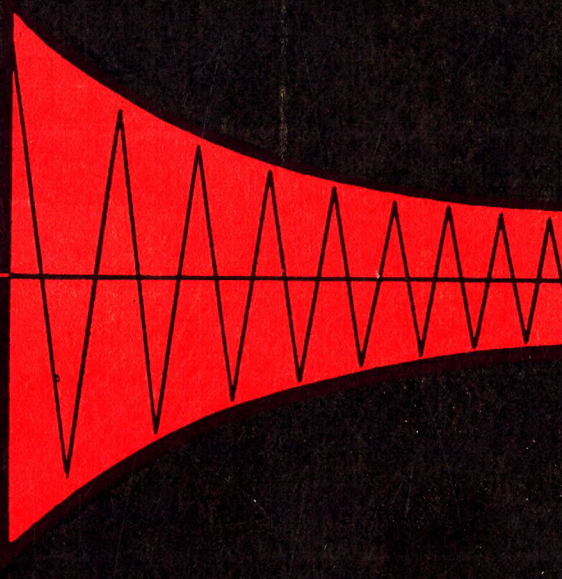
МЕТОДИ ОБЧИСЛЕНЬ

І. П. ГАВРИЛЮК
В. Л. МАКАРОВ

Підручник для вузів

2



І.П. ГАВРИЛЮК
В.Л. МАКАРОВ

МЕТОДИ ОБЧИСЛЕНЬ

У двох частинах

ЧАСТИНА 2

Затверджено Міністерством
освіти України
як підручник для студентів
вузів, які навчаються
за спеціальністю
«Прикладна математика»

КИЇВ
«ВИЩА ШКОЛА»
1995

ББК 22.193я73
Г12
УДК 51 (075.8)

Рецензенти: д-р фіз.-мат. наук *В. Я. Скоробогатько* (Інститут прикладних проблем механіки і математики ім. Я. С. Підстригача НАН України), д-р фіз.-мат. наук проф. *М. Ф. Кириченко* (Чернівецький університет)

Редакція літератури з математики, фізики, інформатики

Редактори: *Є. В. Бондарчук, Г. Г. Рубан*

Гаврилюк І. П., Макаров В. Л.

Г12 Методи обчислень: Підручник : У 2 ч.— К. : Вища шк., 1995.— Ч. 2.— 431 с. : іл.
ISBN 5-11-004412-0 (ч. 2)
ISBN 5-11-004029-X

Викладено найпоширеніші методи розв'язування операторних рівнянь, таких як задача Коші та крайові задачі для звичайних диференціальних рівнянь, крайові задачі для рівнянь еліптичного типу, початково-крайові задачі для нестаціонарних рівнянь (параболічного та гіперболічного типів). При дослідженні похибок методів на різних класах функцій широко використовується лема Бремблa — Гільберта. Даються початкові відомості про такі сучасні методи, як багатосітковий метод, метод фіктивних областей, метод граничних елементів та деякі інші.

Для студентів вузів, які навчаються за спеціальністю «Прикладна математика».

Г $\frac{1602012000-010}{211-95}$ ББК 22.193я73

ISBN 5-11-004412-0 (ч. 2)
ISBN 5-11-004029-X

ГЛАВА 1

ПРОЕКЦІЙНО-ВАРІАЦІЙНІ МЕТОДИ РОЗВ'ЯЗУВАННЯ ОПЕРАТОРНИХ РІВНЯНЬ

Якщо A — оператор (не обов'язково лінійний), що відображає свою область визначення $D(A)$ з деякого нормованого простору E в область значень $R(A)$ з нормованого простору F, то можна розглянути задачу, яка в певній мірі є оберненою до задачі, що розглядалася у першій частині, а саме: за заданим $f \in B_2$ знайти елемент $u \in B_1$ такий, що

$$Au = f.$$

Таке рівняння називається *операторним*. Далі розглядатимемо наближені методи розв'язування конкретних видів цих рівнянь.

У практиці інженерних розрахунків поширені методи, які вкладаються в загальну схему так званих проекційних та варіаційних методів. Хоча початкові ідеї цих методів різні, проте алгоритми багатьох варіаційних методів можна дістати як окремі випадки алгоритмів проекційних методів. Тому часто ці методи об'єднують в один клас проекційно-варіаційних методів.

1.1. Проекційні методи

Нехай $A : E \to F$ — лінійний оператор, що діє з нормованого простору E в нормований простір F, $D(A)$ — область визначення оператора A, а $R(A)$ — область його значень. Розглянемо рівняння

$$Au = f. \tag{1}$$

Розв'язуючи рівняння (1) проекційним методом, треба: 1) задати дві послідовності підпросторів $\{E_n\}$ та $\{F_n\}$, $E_n \subset D(A) \subset E$, $F_n \subset R(A) \subset F$, а також лінійні проекційні оператори (проектори) $p_n : F \to F_n$ такі, що $p_n^2 = p_n$, $p_n f \in F_n \ \forall f \in F$; 2) замість рівняння (1) розв'язати рівняння

$$p_n(Au_n - f) = 0 \tag{2}$$

і його розв'язок $u_n \in E_n$ взяти за наближений розв'язок рівняння (1). Залежно від вибору підпросторів E_n, F_n і проекційних операторів p дістаємо ті чи інші методи. Якщо простори E та F є просторами зі скалярним добутком, то підпростори E_n, F_n задають таким чином. Вибирають в областях $D(A)$ та F дві послідовності лінійно незалежних елементів $\{\varphi_i\}_{i=1}^{\infty}$, $\{\psi_i\}_{i=1}^{\infty}$, $\varphi_i \in D(A)$, $\psi_i \in F$. Перша

з них називається *координатною системою*, а друга — *проекційною*. За E_n та F_n вибирають лінійні оболонки елементів φ_i та ψ_i, $i = \overline{1, n}$, відповідно, а за p_n — оператор ортогонального проектування F на F_n, тобто оператор, який будь-якому $f \in F$ ставить у відповідність його ортогональну проекцію h на F, тобто $(f - h, t) = 0 \ \forall\, t \in F_n$. Як випливає з результатів п. 3.2.2 з ч. 1, елемент h є елементом найкращого наближення для f у підпросторі F_n. Оскільки $t = \sum_{j=1}^{n} a_j \psi_j$, де a_j — дійсні коефіцієнти, то рівність $(f - h, t) = 0 \ \forall\, t \in F_n$ еквівалентна системі рівностей

$$(f - h, \psi_j) = 0, \quad j = \overline{1, n}, \quad h \in F_n. \tag{3}$$

Таким чином, рівняння (2) еквівалентне системі

$$(Au_n - f, \psi_j) = 0, \quad j = \overline{1, n}, \quad u_n \in E_n. \tag{4}$$

Оскільки u_n можна подати у вигляді

$$u_n = \sum_{i=1}^{n} c_i \varphi_i, \tag{5}$$

де c_j — дійсні коефіцієнти, то з (4) дістаємо співвідношення

$$\sum_{i=1}^{n} c_i (A\varphi_i, \psi_j) = (f, \psi_j), \quad j = \overline{1, n}. \tag{6}$$

Відносно невідомих c_i (6) є системою лінійних алгебраїчних рівнянь. Якщо вона має розв'язок, то шуканий елемент u_n знаходимо у вигляді (5).

Описаний метод для просторів E, F зі скалярним добутком називається *методом моментів* або *методом Бубнова — Гальоркіна — Петрова*.

Якщо простори зі скалярним добутком E і F збігаються і координатна система збігається з проекційною, то метод моментів називають *методом Гальоркіна*.

Якщо ж елементи координатної та проекційної системи пов'язані рівністю $\psi_j = A\varphi_j$, $j = \overline{1, n}$, то метод моментів називається *методом найменших квадратів* і збігається з варіаційним методом найменших квадратів (див. п. 1.2.1).

Нехай E та F — простори функцій, неперервних на компакті $\bar{\Omega}$, $E_n \subset D(A)$ і F_n — підпростори E та F відповідно, натягнуті на системи лінійно незалежних функцій $\varphi_i(x)$ та $\psi_i(x)$, $i = \overline{1, n}$, причому $\psi_j(x)$ — система Чебишева (див. п. 1.2.1). Виберемо в $\bar{\Omega}$ різні точки $x_j^{(n)}$, $j = \overline{1, n}$.

Шукатимемо u_n у вигляді узагальненого многочлена

$$u_n(x) = \sum_{i=1}^{n} c_i \varphi_i(x),$$

а коефіцієнти c_i — з умови

$$Au_n(x_i^{(n)}) - f(x_j^{(n)}) = 0, \quad j = \overline{1, n}. \tag{7}$$

Нехай $p_n : F \to F_n$ — оператор, який будь-якому елементу $f(x) \in F$ ставить у відповідність його узагальнений інтерполяційний многочлен $p(x; f) = \sum_{j=1}^{n} a_j \psi_j(x)$, $p(x_j^{(n)}; f) = f(x_j^{(n)})$, $j = \overline{1, n}$. Неважко помітити, що визначений таким чином оператор p_n є проектором. Тоді, очевидно, (7) еквівалентне рівнянню

$$p_n(Au_n(x) - f(x)) = 0,$$

тобто описаний метод є проекційним. Він називається *методом колокації*. Іноді його також називають *інтерполяційним методом* або *методом збігу*.

Для повного обгрунтування проекційних методів ми маємо вказати умови, за яких: 1) система (2) має єдиний розв'язок (див. також системи (6), (7)); 2) $\lim\limits_{n\to\infty} \| u_n - u \|_E = 0$, де u_n — розв'язок рівняння (2), а u — розв'язок рівняння (1). На жаль, загальний випадок виходить за рамки нашої книги.

Приклад 1. Методом колокації знайти наближений розв'язок задачі

$$u'' + (1 + x^2)u + 1 = x^2, \qquad u(-1) = u(1) = 2. \tag{8}$$

Розв'язок.

$$u_2(x) = 2 + \frac{563}{185}(1 - x^2) + \frac{11}{185} x^2 (1 - x^2).$$

Приклад 2. Методом Бубнова — Гальоркіна — Петрова знайти наближений розв'язок задачі

$$Lu \equiv u'' - u' \cos x + u \sin x = \sin x, \qquad u(-\pi) = u(\pi) = 2,$$

взявши чотири координатних функції.

Розв'язок.

$$u_4(x) = 2 + \frac{8}{7} \sin x + \frac{4}{7} \sin^2 x.$$

Наведемо для порівняння значення наближеного і точного розв'язків $u(x) = e^{\sin x} + 1$ у деяких точках:

x_i	$-\pi/2$	0	$\pi/2$
$u_4(x_i)$	1,429	2	3,714
$u(x_i)$	1,368	2	3,718

1.2. Варіаційні методи

Нехай рівняння

$$Au = f$$

має єдиний розв'язок u_* і відомий функціонал $\Phi : E \to R^1$ (R^1 — простір дійсних чисел), мінімум якого досягається в єдиній точці u^*. Тоді розв'язування рівняння (1) можна замінити задачею мінімізації функціоналу $\Phi(u)$. Методи, що грунтуються на цій ідеї, називаються *варіаційними*.

Розглянемо далі питання про мінімізацію квадратичних функціоналів, а пізніше вкажемо конкретні квадратичні функціонали, мінімізація яких дає варіаційні методи Рітца та найменших квадратів наближеного розв'язування рівняння (1).

Нехай $G(u, v)$ — симетрична білінійна форма, тобто

$$G(\alpha_1 u_1 + \alpha_2 u_2, v) = \alpha_1 G(u_1, v) + \alpha_2 G(u_2, v),$$

$$G(u, \alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 G(u, v_1) + \alpha_2 G(u, v_2),$$

$$G(u, v) = G(v, u),$$

де $u, v, u_i, v_i \in D(G) \subset H$, $i = \overline{1, 2}$; $D(G)$ — область визначення білінійної форми; H — простір зі скалярним добутком $(\cdot, \cdot)$. Із формою $G(u, v)$ пов'язаний квадратичний функціонал $G(u, u)$ з тією самою областю визначення, який вважатимемо додатним, тобто

$$G(u, u) > 0 \quad \forall u \neq 0.$$

Розглянемо функціонал

$$\Phi(u) = G(u, u) - 2l(u) + C, \quad u \in D(G), \tag{2}$$

де $l(u)$ — лінійний функціонал з областю визначення $D(l) \supset D(G)$, C — дійсна стала. Очевидно, область визначення цього функціонала $D(\Phi)$ збігається з $D(G)$.

Лема 1. *Якщо квадратичний функціонал $G(u, u)$ додатно визначений, тобто*

$$G(u, u) > \mu(u, u) \quad \forall u \in D(G),$$

де μ — додатна стала, а $l(u)$ — обмежений, то функціонал $\Phi(u)$ обмежений знизу, тобто існує число $\Phi_0 = \inf\limits_{u \in D(\Phi)} \Phi(u)$.

Доведення леми очевидне, і читач легко проведе його сам.

Введемо таке поняття.

Означення 1. Послідовність $\{u_n\}_{n=1}^{\infty} \subset D(\Phi)$ називається *мінімізуючою* для функціонала $\Phi(u)$, якщо

$$\lim_{n \to \infty} \Phi(u_n) = \Phi_0 = \inf_{u \in D(\Phi)} \Phi(u). \tag{3}$$

Зазначимо, що коли існує елемент $u_* \in D(\Phi)$ такий, що $\Phi_0 = \Phi(u_*) = \inf\limits_{u \in D(\Phi)} \Phi(u)$, то з означення мінімізуючої послідовності ще не випливає рівність $\lim\limits_{n \to \infty} \| u_n - u_* \| = 0$. Надалі вивчимо умови, за яких вона має місце.

Введемо у просторі $D(G)$ віддаль між двома елементами за формулою

$$\rho_G(u, v) = [G(u - v, u - v)]^{1/2}, \quad u, v \in D(G). \tag{4}$$

Вправа 1. Перевірити виконання аксіом віддалі.

Визначимо спочатку умови, за яких мінімізуюча послідовність збігається до u_* у тому розумінні, що

$$\lim_{n \to \infty} \rho_G(u_n, u_*) = 0.$$

Теорема 1. *Нехай $\{u_n\}$ — мінімізуюча послідовність для функціонала $\Phi(u)$. Тоді мають місце такі твердження: 1) послідовність $\{u_n\}$ фундаментальна в метриці ρ_G, тобто $\forall \varepsilon > 0$ існує $N(\varepsilon)$ таке, що $\rho_G(u_n, u_m) < \varepsilon \; \forall \; n, m > N(\varepsilon)$; 2) якщо існує елемент $u_* \in D(G)$, на якому функціонал $\Phi(u)$ досягає свого мінімального значення Φ_0, то $\lim\limits_{n \to \infty} \rho_G(u_n, u_*) = 0$.*

Доведення. З означення мінімізуючої послідовності для будь-якого $\varepsilon > 0$ випливає існування такого $N = N(\varepsilon)$, що

$$\Phi(u_n) < \Phi_0 + \frac{\varepsilon}{4} \quad \forall n > N.$$

Тому

$$\Phi(u_m) + \Phi(u_n) - 2\Phi\left(\frac{u_m + u_n}{2}\right) <$$

$$< 2\Phi_0 + 2\frac{\varepsilon}{4} - 2\Phi_0 < \frac{\varepsilon}{2} \quad \forall n, m > N. \tag{5}$$

Користуючись властивостями білінійної форми $G(u, v)$, маємо

$$\Phi(u) + \Phi(v) - 2\Phi\left(\frac{u+v}{2}\right) = G(u, u) + G(v, v) -$$

$$- 2G\left(\frac{u+v}{2}, \frac{u+v}{2}\right) = \frac{1}{2} G(u, u) + \frac{1}{2} G(v, v) -$$

$$- G(u, v) = \frac{1}{2} G(u - v, u - v) = \frac{1}{2} \rho_G^2(u, v)$$

і тому з виразу (5) одразу дістаємо перше твердження теореми. Друге твердження випливає з першого, якщо взяти до уваги, що послідовність

$$u_1, u_*, u_2, u_*, \ldots$$

також є мінімізуючою. Теорему доведено.

