\lambda}_0\}+[\mathbf{P}(\mathbf{Q})]\{\tilde{\boldsymbol{\lambda}}\}\right)-\{\mathbf{d}\}\right)=\{\mathbf{0}\},\\ [\mathbf{P}(\mathbf{Q})]^{\mathbf{T}}\left([\mathbf{F}]\{\boldsymbol{\lambda}_0\}+[\mathbf{F}][\mathbf{P}(\mathbf{Q})]\{\tilde{\boldsymbol{\lambda}}\}-\{\mathbf{d}\}\right)=\{\mathbf{0}\},\\ [\mathbf{P}(\mathbf{Q})]^{\mathbf{T}}[\mathbf{F}][\mathbf{P}(\mathbf{Q})]\{\tilde{\boldsymbol{\lambda}}\}=[\mathbf{P}(\mathbf{Q})]^{\mathbf{T}}\{\mathbf{d}\}-[\mathbf{P}(\mathbf{Q})]^{\mathbf{T}}[\mathbf{F}]\{\boldsymbol{\lambda}_0\},\\ \left([\mathbf{P}(\mathbf{Q})]^{\mathbf{T}}[\mathbf{F}][\mathbf{P}(\mathbf{Q})]\right)\{\tilde{\boldsymbol{\lambda}}\}=[\mathbf{P}(\mathbf{Q})]^{\mathbf{T}}\left(\{\mathbf{d}\}-[\mathbf{F}]\{\boldsymbol{\lambda}_0\}\right).\end{gathered} \tag{6.72}$$

Отриману систему можна розв'язати прямим методом. Зокрема, прийнявши $\{\tilde{\boldsymbol{\lambda}}\}=[\mathbf{F}]^{-1}\{\mathbf{d}\}$ та $\{\mathbf{Q}\}=[\mathbf{F}]^{-1}$, її можна звести [35], [36] до:

$$\begin{gathered}\{\boldsymbol{\alpha}\}=\left([\mathbf{G}]^{\mathbf{T}}[\mathbf{F}]^{-1}[\mathbf{G}]\right)^{-1}\left([\mathbf{G}]^{\mathbf{T}}[\mathbf{F}]^{-1}\{\mathbf{d}\}-\{\mathbf{e}\}\right),\\ \{\boldsymbol{\lambda}\}=[\mathbf{F}]^{-1}\left(\{\mathbf{d}\}-[\mathbf{G}]\{\boldsymbol{\alpha}\}\right).\end{gathered} \tag{6.73}$$

З іншою сторони, рішення системи (6.72) зручно шукати ітераційно, за допомогою модифікованого методу *спряжених градієнтів* (conjugate gradients, CG) [20]. Класично цей метод застосовується для пошуку локального мінімуму при симетричному позитивно визначеному операторі чи матриці. Він відрізняється від методу найскорішого спуску тим, що здійснює спуск не в напрямку, протилежному до зростання функції, а в напрямку, що спряжений з напрямком попередньої ітерації. Під спряженістю розуміється рівність нулю енергетичного добутку, тобто:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \langle \mathbf{Ax}, \mathbf{y} \rangle = \{\mathbf{x}\}^{\mathbf{T}}[\mathbf{A}]^{\mathbf{T}}\{\mathbf{y}\} = 0. \qquad (6.74)$$

Такий підхід дозволяє знайти рішення швидше, ніж з допомогою методу найскорішого спуску, за умови симетричності та позитивної визначеності оператора чи матриці.

Алгоритм модифікованого методу спряжених градієнтів для рішення систем методу скінченних елементів розривів і з'єднань

Вхідні дані: Матриці [**F**], [**G**], [**Q**];
вектори {**d**}, {**e**};
константа $0 < \varepsilon << 1$;
максимальна кількість ітерацій N.

1: Обчислити $\{\boldsymbol{\lambda}_0\} = [\mathbf{Q}][\mathbf{G}]\left([\mathbf{G}]^{\mathbf{T}}[\mathbf{Q}][\mathbf{G}]\right)^{-1}\{\mathbf{e}\}$;
обчислити $[\mathbf{P(Q)}] = [\mathbf{E}] - [\mathbf{Q}][\mathbf{G}]\left([\mathbf{G}]^{\mathbf{T}}[\mathbf{Q}][\mathbf{G}]\right)^{-1}[\mathbf{G}]^{\mathbf{T}}$.

2: Прийняти $k = 0$;

3: Обчислити $\{\mathbf{r}_0\} = [\mathbf{P(Q)}]^{\mathbf{T}}\left(\{\mathbf{d}\} - [\mathbf{F}]\{\boldsymbol{\lambda}_0\}\right)$;
обчислити $\{\mathbf{z}_0\} = [\mathbf{P(Q)}]\{\mathbf{r}_0\}$;
прийняти $\{\mathbf{s}_0\} = \{\mathbf{z}_0\}$;
обчислити $\beta_0 = \{\mathbf{r}_0\}^{\mathbf{T}}\{\mathbf{z}_0\}$.

4: Обчислити $\{\mathbf{x}_k\} = [\mathbf{P(Q)}]^{\mathbf{T}}[\mathbf{F}]\{\mathbf{s}_k\}$;
обчислити $\alpha_k = \{\mathbf{x}_k\}^{\mathbf{T}}\{\mathbf{z}_k\}$;
обчислити $\alpha = \beta_k / \alpha_k$;
обчислити $\{\boldsymbol{\lambda}_{k+1}\} = \{\boldsymbol{\lambda}_k\} + \alpha\{\mathbf{s}_k\}$;
обчислити $\{\mathbf{r}_{k+1}\} = \{\mathbf{r}_k\} - \alpha\{\mathbf{x}_k\}$;
обчислити $\{\mathbf{z}_{k+1}\} = [\mathbf{P(Q)}]\{\mathbf{r}_{k+1}\}$;
обчислити $\beta_{k+1} = \{\mathbf{r}_{k+1}\}^{\mathbf{T}}\{\mathbf{z}_{k+1}\}$.

5: Якщо $\beta_{k+1} \geq \beta_0 \cdot \varepsilon$, перейти до кроку 9.

6: Обчислити $\beta = \beta_{k+1} / \beta_k$;
обчислити $\{\mathbf{s}_{k+1}\} = \{\mathbf{z}_{k+1}\} + \beta\{\mathbf{s}_k\}$.

7: Якщо $(k+1) \geq N$, перейти до кроку 9.

8: Прийняти $k = k + 1$;
перейти до кроку 4.

9: Обчислити $\{\boldsymbol{\alpha}\} = -\left([\mathbf{G}]^{\mathbf{T}}[\mathbf{Q}][\mathbf{G}]\right)^{-1}[\mathbf{G}]^{\mathbf{T}}[\mathbf{Q}]\left([\mathbf{F}]\{\boldsymbol{\lambda}_k\} - \{\mathbf{d}\}\right)$;
повернути $\{\boldsymbol{\lambda}_k\}$ та $\{\boldsymbol{\alpha}\}$.

Вихідні дані: Наближені рішення $\{\boldsymbol{\lambda}\}$ та $\{\boldsymbol{\alpha}\}$.

Модифікація методу спряжених градієнтів полягає в проектуванні градієнту в нуль-простір $[\mathbf{G}]^{\mathbf{T}}$, тобто у виконанні на кожній ітерації другого рівняння (6.69). За умови симетричності та позитивної визначеності локальних матриць $[\mathbf{K}]_i$, матриці $[\mathbf{F}]_i$ та глобальна матриця [**F**] також відповідатимуть

цим критеріям[1], тому, при початковому наближенні $\{\boldsymbol{\lambda}_0\}$, в результаті отримаємо наближене рішення системи.

Слід зауважити, що починаючи з (6.55), в разі використання не псевдооберненої $[\mathbf{K}]^{\dagger}$, а {1}-оберненої матриці $[\mathbf{K}]^{\{1\}}$, побудова матриці $[\mathbf{R}]$ зводиться до:

$$[\mathbf{R}]=\begin{bmatrix}-[\mathbf{K}_{11}]^{-1}[\mathbf{K}_{12}]\\ [\mathbf{E}_{22}]\end{bmatrix}. \qquad (6.75)$$

Кількість лінійно незалежних стовпців проектора $[\mathbf{E}]-[\mathbf{K}]^{\dagger}[\mathbf{K}]$ буде рівна кількості степенів свободи у вузлах дискретизації, за умови симетричності та позитивної визначеності оператора чи матриці задачі. Наприклад, для задачі стаціонарної теплопровідності вона буде рівна одиниці, звідки оператор $[\mathbf{R}]$ перетворюється у вектор, заповнений одиницями. Це пояснюється також тим, що в матрицях жорсткості доменів цієї задачі, де не задані крайові умови Діріхле, завжди присутній один лінійно залежний стовпець та рядок. Тому, для знаходження {1}-оберненої матриці за (6.51), у якості $[\mathbf{K}_{11}]$ можна взяти матрицю $[\mathbf{K}]$ без останнього рядка та стовпця.

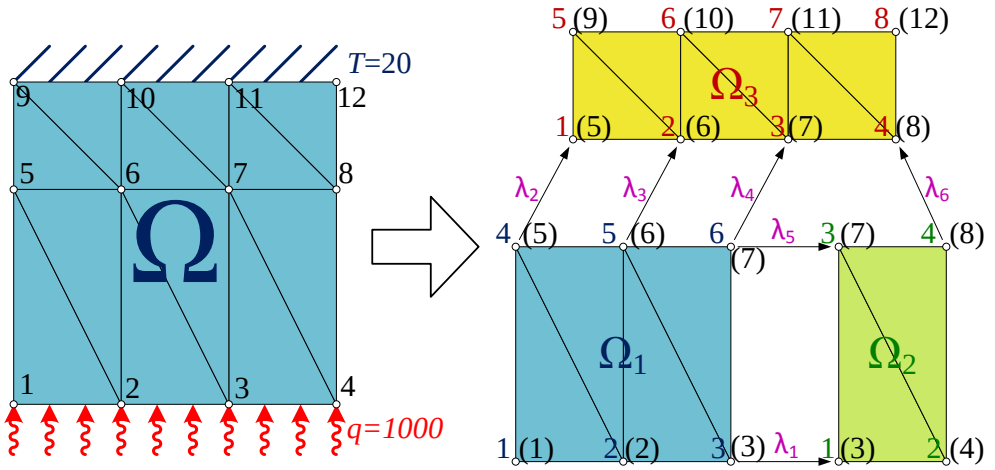


Рис. 6.9 Зображення умов двовимірної задачі стаціонарної теплопровідності, дискретизації області скінченними елементами та її декомпозиції на три домени

Підсумуємо даний розділ прикладом застосування методу скінченних елементів розривів і з'єднань для рішення задачі стаціонарної теплопровідності у двовимірному випадку. Нехай коефіцієнт теплопровідності матеріалу $\eta = 250$ Вт/м°С (класичне позначення змінено, щоб не плутати з Лагранжевими множниками), матеріал займає квадратну область $0 \le x, y \le 0{,}03$ м. На стороні

[1] У протилежному випадку можна використати іншу модифікацію – метод біспряжених градієнтів.

$y = 0,03$ підтримується постійна температура 20°C, тоді як через сторону $y = 0$ подається тепло з швидкістю 10^5 Вт/м^2°C на одиницю довжини (*Рис. 6.9*).

Побудуємо дискретизацію області регулярною сіткою, так, як це зображено на *Рис. 6.9*. "Розірвемо" побудовану дискретизацію на три домени різних розмірів та "з'єднаємо" їх за допомогою Лагранжевих множників.

Матриці координат вузлів, індексів елементів та оператора стрибку, для першого домену рівні:

$$Crds_1 = \begin{bmatrix} 0,00 & 0,00 \\ 0,01 & 0,00 \\ 0,02 & 0,00 \\ 0,00 & 0,02 \\ 0,01 & 0,02 \\ 0,02 & 0,02 \end{bmatrix}, \quad Elms_1 = \begin{bmatrix} 0 & 1 & 3 \\ 1 & 4 & 3 \\ 1 & 2 & 4 \\ 2 & 5 & 4 \end{bmatrix}, \quad [\mathbf{B}]_1 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{6.76}$$

для другого домену:

$$Crds_2 = \begin{bmatrix} 0,02 & 0,00 \\ 0,03 & 0,00 \\ 0,02 & 0,02 \\ 0,03 & 0,02 \end{bmatrix}, \quad Elms_2 = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 3 & 2 \end{bmatrix}, \quad [\mathbf{B}]_2 = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{6.77}$$

для третього домену:

$$Crds_3 = \begin{bmatrix} 0,00 & 0,02 \\ 0,01 & 0,02 \\ 0,02 & 0,02 \\ 0,03 & 0,02 \\ 0,00 & 0,03 \\ 0,01 & 0,03 \\ 0,02 & 0,03 \\ 0,03 & 0,03 \end{bmatrix}, \quad Elms_3 = \begin{bmatrix} 0 & 1 & 4 \\ 1 & 5 & 4 \\ 1 & 2 & 5 \\ 2 & 6 & 5 \\ 2 & 3 & 6 \\ 3 & 7 & 6 \end{bmatrix}, \quad [\mathbf{B}]_3 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{6.78}$$

Знайдемо матриці жорсткості доменів. Кожна матриця будується незалежно (паралельно) одна від одної:

$$[\mathbf{K}]_1 = \begin{bmatrix} 312,5 & -250 & 0,00 & -62,5 & 0,00 & 0,00 \\ -250 & 625 & -250 & 0,00 & -125 & 0,00 \\ 0,00 & -250 & 312,5 & 0,00 & 0,00 & -62,5 \\ -62,5 & 0,00 & 0,00 & 312,5 & -250 & 0,00 \\ 0,00 & -125 & 0,00 & -250 & 625 & -250 \\ 0,00 & 0,00 & -62,5 & 0,00 & -250 & 312,5 \end{bmatrix}, \quad [\mathbf{K}]_2 = \begin{bmatrix} 312,5 & -250 & -62,5 & 0,00 \\ -250 & 312,5 & 0,00 & -62,5 \\ -62,5 & 0,00 & 312,5 & -250 \\ 0,00 & -62,5 & -250 & 312,5 \end{bmatrix}, \tag{6.79}$$

$$[\mathbf{K}]_3 = \begin{bmatrix} 250 & -250 & 0,00 & 0,00 & -125 & 0,00 & 0,00 & 0,00 \\ -250 & 500 & -125 & 0,00 & 0,00 & -250 & 0,00 & 0,00 \\ 0,00 & -125 & 500 & -125 & 0,00 & 0,00 & -250 & 0,00 \\ 0,00 & 0,00 & -125 & 250 & 0,00 & 0,00 & 0,00 & -125 \\ -125 & 0,00 & 0,00 & 0,00 & 250 & -125 & 0,00 & 0,00 \\ 0,00 & -250 & 0,00 & 0,00 & -125 & 500 & -125 & 0,00 \\ 0,00 & 0,00 & -250 & 0,00 & 0,00 & -125 & 500 & -125 \\ 0,00 & 0,00 & 0,00 & -125 & 0,00 & 0,00 & -125 & 250 \end{bmatrix}.$$

Модифікуємо побудовані матриці з врахуванням крайових умов. Оскільки крайова умова Діріхле (стала температура границі) присутня тільки в третьому домені, отримаємо:

$$[\mathbf{K}]_3 = \begin{bmatrix} 250 & -125 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ -125 & 500 & -125 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & -125 & 500 & -125 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & -125 & 250 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 250 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 500 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 500 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 250 \end{bmatrix}, \quad \{\mathbf{f}\}_3 = \begin{Bmatrix} 2500 \\ 500 \\ 500 \\ 2500 \\ 500 \\ 10000 \\ 10000 \\ 5000 \end{Bmatrix}. \tag{6.80}$$

Для першого і другого доменів задана тільки крайова умова Неймана (сталий потік), тому їх матриці жорсткості залишаться незмінними, а вектори навантаження приймуть вигляд:

$$\{\mathbf{f}\}_1 = \{500 \quad 1000 \quad 500 \quad 0{,}00 \quad 0{,}00 \quad 0{,}00\}^{\mathbf{T}}, \qquad \{\mathbf{f}\}_2 = \{500 \quad 500 \quad 0{,}00 \quad 0{,}00\}^{\mathbf{T}}. \tag{6.81}$$

Обчислимо обернені та псевдообернені матриці за (6.51). При цьому, у якості лінійно незалежних рядків і стовпців беремо всі крім останніх:

$$[\mathbf{K}]_1^{\dagger} = 10^{-4} \cdot \begin{bmatrix} 88 & 60 & 48 & 40 & 28 & 0{,}00 \\ 60 & 66{,}67 & 53{,}33 & 33{,}33 & 26{,}67 & 0{,}00 \\ 48 & 53{,}33 & 74{,}67 & 26{,}67 & 21{,}33 & 0{,}00 \\ 40 & 33{,}33 & 26{,}67 & 66.67 & 33{,}33 & 0{,}00 \\ 28 & 26{,}67 & 21{,}33 & 33{,}33 & 34{,}67 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \end{bmatrix}, \quad [\mathbf{K}]_2^{\dagger} = 10^{-4} \cdot \begin{bmatrix} 100 & 80 & 20 & 0{,}00 \\ 80 & 96 & 16 & 0{,}00 \\ 20 & 16 & 36 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \end{bmatrix}, \tag{6.82}$$

$$[\mathbf{K}]_3^{-1} = 10^{-4} \cdot \begin{bmatrix} 46{,}22 & 12{,}44 & 3{,}56 & 1{,}78 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 12{,}44 & 24{,}89 & 7{,}11 & 3{,}56 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 3{,}56 & 7{,}11 & 24{,}89 & 12{,}44 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 1{,}78 & 3{,}56 & 12{,}44 & 46{,}22 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 40 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 20 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 20 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 40 \end{bmatrix}.$$

Обчислимо матриці $[\mathbf{F}]$ та вектори $\{\mathbf{d}\}$:

$$[\mathbf{F}]_1 = 10^{-4} \cdot \begin{bmatrix} 74{,}67 & 26{,}67 & 21{,}33 & 0{,}00 & 0{,}00 & 0{,}00 \\ 26{,}67 & 66.67 & 33{,}33 & 0{,}00 & 0{,}00 & 0{,}00 \\ 21{,}33 & 33{,}33 & 34{,}67 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \end{bmatrix}, \quad [\mathbf{F}]_2 = 10^{-4} \cdot \begin{bmatrix} 100 & 0{,}00 & 0{,}00 & 0{,}00 & 20 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 20 & 0{,}00 & 0{,}00 & 0{,}00 & 36 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \end{bmatrix}, \tag{6.83}$$

$$[\mathbf{F}]_3 = 10^{-4} \cdot \begin{bmatrix} 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 46{,}22 & 12{,}44 & 3{,}56 & 0{,}00 & 1{,}78 \\ 0{,}00 & 12{,}44 & 24{,}88 & 7{,}11 & 0{,}00 & 3{,}56 \\ 0{,}00 & 3{,}56 & 7{,}11 & 24{,}89 & 0{,}00 & 12{,}44 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 1{,}78 & 3{,}56 & 12{,}44 & 0{,}00 & 46{,}22 \end{bmatrix}, \quad \{\mathbf{d}\}_1 = \begin{Bmatrix} 11{,}47 \\ 6{,}67 \\ 5.13 \\ 0{,}00 \\ 0{,}00 \\ 0{,}00 \end{Bmatrix}, \quad \{\mathbf{d}\}_2 = \begin{Bmatrix} -9 \\ 0{,}00 \\ 0{,}00 \\ 0{,}00 \\ -1{,}8 \\ 0{,}00 \end{Bmatrix}, \quad \{\mathbf{d}\}_3 = \begin{Bmatrix} 0{,}00 \\ -20 \\ -20 \\ -20 \\ 0{,}00 \\ -20 \end{Bmatrix}.$$

Для плаваючих доменів, за допомогою (6.75) знайдемо $[\mathbf{R}]$, $[\mathbf{G}]$ та $\{\mathbf{e}\}$:

$$\begin{gathered}[\mathbf{R}]_1=[1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1]^{\mathbf{T}},\quad [\mathbf{G}]_1=[-1\ \ -1\ \ -1\ \ -1\ \ -1\ \ 0]^{\mathbf{T}},\quad \{\mathbf{e}\}_1=\{-2000\},\\ [\mathbf{R}]_2=[1\ \ 1\ \ 1\ \ 1]^{\mathbf{T}},\quad [\mathbf{G}]_2=[1\ \ 0\ \ 0\ \ 0\ \ 1\ \ -1]^{\mathbf{T}},\quad \{\mathbf{e}\}_2=\{-1000\}.\end{gathered} \tag{6.84}$$

Тепер можна побудувати інтерфейсну систему (6.64):

$$[\mathbf{F}]=10^{-4}\cdot\begin{bmatrix}174{,}67 & 26{,}67 & 21{,}33 & 0{,}00 & 20 & 0{,}00\\ 26{,}67 & 112{,}89 & 45{,}78 & 3{,}56 & 0{,}00 & 1{,}78\\ 21{,}33 & 45{,}78 & 59{,}56 & 7{,}11 & 0{,}00 & 3{,}56\\ 0{,}00 & 3{,}56 & 7{,}11 & 24{,}89 & 0{,}00 & 12{,}44\\ 20 & 0{,}00 & 0{,}00 & 0{,}00 & 36 & 0{,}00\\ 0{,}00 & 1{,}78 & 3{,}56 & 12{,}44 & 0{,}00 & 46{,}22\end{bmatrix},\quad \{\mathbf{d}\}=\begin{Bmatrix}2{,}47\\ -13{,}33\\ -14{,}87\\ -20\\ -1{,}8\\ -20\end{Bmatrix}, \tag{6.85}$$

$$[\mathbf{G}]=\begin{bmatrix}-1 & -1 & -1 & -1 & -1 & 0\\ 1 & 0 & 0 & 0 & 1 & -1\end{bmatrix}^{\mathbf{T}},\quad \{\mathbf{e}\}=\begin{Bmatrix}-2000\\ -1000\end{Bmatrix}.$$

Приймемо в якості передобумовлювача $[\mathbf{Q}]=[\mathbf{E}]$ та обчислимо $[\mathbf{P}(\mathbf{Q})]$:

$$[\mathbf{P}(\mathbf{Q})]=\begin{bmatrix}0{,}64 & -0{,}09 & -0{,}09 & -0{,}09 & -0{,}36 & 0{,}27\\ -0{,}09 & 0{,}73 & -0{,}27 & -0{,}27 & -0{,}09 & -0{,}18\\ -0{,}09 & -0{,}27 & 0{,}73 & -0{,}25 & -0{,}09 & -0{,}18\\ -0{,}09 & -0{,}27 & -0{,}25 & 0{,}73 & -0{,}09 & -0{,}18\\ -0{,}36 & -0{,}09 & -0{,}09 & -0{,}09 & 0{,}64 & 0{,}27\\ 0{,}27 & -0{,}18 & -0{,}18 & -0{,}18 & 0{,}27 & 0{,}55\end{bmatrix}. \tag{6.86}$$

Можна переконатися, що $[\mathbf{P}(\mathbf{Q})]^{\mathbf{T}}[\mathbf{G}]$ буде рівний нулю.