Наслідок. Якщо квадратичний функціонал $G(u, u)$ є додатно визначеним, то твердження теореми 1 мають місце в нормі простору H, тобто в нормі $\|u\| = (u, u)^{1/2}$.

Теорема 1 дає умови збіжності мінімізуючої послідовності до елемента u_*, при якому досягається мінімум функціонала $\Phi(u)$ (якщо такий елемент u_* існує). Далі природно поставити запитання: як же знайти мінімізуючу послідовність? Наступні твердження вказують алгоритм її побудови.

Теорема 2. *Нехай функціонал $\Phi(u)$ досягає мінімуму в деякій точці $u_* \in D(G)$, тоді:* 1) $G(u_*, v) = l(v)\ \forall v \in D(G)$; 2) $\Phi(u_* + v) = \Phi(u_*) + G(v, v)\ \forall v \in D(G)$, *тобто мінімум досягається на єдиному елементі u_* (кажуть також, що мінімум є строгим).*

Доведення. Для будь-якого $v \in D(G)$ з умов теореми випливає, що

$$\Phi(u_* + v) = G(u_* + v, u_* + v) - 2l(u_* + v) + C = \\ = \Phi(u_*) + 2G(u_*, v) - 2l(v) + G(v, v) \geqslant \Phi(u_*). \qquad (6)$$

Звідси $\forall v \in D(G)$ маємо нерівність

$$2G(u_*, v) - 2l(v) + G(v, v) \geqslant 0,$$

яка виконується також при заміні v на tv, де t — будь-яке дійсне число. Після такої заміни

$$2t[G(u_*, v) - l(v)] + t^2 G(v, v) \geqslant 0, \quad \forall v \in D(G), \\ \forall t \in R^1.$$

Це є квадратна нерівність відносно t, яка виконується при будь-якому t лише тоді, коли дискримінант квадратного тричлена недодатний, тобто

$$4[G(u_*, v) - l(v)]^2 \leqslant 0 \quad \forall v \in D(G).$$

Звідси

$$G(u_*, v) = l(v) \quad \forall v \in D(G),$$

і перше твердження теореми доведено. Неважко помітити, що друге твердження є наслідком першого і рівності (6). Теорему доведено.

Виберемо в $D(G)$ лінійно незалежну систему елементів $\{\varphi_i\}_{i=1}^{\infty}$ і через $H_n \subset D(G)$ позначимо лінійну оболонку перших n елементів φ_i, $i = \overline{1, n}$. Мають місце такі два твердження, які ми пропонуємо довести читачеві.

Лема 2. *Функціонали $G(u, u)$ і $l(u)$ неперервні на кожному з підпросторів H_n, $n = 1, 2, \ldots$.*

Лема 3. *Квадратичний функціонал $G(u, u)$ додатно визначений на кожному з просторів H_n, тобто існують сталі $m_n > 0$ такі, що*

$$G(u, u) \geqslant m_n(u, u) \equiv m_n \|u\|^2 \quad \forall u \in H_n, \quad n = 1, 2, \ldots.$$

Доведемо, що в кожному з підпросторів H_n, $n = 1, 2, \ldots$, існує єдиний елемент, на якому функціонал $\Phi(u)$ досягає свого мінімального значення на H_n.

Лема 4. *Для будь-якого n в H_n існує єдиний елемент u_n такий, що*

$$\Phi_n = \Phi(u_n) = \inf_{u \in H_n} \Phi(u).$$

Доведення. Оскільки обмеженість лінійного оператора є необхідною і достатньою умовою неперервності, то за лемою 2

$$|l(u)| \leqslant d_n \|u\| \quad \forall u \in H_n,$$

і тому

$$\Phi(u) \geqslant m_n \|u\|^2 - 2d_n \|u\| + C = m_n \left[\|u\| - \frac{d_n}{m_n}\right]^2 - \frac{d_n^2}{m_n^2} + C.$$

Звідси випливає, що при $\|u\| > \frac{2d_n}{m_n}$ має місце нерівність $\Phi(u) > C$. Нехай $s_n = \left\{u : \|u\| \leqslant \frac{2d_n}{m_n}\right\}$. Неважко помітити, що

$$\inf_{H_n \setminus s_n} \Phi(u) > C = \Phi(0) \geqslant \inf_{s_n} \Phi(u).$$

Звідси випливає рівність $\inf_{H_n} \Phi(u) = \inf_{S_n} \Phi(u)$, але s_n є компактом, причому функціонал $\Phi(u)$ неперервний на цьому компакті. Тому існує елемент $u_n \in s_n$, для якого $\inf_{u \in H_n} \Phi(u) = \Phi(u_n)$. Єдиність цього елемента випливає з другого твердження теореми 2.

Як же знайти елемент u_n, про який йшлося в попередній лемі? Оскільки $u_n \subset H_n$, то його можна зобразити у вигляді

$$u_n = \sum_{i=1}^{n} a_i^{(n)} \varphi_i, \qquad (7)$$

де $a_i^{(n)}$ — дійсні коефіцієнти, які підлягають визначенню. Далі маємо

$$\Phi(u_n) = \Phi\left(\sum_{i=1}^{n} a_i^{(n)} \varphi_i\right) = \sum_{i,j=1}^{n} a_i^{(n)} a_j^{(n)} G(\varphi_i, \varphi_j) - 2\sum_{i=1}^{n} a_i^{(n)} l(\varphi_i) + C,$$

тобто $\Phi(u_n)$ є функцією змінних $a_i^{(n)}$, $i = \overline{1, n}$. На підставі леми 4 ця функція має єдиний мінімум, необхідною умовою якого є система рівностей

$$\frac{\partial \Phi(u_n)}{\partial a_i^{(n)}} = 2\sum_{j=1}^{n} a_j^{(n)} G(\varphi_i, \varphi_j) - 2l(\varphi_i) = 0, \quad i = \overline{1, n},$$

або

$$\sum_{j=1}^{n} a_j^{(n)} G(\varphi_i, \varphi_j) = l(u_i), \quad i = \overline{1, n}. \tag{8}$$

Неважко помітити, що $G(u, v)$ можна взяти за скалярний добуток (перевірте виконання аксіом скалярного добутку). Тому визначник системи (8) є визначником Грамма системи лінійно незалежних елементів і тому відмінний від нуля. Це означає, що система лінійних алгебраїчних рівнянь (8) має єдиний розв'язок, який разом із (7) визначає u_n. Описаний метод визначення елемента u_n, на якому функціонал $\Phi(u)$ набуває найменшого значення на H_n, називається *процесом Рітца*. Далі виникає запитання: чи буде послідовність $\{u_n\}_{n=1}^{\infty}$, яка визначається процесом Рітца, мінімізуючою? Умови, за яких відповідь на це запитання буде позитивною, визначимо пізніше. Зазначимо, що на практиці, як правило, обмежуються відшуканням u_n при якомусь значенні n. Якщо нас цікавить u_{n_1} при якомусь іншому значенні n_1, то всі обчислення необхідно повторити спочатку (u_n використовувати не можна).

Далі розглянемо найбільш поширені варіаційні методи детальніше.

1. *Метод найменших квадратів.* Розглянемо рівняння (1), в якому оператор $A : H \to H$ діє у просторі зі скалярним добутком, причому існує обернений оператор A^{-1}. У методі найменших квадратів у цьому разі мінімізується функціонал

$$\Phi(u) = \|Au - f\|^2 = (Au - f, Au - f) = (Au, Au) - 2(Au, f) + \|f\|^2. \tag{9}$$

В силу припущень щодо A і вигляду функціонала $\Phi(u)$ маємо:

$$\Phi(u_*) = 0 = \inf_{u \in D(A)} \Phi(u),$$

де u_* — розв'язок рівняння (1). Порівнюючи функціонал (9) з функціоналом (2), помічаємо, що вони збігаються, якщо покласти

$$D(G) = D(A), \quad G(u, v) = (Au, Av), \quad l(u) = (Au, f), \quad C = \|f\|^2.$$

За допомогою процесу Рітца побудуємо послідовність $\{u_n\}_{n=1}^{\infty}$. Система (8) у цьому разі має вигляд

$$\sum_{j=1}^{n} a_j^{(n)} (A\varphi_j, A\varphi_i) = (f, A\varphi_i), \quad i = \overline{1, n}. \tag{10}$$

Неважко помітити, що система (10) однакова з системою (6) (з п. 1.1) методу моментів, і в цьому смислі метод найменших квадратів можна вважати його окремим випадком. Проте підкреслимо ще раз, що ці два методи різні за ідеєю.

Визначимо тепер умови, при яких послідовність $\{u_n\}$, знайдена методом найменших квадратів, є мінімізуючою для функціонала (9). Для цього доведемо спочатку таке допоміжне твердження.

Лема 5. *Нехай v — лінійна оболонка системи лінійно незалежних векторів $\{\varphi_i\}_{i=1}^{\infty} \in H$ у гільбертовому просторі H, H_n — лінійна оболонка елементів φ_i, $i = \overline{1, n}$, $\bar{v}$ — замикання множини v. Тоді: 1) для того щоб елемент $u \in H$ належав також $\bar{v}$, необхідно і достатньо, щоб*

$$\lim_{n\to\infty} p_{H_n} u = \lim_{n\to\infty} u_n = u, \tag{11}$$

де $\{u_n\}$ — послідовність проекцій u на H_n (тобто елементів найкращого наближення для u в H_n); 2) для того щоб $\bar{v} = H$, необхідно і достатньо, щоб співвідношення (11) було справедливим для будь-якого $u \in H$.

Доведення. Доведемо необхідність першого твердження теореми. Нехай $u \in H$ і $u \in \bar{v}$. Доведемо, що має місце (11). Дійсно, оскільки $u \in \bar{v}$, то $\forall \varepsilon > 0$ існує $N = N(\varepsilon)$ таке, що

$$\|u - \tilde{u}^{(N)}\| < \varepsilon,$$

де $\tilde{u}^{(N)} \in H_N$. Нехай $\|u - u_n\| = \inf_{v \in H_n} \|u - v\|$, тобто u_n є елементом найкращого наближення для u в H_n. На основі властивостей елемента найкращого наближення (див. п. 3.2 в ч. 1) маємо

$$(u - u_n, v) = 0 \quad \forall v \in H_n,$$

тобто $u_n = p_{H_n} u$ — проекція u на підпростір H_n. Далі маємо ланцюжок нерівностей:

$$\|u_n - u\| = \|p_{H_n} u - u\| \leqslant \|p_{H_N} u - u\| \leqslant \|\tilde{u}^{(N)} - u\| < \varepsilon \quad \forall n \geqslant N,$$

що і доводить співвідношення (11).

Доведемо достатність першого твердження. Якщо має місце (11), то послідовність елементів $u_n = p_{H_n} u \in H_n \subset \bar{v}$ збігається до деякого елемента u. Оскільки всі граничні елементи послідовностей з v належать $\bar{v}$ (за означенням $\bar{v}$), то $u \in \bar{v}$ і перше твердження теореми повністю доведено. Друге твердження очевидно випливає з першого.

Теорема 3. *Для того щоб послідовність $\{u_n\}_{n=1}^{\infty}$, побудована за допомогою процесу Рітца, у гільбертовому просторі H була мінімізуючою для функціонала (9), необхідно і достатньо, щоб $f \in \bar{v}$, де $\bar{v}$ — замкнена лінійна оболонка системи елементів $\{v_i\}_{i=1}^{\infty} = \{A\varphi_i\}_{i=1}^{\infty}$.*

Доведення. Згідно з означенням послідовність $\{u_n\}_{n=1}^{\infty}$ буде мінімізуючою для функціонала (9), якщо

$$\lim_{n\to\infty} \| Au_n - f \| = 0.$$

Нехай V_n — лінійна оболонка елементів $v_i = A\varphi_i$, $i = \overline{1, n}$. Тоді за побудовою елемент $V^{(n)} = Au_n$ є елементом найкращого наближення для f у підпросторі V_n гільбертового простору H, бо $(f - u_n,\ A\varphi_i) = \left(f - \sum_{i=1}^{n} a_j^{(n)}\varphi_j,\ A\varphi_j\right) = 0$. Отже,

$$p_{V_n} f = v^{(n)} = Au_n$$

і твердження теореми 3 випливає з першого твердження леми 5.

Наслідок 2. Якщо система елементів $\{A\varphi_i\}_{i=1}^{\infty}$ повна в гільбертовому просторі H, то $\forall f \in R(A)$ і має місце рівність

$$\lim_{n\to\infty} \| Au_n - f \| = 0.$$

Виходячи з наслідку 1, доведемо наступне твердження про збіжність мінімізуючої послідовності до розв'язку u_* задачі (1) за нормою.

Теорема 4. *Якщо рівняння* (1) *задане у гільбертовому просторі H, $f \in \bar{V}$, і оператор A^*A додатно визначений (A^* — оператор, спряжений до A, тобто $(Au, v) = (u, A^*v)$ $\forall u, v$), то*

$$\lim_{n\to\infty} \| u_n - u_* \| = 0.$$

Доведення. Оскільки f належить замкненій лінійній оболонці $\bar{V}$ системи елементів $\{A\varphi_i\}_{i=1}^{\infty}$, то за теоремою 3 послідовність $\{u_n\}_{n=1}^{\infty}$ є мінімізуючою для функціонала (9). Із додатної визначеності оператора A^*A маємо

$$G(u, u) = (Au, Au) = (A^*Au, u) \geqslant \mu \| u \|^2 \ \forall u \in D(A),$$

де μ — додатна стала. Отже, твердження теореми 4 випливає з наслідку 1.

Зауваження. Неважко помітити, що теорема 4 є справедливою, якщо умову додатної визначеності оператора A^*A замінити умовою існування обмеженого оператора A^{-1}. Це випливає з того, що

$$\| u \|^2 = (u, u) = \| A^{-1}Au \|^2 \leqslant \| A^{-1} \|^2 \| Au \|^2,$$

$$\| u \|^2 \leqslant M^2 \| Au \|^2 = M^2 G(u, u), \quad G(u, u) \geqslant M^{-2} \| u \|^2.$$

2. Метод Рітца. Хоча метод Рітца застосовується при більш жорстких умовах, ніж інші проекційно-варіаційні методи, він також широко використовується.

Отже, нехай потрібно розв'язати рівняння (1) за таких умов: 1) A — самоспряжений оператор у гільбертовому просторі H, тобто

$$(Au, v) = (u, Av) \ \forall u, v \in D(A) \tag{12}$$

(це записують також у вигляді $A^* = A$); 2) A є додатно визначеним оператором, тобто

$$(Au, u) \geqslant \mu \| u \|^2 \quad \forall u \in D(A), \quad \mu > 0. \tag{13}$$

Згідно з ідеєю варіаційних методів замінимо задачу розв'язування рівняння (1) задачею мінімізації функціонала

$$\Phi(u) = (Au, u) - 2(f, u), \quad u \in D(A), \tag{14}$$

який називається *функціоналом енергії*. Якщо в (2) покласти

$$D(G) = D(A), \quad G(u, u) = (Au, u), \quad l(u) = (f, u), \qquad C = 0,$$

то дістанемо функціонал (14). Оскільки функціонал $l(u) = (f, u)$ обмежений на $D(G)$, а квадратичний функціонал $G(u, u)$ додатно визначений, то функціонал $\Phi(u)$ обмежений знизу. Доведемо тепер, що задачі розв'язування рівняння (1) за прийнятих припущень і мінімізації функціонала (14) еквівалентні.

Теорема 5. *Нехай $A : H \to H$ — лінійний оператор у просторі зі скалярним добутком H із щільною в H областю визначення $D(A)$ та виконуються умови* (12), (13). *Крім того, рівняння* (1) *має розв'язок $u_* \in D(A)$. Тоді функціонал* (14) *набуває на u_* мінімального значення в $D(A)$, тобто $\Phi(u) \geqslant \Phi(u_*)$ $\forall u \in D(A)$, причому $\Phi(u) = \Phi(u_*)$ тільки при $u = u_*$. З іншого боку, нехай серед всіх елементів $u \in D(A)$ функціонал* (14) *набуває мінімального значення на елементі u_*. Тоді u_* є розв'язком рівняння* (1).

Доведення. Нехай рівняння (1) має розв'язок $u_* \in D(A)$, тобто $f = Au_*$. Підставляючи Au_* замість f в (14), дістанемо

$$\Phi(u) = (Au, u) - 2(Au_*, u) = (Au, u) - (Au_*, u) -$$

$$- (Au_*, u) = (Au, u) - (Au_*, u) - (u_*, Au) =$$

$$= (Au, u) - (Au_*, u) - (Au, u_*) + (Au_*, u_*) -$$

$$- (Au_*, u_*) = (A(u - u_*), u - u_*) - (Au_*, u_*) \quad \forall u \in D(A).$$

Оскільки оператор A додатно визначений, то $\Phi(u) > -(Au_*, u_*)$, причому при $u = u_*$ маємо $\Phi(u) = -(Au_*, u_*)$, що і доводить першу частину теореми.

Нехай тепер u_* — елемент із $D(A)$, на якому $\Phi(u)$ набуває мінімального значення на множині $D(A)$. Тоді для будь-якого дійсного t

$$\Phi(u_* + tv) > \Phi(u_*) \quad \forall v, t. \tag{15}$$

Неважко помітити, що за умов теореми

$$\Phi(u_* + tv) = (A(u_* + tv), u_* + tv) - 2(f, u_* + tv) =$$
$$= (Au_*, u_*) + 2t(Au_*, v) + t^2(Av, v) - 2t(f, v) - 2(f, u_*)$$
$$\forall v \in D(A), \quad t \in R^1.$$

Враховуючи (15), дістаємо

$$(Au_*, u_*) + 2t(Au_*, v) + t^2(Av, v) - 2t(f, v) -$$
$$- 2(f, u_*) \geqslant (Au_*, u_*) - 2(f, u_*),$$

або

$$2t[(Au_*, v) - (f, v)] + t^2(Av, v) \geqslant 0$$
$$\forall v \in D(A), \quad t \in R^1.$$

Для того щоб квадратний тричлен відносно t із додатним старшим коефіцієнтом $(Av, v) \geqslant \mu \| v \|^2$ був невід'ємним $\forall t \in R^1$, потрібно, щоб його дискримінант був недодатним, тобто

$$[(Au_*, v) - (f, v)]^2 \leqslant 0 \quad \forall v \in D(A).$$

Звідси

$$(Au_* - f, v) = 0 \quad \forall v \in D(A).$$

Оскільки множина $D(A)$ щільна в H, то

$$Au_* = f,$$

що і треба було довести.

Далі застосуємо *процес Рітца*. Для цього виберемо координатну систему $\{\varphi_i\}_{i=1}^{\infty} \subset D(A)$ і побудуємо послідовність підпросторів H_n, які є лінійними оболонками елементів φ_i, $i = \overline{1, n}$. Елементи $u_n = \sum_{i=1}^{n} a_i^{(n)} \varphi_i \in H_n$, на яких досягається $\inf_{u \in H_n} \Phi(u)$, знаходимо за допомогою процесу Рітца. Згідно з (8) для коефіцієнтів $a_i^{(n)}$ дістаємо систему лінійних алгебраїчних рівнянь

$$\sum_{i=1}^{n} a_i^{(n)} (A\varphi_i, \varphi_j) = (f, \varphi_j), \quad j = \overline{1, n}, \tag{16}$$

причому з леми 4 випливає існування і єдиність розв'язку цієї системи і елемента u_n.

Порівнюючи (16) із системою (6) з п. 1.1 методу моментів, помічаємо, що вони будуть однаковими, якщо в (6) з п. 1.1 вибрати $\varphi_i = \psi_i$, $i = \overline{1, n}$. У цьому розумінні метод Рітца можна вважати окремим випадком методу моментів.

Для дослідження збіжності методу Рітца введемо в $D(A)$ скалярний добуток за формулою

$$(u, v)_A = (Au, v) \quad \forall u, v \in D(A)$$

(перевірку аксіом скалярного добутку читач легко може здійснити самостійно) і відповідну норму

$$\| u \|_A = [(u, u)_A]^{1/2} = [(Au, u)]^{1/2}.$$

Такі скалярний добуток і норму називають ще *енергетичним скалярним добутком* і *енергетичною нормою*. Поповнення $D(A)$ (див. п. 1.2) у нормі $\| \cdot \|_A$ позначимо через H_A і називатимемо *енергетичним простором* оператора A. Із означення H_A випливає, що $D(A)$ є всюди щільною в H_A множиною, а з додатної визначеності оператора A випливає нерівність

$$\| u \| \leqslant \frac{\| u \|_A}{\sqrt{\mu}} \quad \forall u \in H_A. \tag{17}$$

З нерівності Коші — Буняковського

$$\| u \|_A^2 = (Au, u) \leqslant \| Au \| \| u \|$$

і (17) дістанемо також

$$\| u \|_A \leqslant \frac{\| Au \|}{\sqrt{\mu}} \quad \forall u \in D(A). \tag{18}$$

Покажемо, що елемент u_n, визначений у процесі Рітца, є проекцією розв'язку рівняння (1) u_* на H_n в енергетичному просторі H_A. Дійсно,

$$\Phi(u) = (Au, u) - 2(f, u) = \| u \|_A^2 - 2(Au_*, u) =$$
$$= \| u \|_A^2 - 2(u_*, u)_A \quad \forall u \in D(A),$$

або

$$\Phi(u) = \| u - u_* \|_A^2 - \| u_* \|_A^2 \quad \forall u \in D(A).$$

Оскільки значення u_n забезпечує мінімум функціонала $\Phi(u)$ на підпросторі H_n, то

$$u_n = p_{H_n}^{H_A} u^* = \inf_{u \in H_n} \| u - u_* \|_A,$$

тобто u_n є проекцією u_* на H_n у просторі H_A.

Доведемо тепер теорему про умови, за яких послідовність $\{u_n\}$ буде мінімізуючою.

Теорема 6. *Для того щоб послідовність $\{u_n\}$ була мінімізуючою, необхідно і достатньо, щоб $u_* \in K^{(\infty)}$, де $K^{(\infty)}$ — замикання в нормі H_A лінійної оболонки координатної системи $\{\varphi_i\}_{i=1}^{\infty}$.*

Доведення. Для того щоб u_* належало K^∞, необхідно і достатньо виконання умови

$$\lim_{n\to\infty} p_{H_n}^{H_A} u_* = u_*.$$

Оскільки $P_{H_n}^{H_A} u_* = u_n$, то теорему доведено.

Зауваження. Із збіжності послідовності $\{u_n\}$ в H_A і нерівності (17) випливає збіжність $\{u_n\}$ до u_* у нормі H.

Розглянемо, нарешті, умови, за яких мінімізуюча послідовність збігатиметься до u_*.

Теорема 7. *Нехай координатна система* $\{\varphi_i\}_{i=1}^\infty$ *є повною в* H_A, *тоді*

$$\lim_{n\to\infty} \| u_n - u_* \|_A = \lim_{n\to\infty} \| u_n - u_* \| = 0.$$

Доведення випливає з того, що за умовою теореми 7 замикання системи $\{\varphi_i\}_{i=1}^\infty$ у нормі H_A збігається з H_A, а також з попередньої теореми.

Зауваження. Метод Рітца для рівняння

$$A^*Au = A^*f$$

є еквівалентом методу найменших квадратів. Дійсно, функціонал методу Рітца для цього рівняння має вигляд

$$\Phi(u) = (A^*Au, u) - 2(A^*f, u) = (Au, Au) - 2(f, Au) =$$
$$= (Au - f, Au - f) - (f, f) = \|Au - f\|^2 - \|f\|^2,$$

що лише на сталу відрізняється від функціонала методу найменших квадратів.

Вправа. За методом Рітца знайти наближений розв'язок задачі

$$-u'' + u(\sin x + 2) = \sin x,$$
$$u(-\pi) = u(\pi) = 2. \tag{19}$$

ГЛАВА 2

МЕТОД СІТОК РОЗВ'ЯЗУВАННЯ ОПЕРАТОРНИХ РІВНЯНЬ

Метод сіток набув поширення внаслідок своєї простоти й універсальності. Ще однією його перевагою є спеціальний вигляд сіткових рівнянь, що дає змогу ефективно реалізувати їх на ЕОМ.

У цій главі розглянемо основні положення методу сіток. Вони будуть використані у наступних главах при побудові і дослідженні сіткових (різницевих) схем для різних задач математичної фізики — задачі Коші для звичайних диференціальних рівнянь, крайових задач для еліптичних рівнянь з частинними похідними, а також початково-крайових задач для нестаціонарних (параболічних та гіперболічних) рівнянь з частинними похідними.

2.1. Основні поняття методу сіток. Теорема про умови збіжності методу сіток (теорема Лакса — Філіпова)

Нехай $\tilde{B}_1$, $\tilde{B}_2$ — деякі нормовані простори функцій, заданих в області $\bar{\Omega} = \Omega \cup \partial\Omega$ з границею $\partial\Omega$. Норми у просторах $\tilde{B}_\alpha$, $\alpha = 1, 2$, позначимо через $\|\cdot\|_\alpha$, $\alpha = 1, 2$. Нехай A — лінійний оператор з областю визначення $D(A) \equiv B_1 \subset \tilde{B}_1$ і областю значень $R(A) \equiv B_2 \subset \tilde{B}_2$.

Розглянемо задачу визначення елемента $u \in B_1$ за заданим елементом $f \in B_2$ з операторного рівняння

$$Au = f. \tag{1}$$

Оператор A визначається, по-перше, своєю областю визначення $D(A)$ і, по-друге, правилом, за яким кожному елементу з $D(A)$ ставиться у відповідність єдиний елемент з $R(A)$. При постановці задач математичної фізики часто буває корисним розділити ці дві частини означення оператора A. Тоді постановка задачі (1) набирає вигляду

$$Au = f(x), \quad x \in \Omega,$$
$$au = g(x), \quad x \in \partial\Omega, \tag{2}$$

де a — оператор граничної умови.

Приклад 1. Крайова задача

$$-u''(x) = f(x), \quad x \in (0, 1) \equiv \Omega, \tag{3}$$
$$u(0) = 0, \quad u'(1) = 1 \tag{4}$$

записується у вигляді (1), якщо оператор A задати: 1) його областю визначення $D(A)$, яка складається з двічі неперервно диференційовних функцій, що задовольняють умови (4); 2) за правилом $Au \equiv -u''(x)$, $u \in D(A)$. Ця задача записується у вигляді (2), якщо позначити $Au \equiv -u''(x)$, $u \in C^2(0, 1)$,

$$au = \begin{cases} u(x), & x = 0, \\ u'(x), & x = 1. \end{cases}$$

Вважатимемо, що задача (1) або (2) поставлена коректно, тобто для будь-якого $f \in R(A)$ вона має єдиний розв'язок, що неперервно залежить від f. Останнє означає, що коли $Au_1 = f_1$, $Au_2 = f_2$, то $\|u_1 - u_2\|_1 \leqslant C\|f_1 - f_2\|_2$, де стала C не залежить від u_1, u_2, f_1, f_2.

Щоб побудувати для задачі (1), (2) сітковий метод розв'язування, треба виконати наступні три кроки: 1) неперервні області зміни

функцій $\overline{\Omega}$ $(\Omega, \partial\Omega)$ замінити деякими дискретними (сітковими) областями $\overline{\omega}_h$ (ω_h, γ_h); 2) простори функцій неперервного аргументу B_1, B_2, B_1, B_2 замінити просторами сіткових функцій $\tilde{B}_{1h}$, $\tilde{B}_{2h}$, B_{1h}, B_{2h} з нормами $\|\cdot\|_{1h}$, $\|\cdot\|_{2h}$, заданих на сітці $\overline{\omega}_h$; 3) оператор A (або A, a в (2)) замінити деякими сітковими операторами A_h $(A_h,\ a_h)$. В результаті приходимо до сім'ї операторних рівнянь

$$A_h v_h = \varphi_h, \quad v_h \in B_{1h}, \quad \varphi_h \in B_{2h}, \tag{5}$$

яка залежить від параметра h, або для постановки (2)

$$\begin{aligned} A_h v_h &= \varphi_h(x), \quad x \in \omega_h, \\ a_h v_h &= g_h(x), \quad x \in \gamma_h. \end{aligned} \tag{6}$$

Сім'я операторних рівнянь (5) або (6) називається сітковою схемою для задачі (1) або (2).

Роль параметра h в (5), (6) відіграє деяка функція від параметрів, що визначають сітки $\overline{\omega}_h$, ω_h, γ_h. Якщо, наприклад, $\overline{\omega}_h = \{x_i = ih : i = \overline{0, N},\ h = 1/N\}$ — рівномірна сітка, яка покриває відрізок $[0, 1]$, то цим параметром є крок сітки h. Для нерівномірної сітки $\overline{\omega}_h = \{x_i : i = \overline{0, N},\ x_0 = 0,\ x_N = 1,\ x_i - x_{i-1} = h_i,\ \sum_{i=1}^{N} h_i = 1\}$ роль h може відігравати деяка норма вектора кроків сітки $(h_1, h_2, \ldots h_N)$, наприклад $h = \max\limits_{i=\overline{1,N}} h_i$. Для рівномірної сітки по кожному напрямку сітки $\overline{\omega}_h = \overline{\omega}_{h_1} \times \overline{\omega}_{h_2}$, $\overline{\omega}_{h_\alpha} = \{x_\alpha = i_\alpha h_\alpha : i_\alpha = \overline{0, N_\alpha} : h_\alpha = l_\alpha/N_\alpha\}$, $\alpha = 1, 2$, яка покриває прямокутник $\overline{\Omega} = \{(x_1, x_2) : 0 \leqslant x_\alpha \leqslant l_\alpha,\ \alpha = 1, 2\}$, за h можна взяти $h = \max(h_1, h_2)$ або $h = \sqrt{h_1^2 + h_2^2}$ і т. д.