Рішення інтерфейсної проблеми знайдемо за допомогою описаного модифікованого методу спряжених градієнтів. При початковому наближенні:

$$\{\boldsymbol{\lambda}_0\}=\{-90{,}90\ \ 727{,}27\ \ 727{,}27\ \ 727{,}27\ \ -90{,}90\ \ 818{,}18\}^{\mathbf{T}}, \tag{6.87}$$

ітераційний процес приведе до рішення:

$$\{\boldsymbol{\lambda}\}=\{0{,}00\ \ 500\ \ 1000\ \ 1000\ \ -500\ \ 500\}^{\mathbf{T}}, \tag{6.88}$$

з точністю $\varepsilon=10^{-6}$, вже на третій ітерації. Згідно (6.70) отримаємо значення для $\{\boldsymbol{\alpha}\}$:

$$\{\boldsymbol{\alpha}\}=\begin{Bmatrix}24\\ 24\end{Bmatrix}. \tag{6.89}$$

Залишилося обчислити значення температури у вузлах окремо для кожного з доменів за допомогою (6.55), підставляючи відповідні для кожного з доменів значення:

$$\begin{gathered}\{\mathbf{u}\}_i=[\mathbf{K}]_i^{\dagger}\left(\{\mathbf{f}\}_i-[\mathbf{B}]_i^{\mathbf{T}}\{\boldsymbol{\lambda}\}\right)+[\mathbf{R}]_i\{\boldsymbol{\alpha}\}_i,\quad \left|[\mathbf{K}]\right|\neq 0\Leftrightarrow[\mathbf{R}]_i=0,\\ \{\mathbf{u}\}_1=\{32\ \ 32\ \ 32\ \ 24\ \ 24\ \ 24\}^{\mathbf{T}},\\ \{\mathbf{u}\}_2=\{32\ \ 32\ \ 24\ \ 24\}^{\mathbf{T}},\\ \{\mathbf{u}\}_3=\{24\ \ 24\ \ 24\ \ 24\ \ 20\ \ 20\ \ 20\ \ 20\}^{\mathbf{T}}.\end{gathered} \tag{6.90}$$

Перевіримо правильність отриманого рішення. Знаючи, що для однорідного тіла, різниця температур при нагріванні виражається залежністю $\Delta T = d \cdot q/\eta$, де d – довжина тіла, отримаємо: $\Delta T = 0.03 \cdot 10^5 / 250 = 12$. Знайдена за допомогою FETI різниця температур становить $\{\mathbf{u}\}_{1,1} - \{\mathbf{u}\}_{3,5} = 12$. Отже, отримане рішення є правильним.

У цьому розділі описано лише основи методу скінченних елементів розривів і з'єднань. Основний акцент зроблено на висвітлення геометричної інтерпретації методу, шляхом розгляду взаємозв'язків просторів лінійних операторів та векторів, що в них лежать. Так, введені для з'єднання розірваних доменів множники Лагранжа є еквівалентами неявних допоміжних крайових умов, утворених при декомпозиції. Оператори обмежень, що виникають в плаваючих доменах, у загальному випадку відображають можливу зміну шуканих потенціалів даного домену, тобто є загальними рішеннями сумісних систем. Починаючи з задач механіки, вони отримують класичну фізичну інтерпретацію – це можливі напрямки переміщень всього домену. Коефіцієнти при операторах обмеження, це ніщо інше, ніж знайдені амплітуди таких переміщень. З формулювання методу видно, що розподілення обчислень та їх розпаралелювання закладено в нього на фундаментальному рівні через парадигму "розділяй і володарюй". Саме тому він може бути успішно застосований для пришвидшеного рішення надвеликих задач та декомпозиції обчислень на компонентному рівні проектування МЕМС.

Очевидно, що на даний момент з'явилося багато модифікацій, які значно підвищують ефективність методу скінченних елементів розривів і з'єднань. Наприклад, при побудові інтерфейсної проблеми, не обов'язково розглядати внутрішні для доменів вузли. Використовуючи розклад Шура, можна таким чином, значно зменшити розміри матриць. Оскільки всі оператори стрибку є сильно розрідженими булевими матрицями, при обчисленнях немає необхідності безпосередньо здійснювати операції їх множення. Тим більше, непотрібно зберігати ці оператори в явному вигляді. Те саме стосується і операторів обмеження.

Без розгляду також залишені питання реалізації розпаралелювання обчислень та їх ефективності на суперкомп'ютерах чи розподілених системах, оскільки це окремі й надто великі питання, що не вміщаються в рамки даної роботи.

За необхідності, їх розгляд можна знайти в приведених джерелах інформації. Відправною точкою може бути присвоєння обчислень одного демена одному процесору. З іншої сторони, завжди можлива ієрархічна декомпозиція, з врахуванням швидкодії окремих процесорів обчислювальної системи. Або навіть динамічне балансування обчислювального навантаження в складних мультизадачних системах. При цьому слід відштовхуватися від того, що оцінка складності та відповідного прискорення реалізованого методу є пропорційна відношенню кількості інтерфейсів домену до кількості елементів, з яких він складається.

6.6. Список використаної літератури до розділу 6

[1] Таненбаум Э., ван Стеен М. – Распределенные системы. Принципы и парадигмы / Tanenbaum A., van Steen M. – Distributed systems. Principles and paradigms // Санкт-Петербург: Питер, 2003.

[2] Jaworski N., Lobur M., Farmaga I., Kurzydlowski K. – Architecture of the Composite Materials Distributed Heterogeneous Computer-Aided Design System. // Proc. of the XII Intern. Conf. on The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM'2013), pp 440-442. - February 19–23, Polyana-Svalyava, Ukraine, 2013.

[3] Hughes C., Hughes T. – Parallel and Distributed Programming Using C++ // Addison-Wesley Professional, 2003.

[4] Breshears C. – The Art of Concurrency // O'Reilly Media Inc, 2009.

[5] Немнюгин С., Стесик О. – Параллельное программирование для многопроцессорных вычислительных систем // СПб: "БХВ", 2002.

[6] [Electronic resource] Foster's Methodology: Application Examples // Department of Computer Science and Engineering (DEI) Instituto Superior Tecnico. 2013 – https://dspace.ist.utl.pt/bitstream/2295/ 1021381/1/cpd-11.pdf.

[7] Nelson P. – Parallel Programming Paradigms // Washington, 1987.

[8] Jorge L., Arjona O., Roberts G. – Architectural Patterns for Parallel Programming // London, 1998.

[9] Keyes D. – Domain Decomposition Methods for Partial Differential Equations // New-York: NCAR, 2008.

[10] Pechstein C. – Finite and Boundary Element Tearing and Interconnecting Solvers for Multiscale Problems // Springer LNCSE Series, Vol. 90, 2013.

[11] Багдасаров Г., Дьяченко С., Ольховская О. – Измерение производительности и масштабируемости программного комплекса MARPLE3D // Препринт ИПМ им. М.В.Келдыша, № 37. 23с, Москва, 2012.

[12] Saad Y. – Iterative methods for sparse linear systems. 2-nd ed. // Philadelphia: Society for Industrial and Applied Mathematics, 2003.

[13] Pechstein C. – Finite and Boundary Element Tearing and Interconnecting Methods for Multiscale Elliptic Partial Differential Equations // preprint. Linz: Johannes Kepler University, 2008.

[14] Дияк І., Заяць М., Макар І. – Чисельна реалізація однорівневого МСЕРЗ (FETI) методу для плоскої задачі теорії пружності. // Відбір і оброб. інформації: Міжвід. зб. наук. пр., вип. 32(108), сс. 50-55, 2010.

[15] Farhat C., Mandel J. – Scalable Substructuring by Lagrange Multipliers in Theory and Practice // Proceedings of the 9th International Conference on Domain Decomposition Methods in Science and Engineering, pp 22-30, 1998.

[16] Farhat C., Lesoinne M., LeTallec P., Pierson K., Rixen D. – FETI-DP a dual-primal unified FETI method. Part I: A faster alternative to the two-level FETI method // Int. J. Numer. Meth. Engng, No. 50, pp 1523-1544, 2001.

[17] Kabelikova P. – Implementation of Non-Overlapping Domain Decomposition Techniques for FETI Methods // Ostrava, 2012.

[18] Dostal Z., D. Horak, and R. Kucera – Total FETI – an easier implementable variant of the FETI method for numerical solution of elliptic PDE // Commun. in Num. Meth. in Eng., No. 22, pp. 1155–1162, 2006.

[19] Stefanica D. – Domain Decomposition Methods for Mortar Finite Elements // New-York, 1999.

[20] Himmelblau D. – Applied nonlinear programming / Прикладное нелинейное программирование // Москва: "Мир", 1975.
[21] Нинул А. – Оптимизация целевых функций. Аналитика. Численные методы. Планирование эксперимента // Москва: ФИЗМАТЛИТ, 2009.
[22] Fragakis Y., Papadrakakis M. – A unified framework for formulating Domain Decomposition Methods in Structural Mechanics // Technical Report, NTUA, Athens, Greece, 2002.
[23] Гантмахер Ф. – Теория матриц. 2-изд. доп. // Москва: "Наука", 1966.
[24] Strang G. – Linear Algebra and Its Applications / Линейная алгебра и ее применения // Москва: "Мир", 1980.
[25] Ben-Israel A., Greville T. – Generalized Inverses: Theory and Applications. 2-nd ed. (CMS Books in Mathematics) // Springer, 2003.
[26] Miljkovic S. – Iterative methods for computing generalized inverses of matrices // Nis, 2012.
[27] Suarez A., Conzalez L. – A generalization of the Moore–Penrose inverse related to matrix subspaces of C nxm // Applied Mathematics and Computation, No 216, pp 514–522, 2010.
[28] Hogben L. – Handbook of Linear Algebra (Discrete Mathematics and Its Applications). 1-t ed. // Chapman and Hall/CRC, 2006.
[29] Albert A. – Regression and the Moore-Penrose Pseudoinverse // New York: Academic Press, 1972.
[30] Golub G., Van Loan C.– Matrix Computations, 3-d ed. / Матричные вычисления // Москва: "Мир", 1999.
[31] Courrieu P. – Fast Computation of Moore-Penrose Inverse Matrices // Neural Information Processing – Letters and Reviews, Vol.8, No.2, pp. 25-29, 2005.
[32] Rao C. – A note on generalized inverse of a matrix with applications to problems in mathematical statistics // Journal of the Royal Statistical Society, Series B 24, pp. 152–158, 1962.
[33] Dostal Z., Kozubek T., Markopoulos A. Mensik M. – Cholesky factorization and a generalized inverse of the stiffness matrix of a floating structure with known null space // Int. J. Numer. Meth. Engng, 2000.
[34] Kruis J., Bittnar Z. – Reinforcement-matrix interaction modeled by FETI method // Domain Decomposition Methods in Science and Engineering XVII, Lecture Notes in Computational Science and Engineering, vol. 60, Springer Verlag, pp. 567–573, 2008.
[35] Gosselet P., Rey C., Rixen D. – On the initial estimate of interface forces in FETI methods // Comput. Methods Appl. Mech. Engrg., vol. 192, pp. 2749–2764, 2003.
[36] Marcsa D., Kuczmann M. – Finite Element Tearing and Interconnecting Method and its Algorithms for Parallel Solution of Magnetic Field Problems // Electrical, Control and Communication Engineering, Vol. 3, Is. 1, pp. 25–30, 2013.
[37] Mandel J., Tezaur R., Farhat C. – A scalable substructuring method by lagrange multipliers for plate bending problems // SIAM J. Numer. Anal., 36(5), pp. 1370–1391, 1997.

7. Моделі і методи аналізу та діагностування МЕМС

7.1. Класифікація дефектів та несправностей при діагностуванні МЕМС

Підходи до моделювання несправностей, діагностування, тестування та забезпечення відмовостійкості мікроелектромеханічних систем, що в даному випадку розглядаються як цифрові схеми, мають важливе значення для створення надійних виробів [1]–[49]. З розвитком VLSI технології різко істотно збільшилася кількість компонентів, що розміщуються на одному кристалі і, як результат – зросла степінь інтеграції і кількість несправностей.

У процесі проектування виробу необхідно періодично перевіряти (шляхом тестування) коректність функціонування схеми і відсутність несправностей, а також забезпечити коректну роботу схеми в разі появи несправностей (відмовостійкість) [2].

Під несправністю МЕМС розуміють фізичний дефект одного або більше її компонентів [3], [38], [39]. Розрізняють постійні та нерегулярні несправності. Постійні (жорсткі) несправності можуть бути наслідком руйнування або зносу компонента. Нерегулярні (м'які) дефекти виявляються в певні проміжки часу і можуть бути короткочасними (transient) і перемежованими (intermittent) [4]–[7].

Несправності можуть бути логічними і параметричними. Логічна несправність змінює булеву функцію, яка реалізується системою. Параметрична – змінює значення параметра системи (струм, напруга). До параметричних дефектів відносять несправності затримки, пов'язані з різним часом проходження сигналу через логічні вентилі, що призводить до перегонів (змагань) сигналів [8].

Моделювання великої кількості фізичних дефектів може бути засноване на використанні однієї моделі логічної несправності, що дозволяє істотно зменшити складність моделювання. Модель логічної несправності не залежить від технології імплементації проекту, а тести, розроблені для виявлення логічної несправності, можуть застосовуватися також і для виявлення фізичних дефектів.

Модель логічної несправності може бути явною або неявною. Явна модель несправності визначає простір несправностей, в якому кожна несправність може бути ідентифікована, а несправності, підлягають аналізу, можуть бути явно описані. Явна модель несправності практично застосовна, якщо її розмірність не є надто великою. Неявна модель описує простір несправностей шляхом сукупної ідентифікації несправностей певного типу їх характеристичними ознаками. Моделювання несправностей тісно пов'язане з моделюванням системи. Несправності, які визначаються в поєднанні зі структурною моделлю, відносяться до структурних несправностей, що виявляється у зміні структури з’єднань компонентів. Функціональні несправності визначаються в поєднанні з функціональною моделлю. Так, наприклад, функціональна несправність може проявлятися у зміні таблиці істинності компонента або спотворенні RTL (Register Transfer Level) операції.

Існує три класи логічних несправностей: константна несправність (stuck-at-fault), місткова несправність (bridging fault) і несправність затримки (delay fault).

Найбільш поширена модель константної несправності – одинична константна несправність. Сутність моделі полягає в тому, що несправність логічного вентиля призводить до "залипання" логічного 0 (константа 0, sa-0) або логічної 1 (константа 1, sa-1) на одному з його входів або виходів [5]–[7].

Модель константної несправності також використовується для зображення кратних несправностей у системі. При цьому передбачається, що більш, ніж на одній лінії схеми є константна несправність sa-0 або sa-1. Іншими словами, сукупність константних несправностей існує в схемі в один і той же час. Різновидом кратної несправності є односпрямована несправність – коли всі складові частини несправності являють собою sa-0 або sa-1, але не обидві одночасно.

Модель константної несправності не є ефективною при моделюванні МЕМС та надвеликих інтегральних схем (НВІС / VLSI), побудованих за комплементарною метал-оксидно-напівпровідниковою (КМОН / CMOS) технологією. Несправності в CMOS схемах не обов'язково є логічними дефектами, які можуть бути описані моделлю константної несправності. Дефекти CMOS-схем відображаються також моделями стійких обривів транзисторів SOP (stuck-open) і стійких замикань транзисторів SON (stuck-on) [5].

Місткові несправності типу "коротке замикання" являють собою постійні дефекти, які не можуть бути змодельовані константною несправністю. Коротке замикання виникає, коли дві або більше сигнальних ліній схеми електрично пов'язані одна з одною. Місткові несправності вентильного рівня класифікуються наступним чином: вхідні – викликані коротким замиканням входів логічного елемента, несправності типу зворотного зв'язку – викликані замиканням вхідної і вихідної ліній, а також несправності без зворотного зв'язку, які не відносяться до перших двох типів. У теорії моделювання місткових несправностей робиться припущення, що ймовірність замикання більше двох ліній є низькою і логіка з’єднань реалізується у вигляді зв'язків. Місткова несправність в позитивній логіці виникає в тому випадку, коли її поведінка описується проводовим AND (0 є домінантним логічним значенням), і в негативній логіці – коли її поведінка описується проводовим OR (1 є домінантним логічним значенням).

Несправності затримки [4], [6]. Невелика кількість дефектів, які можуть викликати розриви і короткі замикання в схемі, мають досить високу ймовірність появи через наявність відхилень параметрів виробничого процесу. Дефекти можуть також призводити до порушень часових параметрів схеми без зміни логіки її роботи: затримка перемикання сигналу з 0 в 1 і навпаки. Існує два види несправностей затримки: несправність затримки вентиля і несправність затримки шляху. Затримка вентиля використовується для моделювання дефектів, при яких час проходження сигналу через вентиль перевищує гранично-допустимий. Дана модель може бути використана тільки для ізольованих, не транспортованих дефектів, наприклад, кілька малих затримок. Модель затримки шляху може бути використана як для ізольованих,

так і для транспортованих дефектів. При цьому передбачається, що несправність проявляється у випадку, якщо затримка поширення сигналу уздовж лінії схеми перевищує припустиме значення.

Перемежовані несправності розглядаються як часові дефекти. Основна частина несправностей цифрових схем викликана саме часовими дефектами, які характеризуються складністю виявлення та усунення. Перемежовані дефекти є неповторюваними і викликаються, як правило, флуктуаціями напруги живлення або впливом радіаційного випромінювання. Вони є основною причиною відмови елементів пам'яті систем на кристалах.

Перемежовані несправності можуть з'являтися в результаті порушення з'єднань, застосування дефектних компонентів, впливу зовнішніх факторів (температура, вологість, вібрація) або бути наслідком помилок проектування. Перемежовані несправності виникають випадковим чином і моделюються за допомогою імовірнісних методів (Марківські моделі).

7.2. Методи генерації тестів

Безперервне вдосконалення технологій проектування і виробництва МЕМС призводить до збільшення щільності компонування і складності пристроїв, досягнення необхідного рівня надійності яких забезпечується тестуванням. Для вирішення завдання тестування сучасних надскладних електронних пристроїв необхідні нові, більш ефективні методи побудови тестів [9]–[14]. Виробництво систем на кристалах (system-on-chip, SoC) з використанням технології глибокого субмікронного (Deep Submicron, DSM) дозволяє знизити витрати, але при цьому вартість тестування залишається незмінною і являє собою значну частину загальної вартості проекту. Зменшити витрати на тестування виробу можна шляхом повторного використання блоків інтелектуальної власності (IP cores), а також розробки моделей і методів тестування SoC на високому рівні ієрархії [15]. Високорівневі модулі системи на кристалі описуються в термінах поведінки функціональних компонентів, що не дозволяє використовувати для їх тестування готові технічні рішення вентильного рівня. Класична модель одиночної константної несправності (stuck-at fault), що представляє внутрішні логічні вентилі або їх міжз'єднання, не може бути застосована для використання на системному рівні. Структурне високорівневе тестування не може бути виконано з використанням готових тестових рішень, оскільки генерація тесту виконується після структурного синтезу. Реалізація тестування залежить від технології виготовлення SoC і змінюється в процесі життєвого циклу виробу [2].

Для забезпечення можливості багаторазового використання тестів у нових проектах необхідно розробити таку модель несправності, яка є незалежною від реалізації системи на кристалі. Слід також знайти відповіді на запитання: 1) Чи може тест, побудований на базі функціональної моделі несправності, бути ефективно використаний для непокриваємих тестом фізичних дефектів? 2) Як ефективність тесту залежить від синтезованої структури? Ці запитання є важливими не тільки з точки зору повторного використання тестів, але також

через те, що програмні модулі можуть бути синтезовані за допомогою існуючих систем автоматизованого проектування і збережені в бібліотеках IP модулів.

Для успішного проектування і виробництва виробу необхідні методологія тестування і моделі несправностей, які забезпечують високорівневу валідацію проекту [2].

Дефект-орієнтоване тестування, засноване на генерації тестів на транзисторному рівні і використанні струмових моделей (IDDQ), ефективно використовується в технології глибокого субмікронного. IDDQ метод тестування заснований на вимірюванні струму і добре працює, коли середній струм схеми з несправністю більше, ніж струм справного пристрою. Дефект-орієнтоване тестування починається після реалізації етапу розміщення і трасування (place and route) (*Рис. 6.1*) [2].

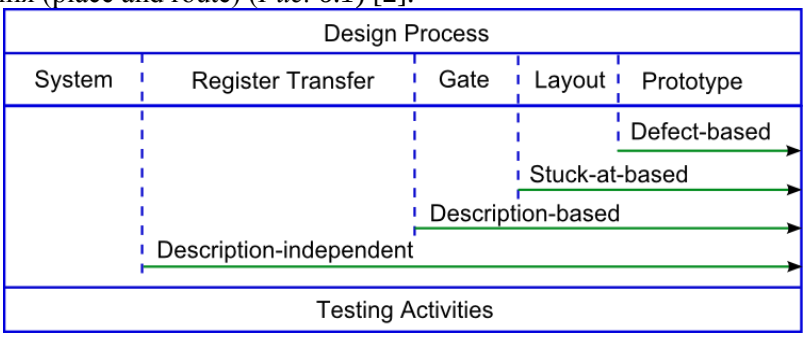


Рис. 7.1 Генерація тестів в процесі проектування

Сутністю константної моделі несправностей є абстракція реального дефекту. Модель є основою для автоматичної генерації тестових наборів і формування алгоритмів моделювання несправностей. Умовою досягнення високого рівня покриття несправностей є те, що тест повинен транспортувати деяке конкретне значення (або їх сукупність) в дефектну область від керованих точок введення і далі до спостережуваних виходів з метою виявлення несправної поведінки. Дана модель найбільш ефективна для тестування на кристалі (post-silicon testing). Генерація тестів для константних несправностей виконується перед розміщенням і трасуванням проекту (*Рис. 6.1*). Тести для константних несправностей покривають реальні дефекти топології тільки примітивних вентилів. У цьому випадку говорять про генерації тестів, не залежних від топології кристала.

Відомі кілька підходів до генерації тестових наборів на рівні регістрових передач (Register Transfer Level, RTL) [2]: використання бінарних дерев рішень, виявлення спотворень в RTL описі схеми, об'єднання статичного аналізу з симуляцією. Більшість з них дозволяють генерувати тестові послідовності задовільної якості, що сумісні із засобами ATPG (Automatic Test Pattern Generation) вентильного рівня. Головною перевагою тестування пристрою на RTL рівні є те, що розмірність опису схеми тут набагато менше, ніж на логічному рівні. При генерації тестів на рівні регістрових передач набір тестових послідовностей створюється для всіх можливих реалізацій, і можна

говорити про генерацію тестів, не залежних від імплементації. Завдання генерації тестів може виконуватися паралельно з синтезом схеми на вентильному рівні (див. верхню частину *Рис. 7.2*).

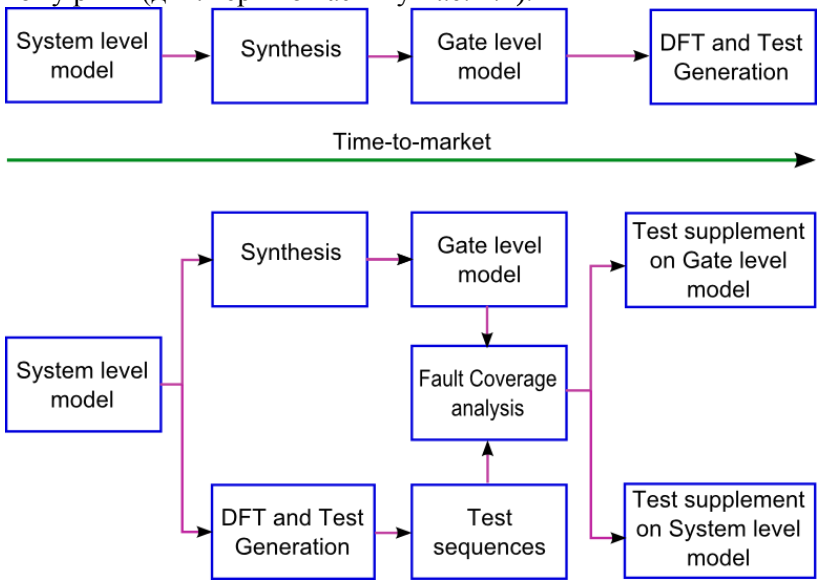


Рис. 7.2 Час time-to-market для різних стратегій проектування

Генерація тестів на системному рівні залежить від використовуваної моделі несправностей. У цьому випадку не тільки реалізація, а й синтезований поведінковий опис не відомі. Задача генерації тестів в цьому випадку є більш складною, але може вирішуватися одночасно з формуванням синтезованого опису та синтезом схеми на вентильному рівні (див. *Рис. 7.2*) [2].

Перед відправкою проекту у виробництво необхідно сформувати тестові набори. Верифікація пристрою передбачає запуск тестової програми на робочій моделі кристала. Тестери є дорогими компонентами і можуть використовуватися протягом тривалого часу. Початок розробки тестів наприкінці процесу проектування значно збільшує час виходу виробу на ринок. Якщо використовується методологія проектування зверху вниз, то системна модель виробу на кристалі формується на самому початку процесу проектування і може бути використана при розробці тестової програми. Таким чином, інженер-тестувальник може стати учасником процесу проектування на ранніх стадіях і використовувати віртуальний прототип пристрою у вигляді системної моделі. Це дозволить суттєво зменшити час проектування і вартість виробу.

Методологія проектування зверху вниз орієнтована на автоматичний синтез списку з'єднань вентильного рівня з використанням поведінкового опису або системної моделі. Час виходу виробу на ринок (time-to-market) залежить від тривалості процедури логічного синтезу, тривалості тестопридатного проектування і генерації тестів (див. *Рис. 7.2*). Тестопридатність проектування і

генерація тестів на основі системної моделі дозволить скоротити час виходу на ринок [2]. Аналіз тестового покриття на вентильному рівні не є часовитратною процедурою і дає можливість зменшити довжину тестових послідовностей, отриманих на системному рівні.

Генерація тестів на системному рівні не може гарантувати 100% покриття несправностей вентильного рівня для кожної можливої імплементації. Формування тестових послідовностей необхідно виконувати паралельно на системному і на вентильному рівнях, оскільки ймовірність генерації тестів для несправностей, що важко виявляються, може бути різною на кожному рівні опису.

7.3. Класифікація моделей функціональних несправностей

Все більш широке використання програмно-апаратних систем в критичних додатках привело до підвищення значущості верифікації та тестування програмних і апаратних модулів. В даний час існує ряд проблем верифікації, пов'язаних з високою складністю програмно-апаратних додатків і їх гетерогенною структурою [10]–[15]. Вартість верифікації системи збільшилася до такої міри, що іноді навіть перевищує вартість проекту. Формальні методи верифікації дозволяють перевірити функціональність за допомогою формальних методів (перевірка моделей, перевірка еквівалентності, автоматичне доведення теорем). Для управління складністю задачі верифікації, запропоновані методи, що засновані на симуляції (емуляції) опису системи заданої вхідної послідовністю.

Функціональні несправності спотворюють простір станів цифрового виробу, що представлено специфікацією. Дефект проектування являє собою неправильну деталь проекту, сформовану розробником. Дефекти проектування є наслідком синтаксичних (семантичних) помилок в описі пристрою або фундаментального нерозуміння функціональності, описаної проектною специфікацією. Кількість потенційних дефектів проектування може бути занадто великим, щоб з ними можна було боротися автоматично або вручну, тому необхідно застосовувати способи зменшення складності проекту без шкоди для точності результатів. Модель проектної несправності описує поведінку деякої множини дефектів проектування. Модель функціональної несправності описує фізичні та проектні дефекти апаратних і програмних модулів. Модель функціональної несправності можна оцінити точністю моделювання проектних дефектів і ефективністю.

Більшість апаратних систем розробляються на основі методології проектування зверху вниз, яка починається з поведінкового опису системи. Як результат, більшість моделей функціональних несправностей є моделями поведінкового або алгоритмічного рівня. Існуючі моделі функціональних несправностей можуть бути класифіковані за стилем поведінкового опису, на якому вони базуються. Системна поведінка описується на мовах програмування (System C) або опису апаратури (VHDL, Verilog) і перетворюється у внутрішній формат для використання в процесі симуляції.

- Текстові (семантичні) моделі несправностей.

Текстова (семантична) модель несправностей використовується для вихідного текстового поведінкового опису проекту [16]. Найпростішою текстовою моделлю є метрика покриття інструкцій (statement coverage metric), яка використовується при тестуванні програмного забезпечення, що пов'язує потенційну помилку з кожним рядком коду, і вимагає, щоб кожен оператор поведінкового опису виконувався під час тестування [16]. Ця модель не дуже ефективна, оскільки кількість можливих несправностей дорівнює числу рядків коду. Обмеження точності покриття інструкцій дозволяє в поєднанні з іншими моделями несправностей підвищити ефективність тестування.

Ряд моделей функціональних несправностей базується на обході шляхів графа потоків управління (Control-Data Flow Graph, CDFG), що описує поведінку системи [16]. Ранні моделі несправностей CDFG ґрунтувалися на покритті гілок і шляхів графа. Покриття гілок припускає, що багато всіх перевірених шляхів графа CDFG охоплює два напрямки реалізації всіх бінарних умов. Покриття гілок широко використовується при тестуванні програмного і апаратного забезпечення, однак використання тільки даної моделі не дозволяє отримати повну гарантію коректності коду.

Метрика покриття шляхів є більш ефективною у порівнянні з метрикою покриття гілок, оскільки вона відображає кількість шляхів графа потоків управління. Передбачається, що дефект пов'язаний з деяким шляхом графа потоків управління і, отже, для гарантованого виявлення всіх несправностей повинні бути виконані всі шляхи потоку управління. Кількість шляхів управління може бути нескінченною, якщо граф CDFG містить цикл. Тому, метрика обходу шляхів може бути обмежена довжиною шляху. Оскільки загальне число шляхів потоку управління зростає експоненціально з кількістю умовних операторів, можна вибрати підмножину всіх шляхів потоку управління, необхідну і достатню для тестування. Одним з критеріїв вибору шляху може бути базисний набір шляхів або підмножина шляхів, які лінійно незалежні і можуть утворювати будь-який інший шлях. При тестуванні потоків даних поява кожної змінної розглядається або як опис змінної, або як її використання. При виборі шляху розглядаються такі, які пов'язують визначення змінної з її використанням. Критерії тестування потоків також застосовуються для перевірки поведінкового опису апаратних модулів.

Більшість моделей несправностей графа потоків управління розглядають шляхи без обмеження значень змінних і сигналів. На противагу їм, існують моделі несправностей, орієнтовані на змінні/сигнали, які включають більш жорсткі обмеження на величину сигналу з метою виявлення несправностей. Методика аналізу домену при тестуванні програмного забезпечення розглядає не тільки шлях потоку управління, але і значення змінних і сигналів під час виконання. Домен являє собою підмножину простору вхідних елементів програми, в якому кожен елемент активізує виконання програми за деяким шляхом. Несправність домену викликає виконання програми, наслідком якого є перехід в неправильну область.

Багато моделей несправностей графа потоків управління містять вимоги до активізації несправностей не залежно від значення спостережуваності. Для усунення цього недоліку запропоновані поведінкові моделі несправностей, застосовні для тестування software і hardware модулів. Підхід OCCOM [16], [17] базується на додаванні несправностей, званих тегами, до кожного визначення змінної, що представляють позитивне чи негативне зміщення від правильного значення сигналу. Знак помилки відомий, але величина – ні. Аналіз спостережуваності уздовж шляху потоку управління робиться імовірнісно, з використанням алгебраїчних властивостей операцій і даних моделювання. Тег буде поширюватися за допомогою поведінкової операції, якщо будуть виконані дві умови: 1) співпаде знак; 2) інші входи в процесі виконання операції не контролюються. Розроблено також точний метод визначення спостережуваності, в якому константна несправність вводиться на внутрішніх змінних, і її поширення забезпечується поведінкою об'єкта. Оскільки аналіз спостережуваності є точним, обчислювальна складність при цьому зростає.

- Мутаційні моделі несправностей.

Мутаційне тестування засноване на штучному внесенні несправностей в код програми і застосовується для тестування програмного і апаратного забезпечень [16]. Основна ідея полягає в імітації типових помилок програміста і створення спеціальних тестів для їх виявлення (тестів, які б виявляли несправності, якби вони були присутні). Несправності вводяться в оригінальну програму і створюється багато несправних версій програми. Кожна з них містить одну помилку. Помилкові програми називаються мутантами оригінальної програми. Метою генерації тестів є розмежування оригінальної програми та всіх її мутантів. Оригінальна програма і всі її мутації тестуються на одному і тому ж наборі тестів. Якщо на цьому наборі підтверджується правильність програми і виявляються всі помилки в програмах-мутантів, то оригінальна програма оголошується правильною. Набір мутаційних операторів мови VHDL включає зміни наступних об'єктів: арифметичні оператори, визначення і зміна абсолютного значення і константи, логічні оператори, реляційні оператори, додавання унарного оператора. Кожен оператор зображує певний клас несправностей. Всі можливі зміни в програмі не можуть розглядатися через їх непомірно велику кількості. Зміни можуть бути обмежені до прийнятного набору на основі двох гіпотез: ефект зчеплення і компетентний програміст. Ефект зчеплення свідчить, що складні несправності можуть поєднуватися з простими несправностями, таким чином, тестовий набір, який виявляє всі прості помилки в програмі, буде виявляти також і складні несправності. Гіпотеза "компетентний програміст" стверджує, що компетентний програміст прагне писати програми, які практично є правильними. Іншими словами, програма, написана компетентним програмістом може бути неправильною, але вона буде відрізнятися від правильної версії відносно простими помилками. Недоліком мутаційних моделей є локальний характер мутацій, що обмежує застосування моделей для опису великого набору дефектів проектування.

- Моделі несправностей кінцевого автомата.

Кінцеві автомати (КА) є класичним способом опису поведінки послідовних схем, і для них визначені моделі несправностей [16]. Найбільш поширеною є модель покриття станів, заснована на вимозі покриття всіх можливих станів і виконання всіх можливих переходів в процесі тестування. Дефекти КА не виводять його за простір станів, заданий специфікацією. Проблема використання моделей несправностей кінцевого автомата полягає у високій складності вирішення завдання тестування, яка обумовлена великою розмірністю простору станів типової обчислювальної системи. Рішенням зазначеної проблеми може бути виявлення підмножини станів кінцевого автомата, що мають вирішальне значення для його коректного функціонування. Моделі розширеного кінцевого автомата (Extended Finite State Machine, EFSM [18], [19]) і машини контролю (Extracted Control Flow Machine, ECFM) дозволяють зменшити кінцевий автомат шляхом його розподілу на простір станів і простір даних. Зменшений кінцевий автомат генерується шляхом проектування оригінального кінцевого автомата на множину станів, що мають найбільше значення для процесу валідації.

7.4. Методи моделювання несправностей

Логічне моделювання є формою верифікаційного тестування проекту з використанням моделі проектованої системи. На вхід моделі подаються вхідні стимули, виконується побудова та аналіз часових діаграм для зовнішніх входів, виходів схеми і внутрішніх ліній [49].

Верифікація проекту дозволяє перевірити функціональні режими щодо специфікації. Перевірка здійснюється на кожній стадії перетворення моделі від системного рівня до рівня імплементації проекту в кристал шляхом порівняння результатів, отриманих в процесі моделювання, і еталонних результатів, передбачених специфікацією.

Завдання, які вирішуються в процесі логічного моделювання: 1. Перевірка правильності функціонування цифрової схеми. 2. Дослідження часових параметрів схеми (швидкодія, час виконання операцій, тактова частота). Виявлення змагань, ризиків збою, аналіз затримок. 3. Оптимізація проектних рішень. 4. Генерація часових діаграм. 5. Генерація тестових послідовностей. 6. Моделювання несправностей.

Для верифікації проекту необхідний прототип пристрою, що функціонує на заданій робочій частоті, однак створення прототипу є дорогим і трудомістким процесом. Заміна прототипу програмної моделлю називається симуляцією. Верифікація проекту з використанням симулятора має такі переваги:

- перевірка помилкових умов (наприклад, конфлікти шин);
- можливість зміни затримок в моделі для перевірки граничних часових параметрів;
- перевірка заданих користувачем значень параметрів схеми;
- можливість початку моделювання схеми на будь-якому етапі

проектування;
- точний контроль таймінгу асинхронних подій (наприклад, переривань);
- можливість формування автоматизованого тестового оточення модельованої схеми, використання RTL моделі для управління і спостереження за поведінкою схеми в процесі моделювання.

Технології використання апаратного прототипування за допомогою PLD мають ряд істотних переваг в порівнянні з програмним симулятором: швидкодія і можливість корекції проекту.

Методи логічного моделювання можна класифікувати за такими ознаками: 1) залежно від способу обліку часу поширення сигналу - синхронні (без урахування затримок в елементах схеми) і асинхронні (з урахуванням затримок); 2) залежно від способу подання сигналів – двійкові і багатозначні (троїчні, п'ятизначні); 3) за способом організації роботи програми –компілятивні та інтерпретативні; 4) залежно від організації черговості моделювання – покрокові і подієві.

Синхронне моделювання призначене для аналізу перехідних процесів в цифрових пристроях вентильного та функціонального рівнів опису на основі моделей елементів, представлених їх логічними функціями без урахування затримок сигналів. У процесі моделювання обчислюють значення сигналів на виходах логічних елементів схеми за заданими вхідними сигналами. При цьому передбачається, що час існування перехідного процесу набагато більше номінальної затримки схеми. Синхронне моделювання найбільш ефективно використовується для аналізу роботи комбінаційних схем в сталому режимі. Результат моделювання в цьому випадку найбільш точно відповідає реальному режиму роботи пристрою. До методів синхронного моделювання відносять:
- Метод Ейхельбергера, призначений для синхронного аналізу перехідних процесів в цифрових пристроях вентильного рівня опису;
- Багатозначне синхронне моделювання, що дозволяє виявляти всі реальні змагання в схемі. Цей метод іноді вказує на помилкові змагання, що призводить до додаткових витрат при логічній верифікації цифрових систем.

Рішенням зазначеної вище проблеми є асинхронні методи аналізу цифрових схем. Їх різноманітність визначається значністю алфавіту моделювання та ступенем адекватності моделей за реальними часовими параметрами.

Асинхронне моделювання застосовується для аналізу перехідних процесів в логічних схемах з урахуванням часу поширення сигналів в елементах і сполучних ланцюгах схеми. Кожен компонент схеми характеризується деякою середньої затримкою, значення якої може змінюватися залежно від режиму роботи компонента, комбінації вхідних сигналів, температури, відхилень в технології виготовлення.

До методів асинхронного моделювання відносять:
- Δ-Троїчне моделювання, яке усуває недоліки двійкового асинхронного і троїчного синхронного методів;

- Асинхронне троїчне моделювання з наростаючою невизначеністю, яке усуває детермінізм в модельній затримці компонента схеми, укладаючи її в деякий інтервал.

Сутність моделювання несправностей полягає у визначенні впливу одного або декількох дефектів на стани ліній об'єкта при подачі тестових послідовностей [20], [21]. Методи моделювання несправностей можна класифікувати наступним чином: одиночне, паралельне, дедуктивне, кубічне і спільне моделювання.

Одиночне моделювання несправностей базується на внесенні однієї одиночної константної несправності еквіпотенційної лінії до схеми. При подачі тестових послідовностей виконується аналіз прояву несправності на зовнішніх виходах об'єкта діагностування. Метод орієнтований на обробку схем нерегістрового рівня і не вимагає значних часових витрат.

Паралельне моделювання несправностей ґрунтується на використанні машинних команд паралельної обробки слів (регістрів): логічне додавання, множення, інверсія, виключне АБО. Метод відноситься до компілятивного моделювання, оскільки поведінка примітивів схеми описується за допомогою алгоритмічних мов або асемблерів. У процесі моделювання одночасно виконується аналіз P несправностей на вхідному наборі, де P – розрядність машинного слова, доступного для паралельної обробки. До недоліків методу відносять складність проектування моделей і їх орієнтацію на конкретну обчислювальну платфрму. Швидкодія методу в P разів вище одиночного моделювання несправностей. Ідея паралельної обробки бінарного вектора за допомогою тільки логічних операцій може бути використана для істотного збільшення швидкості моделювання.

Дедуктивне моделювання несправностей полягає в одночасній обробці всіх одиночних константних несправностей схеми на одному вхідному наборі і виділенні при цьому підмножини перевірюваних дефектів. Метод орієнтований на вентильний рівень опису моделі проектованого об'єкта в базисі І-АБО-НЕ. Необхідність отримання аналітичних формул для кожного типу примітивного елемента і великі витрати пам'яті для зберігання списків несправностей ускладнюють практичну реалізацію методу.

У спільному (конкурентному) моделюванні, як і в дедуктивному, виявляються відразу всі перевірювані несправності для даного вхідного набору. Метод орієнтований на обробку різних типів моделей схем, несправностей, затримок і сигналів. На відміну від дедуктивного методу, де дефекти моделюються неявно, конкурентний алгоритм аналізує явно справну роботу і ті несправності, які модифікують стани входів або виходів схеми, що використовуються ефективні моделі елементів, такі як табличні та функціональні.

Дедуктивно-паралельне моделювання несправностей цифрових систем ґрунтується на використанні переваг дедуктивного і паралельного алгоритмів [20] і дозволяє за одну ітерацію обробки модельованої схеми виявити всі несправності, що перевіряються на тест-векторі. Метод дозволяє істотно

підвищити швидкодію моделювання одиночних константних несправностей для оцінки якості синтезованих тестів цифрових систем, імплементованих в ПЛІС, що містять мільйони вентилів.

7.5. Методи діагностування несправностей

Модель об'єкта діагностування – це сукупність гетерогенних компонентів, взаємопов'язаних у часі та просторі, що із заданою адекватністю описують певний процес або явище. Модель може бути подана в аналітичній, табличній, векторній, графічній або іншій формі і задана в явному або неявному вигляді [3], [38], [39].

Явна модель об'єкта діагностування складається з описів його справної і всіх несправних модифікацій. Неявна модель містить опис справного об'єкта, моделі його фізичних несправностей і правила отримання за ними всіх несправних модифікацій об'єкта. Універсальною математичною моделлю об'єкта діагностування є таблиця функцій несправностей (ТФН). Кожний несправний стан об'єкта діагностування відповідає одній несправності (одиночній або кратній) із заданого класу несправностей. Недоліком ТФН є її великі розміри. Модель дискретної системи може бути зображена у вигляді таблиці істинності, логічної мережі, альтернативного графа, еквівалентної нормальної форми подання булевих функцій, таблиці переходів-виходів багатотактної схеми. Вибір моделі впливає на глибину і трудомісткість процесу діагностування.