До сіткового методу розв'язування задачі (1) або (2) слід віднести також спосіб порівняння елементів просторів B_1 та B_{1h}, адже нас цікавить точність, з якою розв'язок рівняння (5) або (6) наближає розв'язок рівняння (1) або (2).

Спосіб порівняння елементів B_1 і B_{1h} визначається за допомогою сім'ї лінійних операторів $P_{1h} : B_1 \to B_{1h}$ таких, що

$$\begin{aligned} &P_{1h} u = u_h \in B_{1h} \quad \forall\, u \in B_1, \\ &\lim_{h \to 0} \|P_{1h} u\|_{1h} = \|u\|_1. \end{aligned} \tag{7}$$

Ця умова називається *умовою узгодження норм* $\|\cdot\|_1$ і $\|\cdot\|_{1,h}$. Якщо оператори P_{1h} визначені, то можна говорити про збіжність елементів $v_h \in B_{1h}$ до деякого елемента $u \in B_1$ у такому розумінні:

$$\|z_h\|_{1h} = \|v_h - P_{1h} u\|_{1h} = \|v_h - u_h\|_{1h} \to 0 \quad \text{при } h \to 0.$$

Якщо u — розв'язок (1) або (2), v_h — розв'язок (5) або (6), то величину

$$z_h = v_h - P_{1h} u = v_h - u_h$$

називають *похибкою сіткової схеми* (5) або (6).

Кажуть, що розв'язок сіткової схеми (5) або (6) збігається до розв'язку задачі (1) або (2) з швидкістю $O(h^m)$, або має порядок точності $m > 0$, якщо при досить малих $h < h_0$ виконується рівність

$$\|z_h\|_{1h} = \|v_h - u_h\|_{1h} = O(h^m),$$

або, що те саме, нерівність

$$\|z_h\|_{1h} \leqslant c h^m,$$

де $c > 0$ — стала, яка не залежить від h.

Зазначимо, що існують й інші способи порівняння елементів з B_1 і B_{1h}. Наприклад, функцію v_h можна доозначити в усіх точках області Ω, яка покривається сіткою ω_h, і знайти деяку функцію неперервного аргументу $v(x)$, $x \in \Omega$. Нехай B_1 — нормований простір функцій неперервного аргументу x, заданих в області Ω, і $v(x) - u(x) \in B_1$. Тоді можна говорити про збіжність елементів $v_h \in B_{1h}$ до елемента $u \in B_1$ у такому розумінні: $\|v(x) - u(x)\|_1 \to 0$ при $h \to 0$.

Сіткова схема (5) називається коректною, якщо вона має єдиний розв'язок і є стійкою, тобто: 1) для будь-якого $\varphi_h \in B_{2h}$ існує єдиний розв'язок v_h рівняння (5); 2) розв'язок v_h рівняння (5) неперервно залежить від φ_h. Однозначна розв'язність схеми (5) означає, що існує оператор A_h^{-1} і $v_h = A^{-1}\varphi_h$, а стійкість означає, що оператор $A_h^{-1} : B_{2h} \to B_{1h}$ рівномірно по h обмежений, тобто $\|A_h^{-1}\| \leqslant \varkappa$, де $\varkappa > 0$ — стала, незалежна від h. Таким чином, для коректних сіткових схем

$$\|v_h\|_{1h} \leqslant \|A_h^{-1}\| \, \|\varphi_h\|_{2h} \leqslant \varkappa \|\varphi_h\|_{2h},$$

або

$$\|v_h\|_{1h} \leqslant \varkappa \|\varphi_h\|_{2h}. \tag{8}$$

Ця нерівність називається апріорною оцінкою для схеми (5) або нерівністю коректності. Визначення таких оцінок є метою теорії стійкості різницевих схем. Часто нерівність (8) називається також нерівністю стійкості схеми (5). Схема, для якої виконується (8), є стійкою. Стійкість схеми можна тлумачити як неперервну залежність її розв'язку від вхідних даних. Дійсно, якщо $A_h v_h^{(1)} = \varphi_h^{(1)}$, $A_h v_h^{(2)} = \varphi_h^{(2)}$, то $A_h(v_h^{(1)} - v_h^{(2)}) = \varphi_h^{(1)} - \varphi_h^{(2)}$ і з (8) випливає, що

$$\|v_h^{(1)} - v_h^{(2)}\|_{1h} \leqslant \varkappa \|\varphi_h^{(1)} - \varphi_h^{(2)}\|_{2h}.$$

Стійкість сіткової схеми (6) визначається нерівністю

$$\|v_h\|_{\omega_h} \leqslant \varkappa_1 \|\varphi_h\|_{\omega_h} + \varkappa_2 \|g_h\|_{\gamma_h},$$

де сталі $\varkappa_1$, $\varkappa_2$ не залежать від h, а $\|\cdot\|_{\bar{\omega}_h}$, $\|\cdot\|_{\omega_h}$, $\|\cdot\|_{\gamma_h}$ означають норми у просторах сіткових функцій, заданих відповідно на $\bar{\omega}_h$, ω_h, γ_h.

Для похибки $z_h = v_h - u_h$ маємо рівняння

$$A_h z_h = \psi_h, \quad z_h \in B_{1h}, \quad \psi_h \in B_{2h}, \tag{9}$$

де $\psi_h = A_h v_h - A_h u_h = \varphi_h - A_h u_h$ називається похибкою апроксимації рівняння (1) сітковою схемою (5) на розв'язку u вихідної задачі (1).

У випадку схеми (6)

$$\begin{aligned} A_h z_h &= \psi_{\omega_h}(x), \quad x \in \omega_h, \\ a_h z_h &= \psi_{\gamma_h}(x), \quad x \in \gamma_h, \end{aligned} \tag{9'}$$

де $\psi_{\omega_h} = A_h v_h - A_h u_h = \varphi_h - A_h u_h$, $\psi_{\gamma_h} = g_h - a u_h = a v_h - a u_h$ — відповідно похибки апроксимації рівняння та крайових умов.

Сіткова схема (5) має m-й порядок апроксимації на елементі $u \in B_1$, якщо при $h < h_0$

$$\|\psi_h\|_{2h} = \|\varphi_h - A_h u_h\|_{2h} = O(h^m), \tag{10}$$

або

$$\|\psi_h\|_{2h} \leqslant M h^m,$$

де стала M не залежить від h. Якщо $\psi_h(x) = O(h^m)$ або $|\psi(x)| \leqslant M(x)\, h^m$, $x \in \omega_h$, то кажуть, що схема має m-й порядок локальної апроксимації.

Відповідно задача (6) апроксимує задачу (2) з порядком m на розв'язку u, якщо існують додатні сталі h_0, M_1, M_2 такі, що для всіх $h < h_0$ виконуються нерівності

$$\begin{aligned} \|\varphi_{\omega_h}\|_{\omega_h} &= \|A_h u_h - \varphi_h\|_{\omega_h} \leqslant M_1 h^{n_1}, \\ \|\psi_{\gamma_h}\|_{\gamma_h} &= \|a_h u_h - g_h\|_{\gamma_h} \leqslant M_2 h^{n_2} \end{aligned} \tag{10'}$$

і $m = \min(n_1, n_2)$ (n_1, n_2 — відповідно порядки апроксимації рівняння та крайових умов).

Нехай визначено оператори $P_{2h} : B_2 \to B_{2h}$ такі, що

$$P_{2h} f = f_h \in B_{2h},$$

і виконана умова узгодженості норм

$$\lim_{h \to 0} \|P_{2h} f\|_{2h} = \|f\|_2.$$

Якщо для будь-якого $u \in B_1$ виконується оцінка

$$\|A_h(P_{1h}u) - P_{2h}(Au)\|_{2h} = O(h^m), \tag{11}$$

то кажуть, що оператор A_h апроксимує оператор A з порядком $m > 0$.

Якщо φ_h апроксимує $f \in B_2$ з порядком $m > 0$, то при всіх достатньо малих $h < h_0$ справджується рівність

$$\|\varphi_h - P_{2h} f\|_{2h} = O(h^m). \tag{12}$$

Неважко помітити, що коли виконуються умови (11) і (12), то різницева схема (5) має m-й порядок апроксимації на розв'язку u рівняння (1). Справді,

$$\begin{aligned} \|\psi_h\|_{2h} = \|\varphi_h - A_h u_h\| = \|\varphi_h - A_h u_h + P_{2h}(Au) - P_{2h} f\| \leqslant \\ \leqslant \|\varphi_h - P_{2h} f\|_{2h} + \|A_h(P_{1h}u) - P_{2h}(Au)\|_{2h} = O(h^m). \end{aligned}$$

Аналогічні означення можна дати і для так званих нестаціонарних задач, наприклад:

$$\frac{\partial u}{\partial t} + Au = f, \quad u|_{t=0} = \varphi^{(0)} \tag{13}$$

або (аналогічно (2))

$$\begin{aligned} &\frac{\partial u}{\partial t} + Au = f(x, t), \quad (x,t) \in \Omega \times \Omega_t, \\ &au = g(x), \quad x \in \partial\Omega \times \Omega_t, \\ &u(x, 0) = \varphi^{(0)}(x), \quad x \in \Omega. \end{aligned} \tag{13'}$$

Приклад 2. Якщо оператор A задати першим способом з прикладу 1, то початково-крайову задачу для рівняння теплопровідності запишемо у вигляді (13)

$$\begin{aligned} &\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(x, t), \quad x \in \Omega \equiv (0, 1), \quad t \in \Omega_t \equiv (0, T], \\ &u(0, t) = u(1, t) = 0, \quad t \in \Omega_t, \\ &u(x, 0) = \varphi^{(0)}(x), \quad x \in \Omega. \end{aligned}$$

Якщо A та a означити другим способом прикладу 1, то дістанемо постановку (13′).

Апроксимацію задачі (13) або (13′) можна здійснити в два етапи. Спочатку цю задачу апроксимуємо в області $\bar{\omega}_h \times \Omega_t$ ($(\omega_h \cup \gamma_h) \times \Omega_t$) за просторовими змінними (тобто за всіма змінними, крім часу). В результаті приходимо до деякої диференціально-сіткової задачі

$$\frac{\partial v_h}{\partial t} + A_h v_h = \varphi_h, \quad v_h|_{t=0} = \varphi_h^{(0)} \tag{14}$$

(аналогічно для (13′)), яка є системою звичайних диференціальних

рівнянь для компонент сіткової функції v_h. Такий метод апроксимації називається ще *методом прямих* (див. гл. 6).

Далі для спрощення опускатимемо індекс h і вважатимемо, що оператор $A_h \equiv \Lambda$ не залежить від часу. Щоб апроксимувати задачу (14) за часом, введемо сітку ω_τ, тобто множину вузлів $t_j = j\tau$, $j = 1, 2, \ldots$, і знову заради спрощення покладемо $\varphi^j = \varphi(t_j)$. Тоді (14) можна апроксимувати, наприклад, явною схемою першого порядку апроксимації за часом

$$\frac{v^{j+1} - v^j}{\tau} + \Lambda v^j = \varphi^j, \quad j = 0, 1, \ldots, \quad v^\circ = \varphi^{(0)}, \tag{14'}$$

або неявною схемою

$$\frac{v^{j+1} - v^j}{\tau} + \Lambda v^{j+1} = \varphi^j, \quad j = 0, 1, \ldots,$$
$$v^0 = \varphi^{(0)}; \quad \varphi^j = \varphi(t_{j+1}). \tag{14''}$$

Розв'язуючи ці схеми відносно невідомого v^{j+1}, приходимо до рекурентного співвідношення

$$v^{j+1} = Tv^j + \tau s\varphi^j, \quad j = 0, 1, \ldots; \quad v^0 = \varphi^{(0)}, \tag{15}$$

де T називається оператором кроку, а s — оператором джерела. Для схеми (14′) маємо

$$T = I - \tau\Lambda, \quad s = I,$$

а для схеми (14″) —

$$T = (I + \tau\Lambda)^{-1}, \quad s = T.$$

Різницеві схеми вигляду (15) для еволюційних (залежних від часу) задач називаються двох'ярусними. На практиці часто застосовують схему другого порядку апроксимації (за часом) — схему Кранка — Ніколсона

$$\frac{v^{j+1} - v^j}{\tau} + \Lambda\frac{v^{j+1} + v^j}{2} = \varphi^j, \quad \varphi^j = \varphi\left(t_j + \frac{\tau}{2}\right), \quad v^0 = \varphi^{(0)}, \tag{14'''}$$

яку можна подати у вигляді (15) при

$$T = \left(I + \frac{\tau}{2}\Lambda\right)^{-1}\left(I - \frac{\tau}{2}\Lambda\right), \quad s = \left(I + \frac{\tau}{2}\Lambda\right)^{-1}.$$

У деяких випадках (див., наприклад, гл. 2) сіткові рівняння (14′) — (14‴) зручно записувати у вигляді одного операторного рівняння в сітковій області $\omega_{h\tau} = \omega_h \times (\{0\} \cup \omega_\tau)$

$$A_{h\tau}v_{h\tau} = \varphi_{h\tau}, \tag{16}$$

або у вигляді системи двох рівнянь, з яких одне апроксимує саме

рівняння (13′) на сітці $\omega_{h\tau} = \omega_h \times \omega_\tau$, а друге — граничну умову:

$$A_{h\tau}v_{h\tau} = \varphi_{h\tau} \text{ на } \omega_{h\tau},$$
$$a_{h\tau}v_{h\tau} = g_{h\tau} \text{ на } \gamma_{h\tau} \equiv \omega_h \times \{0\} \cup \gamma_h \times \omega_\tau, \tag{16'}$$

причому

$$\|A_{h\tau}(u)_{h\tau} - \varphi_{h\tau}\|_{\omega_{h\tau}} \leqslant M_1h^n + N_1\tau^p,$$
$$\|a_{h\tau}(u)_{h\tau} - g_{h\tau}\|_{\gamma_{h\tau}} \leqslant M_2h^n + N_2\tau^p, \tag{17}$$

де $\|\cdot\|_{\omega_{h\tau}}$, $\|\cdot\|_{\gamma_{h\tau}}$ — деякі норми у просторі функцій, заданих на сітках $\omega_{h\tau}$, $\gamma_{h\tau}$ відповідно, $(\cdot)_{h\tau}$ — оператор проектування на відповідний сітковий простір. Якщо схема для нестаціонарного рівняння записана у вигляді (16) або (16′), то її стійкість можна означити так, як і раніше.

Якщо схему записано у вигляді (15), то вона називається *стійкою*, коли при будь-якому h і $j \leqslant T/\tau$ справедлива нерівність

$$\|v^j\|_{\omega_h} \leqslant C_1\|\varphi^{(0)}\|_{\omega_h} + C_2\|\varphi\|_{\omega_{h\tau}}, \tag{18}$$

де сталі C_1, C_2 рівномірно обмежені на $0 \leqslant t \leqslant T$ і не залежать від τ, h.