Задачами технічного діагностування є: визначення технічного стану об'єкта, пошук місця і визначення причин відмови. Визначення технічного стану об'єкта здійснюється за допомогою спеціальної тестової послідовності вхідних впливів. Методи формування тестових послідовностей для діагностування несправностей можна умовно розділити на кілька типів, описаних нижче.

1. Розподіл, використання дерев рішень – полягає в моделюванні поведінки справної системи і систем з N заздалегідь визначеними несправностями. Відгук кожної з них на вхідний вплив використовується для формування (N+1) систем. При цьому повинні виконуватися умови:

- справна система повинна бути швидко відокремлена від несправних (виявлення несправностей);
- всі системи є однозначно ідентифікованими (помітними) (виявлення несправностей і визначення їх місця розташування).

Результуючий розподіл або побудова дерева рішень визначає діагностичну тестову послідовність, яка дозволяє однозначно визначити належність тестованої системи однієї з (N+1) категорій.

Для пошуку несправності застосовують послідовний, комбінаційний і послідовно-комбінаційний методи. Послідовний метод полягає в такій побудові процедури пошуку несправностей, при якому інформація про стан окремих тестованих систем вводиться і логічно обробляється послідовно. Реалізація методу полягає в основному у визначенні черговості контролю. Програма

пошуку при цьому може бути жорсткою або гнучкою. Жорстка програма передбачає наявність заздалегідь визначеної послідовності контролю. При гнучкій програмі зміст і порядок подальших перевірок залежать від попередніх результатів. Комбінаційний метод полягає в тому, що на вхід тестованої системи подається фіксований набір тестів. Діагноз формується тільки після того, як будуть отримані відгуки на всі тестові впливи.

2. Активізація одновимірного шляху. Даний метод грунтується на введенні відомої несправності в схему і її транспортуванні на один з первинних виходів за активізованим шляхом. При цьому будь-яка зміна логічного значення в місці несправності призводить до зміни значення на відповідному первинному виході. Описана процедура носить назву прямої фази. Зворотна фаза полягає у визначенні значень інших первинних входів і виходів, таких, щоб задана несправність виявлялася на первинному виході. Метод простий і зручний у використанні, проте у схемі можуть існувати несправності, для перевірки яких необхідно активізувати кілька шляхів (у разі наявності збіжних розгалужень).

3. Використання таблиці функцій несправностей і таблиці несправностей. Таблиця функцій несправностей є спеціальною формою подання поведінки об'єкта діагностування у справному та несправному станах. Таблиця несправностей пов'язує набір тестів і несправностей, що ними перевіряються. Обмеженням даного методу є розмірність зазначених таблиць.

4. Метод булевих похідних. Булева похідна визначається шляхом виконання операції OR над двома булевими функціями, які представляють справний і несправний об'єкт. Якщо булева похідна дорівнює 1, вважається, що проявляється помилка і визначається відповідна тестова послідовність. Тестові набори визначаються шляхом формування булевої похідної для кожної несправності.

5. Метод еквівалентної нормальної форми, що базується на зображенні булевої функції у вигляді еквівалентної нормальної форми, яка описує конкретну реалізацію схеми. Еквівалентна нормальна форма може бути обчислена методом підстановки, з тією різницею, що надлишкові терми не виключаються, так як вони характеризують конкретну реалізацію схеми.

- TAB модель діагностування несправних компонентів SoC

Мета дослідження – розробка матричної моделі TAB (Tests – Assertions – Blocks) і методу діагностування, що дозволяють зменшити час тестування і обсяг пам'яті для зберігання діагностичної інформації за рахунок формування тернарних відносин (тест – монітор – функціональний компонент) в одній таблиці.

Завдання дослідження: 1) розробка HDL-моделі цифрової системи у формі транзакційного графа для діагностування функціональних блоків з використанням набору асерцій [32], [38]–[43], [46]–[48]; 2) розробка методу аналізу TAB-матриці з метою виявлення мінімального набору несправних блоків [44,45]; 3) Синтез логічних функцій для вбудованої процедури діагностування несправностей [32, 48].

Механізм асерцій (Assertion Engine) – це технологічний апарат

тестопридатного проектування HDL-моделей цифрових систем, орієнтований на контроль виконання програмного коду в його критичних точках. Асерції – блоки, що додаються у вихідний код проекту для спостереження та управління поведінкою моделі проекту. Асерції створюються розробником або можуть бути взяті з існуючих бібліотек для перевірки типових функцій. Так, фірма Synopsys поставляє систему моделювання VCS з бібліотекою асерцій System Verilog. SystemVerlog дозволяє реалізувати два види асерцій: миттєві та паралельні. Миттєві запускаються відразу при передачі управління відповідному оператору вихідного коду. Паралельна асерція перевіряє вираз відповідно до кожного імпульсу синхросигнала і спрацьовує в разі фіксації хибності висловлювання. Асерції, додані до коду тестованого пристрою, називаються внутрішніми, до тестового середовища – зовнішніми. Останні стежать за сигналами, що передаються між пристроєм і тестовим середовищем. Мова SystemVerilog має можливості підключення проекту до модуля, який містить зовнішні асерції.

Модель тестування HDL-коду цифрової системи подана наступними xor-відношеннями параметрів <тест–функціональність–несправні блоки B*>:

$$\begin{aligned} &T \oplus B \oplus B^* = 0; \\ &B^* = T \oplus B = \{T \times A\} \oplus B, \end{aligned} \tag{7.1}$$

які перетворюються у відношення компонентів TAB-матриці:

$$M = \{\{T \times A\} \times \{B\}\}, \quad M_{ij} = (T \times A)_i \oplus B_j. \tag{7.2}$$

Тут координати матриці дорівнюють 1, якщо пара тест-монітор $(T \times A)_i$ перевіряє або активує несправності функціонального блоку $B_j \in B$.

Аналітична модель верифікації з використанням темпоральних асерцій (додаткова спостережуваність операторів або ліній), орієнтована на досягнення заданої глибини діагностування:

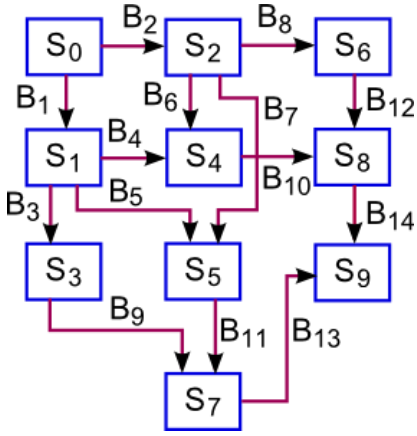
$$\begin{aligned} &\Omega = f(G, A, B, S, T), \\ &G = (A^*B) \times S;\ S = f(T, B); \\ &A = \{A_1, A_2, ..., A_i, ..., A_h\}; \\ &B = \{B_1, B_2, ..., B_i, ..., B_n\}; \\ &S = \{S_1, S_2, ..., S_i, ..., S_m\}; \\ &T = \{T_1, T_2, ..., T_i, ..., T_k\}. \end{aligned} \tag{7.3}$$

Тут $G = (A^*B) \times S$ – функціональність, що зображена CFT графом (Code-Flow Transaction); $S = \{S_1, S_2, ..., S_i, ..., S_m\}$ – вершини або оператори коду програми при симуляції тестових сегментів (наборів). Іншими словами, граф може розглядатися як ABC-граф (Assertion Based Coverage Graph), *Рис. 7.3*. Кожний стан $S_i = \{S_{i1}, S_{i2}, ..., S_{ij}, ..., S_{ip}\}$ визначається значеннями суттєвих змінних проекту (булеві, регістрові змінні, пам'ять). Дуги орієнтованого графа подані множиною програмних блоків:

$$\mathrm{B}=(\mathrm{B}_1,\mathrm{B}_2,...,\mathrm{B}_\mathrm{i},...,\mathrm{B}_\mathrm{n}),\ \bigcup_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{B}_\mathrm{i}=B;\ \bigcap_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{B}_\mathrm{i}=\varnothing \qquad (7.4)$$

де асерції $\mathrm{A}_\mathrm{i}\in\mathrm{A}=\{\mathrm{A}_1,\mathrm{A}_2,...,\mathrm{A}_\mathrm{i},...,\mathrm{A}_\mathrm{n}\}$ можуть бути додані до кожного блоку B_i – послідовність операторів коду, які визначають стан вершини графу $\mathrm{S}_\mathrm{i}=\mathrm{f}(T,B_\mathrm{i})$ у залежності від тестового набору $\mathrm{T}=\{\mathrm{T}_1,\mathrm{T}_2,...,\mathrm{T}_\mathrm{i},...,\mathrm{T}_\mathrm{k}\}$. Асерційний монітор, що об'єднує асерції вхідних дуг $\mathrm{A}(\mathrm{S}_\mathrm{i})=\mathrm{A}_{i1}\vee\mathrm{A}_{\mathrm{i}2}\vee...\vee\mathrm{A}_{\mathrm{ij}}\vee...\vee\mathrm{A}_{\mathrm{iq}}$ може бути встановлений у кожній вершині.

Модель HDL-коду, яка подана у вигляді ABC-графа, описує не тільки структуру програмного коду, але також тестові сегменти функціонального покриття, що формуються з використанням програмних блоків, які входять до розглянутої вершини. Остання визначає простір станів змінних, що досягнуте на тесті, і формує функціональне покриття станів змінних, відповідних розглянутій вершині графа $Q=cardS_i^r\ /\ cardS_i^p$. У сукупності всі вершини графа мають представляти собою повний простір покриття станів змінних програмного коду, яке визначає якість тесту, дорівняний 100%: $Q=card\bigcup_{i=1}^{m}S_i^r\ /\ card\bigcup_{i=1}^{m}S_i^p=1$. Більш того, множина асерцій $<A,S>$, яке існує в графі, дозволяє виконати моніторинг дуг (покриття коду) $B=(B_1,B_2,...,B_n)$ та вершин (функціональне покриття) $S=\{S_1,S_2,...,S_i,...,S_m\}$.



$$\mathrm{B}=(B_1B_3B_9\vee(B_2B_7\vee B_1B_5)B_{11})B_{13}\vee((B_1B_4\vee B_2B_6)B_{10}\vee B_2B_8B_{12})B_{14}=$$
$$=B_1B_3B_9B_{13}\vee B_2B_7B_{11}B_{13}\vee B_1B_5B_{11}B_{13}\vee B_1B_4B_{10}B_{14}\vee B_2B_6B_{10}B_{14}\vee B_2B_8B_{12}B_{14}.$$

Рис. 7.3 Приклад ABC-графа HDL-коду

Асерції на дугах $\mathrm{B}_i\in\mathrm{B}$ графа призначені для діагностування функціональних несправностей в програмних блоках. Асерція у вершинах графа $\mathrm{S}_i\in\mathrm{S}$ несе інформацію про якість тесту та множини асерцій з метою їх покращення або доповнення. ABC-граф дозволяє: 1) оцінювати якість програмного коду шляхом визначення діагнозопридатності проекту

(diagnosability); 2) мінімізувати вартість генерації тестів, діагностування та виправлення функціональних порушень за рахунок використання ассерцій; 3) оптимізувати синтез тестів шляхом покриття всіх дуг і вершин графа мінімальним набором активізованих тестових шляхів. Наприклад, мінімальний тест для наведеного вище АВС-графа містить шість сегментів, які активізують всі наявні дуги і вершини:

$$T=S_0S_1S_3S_7S_9 \vee S_0S_1S_4S_8S_9 \vee S_0S_1S_5S_7S_8 \vee \\ \vee S_0S_2S_4S_8S_9 \vee S_0S_2S_5S_7S_9 \vee S_0S_2S_6S_8S_9. \quad (7.5)$$

В процесі діагностування тестові сегменти $T=\{T_1,T_2,...,T_r,...,T_k\}$ активізують транзакційний шлях графової моделі, який покриває всі вершини та дуги. Як правило, тестова модель подана декартовим добутком $M=<T\times A\times B>$, який відповідно має розмірність $Q=k\times h\times n.$ З метою зменшення кількості діагностичних даних окремий монітор або асерційна точка для візуалізації функціональної активізації блоків призначаються кожному тестовому сегменту. Це дозволяє зменшити розмірність матриці до $Q=n\times k$ та зберегти всі особливості тріади відношень $M=<T\times A\times B>$. Пара «тест – монітор» подається трьома можливими формами:

$$<T_i\to A_j>,<\{T_i,T_r\}\to A_j>,<\{T_i\}\to\{A_j,A_s\}>. \quad (7.6)$$

Метод діагностування функціональних порушень блоків передбачає використання попередньо побудованої ТАВ-матриці (таблиці) $M=[M_{ij}]$, де рядок є відношення між тестовим сегментом та підмножиною блоків, що активізуються:

$$T_i\to A_j\approx(M_{i1},M_{i2},...,M_{ij},...,M_{in}),M_{ij}=\{0,1\} \quad (7.7)$$

які спостерігаються монітором A_j. Стовпець таблиці описує відношення між функціональними блоками і тестовими сегментами щодо монітора $M_j=B_j(T_j,A_j)$.

Для діагностування несправних блоків за допомогою процедури тестування визначається реальний вектор асерційної перевірки $A^*=(A_1^*,A_2^*,...,A_i^*,...,A_n^*)$ на тестовому наборі Т шляхом формування $A_i^*=f(T_i,B_i)$. Виявлення несправних функціональних блоків ґрунтується на хог-операції між реальним вектором ассерційної перевірки і стовпцем ТАВ-матриці $A^*\oplus[M_1(B_1)\vee M_2(B_2)\vee...\vee M_j(B_j)\vee...\vee M_n(B_n)]$. Несправний блок визначається за допомогою вектора B_j за мінімальною кількістю одиничних координат:

$$B=\min_{j=\overline{1,n}}[B_j=\sum_{i=1}^{h}(B_{ij}\oplus A_i^*)]. \quad (7.8)$$

Як доповнення до моделі діагностування слід описати деякі важливі властивості ТАВ-матриці:

$$\mathrm{M}_i = (\mathrm{T}_i \times \mathrm{A}_j);\quad \bigvee_{i=1}^{m} \mathrm{M}_{ij} \to \forall_{j=1}^{n} \mathrm{M}_j = 1;\quad \mathrm{M}_{ij} \bigoplus_{j=1}^{n} \mathrm{M}_{rj} \neq \mathrm{M}_{ij};\quad \mathrm{M}_{ij} \bigoplus_{i=1}^{k} \mathrm{M}_{ir} \neq \mathrm{M}_{ij};$$
$$\log_2 n \le k \leftrightarrow \log_2 |\mathrm{B}| \le |\mathrm{T}|;\quad \mathrm{B}_j = f(\mathrm{T}, \mathrm{A}) \to \mathrm{B} \oplus \mathrm{T} \oplus \mathrm{A} = 0. \tag{7.9}$$

Властивості означають: 1) Кожен рядок матриці являє собою підмножину Декартового добутку тесту і монітора. 2) Диз'юнкція всіх рядків матриці дає результат у вигляді вектора, в якому всі координати дорівнюють 1. 3) Всі рядки матриці різні, що виключає надмірність тесту. 4) Всі стовпці матриці різні, що виключає існування еквівалентних несправних блоків. 5) Кількість рядків матриці повинна бути більше двійкового логарифма числа стовпців, що визначає потенційну діагнозопридатність (діагностованість) кожного блоку. 6) Функція діагнозу кожного блоку залежить від повного тесту і моніторів, які повинні бути мінімізовані без погіршення діагнозопридатності.

Нижче наведені 6 тестових сегментів, які активізують шляхи графа щодо асерційної точки S9:

$$\mathrm{T} = \mathrm{S}_0\mathrm{S}_1\mathrm{S}_3\mathrm{S}_7\mathrm{S}_9 \vee \mathrm{S}_0\mathrm{S}_1\mathrm{S}_4\mathrm{S}_8\mathrm{S}_9 \vee \mathrm{S}_0\mathrm{S}_1\mathrm{S}_5\mathrm{S}_7\mathrm{S}_9 \vee$$
$$\vee \mathrm{S}_0\mathrm{S}_2\mathrm{S}_4\mathrm{S}_8\mathrm{S}_9 \vee \mathrm{S}_0\mathrm{S}_2\mathrm{S}_5\mathrm{S}_7\mathrm{S}_9 \vee \mathrm{S}_0\mathrm{S}_2\mathrm{S}_6\mathrm{S}_8\mathrm{S}_9, \tag{7.10}$$

при цьому досить просто вивити всі функціональні блоки з можливими порушеннями:

$$\mathrm{B} = B_1B_3B_9B_{13} \vee B_2B_7B_{11}B_{13} \vee B_1B_5B_{11}B_{13} \vee$$
$$\vee B_1B_4B_{10}B_{14} \vee B_2B_6B_{10}B_{14} \vee B_2B_8B_{12}B_{14}. \tag{7.11}$$

Механізм асерцій може бути поданий трьома групами компонентів, які формують логічні вирази для моніторингу програмних і апаратних блоків HDL коду функціональності, які засновані на візуальних точках $\{\mathrm{A}_9 \subseteq S_9, \mathrm{A}_3 \subseteq S_3, \mathrm{A}_6 \subseteq S_6\}$:

$$\mathrm{A}_9 = T_1(B_1B_3B_9B_{13}) \vee T_2(B_2B_7B_{11}B_{13}) \vee T_3(B_1B_5B_{11}B_{13}) \vee T_4(B_1B_4B_{10}B_{14}) \vee$$
$$\vee T_5(B_2B_6B_{10}B_{14}) \vee T_6(B_2B_8B_{12}B_{14});\quad \mathrm{A}_3 = T_1(B_1B_3);\ \mathrm{A}_6 = T_6(B_2B_8). \tag{7.12}$$

На наступному кроці формуються 6 рядків ТАВ-матриці $\mathrm{M}_{ij}(G_1)$ у вигляді відношень між тестовими сегментами та блоками, активізуються:

$\mathrm{M}_{ij}(G_1)$	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	B_{11}	B_{12}	B_{13}	B_{14}
$T_1 \to S_9$	1	.	1	.	.	.	.	.	1	.	.	.	1	.
$T_2 \to S_9$	1	.	.	1	.	.	.	.	.	1	.	.	.	1
$T_3 \to S_9$	1	.	.	.	1	.	.	.	.	.	1	.	1	.
$T_4 \to S_9$	.	1	.	.	.	1	.	.	.	1	.	.	.	1
$T_5 \to S_9$	.	1	.	.	.	.	1	.	.	.	1	.	1	.
$T_6 \to S_9$	.	1	.	.	.	.	.	1	.	.	.	1	.	1
$T_1 \to S_3$	1	.	1	.	.	.	.	.	.	.	.	.	.	.
$T_6 \to S_6$	.	1	.	.	.	.	.	1	.	.	.	.	.	.

ТАВ-матриця шляхів активізації показує існування еквівалентних порушень блоків 3 і 9, 8 і 12 на 6 тестових сегментах з однією асерційною точкою у вершині графа 9. Стовпці 3 і 9, 8 і 12 еквівалентні. Для усунення нерозрізненості двох пар несправних блоків необхідно створити два додаткових монітора у вузлах S3 і S6 для тестових сегментів T1 і T6 відповідно. В результаті три додаткові асерції у вузлах $A=(S_9,S_3,S_6)$ дозволять розрізнити всі несправні блоки програмного HDL-коду. Таким чином, граф дозволяє не тільки синтезувати оптимальний тест, а й визначити мінімальну кількість асерційних моніторів у вузлах графа, необхідну для пошуку несправних блоків із заданою глибиною діагностування.

Процедура діагностування з використанням запропонованої матриці визначається наступним виразом на основі векторної хог-операції між реальними вісьмома асерційними значеннями і стовпцем B ТАВ-матриці:

$$\{[\mathrm{A}_9(T_1,T_2,T_3,T_4,T_5,T_6),\mathrm{A}_3(T_1),\mathrm{A}_6(T_6)]\oplus B_j=0\}\rightarrow(B_j-\textit{failed}). \qquad (7.13)$$

7.6. Діагнозопридатне проектування

Діагнозопридатність визначається відношенням $\mathrm{N_d}/\mathrm{N}$ кількості виявлених несправних блоків $\mathrm{N_d}$, (якщо немає еквівалентних компонентів або глибина діагностування дорівнює 1), до загальної кількості NHDL-блоків.

Для оцінки витрат E на реалізацію ТАВ-матричної моделі виявлення функціональних порушень можна використовувати ефективність пари тест-асерція для заданої глибини діагностування. Критерій E функціонально залежить від ставлення між ідеальним $|\log_2\mathrm{N}|\times\mathrm{N}$ і реальним $|\mathrm{T}|\times|\mathrm{A}|\times\mathrm{N}$ обсягами необхідної пам'яті або ресурсів (де $|\mathrm{T}|$ – довжина тесту, $|\mathrm{A}|$ – кількість асерцій) для відповідної ТАВ-матриці та являє собою відносне значення в інтервалі від 0 до 1:

$$\mathrm{E}=\frac{|\log_2\mathrm{N}|\times\mathrm{N}}{|\mathrm{T}|\times|\mathrm{A}|\times\mathrm{N}}=\frac{|\log_2\mathrm{N}|}{|\mathrm{T}|\times|\mathrm{A}|}. \qquad (7.14)$$

Узагальнений критерій якості діагностування залежить від витрат E і діагнозопридатності D:

$$\mathrm{Q}=\mathrm{E}\times\mathrm{D}=\frac{|\log_2\mathrm{N}|}{|\mathrm{T}|\times|\mathrm{A}|}\times\frac{\mathrm{N_d}}{\mathrm{N}}. \qquad (7.15)$$

Наприклад, якість діагностування ТАВ-матриці $\mathrm{M}_{ij}(\mathrm{G}_1)$ до і після додавання двох рядків дорівнює:

$$\begin{aligned}\mathrm{Q}_1[\mathrm{M}(6\times1\times14)]&=\frac{|\log_2 14|}{|6|\times|1|}\times\frac{10}{14}=0,47.\\ \mathrm{Q}_1[\mathrm{M}(8\times1\times14)]&=\frac{|\log_2 14|}{|8|\times|1|}\times\frac{14}{14}=0,5.\end{aligned} \qquad (7.16)$$

Це означає, що розмір першої матриці трохи менше, ніж другий, але діагнозопридатність краще для другого варіанту матриці і в цілому вона є більш переважною. У порівнянні з відомим розв'язком [48], коли кожна комірка матриці містить всі існуючі асерції $\left|M_{ij}\right| = \left|A\right|$, другий варіант оцінюється наступним низьким значенням:

$$Q_2[M(6\times 3\times 14)] = \frac{\left|\log_2 14\right|}{\left|6\right|\times\left|3\right|}\times\frac{14}{14} = 0,2. \qquad (7.17)$$

Таким чином, ТАВ-матриця, що сформована для вибраної пари тест-асерція, дозволяє отримати суттєву перевагу з точки зору скорочення обсягу пам'яті в $\left|A\right|-1$ раз при одинаковому значенні діагнозопридатності.

Діагностична якість ТАВ-матриці визначається відношенням кількості бітів, необхідних для ідентифікації (розпізнавання) всіх блоків $\left|\log_2 N\right|$, до реальної кількістї бітів коду, що поданий добутком довжини тесту на кількість асерцій $\left|T\right|\times\left|A\right|$. Якщо перша частина E критерію якості Q дорівнює 1, що означає - кожен блок з функціональними порушеннями виявляється $N_d = N$, то тест і асерція є оптимальними і обумовлюють найкраще значення критерію якості моделі діагностування Q=1.

Мета аналізу АВС-графа - структурна оцінка розміщення асерційних моніторів, що дозволяє отримати максимальну глибину діагностування несправних блоків. Діагнозопридатність АВС-графа є функцією, яка залежить від кількості N_n транзитних не кінцевих вершин, у яких існує тільки дві сусідніх дуги, одна з яких є вхідною, інша – вихідною. Такі дуги утворюють шляхи без збіжних і розбіжних розгалужень (N – загальна кількість дуг в графі):

$$D = \frac{N - N_n}{N}. \qquad (7.18)$$

Оцінка визначається кількістю невиявлених або еквівалентних функціональних блоків. Потенційна установка додаткових моніторів для поліпшення діагностованих несправних блоків рекомендується для транзитних вузлів, що містять N_n. Критерій якості діагностування АВС-графа має вигляд:

$$Q = E\times D = \frac{\left|\log_2 N\right|}{\left|T\right|\times\left|A\right|}\times\frac{N - N_n}{N}. \qquad (7.19)$$

Останній вираз визначає деякі практичні правила для синтезу діагнозопридатного HDL-коду: 1) Тест або testbench повинні створювати мінімальну кількість одновимірних шляхів активізації, що покривають всі вузли і дуги АВС-графа. 2) Базова кількість моніторів дорівнює числу кінцевих вершин графа без вихідних дуг. 3) Додатковий монітор може бути розміщений в кожній не кінцевій вершині, яка має одну вхідну і одну вихідну дугу. 4) Паралельні незалежні блоки коду повинні мати n моніторів і один паралельний тест, або один інтегрований монітор і n послідовних тестів. 5) Послідовно з'єднані блоки мають один тест активації для послідовного шляху і n-1 монітор

або n тестів і n моніторів. 6) Вузли графа, які мають більше однієї вхідної і вихідної дуги, визначають сприятливі умови для діагностованості поточного фрагменту за допомогою тесту активізації одновимірного шляху без установки додаткових моніторів. 7) Тестовий набір або testbench повинен забезпечити 100% функціональне покриття для вершин ABC-графа. 8) Критерій якості діагностування як функція, що залежить від структури графа, теста і асерційних моніторів, завжди може бути приведена до близького до 1 значення. Для цього є два альтернативні способи. Перший – збільшення кількості тестових сегментів шляхом активізації нових шляхів для розпізнавання еквівалентних несправних блоків без збільшення числа асерцій, якщо структура графа програми дозволяє потенційні зв'язки. Другий шлях – додавання асерційних моніторів у транзитних вершинах графа. Можливий також третій гібридний варіант, заснований на спільному застосуванні двох вищезгаданих способів.

7.7. Метод багаторівневого діагностування цифрових систем

На *Рис. 7.4* наведена багаторівнева модель мультидерева B, у якій кожній вершині відповідає компонент цифрової обчислювальної системи. Модель описується тривимірною таблицею (матрицею) активізації функціональних модулів. Вихідні дуги дозволяють перейти на більш низький рівень деталізації в процесі діагностування, коли заміна несправного блоку вимагає значних матеріальних витрат:

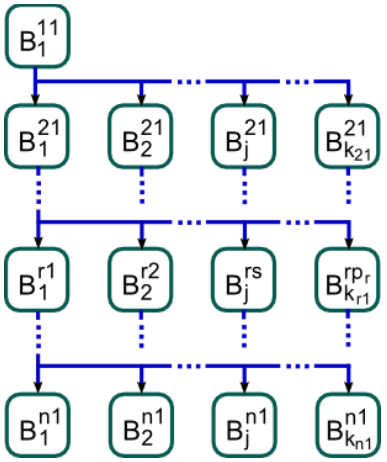


Рис. 7.4 Діагностична модель мультидерева

$$B=[B_{ij}^{rs}],\quad \text{card}B=\sum_{r=1}^{n}\sum_{s=1}^{m_r}\sum_{i=1}^{p_{rs}}\sum_{j=11}^{k_{rs}}B_{ij}^{rs}, \tag{7.20}$$

де n – кількість рівнів діагностування в мультидереві; m_r – кількість функціональних модулів або компонентів на рівні r; k_{rs} (p_{rs}) – кількість компонентів (довжина тесту) в таблиці B^{rs}; $B_{ij}^{rs}=\{0,1\}$ – компонент таблиці активізації, який визначається одиничним значенням, якщо несправна

функціональність виявляється тестовим сегментом T_{i-A_i} відносно спостережуваного монітору A_i. Кожна вершина-таблиця має деяку кількість вихідних дуг, яка дорівнює числу функціональних компонентів, поданих матрицею активізації.

Метод діагностування несправних блоків апаратно-програмної (Hardware-Software) HS-системи, заснований на моделі мультидерева, дозволяє створити універсальний движок у вигляді алгоритму (*Рис. 7.5, блок 6*) для обходу гілок дерева на апріорно заданій глибині:

$$B_{ij}^{rs} \oplus A^{rs} = \begin{cases} 0 \to \{B_j^{r+1,s}, R\}; \\ 0 \to \{B_{j+1}^{rs}, T\}. \end{cases} \qquad (7.21)$$

Тут векторна хог-операція виконується між стовпцями матриці та вектором асерційної (експериментальної) перевірки A^{rs}, який визначається за допомогою хог-операції над реальним (m) і модельним (g) відгуками функціональності на тестові набори: $A^{rs} = m_i^{rs} \oplus g_i^{rs}, i = \overline{1, k_{rs}}$. Якщо всі координати векторної хог-суми $B_j^{rs} \oplus A^{rs} = 0$ дорівнюють нулю, то виконується одно з наступних дій: перехід до матриці активізації більш низького рівня $B_j^{r+1,s}$ або відновлення функціонального блоку $B = B_j^{rs}$.

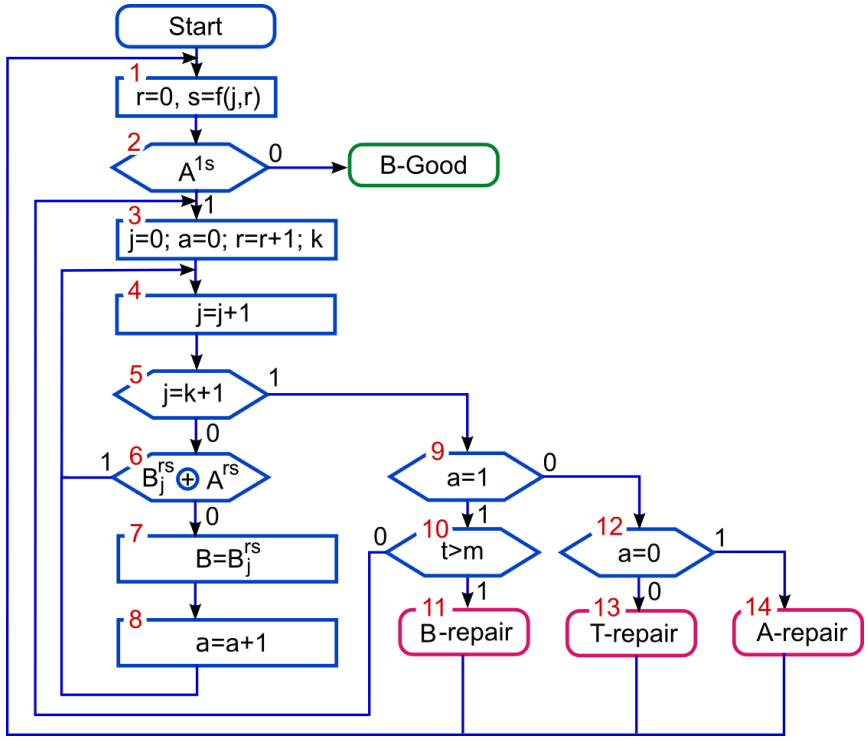


Рис. 7.5 Движок для обходу діагностичного мультидерева

Реалізується один із двох можливих шляхів аналізу у залежності від того, який параметр є найбільш важливим: 1) час (t> m, блок 10) – виконується відновлення несправного блоку; 2) матеріальні витрати (t <m) – здійснюється перехід на більш низький рівень для збільшення глибини діагностування несправностей, оскільки заміна більш дрібного блоку зменшує вартість відновлення. Якщо, принаймні, одна координата результуючого вектора хог-суми дорівнює одиниці $B_j^{rs} \oplus A^{rs} = 1$, здійснюється перехід до наступного стовпця матриці. Якщо всі координати асерційного вектора дорівнюють нулю $A^{1s} = 0$, це свідчить про те, що HS-система знаходиться у справному стані. Якщо всі векторні суми при обробці стовпця TAB-матриці не дорівнюють нулю $B_j^{rs} \oplus A^{rs} \neq 0$, то тест, згенерований для виявлення несправностей даного компонента функціональності, повинен бути скорегованим. Якщо більш, ніж одна векторна сума, отримана при обробці стовпця TAB-матриці дорівнює нулю $B_j^{rs} \oplus A^{rs} = 0$, це означає, що механізм асерцій, створений для діагностування даного компонента функціональності на представленому тесті, повинен бути доповнений ассерційними моніторами. Таким чином, TAB-движок має чотири кінцевих вершини, одна з яких B-good відповідає успішному завершенню тестування. Інші три вершини свідчать про отримання проміжного результату процесу тестування, який необхідно враховувати для збільшення якості тестування та глибини діагностування шляхом формування додаткових ассерцій та/або генерації додаткових тестових сегментів.

Таким чином, граф, наведений на *Рис. 7.4*, дозволяє реалізувати ефективну інфраструктуру сервісного обслуговування для складних технічних систем. Перевагою TAB-движка, інваріантного до рівнів ієрархії, є простота підготовки та подання діагностичної інформації у вигляді мінімізованої таблиці активізації функціонального блоку на тестових сегментах.

Технологічна модель інфраструктури вбудованого тестування, діагностування та відновлення несправних блоків (*Рис. 7.6*) включає три компоненти:

1. Тестування блоків (Unit Under Test – UUT) з використанням еталонної моделі (Model Under Test - MUT) для генерації вектора асерційної перевірки (assertion response vector), розмір якого відповідає кількості тестових наборів.

2. Пошук несправних блоків на основі аналізу TAB-матриці.

3. Відновлення несправних блоків шляхом заміни їх на справні компоненти з наявного резерву.

Процес-модель вбудованого сервісного обслуговування функціонує у реальному часі і дозволяє підтримувати працездатний стан HS-системи без безпосередньої участі людини. Запропонований алгоритм або TAB-движок для аналізу TAB-матриці, а також введений критерій якості діагностування дозволяють вирішувати завдання квазі-оптимального покриття програмних і апаратних блоків тестами і асерціями. Модель, що зображена на *Рис. 7.6,* дозволяє забезпечити ефективне сервісне обслуговування складних HS-систем.

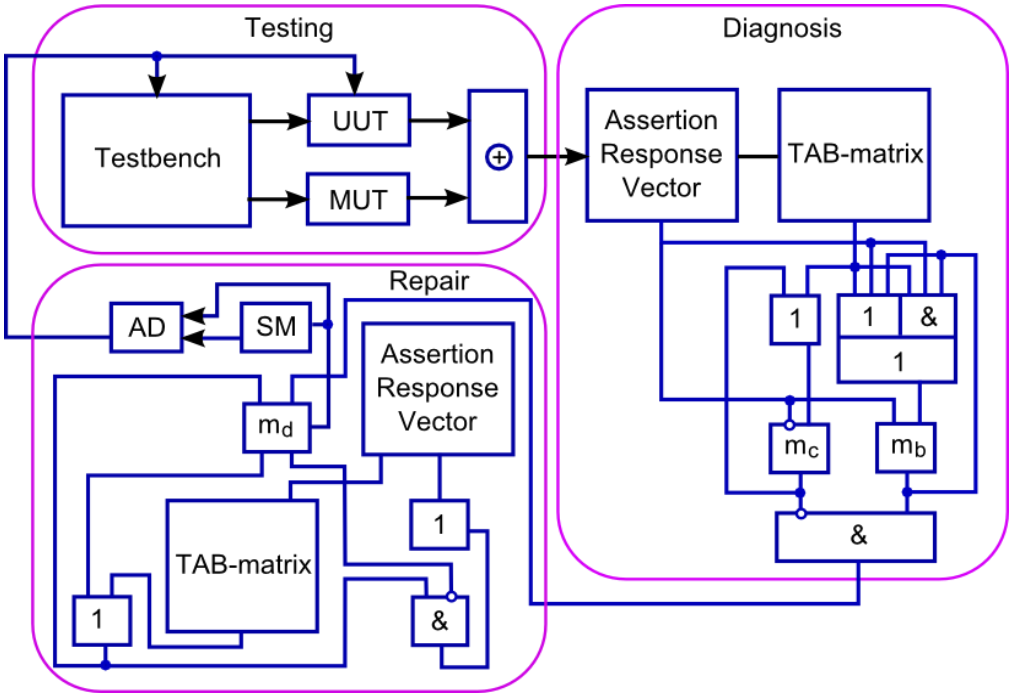


Рис. 7.6 Модель вбудованого тестування HS-компонентів

Виграш у часі виходить за рахунок введення в проект додаткової інфраструктури, *Рис. 7.6*, яка дозволяє виконувати вибіркове тестування і діагностування, а також перепрограмування окремих модулів несправних блоків.

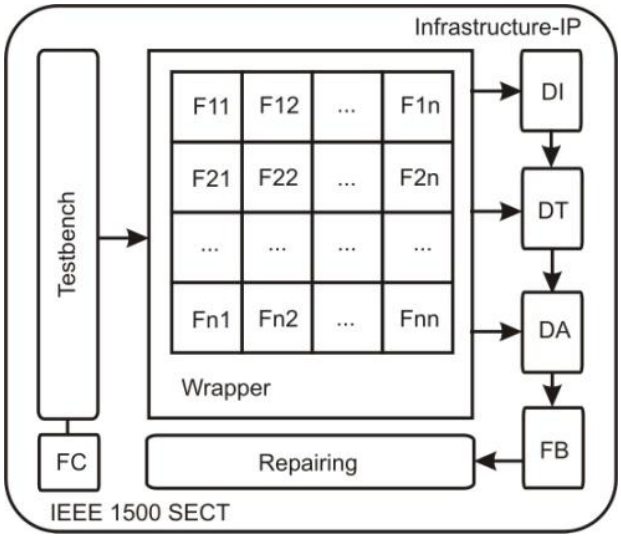


Рис. 7.7 Інфраструктура для тестування обчислювальних систем

На *Рис. 7.7* подані наступні блоки: Testbench – тести для функціональних блоків; FC – функціональне тестове покриття; F – функціональні блоки; DI – діагностична інформація у вигляді таблиць несправностей блоків; DT – методи та засоби діагностування; DA – результати аналізу діагностичної інформації; FB – несправні функціональні модулі; Repairing – відновлення функціональних модулів. Комірка граничного сканування, що представлена на *Рис. 7.8*, забезпечує сервісне обслуговування одного функціонального модуля.

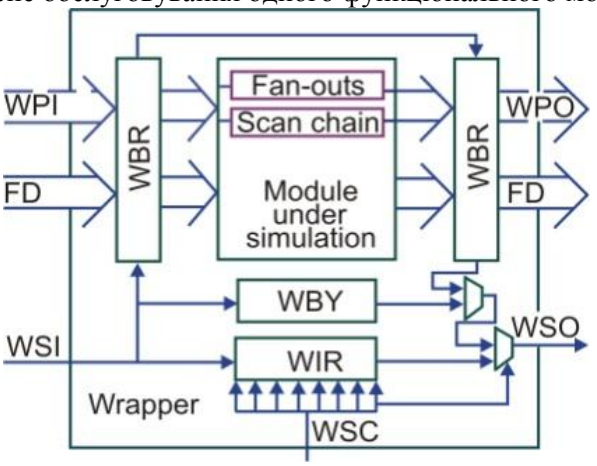


Рис. 7.8 Комірка граничного сканування

7.8. Приклад розв'язання задачі діагностування

Для ілюстрації ефективності запропонованої моделі та методу нижче використовується функціональність у вигляді трьох модулів цифрового фільтра Добеші [44].

Як другий контрольний приклад практичного використання запропонованої моделі активізації та xor-методу аналізу TAB-матриці для пошуку несправних блоків далі представлений синтез діагностичної матриці для графа основного фільтра, *Рис. 7.9*.

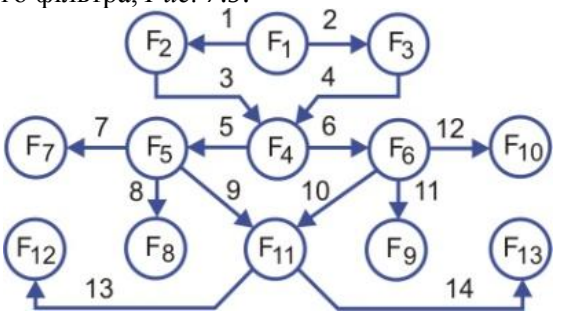


Рис. 7.9 Транзакційний граф main-TL

Граф пов'язаний з діагностичною TAB-матрицею, де мається 6 активізаційних тестових сегментів та 8 асерцій:

$M_{ij}(TL)$	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	B_{11}	B_{12}	B_{13}	B_{14}
$T_1 \to F_7$	1	.	1	.	1	.	1	.	.	.	.	.	.	.
$T_2 \to F_8$	.	1	.	1	1	.	.	1	.	.	.	.	.	.
$T_3 \to F_9$	1	.	1	.	.	1	.	.	.	.	1	.	.	.
$T_4 \to F_{10}$	.	1	.	1	.	1	.	.	.	.	.	1	.	.
$T_5 \to F_{12}$	1	.	1	.	1	.	.	.	1	.	.	.	1	.
$T_6 \to F_{13}$	.	1	.	1	.	1	.	.	.	1	.	.	.	1
$T_1 \to F_2$	1	.	.	.	.	.	.	.	.	.	.	.	.	.
$T_2 \to F_3$	.	1	.	.	.	.	.	.	.	.	.	.	.	.

Система діагностичних функцій для апаратної імплементації частини інфраструктури сервісного обслуговування, що відповідна рядкам або моніторам:

$$F_7(T_1) = B_1^1 B_3^1 B_5^1 B_7^1; \quad F_8(T_2) = B_2^1 B_4^1 B_5^1 B_8^1; \quad F_9(T_3) = B_{11}^1 B_6^1 B_1^1 B_3^1;$$
$$F_{10}(T_4) = B_4^1 B_5^1 B_6^1 B_{12}^1; \quad F_{12}(T_5) = B_1^1 B_3^1 B_5^1 B_9^1 B_{13}^1; \qquad (7.22)$$
$$F_{13}(T_6) = B_2^1 B_4^1 B_6^1 B_{10}^1 B_{14}^1; \quad F_2(T_1) = B_1^1; \quad F_3(T_2) = B_2^1.$$

Синтез діагностичної матриці для одного модуля дискретного косинусного перетворення з бібліотеки Xilinx у вигляді функціонального покриття показаний у *лістингу 7.1*.

Лістинг 7.1 – Фрагмент функціонального покриття

```
c0: coverpoint xin
{
    bins minus_big={[128:235]};
    bins minus_sm={[236:255]};
    bins plus_big={[21:127]};
    bins plus_sm={[1:20]};
    bins zero={0};
}
c1: coverpoint dct_2d
{
    bins minus_big={[128:235]};
    bins minus_sm={[236:255]};
    bins plus_big={[21:127]};
    bins plus_sm={[1:20]};
    bins zero={0};
    bins zero2=(0=>0);
}
endgroup
```

Розроблені також інші 12 модулів транзакційного графа, активізаційні ТАВ-матриці і логічні функції для тестування і виявлення несправностей в модулі дискретного косинусного перетворення. Фрагмент механізму моніторів наведено на *лістингу 7.2*.