Це означення також означає неперервну залежність розв'язку схеми (15) від вихідних даних. Дійсно, нехай v та v_* — розв'язки (15), які відповідають вихідним даним $\varphi^{(0)}$, φ та φ_*^0, φ_* відповідно. Тоді для різниці $\varepsilon^j = v^j - v_*^j$

$$\varepsilon^{j+1} = T\varepsilon^j + \tau S(\varphi^j - \varphi_*^j), \quad j = 0, 1, \ldots,$$
$$\varepsilon^0 = \varphi^{(0)} - \varphi_*^{(0)}.$$

Далі з (18) маємо

$$\|v^j - v_*^j\|_{\omega_h} \leqslant C_1\|\varphi^{(0)} - \varphi_*^{(0)}\|_{\omega_h} + C_2\|\varphi - \varphi_*\|_{\omega_{h\tau}},$$

звідки випливає, що малі варіації вихідних даних $\varphi^{(0)}$ та φ спричиняють і малі варіації розв'язку v.

Якщо схема стійка при довільних $\tau > 0$, $h > 0$, то вона називається *абсолютно стійкою*. Якщо стійкість (тобто (18)) має місце лише при певній залежності τ від h або при інших обмеженнях на τ, то така схема називається *умовно стійкою*.

Зазначимо, що в літературі зустрічаються й інші означення стійкості.

Дослідження збіжності сіткового розв'язку до розв'язку вихідної задачі здійснюється за однією і тією самою схемою і для стаціонарних і нестаціонарних рівнянь. Тому основну ідею розглянемо на прикладі схеми (5).

Якщо різницева схема (5) коректна, а u — розв'язок задачі (1), то з (8) випливає, що

$$\| z_h \|_{1h} \leqslant \varkappa \| \psi_h \|_{2h}.$$

Звідси робимо висновок, який іноді називають *теоремою Лакса — Філіпова.*

Теорема. *Якщо схема (5) коректна і апроксимує задачу* (1), *то розв'язок задачі* (5) *при* $\| h \| \to 0$ *збігається до розв'язку и задачі* (1), *причому порядок швидкості збіжності (порядок точності) схеми збігається з порядком похибки апроксимації на розв'язку и задачі* (1), *або, іншими словами, із стійкості і апроксимації схеми* (5) *випливає її збіжність до розв'язку и рівняння* (1).

А. Ф. Філіпов дав означення стійкості для довільних сіткових задач

$$A_{h\tau} v_{h\tau} = \varphi_{h\tau}$$

як рівномірної обмеженості операторів $(A_{h\tau})^{-1}$ і довів, що із апроксимації і стійкості випливає збіжність розв'язку сіткової задачі до розв'язку диференціальної. П. Лакс запропонував для коректно поставлених нестаціонарних задач означення апроксимації і стійкості, при яких стійкість має місце разом зі збіжністю, якщо справедлива апроксимація.

Таким чином, дослідження збіжності сіткової схеми (5) і здобуття оцінок швидкості збіжності можна розбити на два етапи:

1) дослідження коректності сіткової схеми (5), тобто встановлення апріорної оцінки (8); 2) дослідження порядку апроксимації. Найчастіше B_1, B_2 являють собою нормовані простори функцій неперервного аргументу x, що змінюється в деякій області Ω. Для побудови просторів B_{1h}, B_{2h} область Ω покривають деякою сіткою ω_h і за B_{1h}, B_{2h} вибирають простори сіткових функцій дискретного аргументу $x \in \omega_h$ з відповідною нормою.

Як приклади сіток, що покривають область $\bar{\Omega} = [a, b]$, ми вже розглядали рівномірну сітку $\bar{\omega}_h = \{x_i : x_i = a + ih;\ i = \overline{0, N},\ h = (b-a)/N\}$ і нерівномірну сітку $\hat{\bar{\omega}}_h = \left\{x_i : x_i = x_{i-1} + h_i,\ x_0 = a,\ i = \overline{0, N},\ h_i = x_i - x_{i-1},\ \sum_{i=1}^{N} h_i = b - a\right\}$. Розглянемо приклади вибору просторів B_{1h}, B_{2h} і операторів P_{ih}, $i = 1, 2$.

Приклад 3. Нехай $B_1 = C[a, b]$ і $\hat{\bar{\omega}}_h$ — нерівномірна сітка, яка покриває $[a, b]$. За B_{1h} виберемо нормований простір сіткових функцій $C(\hat{\bar{\omega}}_h)$ з нормою $\| v \|_{C(\hat{\bar{\omega}}_h)} = \max\limits_{x \in \omega_h} | v(x) |$. Оператор P_{1h} визначимо рівністю

$$P_{1h} u \equiv u_h = u(x), \quad x \in \hat{\bar{\omega}}_h. \tag{19}$$

Неважко помітити, що умова узгодження норм (7) виконується.

Приклад 4. Нехай $B_1 = C^1[a, b]$ з нормою $\| u \|_1 = \max\limits_{x \in [a,b]} | u(x) | + \max\limits_{x \in [a,b]} | u'(x) |$ і рівномірна сітка $\bar{\omega}_h$ покриває $[a, b]$. За B_{1h} виберемо простір сіткових функцій з нормою

$$\| v \|_{1h} = \max_{x \in \bar{\omega}_h} | v(x) | + \max_{x \in \omega_h^-} | (v(x+h) - v(x))/h |,$$

де $\omega_h^- = \{x_i = a + ih,\ i = \overline{0, N-1},\ h = (b-a)/N\}$. Оператор P_{1h} визначимо формулою (19). З результатів п. 2.9 випливає, що у випадку $u \in C^2[a, b] \equiv B_1$

$$\max_{x \in \omega_h^-} | (u(x+h) - u(x))/h | = \max_{x \in \omega_h^-} | u_x(x) | = \max_{x \in [a,b]} | u'(x) | \quad \text{при} \quad h \to 0,$$

тому умова узгодження норм (7) виконується.

Приклад 5. Нехай $B_2 = L_2[a, b]$ — простір неперервних функцій з нормою

$$\| f \|_2 = \left[\int_a^b f^2(t)\, dt \right]^{1/2}.$$

Покриваємо $[a, b]$ рівномірною сіткою ω_h і за B_{2h} виберемо простір сіткових функцій з нормою

$$\| v \|_{2h} = \left(\sum_{x \in \omega_h} h v^2(x) \right)^{1/2},$$

а за оператор P_{2h} — оператор, що визначається формулою

$$P_{2h} f \equiv f_h = h^{-1} \int_{x - h/2}^{x + h/2} f(t)\, dt, \quad x \in \omega_h.$$

Неважко помітити, що умова узгодження норм (7) виконується.

Часто норми в B_1, B_{1h} вводяться з урахуванням властивостей оператора A. Якщо A — додатний і самоспряжений оператор, то на $D(A) = B_1$ можна побудувати енергетичний гільбертів простір H_A, замкнувши $D(A)$ за нормою $\| u \|_A = \sqrt{(Au, u)}$. Можна також побудувати сітковий аналог цієї норми з урахуванням властивостей оператора A_h таким чином, що буде виконуватися умова узгодження. Якщо оператор A_h додатний, самоспряжений і існує A_h^{-1}, то для дослідження коректності сіткової задачі часто використовують негативну норму $\| v \|_{A_h^{-1}} = \sqrt{(A_h^{-1} v, v)}$. Вибір норми в усіх випадках визначається з однією метою — визначити збіжність і порядок швидкості збіжності сіткової схеми. Останнє залежить як від вибору B_{1h}, так і B_{2h}, а також норм в них. Вибирати їх треба таким чином, щоб зберігалася стійкість і мала місце апроксимація якомога вищого порядку.

Розглянемо приклад стійкої схеми.

Приклад 6. Нехай задано різницеву крайову задачу

$$y_{\bar{x}x,i} = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = -\varphi_i, \quad i = \overline{1, N-1}, \quad y_0 = 0, \quad y_N = 0. \tag{20}$$

Нехай $B_{1h} \equiv H_h$ — простір сіткових функцій, заданих на сітці $\bar{\omega}_h = \{x_i = ih : i = \overline{0, N}\}$ і рівних нулю на границі $\bar{\omega}_h$, тобто $y_0 = y_N = 0$. Визначимо оператор A_h за допомогою рівності

$$(A_h y)_i = -y_{\bar{x}x,i}, \quad i = \overline{1, N-1}; \quad y \in H_h.$$

Тоді вихідну задачу запишемо у вигляді операторного рівняння

$$A_h y_h = \varphi_h. \tag{21}$$

У просторі H_h введемо скалярний добуток

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h.$$

Оператор A_h в H_h самоспряжений і додатно визначений, причому

$$\delta \| y \|^2 \leqslant (A_h y, y) \in \Delta \| y \|^2 \quad \forall\, y \in H_h,$$

де δ і Δ — найменше і найбільше власні значення оператора A_h, що дорівнюють

$$\delta = \frac{4}{h^2} \sin^2 \frac{\pi h}{2}, \quad \Delta = \| A_h \| = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}$$

(див. п. 1.4.4, ч. 1). Звідси випливає, що

$$\Delta^{-1} \| y \|^2 \leqslant (A_h^{-1} y, y) \leqslant \delta^{-1} \| y \|^2, \quad \| A_h^{-1} \| \leqslant \delta^{-1}.$$

Це означає, що норма оберненого оператора рівномірно по h обмежена, тобто $\| A_h^{-1} \| \leqslant \delta^{-1} < \frac{1}{8}$, і має місце апріорна оцінка

$$\| y_h \| \leqslant \delta^{-1} \| \varphi_h \| \leqslant \frac{1}{8} \| \varphi_h \|,$$

яка виражає стійкість схеми (21).

Цю саму оцінку можна дістати методом енергетичних нерівностей. Помножимо рівняння (21) скалярно на y_h: $(A_h y_h, y_h) = (\varphi_h, y_h)$. Застосовуючи формулу підсумовування частинами, маємо $(A_h y_h, y_h) = | y_h |^2_{\mathring{W}_2^1(\omega_h)}$. Крім того, використовуючи результати п. 1.5, ч. 1, дістанемо

$$| (\varphi_h, y_h) | \leqslant \| \varphi_h \|_{L_2(\omega_h)} \| y_h \|_{L_2(\omega_h)} \leqslant \frac{1}{\sqrt{8}} \| \varphi_h \|_{L_2(\omega_h)} | y_h |_{\mathring{W}_2^1(\omega_h)}.$$

Таким чином,

$$| y_h |_{\mathring{W}_2^1(\omega_h)} \leqslant \frac{1}{\sqrt{8}} \| \varphi_h \|_{L_2(\omega_h)}. \tag{22}$$

Враховуючи, що $\| \varphi_h \|_{L_2(\omega_h)} \leqslant \| \varphi_h \|_{C(\omega_h)}$, і теорему вкладення (див. лему 2 з п. 1.5, ч. 1) $\| y_h \|_{C(\omega_h)} \leqslant \frac{1}{2} | y_h |_{\mathring{W}_2^1(\omega_h)}$, дістанемо оцінку

$$\| y_h \|_{C(\omega_h)} \leqslant \frac{1}{4\sqrt{2}} \| \varphi_h \|_{C(\omega_h)}. \tag{23}$$

Оцінки (22) і (23) виражають стійкість схеми (21) при різному виборі просторів B_{1h} і B_{2h}. У першому випадку $B_{1h} = \mathring{W}_2^1(\omega_h)$ — простір сіткових функцій, заданих на $\bar{\omega}_h$ і рівних нулю на границі з нормою $\| \cdot \|_{\mathring{W}_2^1(\omega_h)} = | \cdot |_{\mathring{W}_2^1(\omega_h)}$, в другому випадку $B_{1h} = B_{2h} = C(\bar{\omega}_h)$.

Приклад 7. Розглянемо приклад некоректної схеми. Нехай в (21) задано y_h і треба знайти φ_h. Цю задачу можна записати у вигляді операторного рівняння

$$B_h \varphi_h = y_h, \quad B_h = A_h^{-1}. \tag{24}$$

Оскільки $\| B_h^{-1} \| = \| (A_h^{-1})^{-1} \| = \| A_h \| = \Delta = \frac{4}{h^2} \cos^2 \frac{\pi h}{2} \to \infty$ при $h \to \infty$, то довести нерівність (8) з незалежною від h сталою $\varkappa$ при $h \to 0$ не можна, хоча

$$\| \tilde{\varphi}_h - \varphi_h \| \leqslant \| B_h^{-1} \| \| \tilde{y}_h - y_h \|,$$

де $\tilde{\varphi}_h$ — розв'язок рівняння $B_h \tilde{\varphi}_h = \tilde{y}_h$. Оскільки $\| B_h^{-1} \| \to \infty$ при $h \to 0$, то $\forall\, M$, не залежного від h, можна вказати $h_0(M)$ таке, що $\| B_h^{-1} \| > M$ при $h \leqslant h_0(M)$ і $\| B_h^{-1} \| \leqslant M$ при $h > h_0(M)$. Отже, при $h \geqslant h_0$

$$\| \tilde{\varphi}_h - \varphi_h \| \leqslant M \| \tilde{y}_h - y_h \|.$$

Схема, для якої виконується така умова, називається *квазістійкою*. Покажемо, що вона не завжди придатна для практичного знаходження φ_h із заданою точністю ε, якщо y_h задане з точністю ε_0, тобто $\| \tilde{y}_h - y_h \| \leqslant \varepsilon_0$. Досягти необхідної точності можна, якщо $\| B_h^{-1} \| \varepsilon_0 \leqslant \varepsilon$. Звідси дістанемо допустимий крок сітки $h \geqslant h_0(\varepsilon/\varepsilon_0)$. Проте може статися, що така умова не може бути виконана. Дійсно, для задачі (21) умова $\| B_h^{-1} \| \varepsilon_0 = \Delta \varepsilon_0 \leqslant \varepsilon$ виконується, якщо $4\varepsilon_0/h^2 \leqslant \varepsilon$, або $h \geqslant h_0 = 2\sqrt{\varepsilon_0/\varepsilon}$. Якщо $\varepsilon_0 = 1/4 \cdot 10^{-4}$, $\varepsilon = 10^{-4}$, то $h_0 = 1$, і оскільки розв'язок шукають на відрізку $[0, 1]$, то точності $\varepsilon = 10^{-4}$ не можна досягти на жодній сітці.

Наведемо приклади оцінок похибки апроксимації і точності сіткової схеми.