Лістинг 7.2 – Фрагмент коду механізму моніторів

```
sequence first( reg[7:0] a, reg[7:0]b);
 reg[7:0] d;
 (!RST,d=a)
##7 (b==d);
 endsequence
property f(a,b);
 @(posedge CLK)
 // disable iff(RST||$isunknown(a)) first(a,b);
!RST |=> first(a,b);
endproperty
odin:assert property (f(xin,xa7_in))
 // $display("Very good");
else $error("The end, xin =%b,xa7_in=%b", $past(xin, 7),xa7_in);
```

Тестування дискретного косинусного перетворення у середовищі Riviera фірми Aldec дозволило виявити некоректності у семи рядках HDL-моделі:

```
//add_sub1a <= xa7_reg + xa0_reg;//
```

Наступна коректировка коду дозволила отримати наступний фрагмент (*лістинг 7.3*).

Лістинг 7.3 – Скоригований фрагмент коду

```
add_sub1a <= ({xa7_reg[8],xa7_reg} + {xa0_reg[8],xa0_reg});
add_sub2a <= ({xa6_reg[8],xa6_reg} +{xa1_reg[8],xa1_reg});
add_sub3a <= ({xa5_reg[8],xa5_reg} +{xa2_reg[8],xa2_reg});
add_sub4a <= ({xa4_reg[8],xa4_reg} + {xa3_reg[8],xa3_reg});
end
else if (toggleA == 1'b0)
begin
add_sub1a <= ({xa7_reg[8],xa7_reg} - {xa0_reg[8],xa0_reg});
add_sub2a <= ({xa6_reg[8],xa6_reg} - {xa1_reg[8],xa1_reg});
add_sub3a <= ({xa5_reg[8],xa5_reg} - {xa2_reg[8],xa2_reg});
add_sub4a <= ({xa4_reg[8],xa4_reg} - {xa3_reg[8],xa3_reg});
```

Практична імплементація моделей і методів верифікації інтегрована у середовище моделювання Riviera фірми Aldec Inc, *Рис. 7.10*. Нові асерції і модулі діагностування, додані до системи, дозволяють поліпшити існуючий процес верифікації, що дозволяє на 15% скоротити час розробки цифрового продукту.

Застосування асерцій дає можливість зменшити довжину testbench і значно скоротити (x3) час проектування (*Рис. 7.11*), яке є найбільш витратним. Механізм асерцій дозволяє збільшити глибину діагностування функціональних порушень в програмних блоках до рівня 10-20 операторів HDL-коду.

Завдяки взаємодії засобів моделювання та механізму асерцій, автоматично розміщених всередині HDL-коду, з'являється доступ засобів діагностування до значень всіх внутрішніх сигналів. Це дозволяє швидко визначити місце розташування і тип функціонального порушення, а також скоротити час

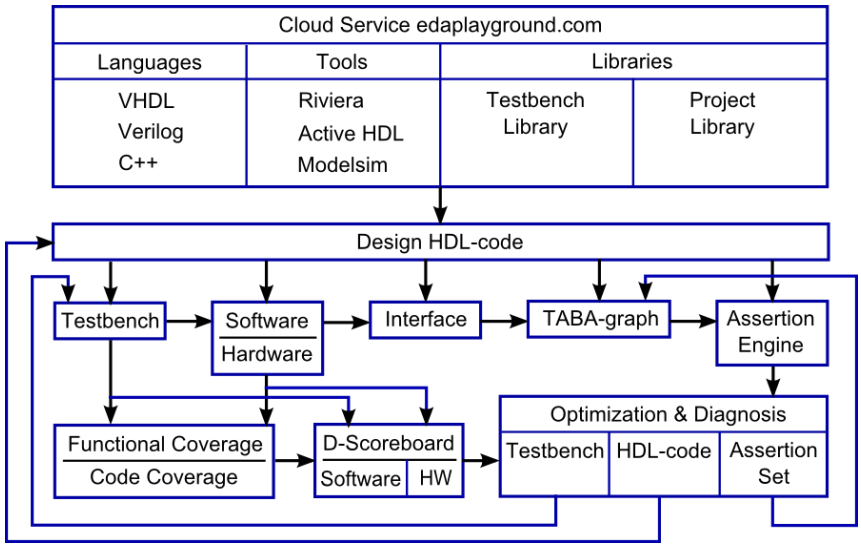


Рис. 7.10 Імплементація результатів у систему Riviera

виявлення помилок при використанні методології проектування зверху вниз. Застосування асерцій для 50 реальних проектів (від 5000 до 5000000 вентилів) дозволило отримати сотні спеціалізованих рішень, включених до бібліотеки верифікаційних шаблонів VTL, які є узагальненням найбільш популярних на ринку EDA (Electronic Design Automation) обмежень темпоральной верифікації для широкого класу цифрових продуктів. Програмна реалізація запропонованої системи аналізу асерцій та діагностування HDL-коду є частиною багатофункціонального інтегрованого середовища Aldec Riviera для моделювання та верифікації HDL-моделей.

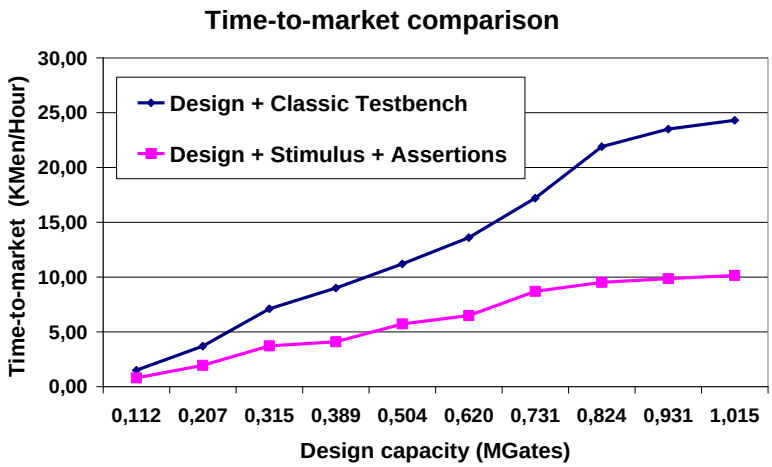


Рис. 7.11 Порівняльний аналіз методів верифікації

Висока продуктивність і технологічність поєднання системи аналізу асерцій і HDL-симулятора компанії Aldec в значній мірі досягається за рахунок інтеграції з внутрішніми компонентами симулятора, в тому числі, компіляторами HDL-мов. Обробка результатів системи аналізу асерцій забезпечується набором візуальних засобів системи Riviera, що дозволяють полегшити діагностування та усунення функціональних порушень. Модель аналізу асерцій може бути імплементована також в апаратні засоби з певними обмеженнями на підмножину підтримуваних мовних структур. Продукти Riviera, включаючи компоненти темпоральної верифікації асерцій, які дозволяють поліпшити якість проекту на 3-5%, в даний час займають провідні позиції на світовому IT-ринку з кількістю системних інсталяцій 5000 на рік у 200 компаніях і університетах більш ніж 20 країн світу.

7.9. Теоретичні основи дедуктивного аналізу дефектів

Пропонується дедуктивно-паралельний метод моделювання несправностей [22]–[35], орієнтований на обробку цифрових проектів великої розмірності вентильного і реєстрового рівнів опису з метою отримання таблиці несправностей і оцінки якості покриття тестом дефектів заданого класу. Об'єкт тестування представлений у формі структур, таблиць, булевих рівнянь, кубічних покриттів і реалізують складну цифрову систему, яка імплементується в кристали SoC. Пропонований метод моделювання несправностей представляє поєднання переваг дедуктивного визначення списків несправностей, ефективного з позиції математики, і виконання паралельних процедур, орієнтованих на високошвидкісну обробку цифрових пристроїв вентильного, системного і реєстрового рівнів опису.

Мета – створення швидкодіючого методу моделювання одиночних константних несправностей для оцінки якості синтезованих тестів цифрових систем, які імплементуються в кристали, що містять мільйони вентилів.

Основа (Backtraced Deductive-Parallel) дедуктивно-паралельного моделювання несправностей - методи підвищення швидкодії аналізу несправностей [22], [23], [35], дедуктивна модель транспортування несправностей [36], [37], паралельний метод обробки списків дефектів функціонального елемента [36] і алгоритм зворотного простежування примітивів [28] при обробці цифрового пристрою.

Дедукція є умовивід в системі доказів від загального до конкретного. У разі її застосування до аналізу дефектів мається на увазі знаходження таких алгебрологічних закономірностей, які дозволяють використовувати одного разу отримані складні моделі, багаторазово використовувані для обробки цифрових систем з метою моделювання несправностей. При цьому кожен дефект повинен бути спочатку описаний за допомогою таблиці істинності, булева рівняння, графа переходів. Фактично модель дедуктивного аналізу несправностей довільної цифрової функціональності дозволяє за одну ітерацію (кілька – для послідовних схем з глобальними зворотними зв'язками) обробки схеми обчислювати всі дефекти, що перевіряються на тест-векторі. Математична

модель дедуктивного аналізу дефектів цифрових систем може бути подана матричним рівнянням [25], [29]–[31], [33]–[35]:

$$(T_1,T_2,...,T_t,...,T_n)\oplus\begin{bmatrix} C_{11},C_{12},...,C_{1i},...,C_{1n} \\ ..... \\ C_{t1},C_{t2},...,C_{ti},...,C_{tn} \\ ..... \\ C_{k1},C_{k2},...,C_{ki},...,C_{kn} \end{bmatrix}=\begin{bmatrix} L_{11},L_{12},...,L_{1i},...,L_{1n} \\ ..... \\ L_{t1},L_{t2},...,L_{ti},...,L_{tn} \\ ..... \\ L_{k1},L_{k2},...,L_{ki},...,L_{kn} \end{bmatrix} \quad (7.23)$$

де C – кубічне покриття справної поведінки пристрою, що мають n ліній; $T=(T_1,T_2,...,T_t,...,T_n)$ – тест-вектор для перевірки дефектів, що спотворюють роботу функціональності C, який довизначений в процесі справного моделювання на множині вхідних, внутрішніх і вихідних ліній; координата матриці дефектів визначається на основі виконання логічної операції XOR: $L_{ti}=T_t\oplus C_{ti}$; матриця $L=\left|L_{ti}\right|$ – дедуктивна функція (ДФ) моделювання несправностей на тест-векторі T, що відповідає справному елементу з покриттям C, яка дає можливість обчислювати список вхідних несправностей, що транспортуються на виходи елементу [36].

У загальному випадку функція цифрового пристрою зображена таблицею істинності, а застосування дедуктивної формули (7.23) дозволяє отримати для заданого тест-вектора T таблицю перевірки несправностей, за якою можна записати аналітичну формулу моделювання дефектів. Приклади отримання таких функцій представлені нижче у вигляді (тест-вектор, таблиця істинності, таблиця перевірки дефектів):

X_1	X_2	Y_1
0	1	0

$\oplus$

X_1	X_2	Y_1
0	0	0
0	1	0
1	0	1
1	1	1

$=$

X_1	X_2	L_1
0	1	0
0	0	0
1	1	1
1	0	1

$\rightarrow L_1=X_1X_2\vee X_1\overline{X}_2;$

X_1	X_2	Y_2
1	1	1

$\oplus$

X_1	X_2	Y_2
0	0	0
0	1	0
1	0	0
1	1	1

$=$

X_1	X_2	L_2
1	1	1
1	0	1
0	1	1
0	0	0

$\rightarrow L_2=X_1X_2\vee X_1\overline{X}_2\vee\overline{X}_1X_2.$ (7.24)

Дедуктивні функції записані у вигляді диз'юнктивної нормальної форми за конституентами одиниць таблиць перевірки дефектів. Якщо модель виробу подана у вигляді структури логічних елементів або більших компонентів, то дедуктивний аналіз кожного примітиву цифрової схеми здійснюється у відповідності з виразом:

$$L_{ti}=T_t\oplus F_i=f_{ti}[(X_{i1}\oplus T_{t1}),(X_{i2}\oplus T_{t2}),...,(X_{ij}\oplus T_{tj}),...,(X_{in_i}\oplus T_{tn_i})]\oplus T_{ti}, \quad (7.25)$$

яке ізоморфно формулі (7.23). Таким чином, формули (7.23) і (7.25) покривають всі цифрові системи, описи яких представлені як на високому (системний, регістровий), так і на низькому (вентильному) рівнях.

7.10. Синтез дедуктивних компонентів для функцій SoC

Вентильний рівень опису схеми характеризується логічними елементами, функціонування яких задається таблицями істинності, кубічними покриттями або логічними рівняннями. В даному випадку технологічно розглядати процедури синтезу на основі використання аналітичної форми. При цьому двохвходовий логічний елемент трансформується в чотиривходовий, де два додаткових входи (a, b) є регістровими і служать для транспортування списків несправностей. При цьому булеві входи (x, y), по суті, є керуючими для виконання операцій над зовнішніми списками дефектів. Нехай є логічний елемент And, для якого дедуктивна функція представлена у вигляді карти Карно [38]:

$L = f(x, y, a, b) =$

$(x, y) \setminus (a, b)$	00	01	11	10
00	0	0	1	0
01	0	0	0	1
11	0	1	1	1
10	0	1	0	0

(7.26)

Мінімізація примітиву, заданого в (7.26), призводить до трьох варіантів дедуктивної функції з різною обчислювальною складністю (кількість змінних і термів) за Квайном [19], [15] і [17]:

$$1)\ L = f(x, y, a, b) = (\bar{x}\,\bar{y} \wedge ab) \vee (y \wedge a\bar{b}) \vee (x \wedge \bar{a}b) \vee (xy \wedge b) \vee (xy \wedge a) =$$
$$= (\bar{x}\,\bar{y} \wedge ab) \vee (x \wedge \bar{a}b) \vee (y \wedge a\bar{b}) \vee [(xy \wedge (a \vee b)];$$
$$2)\ L = f(x, y, a, b) = (\bar{x}\,\bar{y} \wedge ab) \vee (y \wedge a\bar{b}) \vee (x \wedge \bar{a}b) \vee (xy \wedge a) = \qquad (7.27)$$
$$= (\bar{x}\,\bar{y} \wedge ab) \vee (x \wedge \bar{a}b) \vee [(ya \wedge (x \vee \bar{b})];$$
$$3)\ L = f(x, y, a, b) = (\bar{x}\,\bar{y} \wedge ab) \vee (y \wedge a\bar{b}) \vee (x \wedge \bar{a}b) \vee (xy \wedge b) \vee (xy \wedge a) =$$
$$= (\bar{x}\,\bar{y} \wedge ab) \vee [(ya \wedge (x \vee \bar{b})] \vee [(xb \wedge (y \vee \bar{a})].$$

Вибір кращої з них призводить до реалізації формули під номером 2. Аналогічні перетворення, що застосовні до елементів Or, Not, призводять до синтезу булевих рівнянь і схемним структурам. Елемент Or. Синтез його дедуктивної функції представлений наступними перетвореннями:

$L = f(x, y, a, b) =$

$(x, y) \setminus (a, b)$	00	01	11	10
00	0	1	1	1
01	0	1	0	0
11	0	0	1	0
10	0	0	0	1

$$1)\ L = f(x,y,a,b) = (\bar{x}\,\bar{y} \wedge a) \vee (\bar{y} \wedge a\bar{b}) \vee (\bar{x} \wedge \bar{a}b) \vee (xy \wedge ab) =$$
$$= [\bar{y}a \wedge (\bar{x} \vee \bar{b})] \vee (\bar{x} \wedge \bar{a}b) \vee (xy \wedge ab);$$
$$2)\ L = f(x,y,a,b) = (\bar{x}\,\bar{y} \wedge a) \vee (\bar{x}\,\bar{y} \wedge b) \vee (\bar{y} \wedge a\bar{b}) \vee (\bar{x} \wedge \bar{a}b) \vee (xy \wedge ab) = \quad (7.28)$$
$$= [\bar{x}\,\bar{y}(a \vee b)] \vee (\bar{y} \wedge a\bar{b}) \vee (\bar{x} \wedge \bar{a}b) \vee (xy \wedge ab);$$
$$3)\ L = f(x,y,a,b) = (\bar{x}\,\bar{y} \wedge a) \vee (\bar{y} \wedge a\bar{b}) \vee (\bar{x} \wedge \bar{a}b) \vee (xy \wedge ab).$$

За аналогією виконується синтез дедуктивної функції для елемента Хог:

$L = f(x,y,a,b) =$

$(x,y) \setminus (a,b)$	00	01	11	10
00	0	1	0	1
01	0	1	0	1
11	0	1	0	1
10	0	1	0	1

(7.29)

$$L = f(x,y,a,b) = \bar{a}b \vee a\bar{b}. \quad (7.30)$$

Результати апаратної реалізації мінімальних з точки зору оцінки по Квайну дедуктивних функцій трьох згаданих елементів імплементуються в схеми, представлені на *Рис. 7.12*.

Регістровий рівень опису компонентів цифрової системи відрізняється функціональною складністю, що впливає на розмірність таблиць істинності або кубічних покриттів. Тут розглядаються такі функціональності як: тригери, засувки, лічильники, мультиплексори, регістри, шинні структури. Аналогічні перетворення, спрямовані на синтез дедуктивної функції за допомогою таблиці істинності тригера (три булевих і три регістрових входи) $Q = DC \vee \overline{C\bar{D}}Q(t-1)$, дають результат:

$L = f(T,X) =$

$(T) \setminus (X)$	000	001	011	010	110	111	101	100
000	0	1	1	0	1	1	0	0
001	0	1	1	0	0	0	1	1
011	0	1	1	0	1	1	0	0
010	0	1	1	0	0	0	1	1
110	0	0	1	1	1	0	0	1
111	0	0	1	1	0	1	1	0
101	0	0	1	1	1	0	0	1
100	0	0	1	1	0	1	1	0

(7.31)

$$L = f(c,d,q,C,D,Q) = (\bar{c} \wedge \bar{C}\bar{D}Q) \vee (\bar{C}DQ) \vee (c \wedge \bar{C}D\bar{Q}) \vee$$
$$\vee (\bar{c}d\bar{q} \wedge CD) \vee (\bar{c}\bar{d}q \wedge C\bar{D}) \vee (\bar{c}dq \wedge CD) \vee (\bar{c}d\bar{q} \wedge C\bar{D}) \vee \quad (7.32)$$
$$\vee (cd\bar{q} \wedge C\bar{Q}) \vee (cdq \wedge CQ) \vee (c\bar{d}q \wedge C\bar{Q}) \vee (c\bar{d}\bar{q} \wedge CQ).$$

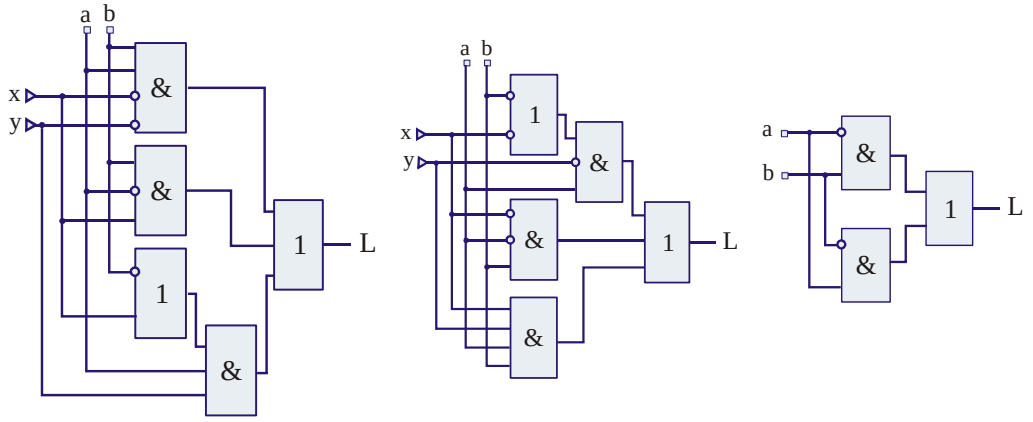


Рис. 7.12 Дедуктивні примітиви логічних елементів (and, or, xor)

Апаратна реалізація (оцінка за Квайном – 62) тригера зображена на *Рис. 7.13*. Дедуктивна функція тригера має більш ніж 10-кратну апаратну надмірність за його порівнянню зі справною функціональністю. Тим не менш, таке уявлення дає можливість вигравати у швидкодії дедуктивного моделювання несправностей в сотні разів.

Що стосується аналізу компонентів системного рівня, то в загальному випадку таблиця істинності (переходів-виходів) є частково або не повністю визначеною. Це означає, що алфавіт опису координат таблиці містить три символи, принаймні (0,1, X). Для даного випадку необхідна модифікація процедури дедуктивного аналізу несправностей, яка в троїчному алфавіті виконання операцій має вигляд:

$$L_r = \bigvee_{\forall i(T_r \oplus C^z_{ir}=1)} [(\bigotimes_{\forall j(T_j \oplus C^x_{ij}=1)} L_j) \setminus (\bigvee_{\forall j[(T_j \oplus C^x_{ij}=0)\&(C^x_{ij}\neq X)]} L_j), \qquad (7.33)$$

де n – число рядків (кубів); m – кількість вхідних ліній; k – число вихідних ліній в пристрої (примітиві); L_r – список несправностей, який формується для виходу r у вигляді дефектів, що транспонуються через примітив або цифровую систему, від зовнішніх входів.

Основні операції в троїчному алфавіті мають вигляд:

$$\begin{aligned} Xor &= [0\oplus 0=0;0\oplus 1=1;1\oplus 0=1;1\oplus 1=0;0\oplus X=X;1\oplus X=X;X\oplus X=X] \\ Or &= [0\vee 0=0;0\vee 1=1;1\vee 0=1;1\vee 1=1;0\vee X=X;1\vee X=1;X\vee X=X] \\ And &= [0\wedge 0=0;0\wedge 1=0;1\wedge 0=0;1\wedge 1=1;0\wedge X=0;1\wedge X=X;X\wedge X=X] \end{aligned} \qquad (7.34)$$

З урахуванням введених визначень далі пропонується синтез дедуктивної функції для функціональності системного рівня, заданої у вигляді граф-схеми алгоритму на *Рис. 7.14*.