Приклад 8. Розглянемо різницеву схему (20), (21), яка апроксимує крайову задачу

$$Lu = -u'' = f(x), \quad 0 < x < 1, \quad u(0) = u(1) = 0. \tag{25}$$

Нехай $f(x) \in C^2[0, 1] \equiv B_2$, тоді $u(x) \in C^4[0, 1] \cap \mathring{C}[0, 1] \equiv B_1$. Крім того, $B_{1h} = \mathring{C}(\bar{\omega}_h)$, де $\mathring{C}(\bar{\omega}_h)$ — множина функцій, заданих на сітці $\bar{\omega}_h$ і рівних нулю при $x = 0, 1$, $B_{2h} = C(\omega_h)$. Оператори P_{1h}, P_{2h} виберемо у вигляді (19). Задача (9) для похибки $z_h = y_h - P_{1h} u \equiv y_h - u_h$ набере вигляду

$$A_h z_h \equiv -(z_h)_{\bar{x}x} = \psi_h(x), \quad x \in \omega_h, \quad z_h(0) = z_h(1) = 0,$$

де

$$\psi_h(x) = \varphi_h(x) - A_h u_h = \varphi + u_{\bar{x}x} = \varphi + \left(u'' + \frac{1}{12} h^2 u^{\mathrm{IV}}\right) =$$

$$= (\varphi + u'') + \frac{h^2}{12} u^{(\mathrm{IV})}(\xi) = \varphi - f + O(h^2),$$

бо $u'' = -f(x)$. Звідси випливає, що $\| \psi_h \|_C = O(h^2)$, коли покласти $\varphi = f$ або $\varphi = f + O(h^2)$. Далі, в силу апріорної оцінки (23), дістанемо

$$\| z_h \|_C = \| y_h - u_h \|_C = O(h^2).$$

Часто похибку апроксимації зручно подати у вигляді

$$\psi_h = \varphi_h - A_h u_h = (\varphi_h - A_h u_h) - (f - Au)_h =$$
$$= (\varphi_h - f_h) - (A_h u_h - (Au)_h) = \psi_h^{(1)} + \psi_h^{(2)}$$

з урахуванням того, що $f - Au = 0$. Тут $\psi_h^{(1)} = -A_h u_h + (Au)_h$ — похибка апроксимації оператора A оператором A_h на розв'язку u задачі (25), $\psi_h^{(2)}$ — похибка апроксимації правої частини. Вимога $\psi_h \|_{2h} = O(h^m)$ виконана, якщо $\| \psi_h^{(1)} \|_{2h} = O(h^m)$ і $\| \psi_h^{(2)} \|_{2h} = O(h^m)$. Однак ці умови не є необхідними для оцінки $\| \psi_h \|_{2h} = O(h^m)$, про що свідчить наступний приклад.

Для задачі (25) знову розглянемо схему вигляду (20), але з іншою правою частиною

$$\varphi_h = f + \frac{1}{12} h^2 f_{\bar{x}x}.$$

Вважатимемо, що $u \in C^6[0, 1] \cap \mathring{C}[0, 1]$, або $f \in C^4[0, 1]$. Тоді

$$\psi_h^{(2)} = \varphi^h - f_h = O(h^2),$$

$$\psi_h^{(1)} = -u_{\bar{x}x} + u'' = \frac{1}{12} h^2 u^{(IV)} + O(h^4) = O(h^2).$$

Разом з тим

$$\psi_h = \psi_h^{(1)} + \psi_h^{(2)} = \varphi - f + \frac{h^2}{12} u^{IV} + O(h^4) = \frac{h^2}{12} (f_{\bar{x}x} + u^{IV}) +$$

$$+ O(h^4) = \frac{h^2}{12} (f'' + u^{IV}) + O(h^4).$$

Оскільки $u^{IV} + f'' = 0$ в силу рівності $u'' + f(x) = 0$, то $\psi_h = O(h^4)$. Тобто схема має четвертий порядок апроксимації.

Якщо в (1) оператор A є нелінійним, то і A_h в (5) буде нелінійним. Тоді

$$A_h v_h - A_h u_h = \psi_h,$$

де $\psi_h = \varphi_h - A_h u_h$ — похибка апроксимації. Звідси

$$\tilde{A}_h z_h = \psi_h, \qquad (9')$$

де $z_h = v_h - u_h$, а лінійний оператор $\tilde{A}_h$ задається формулою

$$\tilde{A}_h = \int_0^1 A_h'(u_h + \theta(v_h - u_h))\, d\theta,$$

де A_h' — похідна Фреше від оператора A_h. Іноді зручніше обчислювати похідну Гато, яка визначається формулою

$$A_h'(x_0) h = \lim_{t \to 0} \frac{A_h(x_0 + th) - A_h(x_0)}{t}.$$

Зазначимо, що з диференційовності нелінійного оператора в розумінні Фреше випливає його диференційовність в розумінні Гато, але не завжди навпаки.

Після знаходження лінійної задачі для похибки (9), (9′) дослідження нелінійної сіткової схеми відбувається за тією самою методикою, що й лінійної.

Означення. Схему (5) з нелінійним оператором A_h називають *стійкою,* якщо для розв'язків задач

$$A_h y_1 = \varphi_1, \quad A_h y_2 = \varphi_2$$

виконується оцінка

$$\| y_1 - y_2 \|_{1h} \leqslant C \| \varphi_1 - \varphi_2 \|_{2h}, \qquad (8')$$

де стала C не залежить від $y_1, y_2, \varphi_1, \varphi_2, h$.

Для лінійного оператора A_h означення стійкості (8) і (8′) еквівалентні.

Враховуючи тотожність $A_h u_h \equiv A_h u_h$ за умови стійкості нелінійної схеми (5), для $z_h = v_h - u_h$ матимемо оцінку

$$\| z_h \|_{1h} \leqslant C \| \varphi_h - A_h u_h \|_{2h} \equiv C \| \psi_h \|_{2h}.$$

Таким чином, і в нелінійному випадку дослідження збіжності і порядку точності схеми (5) зводиться до встановлення стійкості, тобто нерівності (8) або (8′), і дослідження порядку похибки апроксимації. Часто використання рівності (9′) є проміжним кроком до нерівності (8) або (8′).

2.2. Методи відшукання апріорних оцінок (методи дослідження коректності (стійкості) сіткових схем)

Відшукання апріорних оцінок для сіткових схем, як це випливає з результатів п. 2.1, є важливою задачею. На жаль, є небагато стандартних методів, які дозволяють для деяких класів сіткових схем і для певних норм здобути такі оцінки. Наприклад, в п. 4.1 з ч. 1 ми дістали оцінку розв'язку системи лінійних алгебраїчних рівнянь $Ay = f$ (в такому вигляді записується багато сіткових схем) у випадку, коли матриця A має строгу діагональну перевагу. Ще одну оцінку дає теорема Банаха (див. п. 1.4 з ч.1). У багатьох же випадках відшукання апріорних оцінок вимагає досить тонких і нестандартних досліджень, які враховують індивідуальні особливості конкретної задачі і тому не можуть бути універсальними навіть для

класу задач. Нижче ми розглянемо деякі найпростіші і найпоширеніші методи для деяких класів задач і для конкретних норм.

2.2.1. Принцип максимуму за рядками (метод відшукання апріорної оцінки в нормі C (ω_h)). Розглянемо сіткове рівняння (різницеве рівняння другого порядку)

$$Ly_i = a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -\varphi_i, \quad i = \overline{1, N-1},$$
$$y_0 = \mu_1, \quad y_N = \mu_2, \tag{1}$$

задане на деякій сітці $\bar{\omega} = \{x_i : i = \overline{0, N}\}$, і припустимо, що виконуються умови

$$a_i > 0, \quad b_i > 0, \quad c_i \geqslant a_i + b_i, \quad i = \overline{1, N-1}. \tag{2}$$

Для нього має місце твердження, яке називається *принципом максимуму*.

Теорема 1. *Нехай різницевий оператор L визначений формулами (1) і виконуються умови (2). Якщо для функції y_i, заданої на сітці $\bar{\omega}$ і відмінної від сталої при $i = \overline{1, N-1}$, виконується умова $Ly_i \geqslant 0$ ($Ly_i \leqslant 0$) для всіх $i = \overline{1, N-1}$, то ця функція не може набувати найбільшого додатного (найменшого від'ємного) значення у внутрішніх вузлах сітки.*

Доведення. Нехай $Ly_i \geqslant 0$, $i = \overline{1, N-1}$. Припустимо від супротивного, що y_i у деякому внутрішньому вузлі $i = i_*$, $1 \leqslant i_* \leqslant N-1$, досягає найбільшого додатного значення $y_{i_*} = \max\limits_{0 \leqslant i \leqslant N} y_i = M_0 > 0$. Оскільки $y_i \neq \text{const}$, то знайдеться внутрішній вузол i_0 (може статися, що $i_0 = i_*$), в якому $y_{i_0} = y_{i_*} = M_0 > 0$, а в одному із сусідніх вузлів, наприклад у вузлі $i = i_0 - 1$, виконується строга нерівність $y_{i_0-1} < y_{i_0}$. Тоді

$$Ly_{i_0} = a_{i_0} y_{i_0-1} - c_{i_0} y_{i_0} + b_{i_0} y_{i_0+1} = b_{i_0}(y_{i_0+1} - y_{i_0}) -$$
$$- a_{i_0}(y_{i_0} - y_{i_0-1}) - (c_{i_0} - a_{i_0} - b_{i_0}) y_{i_0} < 0,$$

що суперечить умові $Ly_i \geqslant 0$, $i = \overline{1, N-1}$ ($Ly_i \geqslant 0$ в тому числі при $i = i_0$). Це і доводить перше твердження теореми. Друге твердження дістанемо, якщо замінити y_i на $-y_i$ і скористатися щойно доведеним твердженням.

Наслідок 1. Якщо виконуються умови (2), а також $Ly_i \leqslant 0$, $i = \overline{1, N-1}$, $y_0 \geqslant 0$, $y_N \geqslant 0$, то $y_i \geqslant 0$ для всіх $i = \overline{0, N}$. Якщо ж разом з (2) $Ly_i \geqslant 0$, $\mu_0 \leqslant 0$, $\mu_N \leqslant 0$, то $y_i \leqslant 0$, де $i = \overline{0, N}$.

Доведення. Нехай (від супротивного) $Ly_i \leqslant 0$, і $y_i < 0$ в якомусь внутрішньому вузлі $i = i_*$. Тоді y_i досягає найменшого від'ємного значення в якомусь внутрішньому вузлі, що неможливо в силу принципу максимуму.

Звідси безпосередньо випливає таке твердження.

Наслідок 2. Якщо $\varphi_i \geqslant 0$, $\mu_1 \geqslant 0$, $\mu_2 \geqslant 0$, то розв'язок задачі (1) за умов (2) невід'ємний, тобто $y_i \geqslant 0$, $i = \overline{0, N}$.

Наслідок 3. Якщо виконуються умови (2), то задача

$$Ly_i = 0, \quad i = \overline{1, N-1}; \quad y_0 = 0, \quad y_N = 0 \tag{3}$$

має лише нульовий розв'язок, а задача (1) має єдиний розв'язок при будь-яких φ_i, μ_1, μ_2.

Доведення. Припустимо, що розв'язок y_i задачі (3) відмінний від нуля хоча б в одній точці $i = i_*$. Якщо $y_{i_*} > 0$ ($y_{i_*} < 0$), то y_i досягає найбільшого додатного (найменшого від'ємного) значення в деякій внутрішній точці $i = i_0$, що суперечить принципу максимуму. Отже, $y_i \equiv 0$, $i = \overline{0, N}$.

Теорема 2 *(теорема порівняння). Нехай виконуються умови (2) і y_i — розв'язок задачі (1), а $\bar{y}_i$ — розв'язок задачі*

$$L\bar{y}_i = -\bar{\varphi}_i \geqslant 0, \quad i = \overline{1, N-1}; \quad \bar{y}_0 = \bar{\mu}_1 \geqslant 0, \quad \bar{y}_N = \bar{\mu}_2 \geqslant 0,$$

причому $|\varphi_i| \leqslant \bar{\varphi}_i$, $|\mu_1| \leqslant \bar{\mu}_1$, $|\mu_2| \leqslant \bar{\mu}_2$. Тоді справедливою є оцінка

$$|y_i| \leqslant \bar{y}_i, \quad i = \overline{0, N}.$$

Доведення. З наслідку 2 випливає, що $\bar{y}_i \geqslant 0$. Для $\bar{y}_i \pm y_i$ дістанемо задачі вигляду (1) з правими частинами $\bar{\varphi}_i \pm \varphi_i \geqslant 0$, $\bar{\mu}_1 \pm \mu_1 \geqslant 0$, $\bar{\mu}_2 \pm \mu_2 \geqslant 0$. Тоді з наслідку 2 випливає, що $\bar{y}_i \pm y_i \geqslant 0$. Це означає, що $-\bar{y}_i \leqslant y_i \leqslant \bar{y}_i$, тобто $|y_i| \leqslant \bar{y}_i$, а це і треба було довести.

Функція $\bar{y}_i$ називається *мажорантою* для розв'язку задачі (1). Якщо вдається побудувати мажоранту, то це означає, що вдалося відшукати таку оцінку для розв'язку задачі (1): $\|y\|_C \leqslant \|\bar{y}\|_C$.

Наслідок 4. Для розв'язку задачі

$$Ly_i = 0, \quad i = \overline{1, N-1}; \quad y_0 = \mu_1, \quad y_N = \mu_2$$

справедлива оцінка $\|y\|_C = \max\limits_{i=\overline{0,N}} |y_i| \leqslant \max(|\mu_1|, |\mu_2|)$.

Доведення. Розглянемо допоміжну задачу

$$L\bar{y}_i = 0, \quad i = \overline{1, N-1}; \quad \bar{y}_0 = \bar{y}_N = \bar{\mu},$$

де $\bar{\mu} = \max(|\mu_1|, |\mu_2|)$. Принцип максимуму дає нерівність $\|\bar{y}\|_C \leqslant \bar{\mu}$ (оскільки $y_i \geqslant 0$ і може досягати найбільшого додатного значення лише на границі, тобто при $i = 0$ чи $i = N$), а з теореми порівняння дістаємо твердження наслідку 4.

Теорема 3. *Нехай виконуються умови*

$$|a_i| > 0, \quad |b_i| > 0, \quad \bar{d}_i = |c_i| - |a_i| - |b_i| > 0, \quad i = \overline{1, N-1}. \tag{4}$$

Тоді для розв'язку задачі

$$Ly_i = -\varphi_i, \quad i = \overline{1, N-1}; \quad y_0 = y_N = 0 \tag{5}$$

має місце оцінка

$$\|y\|_C \leqslant \|\varphi/\bar{d}\|_C \equiv \max_{i=\overline{1,N-1}} |\varphi_i/\bar{d}_i|.$$

Доведення. Перепишемо (5) у вигляді

$$c_i y_i = a_i y_{i-1} + b_i y_{i+1} + \varphi_i. \tag{6}$$

Нехай $|y_i|$ досягає найбільшого значення $|y_{i_0}| > 0$ при $i = i_0 \in (0, N)$, тобто $|y_i| \leqslant |y_{i_0}|$ для будь-якого $i = \overline{0, N}$. Тоді з (6) при $i = i_0$ маємо

$$|c_{i_0}||y_{i_0}| = |a_{i_0}y_{i_0-1} + b_{i_0}y_{i_0+1} + \varphi_{i_0}| \leqslant |a_{i_0}||y_{i_0-1}| + |b_{i_0}||y_{i_0+1}| + |\varphi_{i_0}| \leqslant (|a_{i_0}| + |b_{i_0}|)|y_{i_0}| + |\varphi_{i_0}|,$$

$$(|c_{i_0}| - |a_{i_0}| - |b_{i_0}|)\, y_{i_0} \leqslant |\varphi_{i_0}|,$$

$$|y_{i_0}| \leqslant |\varphi_{i_0}|/\bar{d}_{i_0} \leqslant \|\varphi/\bar{d}\|_C,$$

що і треба було довести.

Вправа 1. Записавши задачу (5) у матричному вигляді, переконатися, що матриця має строгу діагональну перевагу, і тому твердження теореми 3 випливає з леми 4.1.1 з ч. 1.

Наслідок 5. Нехай для оператора L виконуються умови теореми 3. Тоді для розв'язку задачі (1) виконується апріорна оцінка

$$\|y\|_C \leqslant \|\varphi/d\|_C + \max(|\mu_1|, |\mu_2|).$$

Доведення. Розв'язок задачі (1) можна подати у вигляді $y_i = y_i^{(1)} + y_i^{(2)}$, де $y_i^{(1)}$ — розв'язок неоднорідної задачі (5) з однорідними граничними умовами, а $y_i^{(2)}$ — розв'язок однорідного рівняння з неоднорідними граничними умовами. Оцінки $y_i^{(1)}$ за теоремою 3 і $y_i^{(2)}$ за наслідком 4 дають твердження наслідку 5.

Розглянемо тепер випадок, коли замість строгих нерівностей (4) виконуються нерівності

$$d_i = c_i - a_i - b_i \geqslant 0, \quad a_i > 0, \quad b_i > 0, \quad i = \overline{1, N-1}, \tag{7}$$

тобто d_i можуть при деяких i перетворитися в нуль. Оскільки при доведенні оцінки з наслідку 4 строга нерівність $d_i > 0$ не потрібна, то, очевидно, досить дістати оцінку лише для однорідної задачі

$$Ly_i = -\varphi_i, \quad i = \overline{1, N-1}, \quad y_0 = 0, \quad y_N = 0. \tag{8}$$

Теорема 4. *Якщо умови (7) виконуються, то для розв'язку задачі (8) справедлива оцінка*

$$\|y\|_C \leqslant 2\sum_{k=1}^{N} \frac{1}{A_k} \left| \sum_{j=1}^{k-1} F_j \right|,$$

де $A_k = a_k\eta_k$, $F_j = \varphi_j\eta_j$, $\eta_{i+1} = \prod_{k=1}^{i} (b_k/a_{k+1})$, $h > 0$ — *довільна стала.*

Доведення. Подамо y_i у вигляді суми $y_i = v_i + w_i$, де w_i — розв'язок задачі

$$\mathring{L}w_i = b_i(w_{i+1} - w_i) - a_i(w_i - w_{i-1}) = -\varphi_i, \quad i = \overline{1, N-1}; \quad w_0 = w_N = 0, \tag{9}$$

а v_i — розв'язок задачі

$$Lv_i = b_i(v_{i+1} - v_i) - a_i(v_i - v_{i-1}) - d_i v_i = d_i w_i, \quad i = \overline{1, N-1}, \quad v_0 = v_N = 0. \tag{10}$$

Покажемо, що для розв'язку задачі (10) за умов (7) справедлива оцінка

$$\|v\|_C \leqslant \|w\|_C. \tag{11}$$

Дійсно, якщо $d_i \equiv 0$, то в силу наслідку 3 маємо $v_i \equiv 0$, і оцінка (11) виконується. Нехай $d_i \not\equiv 0$ хоча б в одній точці. Побудуємо мажоранту $\bar{v}_i$ як розв'язок задачі

$$L\bar{v}_i = -d_i|w_i|, \quad i = \overline{1, N-1}, \quad \bar{v}_0 = \bar{v}_N = 0.$$

Оскільки $L\bar{v}_i = -d_i|w_i| \leqslant 0$, то з принципу максимуму випливає, що $\bar{v}_i \geqslant 0$ і $\bar{v}_i$ досягає найбільшого значення при $i = i_0$. Тоді $\bar{v}_{i_0+1} - \bar{v}_{i_0} \leqslant 0$, $\bar{v}_{i_0} - \bar{v}_{i_0-1} \geqslant 0$ і з рівності (10) дістанемо

$$d_{i_0}\bar{v}_{i_0} \leqslant -b_{i_0}(\bar{v}_{i_0+1} - \bar{v}_{i_0}) + a_{i_0}(\bar{v}_{i_0} - \bar{v}_{i_0-1}) + d_{i_0}\bar{v}_{i_0} = d_{i_0}|w_{i_0}|.$$

Якщо $d_{i_0} > 0$, то $\bar{v}_{i_0} < |w_{i_0}|$ і одразу дістанемо оцінку (11), бо $|v_i| \leqslant \bar{v}_i$. Якщо $d_{i_0} = 0$, то рівняння (10) набере вигляду

$$b_{i_0}(\bar{v}_{i_0+1} - \bar{v}_{i_0}) + a_{i_0}(\bar{v}_{i_0} - \bar{v}_{i_0-1}) = 0.$$

З цього рівняння випливає, що $\bar{v}_{i_0+1} = \bar{v}_{i_0} = \bar{v}_{i_0-1}$. Оскільки $\bar{v}_i \not\equiv \text{const}$, то існує таке $i = i_1$, для якого $\bar{v}_{i_1} = \bar{v}_{i_0}$, а, наприклад, для $i = i_1 + 1$ буде $\bar{v}_{i_1+1} < \bar{v}_{i_1}$. Тоді $d_{i_1} \not\equiv 0$ і ми приходимо до розглянутого вище випадку, тобто $\bar{v}_{i_1} = \|\bar{v}\|_C \leqslant |w_{i_0}| \leqslant \|w\|_C$. Отже, оцінку (11) доведено. Знайдемо розв'язок задачі (9) в явному вигляді. Для цього помножимо рівняння (9) на η_i і виберемо його таким чином,

щоб $b_i\eta_i = a_{i+1}\eta_{i+1}$. З останньої рівності маємо

$$\eta_{i+1} = \frac{b_i}{a_{i+1}}\,\eta_i = \eta_1 \prod_{k=1}^{i} \frac{b_k}{a_{k+1}},$$

де η_1 — довільне число. Позначивши $F_i = \varphi_i\eta_i$, $A_i = a_i\eta_i$, замість (9) дістанемо задачу

$$A_{i+1}(w_{i+1} - w_i) - A_i(w_i - w_{i-1}) = -F_i, \quad i = \overline{1, N-1},$$
$$w_0 = w_N = 0, \tag{12}$$

або

$$\Delta(A_i \nabla w_i) = -F_i, \quad i = \overline{1, N-1}, \quad w_0 = w_N = 0, \tag{13}$$

де $\Delta v_i = v_{i+1} - v_i$, $\nabla v_i = v_i - v_{i-1}$.

Нехай $\Delta\mu_i = \mu_{i+1} - \mu_i = F_i$, тобто $\mu_{i+1} = \mu_i + F_i = \sum_{k=1}^{i} F_i + \mu_1$, де μ_1 — довільна стала. Тоді рівняння (13) набере вигляду

$$\Delta(A_i \nabla w_i) = -\Delta\mu_i,$$

або $\Delta(A_i\nabla w_i + \mu_i) = 0$. Звідси $A_i\nabla w_i + \mu_i = c_0$, де c_0 — стала. Далі маємо

$$w_i = w_{i-1} + \frac{c_0 - \mu_i}{A_i} = c_0 \sum_{k=1}^{i} \frac{1}{A_k} - \sum_{k=1}^{i} \frac{\mu_k}{A_k} + w_0,$$

$$0 = w_N = c_0 \sum_{k=1}^{N} \frac{1}{A_k} - \sum_{k=1}^{N} \frac{\mu_k}{A_k},$$

звідки $c_0 = \sum_{k=1}^{N} \frac{\mu_k}{A_k} \Big/ \sum_{k=1}^{N} \frac{1}{A_k}$. Позначивши $\alpha_i = \sum_{k=1}^{i} \frac{1}{A_k} : \sum_{k=1}^{N} \frac{1}{A_k}$, $0 < \alpha_i \leqslant 1$, можемо записати

$$w_i = \alpha_i \sum_{k=1}^{N} \frac{\mu_k}{A_k} - \sum_{k=1}^{i} \frac{\mu_k}{A_k}.$$

Далі знаходимо

$$|w_i| = \left| -(1-\alpha_i) \sum_{k=1}^{i} \frac{\mu_k}{A_k} + \alpha_i \sum_{k=i+1}^{N} \frac{\mu_k}{A_k} \right| \leqslant$$

$$\leqslant (1-\alpha_i) \sum_{h=1}^{i} \frac{|\mu_k|}{A_k} + \alpha_i \sum_{k=i+1}^{N} \frac{|\mu_k|}{A_k} \leqslant \sum_{k=1}^{N} \frac{|\mu_k|}{A_k},$$

тобто

$$\|w\|_C \leqslant \sum_{k=1}^{N} \frac{|\mu_k|}{A_k} = \sum_{k=1}^{N} \frac{1}{A_k} \left| \sum_{j=1}^{k-1} F_j \right| + \mu_1 \sum_{k=1}^{N} \frac{1}{A_k}.$$

Поклавши тут $\mu_1 = 0$, $\eta_1 = 1$ і врахувавши (11), приходимо до твердження теореми 4.

Наслідок 6. Коли $b_i = a_{i+1}$, тобто оператор самоспряжений (див. п. 1.4 з ч. 1) і $a_i \geqslant c_0 > 0$, дістанемо таку оцінку розв'язку задачі (8):

$$\|y\|_C \leqslant \frac{2}{c_0} \sum_{k=1}^{N} \left| \sum_{j=1}^{k-1} \varphi_j \right|. \tag{14}$$

Якщо $\varphi_i = \Delta\mu_i$, то

$$\|y\|_C \leqslant \frac{2}{c_0} \sum_{k=1}^{N} |\mu_k|. \tag{14'}$$

Наслідок 7. За умов (7) та $a_i \geqslant c_0 > 0$ для розв'язку задачі $Ly_i = -\varphi_i$, $y_0 = \mu_1$, $y_N = \mu_2$ маємо

$$\|y\|_C \leqslant \max(|\mu_1|, |\mu_2|) + \frac{2}{c_0} \sum_{k=1}^{N} \left| \sum_{j=1}^{k-1} \varphi_j \right|.$$

Це випливає з наслідків 4 та 6.

2.2.2. Принцип максимуму за стовпчиками (метод відшукання апріорної оцінки в нормі $L_1(\omega_h)$). Розглянемо задачу (8) і припустимо, що

$$a_i > 0, \ b_i > 0, \quad i = \overline{1, N-1}, \quad d_i = c_i - a_{i+1} - b_{i-1} \geqslant 0,$$
$$i = \overline{2, N-2}, \quad d_1 = c_1 - a_2 \geqslant 0, \quad d_{N-1} = c_{N-1} - b_{N-2} \geqslant 0, \tag{15}$$

причому існує i^* таке, що $d_{i^*} > 0$.

Теорема 5. *Якщо виконуються умови* (15) *і* $Ly_i \leqslant 0$ $(Ly_i \geqslant 0)$, $i = \overline{1, N-1}$, *то для розв'язку задачі* (8) *має місце нерівність* $y_i \geqslant 0$ $(y_i \leqslant 0)$, $i = \overline{1, N-1}$.

Доведення. Нехай, наприклад, $Ly_i \leqslant 0$, $i = \overline{1, N-1}$. Припустимо від супротивного, що існують i_1, i_2 такі, що $0 < i_1 \leqslant i_2 < N$ і $y_i < 0$ для $i = \overline{i_1, i_2}$. Розглянемо суму

$$\sum_{i=i_1}^{i_2} Ly_i = \sum_{i=i_1}^{i_2} (a_i y_{i-1} - c_i y_i + b_i y_{i+1}) = -\sum_{i=i_1}^{i_2} d_i y_i +$$

$$+ \sum_{i=i_1}^{i_2} (a_i y_{i-1} - a_{i+1} y_i - b_{i-1} y_i + b_i y_{i+1}) =$$

$$= -\sum_{i=i_1}^{i_2} d_i y_i + a_{i_1} y_{i_1-1} - a_{i_2+1} y_{i_2} - b_{i_1-1} y_{i_1} + b_{i_2} y_{i_2+1}.$$

Неважко помітити, що $-a_{i_2+1}y_{i_2} - b_{i_1-1}y_{i_1} > 0$, а інші складові невід'ємні, тому $\sum_{i=i_1}^{i_2} Ly_i \geqslant 0$ всупереч припущенню. Залишається розглянути випадок $i_1 = 1$, $i_2 = N-1$:

$$\sum_{i=1}^{N-1} Ly_i = -\sum_{i=1}^{N-1} d_i y_i \geqslant -d_{i^*} y_{i^*} > 0.$$

Знову дійшли суперечності, що і доводить теорему.

Наслідок 8. *Нехай для задачі (8) виконуються умови (15), крім того,* $Ly_i \equiv 0$, $i = \overline{1, N-1}$. *Тоді* $y_i = 0$, $i = \overline{1, N-1}$.

Зазначимо, що від умови існування такого i^*, що $d_{i^*} > 0$, можна відмовитися, якщо розглядати розв'язки $y_i \not\equiv 0$.

Наслідок 9. *Нехай для неоднорідної задачі*

$$Ly_i = -\varphi_i, \quad y_0 = \mu_0, \quad y_N = \mu_1 \tag{16}$$

виконуються умови (15). Тоді якщо $Ly_i \leqslant 0$, $i = \overline{1, N-1}$ і $\mu_0 \geqslant 0$, $\mu_N \geqslant 0$, то $y_i \geqslant 0$, $i = \overline{0, N}$.

Доведення цього твердження залишаємо читачеві.

Розглянемо дві неоднорідні задачі

$$Ly_i = -\varphi_i, \quad i = \overline{1, N-1}, \quad y_0 = \mu_0, \quad y_N = \mu_N; \tag{17}$$

$$L\bar{y}_i = -\bar{\varphi}_i, \quad i = \overline{1, N-1}, \quad \bar{y}_0 = \bar{\mu}_0, \quad \bar{y}_N = \bar{\mu}_N, \tag{18}$$

причому

$$|\varphi_i| \leqslant |\bar{\varphi}_i|, \quad |\mu_0| \leqslant |\bar{\mu}_0|, \quad |\mu_1| \leqslant |\bar{\mu}_1|, \quad i = \overline{1, N-1}. \tag{19}$$

Теорема 6 *(теорема порівняння). Нехай виконуються умови (15), (19). Тоді для розв'язків задач (17), (18) мають місце оцінки*

$$\|y\|_{c(\omega)} \leqslant \|\bar{y}\|_{c(\omega)}, \quad \|y\|_{L_1(\omega)} \leqslant \|\bar{y}\|_{L_1(\omega)},$$

де

$$\|y\|_{c(\omega)} = \max_{x \in \omega} |y(x)|, \quad \|y\|_{L_1(\omega)} = \sum_{x \in \omega} h\,|y(x)|, \quad h > 0.$$

Доведення. Розглянемо допоміжні функції

$$u(x) = \bar{y}(x) - y(x), \quad x \in \bar{\omega},$$

$$v(x) = \bar{y}(x) + y(x), \quad x \in \bar{\omega}.$$

Вони є розв'язками таких задач:

$$Lu_i = -(\bar{\varphi}_i - \varphi_i) \leqslant 0, \quad u_0 = \bar{\mu}_0 - \mu_0 \geqslant 0, \quad u_N = \bar{\mu}_1 - \mu_1 \geqslant 0;$$

$$Lv_i = -(\bar{\varphi}_i + \varphi_i) \leqslant 0, \quad v_0 = \bar{\mu}_0 + \mu_0 \geqslant 0, \quad v_N = \bar{\mu}_1 + \mu_1 > 0.$$

З наслідку 9 випливає, що $u_i \geqslant 0$, $v_i \geqslant 0$, $i = \overline{0, N}$, або

$\bar{y}_i - y_i \geqslant 0$, $\bar{y}_i + y_i \geqslant 0$, $i = \overline{0, N}$, тобто $|y_i| \leqslant |\bar{y}_i|$, що і доводить теорему 6.