Матриця переходів абстрактного автомата, що відповідає граф-схемі на *Рис. 7.14*, а також таблиця переходів-виходів структурного автомату з кодованими станами вхідних внутрішніх і вихідних змінних зображені нижче в наступному вигляді:

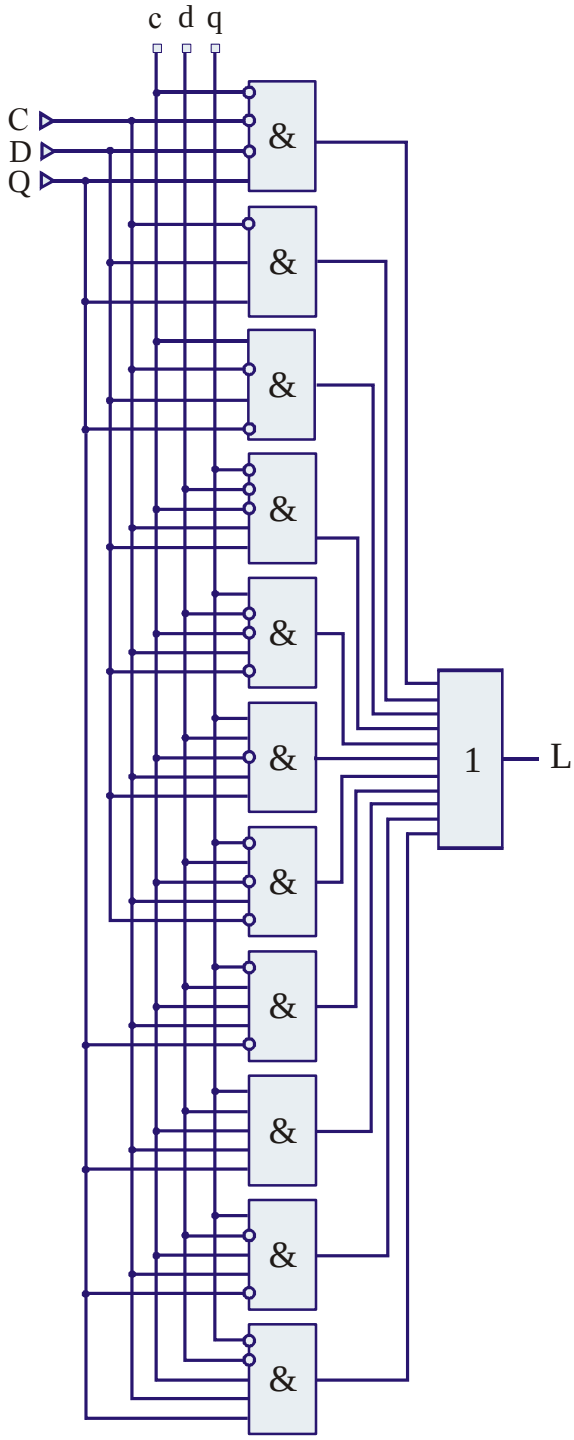


Рис. 7.13 Дедуктивна функція аналізу дефектів для D-тригера

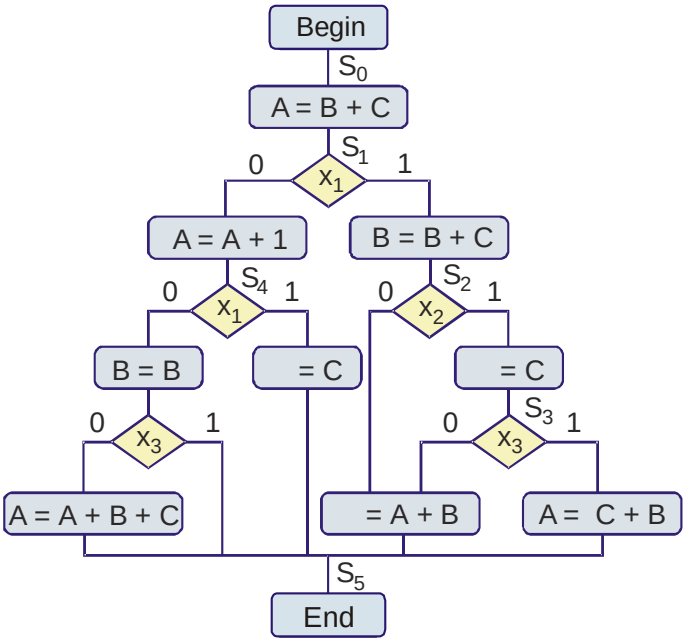


Рис. 7.14 Граф-схема функціональності

$$C = \begin{array}{|c|c|c|c|}\hline X & S_i & S_{i+1} & Y \\ \hline - & S_0 & S_1 & Y_0 : A = B + C \\ X_1 & S_1 & S_2 & Y_1 : B = B + C \\ \overline{X_1} & S_1 & S_4 & Y_2 : A = \overline{A} + 1 \\ X_2 & S_2 & S_3 & Y_3 : C = \overline{C} \\ \overline{X_2} & S_2 & S_6 & Y_4 : C = A + B \\ X_3 & S_3 & S_6 & Y_5 : A = \overline{C} + B \\ \overline{X_3} & S_3 & S_6 & Y_4 : C = A + B \\ X_1 & S_4 & S_6 & Y_3 : C = \overline{C} \\ \overline{X_1} & S_4 & S_5 & Y_6 : B = \overline{B} \\ X_3 & S_5 & S_6 & - \\ \overline{X_3} & S_5 & S_6 & Y_7 : A = A + \overline{B} + C \\ \hline \end{array} = \begin{array}{|c|c|c|c|}\hline X & S_i & S_{i+1} & Y \\ \hline XXX & 000 & 001 & 000 \\ 1XX & 001 & 010 & 001 \\ 0XX & 001 & 100 & 010 \\ X1X & 010 & 011 & 011 \\ X0X & 010 & 110 & 100 \\ XX1 & 011 & 110 & 101 \\ XX0 & 011 & 110 & 100 \\ 1XX & 100 & 110 & 001 \\ 0XX & 100 & 101 & 110 \\ XX1 & 101 & 110 & XXX \\ XX0 & 101 & 110 & 111 \\ \hline \end{array} \quad (7.35)$$

В даному випадку вхідними змінними вважаються вектори, що об’єднані змінними (XSi), вихідними лініями є – (Si + 1Y). Для побудови дедуктивної матриці, яка задає примітив моделювання всіх несправностей, відповідних структурному автомату, необхідно побудувати таблицю істинності на множині рядків або кубів покриття. Вручну дана процедура важко реалізувати. У

комп'ютерному виконанні вона не представляє труднощів. Для одного вхідного вектора матриця дедуктивного аналізу дефектів, яка виходить в результаті виконання операції Xor між вхідною послідовністю і всіма координатами матриці справної поведінки, буде мати вигляд:

$$T \oplus C = L \to (100100\ 110001) \oplus$$

X	S_i	S_{i+1}	Y
XXX	000	001	000
$1XX$	001	010	001
$0XX$	001	100	010
$X1X$	010	011	011
$X0X$	010	110	100
$XX1$	011	110	101
$XX0$	011	110	100
$1XX$	100	110	001
$0XX$	100	101	110
$XX1$	101	110	XXX
$XX0$	101	110	111

$\oplus$

X	S_i	L_{i+1}	L_Y
XXX	100	111	001
$0XX$	101	100	000
$1XX$	101	010	011
$X1X$	110	101	010
$X0X$	110	000	101
$XX1$	111	000	100
$XX0$	111	000	101
$0XX$	000	000	000
$1XX$	000	011	111
$XX1$	001	000	XXX
$XX0$	001	000	110

(7.36)

Природно, що дана дедуктивна модель є структурою регістрового рівня, яку можна реалізовувати в кристалі FPGA, де для завдання функцій використовуються таблиці істинності безпосередньо. Проте, можлива схемна реалізація дедуктивних функцій (Si + 1Y), записаних в ДНФ за констітуентами одиниць у відповідному стовпці:

$$\begin{aligned} L^1_{i+1} &= S^1_i \overline{S^2_i}\,\overline{S^3_i} \vee \overline{X_1} S^1_i \overline{S^2_i} S^3_i \vee X_2 S^1_i S^2_i \overline{S^3_i}; \\ L^2_{i+1} &= S^1_i \overline{S^2_i}\,\overline{S^3_i} \vee X_1 S^1_i \overline{S^2_i} S^3_i \vee X_1 \overline{S^1_i}\,\overline{S^2_i}\,\overline{S^3_i}; \\ L^3_{i+1} &= S^1_i \overline{S^2_i}\,\overline{S^3_i} \vee X_2 S^1_i S^2_i \overline{S^3_i} \vee X_1 \overline{S^1_i}\,\overline{S^2_i}\,\overline{S^3_i}. \end{aligned} \quad (7.37)$$

$$\begin{aligned} L^1_Y &= \overline{X_2} S^1_i S^2_i \overline{S^3_i} \vee X_3 S^1_i S^2_i S^3_i \vee \overline{X_3} S^1_i S^2_i S^3_i \vee X_1 \overline{S^1_i}\,\overline{S^2_i}\,\overline{S^3_i} \vee \overline{X_3}\,\overline{S^1_i}\,\overline{S^2_i}\,\overline{S^3_i}; \\ L^2_Y &= X_1 S^1_i \overline{S^2_i} S^3_i \vee X_2 S^1_i S^2_i \overline{S^3_i} \vee X_1 \overline{S^1_i}\,\overline{S^2_i}\,\overline{S^3_i} \vee \overline{X_3}\,\overline{S^1_i}\,\overline{S^2_i} S^3_i; \\ L^3_Y &= S^1_i \overline{S^2_i}\,\overline{S^3_i} \vee X_1 S^1_i \overline{S^2_i} S^3_i \vee \overline{X_2} S^1_i S^2_i \overline{S^3_i} \vee \overline{X_3} S^1_i S^2_i S^3_i \vee X_1 \overline{S^1_i}\,\overline{S^2_i}\,\overline{S^3_i}. \end{aligned} \quad (7.38)$$

Вирази (7.37) і (7.38) визначають умови формування списків несправностей за шістьома виходами на тест-векторі (100100 110001). Навіть на одному векторі виходить досить складна цифрова схема, зображена на *Рис. 7.15*, апаратурні витрати якої за Квайном мають оцінку 42. Ще більш складний результат у вигляді схеми має функція виходів, яка реалізується за допомогою 84 входів і 17 логічних елементів.

Таким чином, реалізація дедуктивної функції граф-схеми алгоритму на одному вхідному наборі має обчислювальну складність 84 + 42 = 126. Якщо

таку комбінаційну схему мультиплікувати на 2^{12} наборах, то в гіршому випадку апаратурні витрати приведуть до структури, яка визначається оцінкою:

$$Q = Q^t \times 2^{2\times(|X|+|S_i|)} = 126 \times 2^{12} = 516\,096. \qquad (7.39)$$

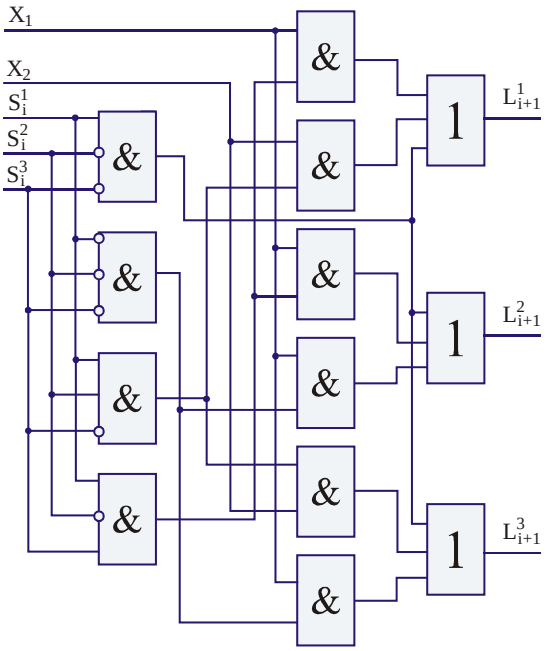


Рис. 7.15 Дедуктивна схема аналізу дефектів

Природно, що півмільйон вентилів є неприйнятною кількістю для моделювання несправностей, нехай навіть зі швидкістю, що перевищує в сотні разів програмний аналог. Виходом, в даному випадку, може служити гібридне рішення - програмно-апаратний комплекс моделювання несправностей, гнучкий по відношенню до тест-векторів. У цьому випадку програмно-орієнтована модель дедуктивного аналізу генерується в реальному масштабі часу, як функція від fault-free поведінки і тестових перевіряючих послідовностей $L = f(T, C)$. У даному випадку автоматна модель процесу аналізу дефектів, розгорнута в часі (X, Z, Y - множина вхідних, внутрішніх і вихідних змінних відповідно), буде мати наступний вигляд:

$$M = \langle L, T, C, X, Z, Y \rangle,$$
$$\begin{cases} L_z^t = f(T_x^t, T_z^{t-1}, C); \\ L_y^t = f(T_x^t, T_z^{t-1}, C). \end{cases} \qquad (7.40)$$

Таким чином, технологія навіть апаратного вбудованого моделювання повертається в поле програмно-орієнтованих рішень. Справді, ринок електронних технологій в найближчі роки має тенденцію до flexible reusable software solutions. Такий напрям має під собою підстави: 1. Реалізація системи

на кристалі стає все більш програмно-орієнтованою, оскільки через 5 років пам'ять на кристалі становитиме 94% від його площі; 2. Для управління обчислювальними процесами, пов'язаними з моделюванням, необхідно мати на кристалі мікропроцесор, виконаний за гнучкою технологією і вбудований в пам'ять, або за жорсткою технологією - виконаний на кристалі.

7.11. Структурні моделі примітивів стимулятора

У загальному випадку отримання дедуктивних примітивів для паралельного моделювання несправностей пов'язано з синтезом функцій на вичерпному тесті. Складність дедуктивних примітивів залежить від рівня представлення функціональностей. Найбільш простими є структури вентильного рівня у вигляді базису логічних елементів And, Or, Not.

За допомогою основного виразу (7.25) синтезу дедуктивних функцій, що транспортують дефекти через логічний елемент, виконується побудова всіх базових компонентів (And, Or, Not):

$$
\begin{aligned}
&L_{And}[T=(00,01,10,11),F=(X_1\wedge X_2)]=L\{(\bar{x}_1\bar{x}_2\vee\bar{x}_1x_2\vee x_1\bar{x}_2\vee x_1x_2)\wedge\\
&\wedge[(X_1\oplus T_{t1}\wedge X_2\oplus T_{t2})\oplus T_{t3})]\}=(\bar{x}_1\bar{x}_2)\{[(X_1\oplus 0)\wedge(X_2\oplus 0)]\oplus 0\}\vee\\
&\vee(\bar{x}_1x_2)\{[(X_1\oplus 0)\wedge(X_2\oplus 1)]\oplus 0\}\vee(x_1\bar{x}_2)\{[(X_1\oplus 1)\wedge(X_2\oplus 0)]\oplus 0\}\vee\\
&\vee(x_1x_2)\{[(X_1\oplus 1)\wedge(X_2\oplus 1)]\oplus 1\}=(\bar{x}_1\bar{x}_2)(X_1\wedge X_2)\vee(\bar{x}_1x_2)(X_1\wedge\bar{X}_2)\vee\\
&\vee(x_1\bar{x}_2)(\bar{X}_1\wedge X_2)\vee(x_1x_2)(X_1\vee X_2);\qquad(7.41)\\
&L_{Or}[T=(00,01,10,11),F=(X_1\vee X_2)]=(\bar{x}_1\bar{x}_2)(X_1\vee X_2)\vee(\bar{x}_1x_2)(\bar{X}_1\wedge X_2)\vee\\
&\vee(x_1\bar{x}_2)(X_1\wedge\bar{X}_2)\vee(x_1x_2)(X_1\wedge X_2);\\
&L_{Not}[T=(0,1),F=\bar{X}_1]=L\{(\bar{x}_1\vee x_1)[(\overline{X_1\oplus T_{t1}})\oplus T_{t2}]\}=\bar{x}_1\wedge\\
&\wedge[(\overline{X_1\oplus 0})\oplus 1]\vee x_1[(\overline{X_1\oplus 1})\oplus 0]=\bar{x}_1\bar{\bar{X}}_1\vee x_1\bar{\bar{X}}_1=\bar{x}_1X_1\vee x_1X_1.
\end{aligned}
$$

У наведених рівняннях $T_t=(T_{t1},T_{t2},T_{t3}),(t=\overline{1,4})$ – тест-вектор, що має 3 координати, де остання з них визначає стан виходів елементів And, Or. Що стосується інвертора, то тут тест-вектор має 2 координати: $T_t=(T_{t1},T_{t2}),(t=\overline{1,2})$, де остання координата є стан виходу елемента. Рівняння для інвертора ілюструє неістотність операції інверсії на виході елемента для транспортування дефектів. Тому дана функція (Not) не присутня на виходах дедуктивних примітивів. Апаратна реалізація дедуктивних функцій [30], [33], [34] для двовходових елементів (And, Or) на вичерпному тесті зображена на *Рис. 7.16* схемою дедуктивно-паралельного аналізу дефектів.

У симуляторі представлені булеві (x1, x2) і регістрові (X1, X2) змінні, сигнал V вибору типу справної функції: V = 0 (And), V = 1 (Or), вихідна регістрова змінна Y. Стани двійкових входів x1, x2 і V формують одну з чотирьох дедуктивних функцій для отримання вектора несправностей Y. Імплементація дедуктивної моделі в HDL-коді подана *лістингом 7.4*.

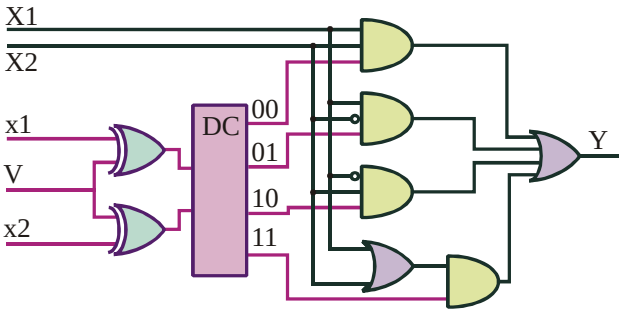


Рис. 7.16 Симулятор несправностей

Лістінг 7.4 – VHDL-модель секвенсора

```
library IEEE;
use IEEE.STD_LOGIC_1164.all;
entity Fub1 is
  port( i0, i1 : in STD_LOGIC;
          o00, o01, o10 , o11 : out STD_LOGIC);
end Fub1;
architecture Fub1 of Fub1 is
begin
  o00 <= not i0 and not i1;
  o01 <= not i0 and i1;
  o10 <= i0 and not i1;
  o11 <= i0 and i1;
end Fub1;
library IEEE;
use IEEE.std_logic_1164.all;
entity sequenstor is
  port( V ,  X1_s, X2_s , x1, x2 : in STD_LOGIC;
         Y : out STD_LOGIC);
end sequenstor;
architecture sequenstor of sequenstor is
component Fub1
        port( i0, i1 : in STD_LOGIC;
          o00, o01, o10 , o11 : out STD_LOGIC);
end component;
signal a0, a1, a2, a3, a4 : STD_LOGIC;
signal o00, o01, o10, o11 : STD_LOGIC;
signal x3, x4 : STD_LOGIC;
begin
 U1 : Fub1  port map(i0 => x3, i1 => x4, 00 => o00, o01 => o01,
o10 => o10, o11 => o11 );
 a0 <= o00 and X2_s and X1_s;
 a1 <= not(X2_s) and o01 and X1_s;
 a2 <= not(X1_s) and X2_s and o10;
 a3 <= X2_s or X1_s;
 a4 <= o11 and a3;
 Y <= a4 or a2 or a1 or a0;
 x3 <= V xor x1;
 x4 <= x2 xor V;
end sequenstor;
```

Робота симулятора демонструється в таблиці паралельного моделювання 8-розрядних вхідних векторів несправностей з метою отримання на Y-виході вектора дефектів для логічних елементів And, Or:

$(V, x1, x2) =$	000	100	011	111	010	110
$X1(RG)$	01110001	01110001	10110110	00111011	00101010	10111001
$X2(RG)$	01111000	01111000	10110101	00110100	10111001	00101010
$Y(RG)$	01110000	01111001	10110111	00110000	10010001	10010001

Застосування такого симулятора дає можливість трансформувати вентильну модель F справної поведінки схеми в дедуктивну L, яка інваріантна в сенсі універсальності тестовим наборам і не припускає в процесі моделювання використовувати модель F. Тому симулятор, як апаратна модель ДФ, орієнтований на створення вбудованих засобів дедуктивно-паралельного моделювання , що підвищують швидкодію аналізу в 10 - 1000 разів у порівнянні з програмною реалізацією. Але при цьому співвідношення обсягів післясинтезних моделей справного моделювання та аналізу несправностей становить 1:16. Апаратний аналіз несправностей спрямований на розширення функціональних можливостей вбудованих засобів справного моделювання (HESTM - Hardware Embedded Simulator) фірми Aldec (www.aldec.com). Обчислювальна складність обробки проекту, що складається з n вентилів, дорівнює $Q = (2n^2\tau)/W$, де τ – час виконання регістрової операції (And, Or, Not); W – розрядність регістра.

Для апаратної реалізації дедуктивно-паралельного моделювання на основі запропонованого симулятора може бути використана обчислювальна структура, що зображена на *Рис. 7.17*.

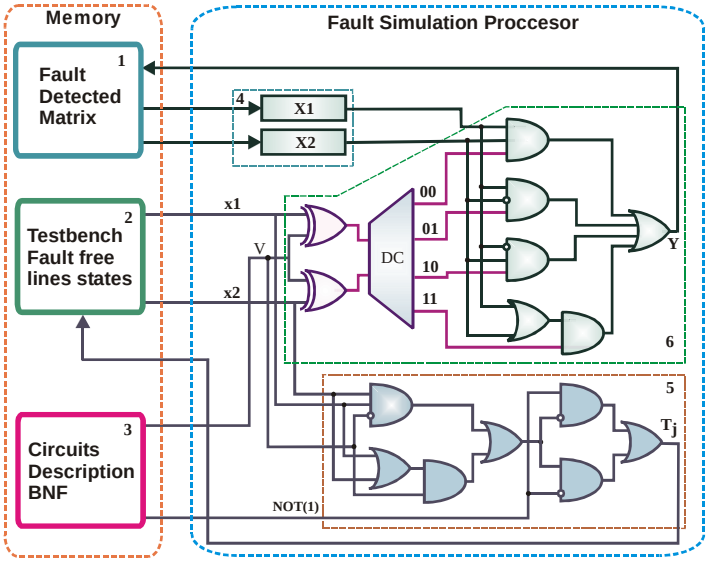


Рис. 7.17 HFS-структура апаратного моделювання

Особливість схемної реалізації полягає у спільному виконанні двох операцій: однобітових – для емуляції функцій логічних елементів And, Or і паралельної – для обробки багаторозрядних векторів несправностей шляхом виконання операцій логічного множення, заперечення і додавання. Функціональне призначення основних блоків (пам'ять і процесор): 1. $M = [M_{ij}]$ – квадратична матриця моделювання несправностей, де i, j = 1, q; q - загальна кількість ліній в оброблюваній схемі. 2. Вектори збереження станів справного моделювання, визначені в моменти часу t-1 і t, необхідні для формування дедуктивних функцій примітивів. 3. Модуль пам'яті для зберігання схемного опису у вигляді структури логічних елементів. 4. Буферні регістри, розмірністю q, для зберігання операндів і виконання регістрових паралельних операцій над векторами несправностей, що зчитані з матриці M. 5. Блок справного моделювання для визначення двійкового стану виходу чергового оброблюваного логічного елемента. 6. Дедуктивно-паралельний симулятор, що обробляє за один такт дві регістрових змінних X1, X2 з метою визначення вектора дефектів, що транспортуються на вихід логічного елемента Y.

Перевага запропонованої структури моделювання несправностей. 1. Суттєве зменшення кількості модельованих дефектів, обумовлених тільки числом збіжних розгалужень, яке становить до 20% від загального числа ліній. 2. Зниження обсягу пам'яті, необхідного для зберігання матриці модельованих несправностей. 3. Простота реалізації Hardware Fault Simulator (HFS) в апаратному виконанні, що дозволяє на порядок збільшити швидкодію моделювання несправностей. 4. Використання HFS як першої фази дедуктивно-топологічного методу, який ґрунтується на результаті обробки збіжних розгалужень для швидкодіючого аналізу деревовидних структур.

Маршрут моделювання цифрових систем на кристалах з попередніми розбиттям моделі пристрою на дві структурні організації (збіжні розгалуження і деревовидні підграфи) зображено на *Рис. 7.18*.

Підсумки запропонованої технології моделювання з попереднім розбиттям схеми на збіжні розгалуження і деревовидні підграфи. Дедуктивно-паралельний аналіз дефектів на основі зворотного простежування несправностей вимагає практично лінійних витрат пам'яті і часу, що залежать від числа ліній схеми. Витрати часу для обробки збіжних розгалужень мають квадратичну залежність від їх числа: $Q = (r^2 / W) + n_r + n_p + (n - r - r^0)$. Тут (r^2 / W) – час моделювання несправностей r збіжних розгалужень, число яких визначається как $r = 0.2 \times n$; $n_r = n -$ час реконфігурування примітивів схеми на вхідном наборі; $n_p = n$ – час пошуку підграфів ліній, що відповідають неперівірюваним збіжним розгалуженням; $(n - r - r^0) = n - 0.2 \times n - 0.4 \times n = 0.4 \times n$ – час виконання суперпозиції рішень на множині ліній схеми без збіжних розгалужень і попередників для неперевірюваних збіжних розгалужень. Враховуючи фактичні значення параметрів у функції від числа ліній схеми, можна отримати оцінку швидкодії дедуктивно-паралельного методу:

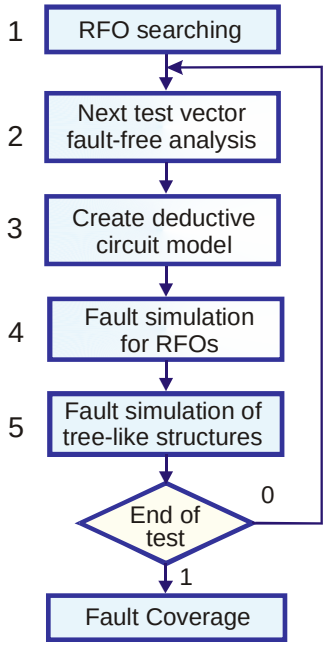


Рис. 7.18 Модель процесу дедуктивно-паралельного моделювання

$$Q = [(0.2 \times n)^2 / W] + n + n + (n - 0.2 \times n - 0.4 \times n) = \\ = [(0.2 \times n)^2 / W] + 2.4 \times n). \qquad (7.42)$$

Таким чином, виграш за швидкодією запропонованого методу тим більше, чим менше відсоток збіжних розгалужень у схемі цифрового пристрою [22], [23], [26], [27].

Для порівняння паралельний алгоритм має обчислювальну складність Ср, яка визначається функціональною залежністю від числа нееквівалентних несправностей (b), довжини комп'ютерного слова (W), кількості еквівалентних вентилів (G): $C_p = (b^2 / W) \times G^3$. Дедуктивний алгоритм має відміни у формулі оцінки швидкодії: $C_d = b^2 \times Q \times G^2 \big|_{Q=G} = b^2 G^3$, де Q – середнє число активізованих несправностями вентилів. Дедуктивно-паралельный метод без розбиття схеми має швидкодію, що визначається виразом: $C_{dp} = G^2 + (b^2 / W) \times G^2$. Перший доданок задає час справного моделювання, другий - аналізу несправностей цифрового пристрою, лінії якого не ранжовані. Для комбінаційної ранжированої схеми швидкодія методу має оцінку $C_{dp}^r = G + (b^2 / W) \times G$. Швидкодія дедуктивно-паралельного методу вище паралельного і дедуктивного $(C_{dp}^r << \{C_p, C_d\})$, завдяки розділенню фаз справного та несправного моделювань.

На основі описаної вище технології створена система **SIGETEST** - (**SI**mulation, **GE**neration of **TEST**) швидкодіюча система моделювання

несправностей і генерації тестів, що використовує моделі проектованих цифрових систем інтерпретативно-компілятивного типу. Об'єктом моделювання може виступати будь-яка цифрова структура, подана у вигляді булевих рівнянь, що реалізуються в кристалах CPLD, FPGA, ASIC. Система обробляє складні цифрові проекти, які налічують сотні тисяч логічних вентилів на стадії після синтезу (gate level description). Система має інтегроване середовище, що реалізує графічний інтерфейс високого рівня. Введення проектів здійснюється у вигляді опису. Підтримувані операції: AND, OR, NOT, XOR. Також підтримуються шинні структури. Компілятор перетворює схемний опис до виду внутрішніх структур даних, зручних для моделювання. Ядро моделювання включає алгоритми справного і несправного моделювання: Parallel, Backtraced Quasi Exact, Deductive-Parallel і Backtraced-Deductive-Parallel. Генератор тестів включає набір алгоритмів (псевдовипадкових, детермінованих, алгоритмічних) для синтезу тестових послідовностей. Результатом роботи програми є test-bench у форматі VHDL. Система надає також інформацію про процеси справного і несправного моделювання, якість покриття несправностей, статистику моделювання. Результати моделювання можна переглянути за допомогою вікна Fault Coverage, що представляє собою багатозначну таблицю несправностей.

SIGETEST має засоби для управління та моніторингу процесу синтезу тестів. Моделювання можна обмежити у часі або задати число тестових наборів, які необхідно промоделювати. Є можливість обмежити знизу відсоток покриття несправностей генерованими наборами. У процесі моделювання система надає інформацію про прогрес моделювання у відсотках від загального числа векторів або наперед заданого часового інтервалу. Система SIGETEST орієнтована на інтеграцію з сучасними засобами синтезу і моделювання, такими як ALDEC Active-HDL, Riviera, SYNOPSYS Design Compiler.

Підведемо підсумки:

1. У розділі представлені інфраструктура і технології аналізу цифрових систем. Запропоновані модель транзакційного графа і метод діагностування цифрових систем на кристаллах, орієнтовані на значне зменшення часу виявлення несправних блоків і пам'яті для зберігання компактної діагностичної матриці, яка описує тернарні відношення у форматі: монітор-орієнтовані тест-сегменти, і призначені для виявлення несправностей функціональних компонентів програмно-апаратних систем.

2. Введено новий критерій якості діагностування, що представляє собою функцію, залежну від структури графа, тестів і асерційних моніторів. Для поліпшення якості діагностування існують два альтернативні шляхи: збільшення набору тестових сегментів для розпізнавання еквівалентних несправних блоків або додавання асерційних моніторів на транзитних вершинах активізаційного графа HDL-коду. Запропонований критерій дозволяє прийняти правильне рішення.

3. Удосконалено TAB-алгоритм виявлення функціональних порушень у програмному або апаратному забезпеченні. Він відрізняється від аналогів використанням xor операції, що дозволяє підвищити продуктивність

діагностування одиночних і кратних несправних блоків за рахунок паралельного аналізу TAB-матриці, застосування граничного сканування на основі стандарту IEEE 1500, а також векторних операцій and, or, xor.

4. Розроблено модель діагностування функціональності системи на кристалі у вигляді мультідерева і метод обходу дерева, імплементований у движок для виявлення несправних блоків із заданою глибиною. Модель і метод дозволяють істотно збільшити продуктивність програмного і апаратного забезпечення інфраструктури IP.

5. Тестова верифікація методу діагностування виконана на трьох реальних прикладах, поданих компонентами SoC фільтра косинусного перетворення, який показав спроможність результатів щодо зменшення часу виявлення несправностей і пам'яті для зберігання діагностичної інформації, а також збільшення глибини діагностування цифрового модуля.

6. Описано дедуктивно-паралельний метод моделювання несправностей, орієнтований на обробку цифрових проектів великої розмірності вентильного і регістрового рівнів з метою отримання таблиці несправностей і оцінки якості покриття тестом дефектів заданого класу.

7. Запропонована технологія програмно-апаратного дедуктивно-паралельного моделювання несправностей орієнтована на створення моделей дедуктивних примітивів вентильного, регістрового і системного рівнів з метою тестування цифрових систем на кристалах, що містять мільйони вентилів. Представлена структурна модель апаратного симулятора і пристрою моделювання в цілому, які орієнтовані на істотне підвищення швидкодії засобів моделювання цифрових виробів великої розмірності, шляхом розподілу функцій справного аналізу та обчислення списків перевірюваних дефектів на вхідних наборах.

7.12. Список використаної літератури до розділу 7

[1] Sziray J. Test Design of Digital Systems / József Sziray.– Széchenyi István University.– 2010.– 160 p.

[2] Електронний ресурс: http://www.scrigroup.com/limba/engleza/92/The-Design-Flow-and-Fault-Mode51775.php

[3] Автоматизация диагностирования электронных устройств/ Ю.В.Малышенко и др./ Под ред. В.П.Чипулиса.– М.: Энергоатомиздат, 1986.– 216с.

[4] Stanisavljevi M. Reliability of Nanoscale Circuits and Systems / M. Stanisavljevi, M. Schmid, Y. Leblebici.– Springer.– 2011.– 240 p.

[5] Fan X. Fault diagnosis of VLSI designs: cell internal faults and volume diagnosis throughput / Xiaoxin Fan // PhD (Doctor of Philosophy) thesis, University of Iowa.– 2012.– 134 p.

[6] Pomeranz I. Transition Path Delay Faults: A New Path Delay Fault Model for Small and Large Delay Defects / I. Pomeranz, S.M. Reddy // IEEE Transactions on Very Large Scale Integration (VLSI) Systems.– 2008.– Vol.16, No.1.– P. 98–107.

[7] Pomeranz I. Selection of a Fault Model for Fault Diagnosis Based on Unique Responses / I. Pomeranz, S.M. Reddy // IEEE Transactions on Very Large Scale Integration (VLSI) Systems. – 2010.– Vol.18, No.11.– P. 1533-1543.

[8] Bareisa E. Functional test generation remote tool / E. Bareisa, V. Jusas, K. Motiejunas, R. Seinauskas // Proceedings 8th Euromicro Conference on Digital System Design.– 2005.– P. 192-195.
[9] Xiaoke Q. Scalable Test Generation by Interleaving Concrete and Symbolic Execution / Qin Xiaoke, P. Mishra // 27th Intern. Conf. on VLSI Design and 13th Intern. Conf. on Embedded Systems.– 2014.– P. 104-109.
[10] Hari S.K.S. Automatic Constraint Based Test Generation for Behavioral HDL Models / S.K.S. Hari, V.V.R. Konda, V. Kamakoti, V.M. Vedula K.S. Maneperambil // IEEE Transactions on Very Large Scale Integration (VLSI) Systems.– 2008.– Vol.16, No.4.– P. 408-421.
[11] Sethuram R. Fault Nodes in Implication Graph for Equivalence/Dominance Collapsing, and Identifying Untestable and Independent Faults / R. Sethuram, M.L. Bushnell, V.D. Agrawal //26th IEEE VLSI Test Symposium.– 2008. – P. 329-335.
[12] Harris I.G. Fault models and test generation for hardware-software covalidation / I.G. Harris // IEEE Design & Test of Computers.– Vol.20, No.4.– P. 40-47.
[13] Yue Jiang. Fault Prediction using Early Lifecycle Data / Jiang Yue, Bojan Cukic, T. Menzies // The 18th IEEE Intern. Symp. on Software Reliability.– 2007.– P. 237-246.
[14] Mathaikutty D.A. Model-driven test generation for system level validation / D.A. Mathaikutty, S. Ahuja, A. Dingankar, S. Shukla // IEEE International High Level Design Validation and Test Workshop.– 2007.– P. 83-90.
[15] Olsen M. A framework for simulation validation coverage / M. Olsen, M. Raunak // Winter Simulation Conference (WSC).– 2013.– P. 1569-1580.
[16] Harris I.G. Hardware-Software Covalidation: Fault Models and Test Generation / Ian G. Harris // Design and Test of Computers.– Vol. 20, Num. 4.– July-August 2003.– 12 p. Электронный ресурс: http://www.ics.uci.edu/~harris/pubdir/hldvt01hwsw.pdf
[17] Fallah F. OCCOM-efficient computation of observability-based code coverage metrics for functional verification / F. Fallah, S. Devadas, K. Keutzer // IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.– 2001.– Vol.20, No.8.– P.1003-1015.
[18] Yin Yongfeng. Real-time embedded software testing method based on extended finite state machine / Yin Yongfeng, Liu Bin, Ni Hongying // Systems Engineering and Electronics.– 2012.– Vol.23, No.2.– P. 276-285.
[19] Batth S.S. Fault Modeling and Detection Capabilities for EFSM Models / S.S. Batth, M.U. Uyar, Yu Wang, M.A. Fecko // IEEE Transactions on Instrumentation and Measurement.– June 2008.– Vol.57, No.6.– P. 1102-1111.
[20] Семенец В.В., Хаханова И.В., Хаханов В.И. Проектирование цифровых систем с использованием языка VHDL.– Харьков: ХНУРЭ.– 2003.– 492 с.
[21] Kato T. A CDFG generating method from C program for LSI design // IEEE Asia Pacific Conference on Circuits and Systems.– 2008. – P. 936-939.
[22] Wang X., Hill F.G., Mi Zh. A sequential circuit fault simulation by surrogate fault propagation // Proc. 1989 IEEE International test conference, IEEE Computer society, 1989. P. 9-18.
[23] Nishida T., Miyamoto S., Kozawa T., Satoh K. RFSIM: Reduced fault simulator // IEEE Transactions on computer-aided design. 1987. Vol. CAD-6, No 3. P. 392-402.
[24] Hahanov V., Kteaman H., Ghribi W., Fomina E. HEDEFS – Hardware embedded deductive fault simulation // Proc. of the 3rd IFAC Workshop, Rydzyna, Poland, 2006. P. 25-29.
[25] Hahanov V., Hahanova I., Obrizan V. High-performance deductive fault simulation

method. Proceedings of the 10 IEEE European test symposium.- Tallinn. Estonia.- May 22-25.- P. 91-96.

[26] Levendel Y.H., Menon P.R. Comparison of fault simulation methods – Treatment of unknown signal values // Journal of digital systems. 1980. Vol. 4. P. 443-459.

[27] Hahanov V.I., Hahanova I.V., Khan S.U., Obrizan V.I. Topological fault simulation method. Proceedings of the 11th International Conference Mixdes Design of Integrated Circuits and Systems. Szczecin. 24-26 June 2004. p.211-214.

[28] Убар Р.Р. Анализ диагностических тестов для комбинационных цифровых схем методом обратного прослеживания неисправностей // Автоматика и телемеханика. 1977. №8. С.168-176.

[29] Hahanov V.I., Babich A.V., Hyduke S.M. Test Generation and Fault Simulation Methods on the Basis of Cubic Algebra for Digital Devices. Proceedings of the Euromicro Symposium on Digital Systems Design DSD 2001. Warsaw, Poland. September, 4-6, 2001. P. 228-235.

[30] Хаханов В.И., Хак Х.М. Джахирул, Масуд М.Д. Мехеди. Модели анализа неисправностей цифровых систем на основе FPGA, CPLD // Технология и конструирование в электронной аппаратуре. 2001. № 2. С. 3-11.

[31] Хаханов В.И., Сысенко И.Ю., Хак Х.М. Джахирул, Масуд М.Д. Мехеди. Кубическое моделирование неисправностейцифровых проектов на основе FPGA, CPLD // Радиоэлектороника, информатика, управление. 2001. № 1. С. 123-129.

[32] Baghdadi Ammar Awni Abbas. Диагностирование HDL-моделей систем на кристаллах / Baghdadi Ammar Awni Abbas, В.И. Хаханов, Е.И. Литвинова, С.А. Зайченко // Радиоэлектроника и информатика. 2013. №4. С.64-72.

[33] Хаханов В.И., Сысенко И.Ю., Колесников К.В. Дедуктивно-параллельный метод моделирования неисправностей на реконфигурируемых моделях цифровых систем // Радиоэлектроника и информатика. 2002. № 1. С. 98-105.

[34] Хаханов В.И., Колесников К.В., Хаханова А.В. BDP-метод моделирования неисправностей для синтеза тестов цифровых проектов // Радиоэлектороника и информатика. 2002. № 2. С. 60-66.

[35] Хаханов В.И., Убар Р.-Й.Р. Технологии проектирования систем на кристаллах. Моделирование неисправностей сверхбольших цифровых проектов // АСУ и приборы автоматики. 2002. Вып.122. С. 16-35.

[36] Abramovici M., Breuer M.A. and Friedman A.D., Digital System Testing and Testable Design, Computer Science Press, 1998. 652 p.

[37] Автоматизированное проектирование цифровых устройств / С.С.Бадулин, Ю.М.Барнаулов и др./ Под ред. С.С. Бадулина. М.: Радио и связь. 1981. 240 с.

[38] Основы технической диагностики / Под. ред. П.П.Пархоменко. М.: Энергия, 1976. – 460 с.

[39] Пархоменко П.П. Основы технической диагностики (Оптимизация алгоритмов диагностирования, аппаратурные средства) / П.П. Пархоменко, Е.С. Согомонян. Под ред. П.П. Пархоменко. М.: Энергия, 1981. – 320 с.

[40] Хаханов В.И., Хаханова И.В., Литвинова Е.И., Гузь О.А. Проектирование и верификация цифровых систем на кристаллах. Verilog & System Verilog: Харьков. – Новое слово, 2010. – 528с.

[41] Da Silva F. The Core Test Wrapper Handbook. Rationale and Application of IEEE Std. 1500™ / F. Da Silva, T. McLaurin, T. Waayers. Springer. – 2006. – XXIX. 276 p.

[42] Marinissen E.J. Guest Editors' Introduction: The Status of IEEE Std 1500 / E.J. Marinissen, Yervant Zorian // IEEE Design & Test of Computers. – 2009. – No26(1). –

P.6-7.
[43] Benso A. IEEE Standard 1500 Compliance Verification for Embedded Cores / A. Benso, S. Di Carlo, P. Prinetto, Y. Zorian // IEEE Transactions on Very Large Scale Integration (VLSI) Systems.– April, 2008.– Vol.16, No.4.– P. 397-407.
[44] Хаханов В.И. Логический ассоциативный вычислитель / В.И. Хаханов, Е.И. Литвинова, С.В. Чумаченко, О.А. Гузь // Электронное моделирование.– 2011.– № 1.– С. 73-90.
[45] Бондаренко М.Ф. Инфраструктура мозгоподобных вычислительных процессов / М.Ф. Бондаренко, О.А. Гузь, В.И. Хаханов, Ю.П. Шабанов-Кушнаренко.– Харьков: Новое Слово.– 2010.– 160 с.
[46] IEEE Standard for Reduced-Pin and Enhanced-Functionality Test Access Port and Boundary-Scan Architecture IEEE Std 1149.7-2009.
[47] Ubar R. Block-Level Fault Model-Free Debug and Diagnosis in Digital Systems / R. Ubar, S. Kostin, J. Raik // 12th Euromicro Conference DSD '09. – 2009.– P. 229-232.
[48] Ngene Christopher Umerah. A diagnostic model for detecting functional violation in HDL-code of SoC / Ngene Christopher Umerah, V. Hahanov // Proc. of IEEE East-West Design and Test Symposium.– Sevastopol, Ukraine.– 19-20 September.– 2011.– P. 299-302.
[49] Feinstein D.Y. Partially Redundant Logic Detection Using Symbolic Equivalence Checking in Reversible and Irreversible Logic Circuits /D.Y. Feinstein, M.A. Thornton, D.M. Miller // Design, Automation and Test in Europe, DATE '08.– 2008. – P. 1378 – 1381.

ЗМІСТ