Теорема 7. *Нехай для задачі (8) виконуються умови*

$$\bar{d}_i = |c_i| - |a_{i+1}| - |b_{i-1}| > 0, \quad i = 2, N-2,$$

$$\bar{d}_1 = |c_1| - |a_2| > 0, \quad \bar{d}_{N-1} = |c_{N-1}| - |b_{N-2}| > 0.$$

Тоді справедлива оцінка

$$\|y\|_{L_1(\omega)} \leqslant \left\| \frac{\varphi}{\bar{d}} \right\|_{L_1(\omega)}. \tag{20}$$

Доведення. Безпосередньо з рівняння випливає

$$|c_i|\,|y_i| \leqslant |a_i|\,|y_{i-1}| + |b_i|\,|y_{i+1}| + |\varphi_i|, \quad i = \overline{1, N-1}.$$

Підсумуємо на сітці ω з вагою $h/\bar{d}$ (x) обидві частини цієї нерівності:

$$\sum_{i=1}^{N-1} \frac{h}{\bar{d}_i} |c_i|\,|y_i| \leqslant \sum_{i=1}^{N-1} \frac{h}{\bar{d}_i} (|a_i|\,|y_{i-1}| + |b_i|\,|y_{i+1}| + |\varphi_i|).$$

Враховуючи умови теореми, маємо

$$\sum_{i=1}^{N-1} hy_i + \sum_{i=2}^{N-2} \frac{h}{\bar{d}_i} (|a_{i+1}|\,|y_i| + |b_{i-1}|\,|y_i|) + \frac{h}{\bar{d}_1} |a_2|\,|y_1| +$$

$$+ \frac{h}{\bar{d}_{N-1}} |b_{N-2}|\,|y_{N-1}| \leqslant \sum_{i=1}^{N-1} \frac{h}{\bar{d}_i} (|a_i|\,|y_{i-1}| +$$

$$+ |b_i|\,|y_{i+1}|) + \sum_{i=1}^{N-1} \frac{h}{\bar{d}_i} |\varphi_i|.$$

Оскільки з крайових умов випливає, що $y_0 = y_N = 0$, то звідси і дістанемо оцінку (20).

Розглянемо сіткову схему

$$Ly_i \equiv a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -\varphi_i, \quad i = \overline{0, N}, \tag{21}$$

яка виникає, наприклад, при різницевій апроксимації задачі з виродженням

$$(1 - x^2) u''(x) + p(x) u'(x) + q(x) u(x) = f(x),$$

$x \in (-1, 1)$, $u(-1) \neq \infty$, $u(+1) \neq \infty$.

Припустимо, що коефіцієнти a_i, c_i, b_i, $i = \overline{0, N}$, задовольняють умови

$$a_i > 0, \quad i = \overline{1, N}, \quad b_i > 0, \quad i = \overline{0, N-1}; \quad d_i = c_i - a_{i+1} - b_{i-1} \geqslant 0,$$

$$i = \overline{1, N-1}, \quad d_0 = c_0 - a_1 \geqslant 0, \quad c_N - b_{N-1} = d_N \geqslant 0; \tag{22}$$

$$\exists\, i \quad d_{i^*} > 0.$$

Теорема 8. *Нехай для задачі* (21) *виконуються умови* (22). *Тоді, якщо* $Ly_i \leqslant 0$ $(Ly_i \geqslant 0)$, $i = \overline{0, N}$, *то* $y_i \geqslant 0$ $(y_i \leqslant 0)$, $i = \overline{0, N}$.

Доведення цієї теореми аналогічне доведенню теореми 5. Необхідно лише розглянути випадки, коли $i_1 = 0$ та $i_2 = N$. Нехай, наприклад, $i_1 = 0$, $i_2 = N$, $Ly_i \leqslant 0$. Тоді

$$\sum_{i=0}^{N} Ly_i = Ly_0 + Ly_N + \sum_{i=1}^{N-1} (a_i y_{i-1} + b_i y_{i+1} - c_i y_i) =$$

$$= -c_0 y_0 + b_0 y_1 + a_N y_{N-1} - c_N y_N + \sum_{i=1}^{N-1} (a_i y_{i-1} + b_i y_{i+1} - a_{i+1} y_i -$$

$$- b_{i-1} y_i) - \sum_{i=1}^{N-1} d_i y_i = -d_0 y_0 - a_1 y_0 + b_0 y_1 + a_N y_{N-1} -$$

$$- d_N y_N - b_{N-1} y_N + a_1 y_0 + b_{N-1} y_N - a_N y_{N-1} - b_0 y_1 -$$

$$- \sum_{i=1}^{N-1} d_i y_i = -\sum_{i=0}^{N} d_i y_i \geqslant d_{i^*} y_{i^*} > 0,$$

тобто дійшли суперечності.

Пропонуємо читачеві самостійно переконатися у справедливості таких тверджень.

1. *Нехай для коефіцієнтів рівняння* (21) *виконуються умови* (22). *Тоді з того, що* $Ly_i \equiv 0$, $i = \overline{0, N}$, виконуються умови $y_i = 0$, $i = \overline{0, N}$ *(наслідок теореми 8).*

2. *Теорема порівняння.* *Якщо* $L\bar{y}_i = -\bar{\varphi}_i$, $i = \overline{0, N}$, $a_0 = b_N = 0$, $|\varphi_i| \leqslant \bar{\varphi}_i$, $i = \overline{0, N}$ *і виконуються умови* (22), *то для розв'язку задачі* (21) *справедливі оцінки*

$$\|y\|_{c(\bar{\omega})} \leqslant \|\bar{y}\|_{c(\bar{\omega})}, \quad \|y\|_{L_1(\omega)} \leqslant \|\bar{y}\|_{c(\bar{\omega})} \tag{23}$$

(наслідок теореми 8).

3. *Нехай для задачі* (21) *виконуються умови* $\bar{d}_i = |c_i| - |a_{i+1}| - |b_{i-1}| > 0$, $i = \overline{1, N-1}$, $\bar{d}_0 = |c_0 - a_1| > 0$, $\bar{d}_N = |c_N| - |b_{n-1}| = 0$. *Тоді має місце апріорна оцінка*

$$\|y\|_{L_1(\bar{\omega})} \leqslant \left\| \frac{\varphi}{\bar{d}} \right\|_{L_1(\bar{\omega})} \tag{24}$$

(доведення аналогічне доведенню теореми 7).

2.2.3. Оцінка оберненого до додатно визначеного оператора у просторі зі скалярним добутком (метод відшукання апріорної оцінки в нормі, що породжується скалярним добутком). Така оцінка вже фактично розглянута в п. 1. 4 з ч.1. Вона стосується випадку, коли оператор A_h діє у деякому просторі сіткових функцій H_h, заданих на

сітці $\bar{\omega}_h$, зі скалярним добутком $(,)$ і відомо, що

$$(A_h y, y) \geqslant \varkappa \|y\|^2, \tag{25}$$

де $\varkappa$ — стала, що не залежить від h, $\|y\| = \sqrt{(y, y)}$. Тоді

$$\|A_h^{-1}\| \equiv \sup_{\|v\| \neq 0} \frac{|(A_h^{-1} v, A_h^{-1} v)|^{1/2}}{\|v\|} = \sup_{z \neq 0} \frac{|(z, z)|^{1/2}}{\|A_h z\|} \leqslant \varkappa^{-1}. \tag{26}$$

Запишемо доведення цієї оцінки:

$$\varkappa^2 \|z\|^4 = \varkappa^2 (z, z)^2 \leqslant (A_h z, z)^2 \leqslant (A_h z, A_h z)(z, z) = \|A_h z\|^2 \|z\|^2.$$

Приклад 1. Переконатися, що виконується нерівність коректності для сіткової схеми

$$A_h y = f,$$

де оператор A_h діє у просторі сіткових функцій $H = H(\bar{\omega}_h)$, заданих на сітці $\bar{\omega}_h = \{x_i = a + ih : i = \overline{0, N}, h = (b-a)/N\}$ зі скалярним добутком

$$(y, v) = \sum_{i=0}^{N} h y_i v_i,$$

причому A_h діє за правилом

$$(A_h v)_i = \begin{cases} -\frac{1}{h} v_{x,0} + \frac{1}{2} g_0 v_0 + \frac{1}{h} \alpha_0 v_0, & i = 0, \\ -v_{\bar{x}x,i} + g_i v_i, & i = \overline{1, N-1}, \\ \frac{1}{h} v_{\bar{x},N} + \frac{1}{2} g_N v_N + \frac{1}{h} \alpha_1 v_N, & i = N, \end{cases}$$

де g_i — задана функція; α_0, α_1 — задані сталі; $\alpha_0 > 0$, $\alpha_1 > 0$, $g(x) \geqslant 0$.

Р о з в' я з а н н я. Доведемо, що оператор A_h додатно визначений. Дійсно,

$$(A_h v, v) = -\sum_{i=1}^{N-1} h v_{\bar{x}x,i} v_i + \sum_{i=1}^{N-1} h g_i v_i^2 - v_{x,0} v_0 + \frac{h}{2} g_0 v_0^2 +$$

$$+ \alpha_0 v_0^2 + v_{\bar{x},N} v_N + \frac{h}{2} g_N v_N^2 + \alpha_1 v_N^2.$$

Після застосування першої різницевої формули Гріна дістанемо

$$(A_h v, v) = (v_{\bar{x}}, v_{\bar{x}})_{\omega_h^+} + \alpha_0 v_0^2 + \alpha_1 v_N^2 + (gv, v)_{\omega_h} + \frac{h}{2} g_N v_N^2 +$$

$$+ \frac{h}{2} g_0 v_0^2 \geqslant (v_{\bar{x}}, v_{\bar{x}})_{\omega_h^+} + \alpha_0 v_0^2 + \alpha_1 v_N^2,$$

де

$$\omega_h^+ = \{x_i = a + ih : i = \overline{1, N}, \ h = (b-a)/N\}, \quad \omega_h = \bar{\omega}_h \setminus \{a, b\},$$

$$(y, v)_M = \sum_{x \in M} h y(x) v(x).$$

Оцінимо величину $(v_{\bar{x}}, v_{\bar{x}})_{\omega_h^+}$. Для цього запишемо тотожність

$$v_i^2 = \left(v_0 + \sum_{j=1}^{i} h v_{\bar{x},j}\right)^2 = \left(v_N - \sum_{j=i+1}^{N} h v_{\bar{x},j}\right)^2$$

і скористаємося нерівністю $(p \pm s)^2 \leqslant 2(p^2 + s^2)$. Маємо

$$v_i^2 \leqslant 2\left[v_0^2 + \left(\sum_{j=1}^{i} h v_{\bar{x},j}\right)^2\right],$$

$$v_i^2 \leqslant 2\left[v_N^2 + \left(\sum_{j=i+1}^{N} h v_{\bar{x},j}\right)^2\right],$$

звідки, застосовуючи нерівність Коші — Буняковського, матимемо

$$v_i^2 \leqslant 2\left[v_0^2 + \left(\sum_{j=1}^{i} h\right)\left(\sum_{j=1}^{i} v_{\bar{x},j}^2\right)\right] \leqslant 2\,[v_0^2 + (b-a)\,(v_{\bar{x}}, v_{\bar{x}})_{\omega_h^+}],$$

$$v_i^2 \leqslant 2\,[v_N^2 + (b-a)\,(v_{\bar{x}}, v_{\bar{x}})_{\omega_h^+}],$$

або

$$v_i^2 \leqslant v_0^2 + v_N^2 + 2(b-a)\,(v_{\bar{x}}, v_{\bar{x}})_{\omega_h^+}.$$

З цієї нерівності знаходимо

$$\|v\|^2 = (v, v) = \sum_{i=0}^{N} h v_i^2 \leqslant [v_0^2 + v_N^2 + 2(b-a)\,(v_{\bar{x}}, v_{\bar{x}})_{\omega_h^+}]\,(b-a+h) \leqslant$$

$$\leqslant 2(b-a)\left[\frac{\alpha_0 v_0^2}{\alpha_0} + \frac{\alpha_1 v_N^2}{\alpha_1} + 2(b-a)\,(v_{\bar{x}}, v_{\bar{x}})_{\omega_h^+}\right] \leqslant$$

$$\leqslant \varkappa\,[\alpha_0 v_0^2 + \alpha_1 v_N^2 + (v_{\bar{x}}, v_{\bar{x}})_{\omega_h^+}] \leqslant \varkappa\,(A_h v, v),$$

де $\varkappa = 2(b-a)\max\{\alpha_0^{-1}, \alpha_1^{-1}, 2(b-a)\}$. З рівняння (26) дістанемо шукану нерівність

$$\|y\| = \sqrt{(y, y)} = \|A_h^{-1} f\| \leqslant \|A_h^{-1}\|\,\|f\| \leqslant \varkappa^{-1}\|f\|.$$

2.2.4. Метод енергетичних оцінок (метод відшукання апріорних оцінок в енергетичній нормі). Цей метод застосовується для доведення коректності сіткових схем

$$A_h y = f$$

в додатно визначеним оператором A_h, $(A_h v, v) \geqslant \varkappa \|v\|^2$, у деякому просторі сіткових функцій $H(\bar{\omega}_h)$, заданих на сітці $\bar{\omega}_h$, і скалярним добутком $(,)$. У цьому просторі можна ввести так звану енергетичну норму (перевірте аксіоми норми)

$$\|v\|_A = \sqrt{(A_h v, v)}. \tag{27}$$

Для певних операторів A_h енергетична норма може бути еквівалентна деяким сітковим нормам. Нехай для визначеності

$$c_1 \|v\|_k^2 \leqslant \|v\|_A^2 \leqslant c_2 \|v\|_k^2, \tag{28}$$

де $\|v\|_k = \|v\|_{W_2^k(\omega_h)}$, $k \geqslant 0$; c_1, c_2 — сталі, що не залежать від v. Помножимо обидві частини рівняння $A_h y = f$ скалярно на y і застосуємо в лівій частині нерівність еквівалентності норм, а в правій частині — нерівність Коші — Буняковського. В результаті

$$c_1 \|y\|_k^2 \leqslant \|f\|\,\|y\|.$$

Якщо справджується вкладення $\overset{\circ}{w}{}_2^k(\omega_h) \subset H(\bar{\omega}_h)$, тобто $\forall v \in \overset{\circ}{w}{}_2^k(\omega_h)$, $\|v\| \leqslant c_3 \|v\|_k$, де c_3 — стала, що не залежить від v, то, скоротивши на $\|y\|_k$, дістанемо нерівність коректності сіткової схеми

$$\|y\|_k \leqslant c_1^{-1} c_3 \|f\|. \tag{29}$$

Приклад 2. Нехай $A_h = I$, де I — одиничний оператор. Тоді за методом енергетичних оцінок для рівняння $Iy = f$ маємо

$$\|y\|_k = (f, y) \leqslant \|f\|\,\|y\|,$$

$$\|y\| \leqslant \|f\|.$$

Приклад 3. Методом енергетичних оцінок знайти нерівність коректності для схеми

$$A_h y = f. \tag{30}$$

У просторі $\overset{\circ}{H}(\bar{\omega}_h)$, $\bar{\omega}_h = \{x_i = ih : i = \overline{0, N},\ h = 1/N\}$, сіткових функцій, що дорівнюють нулю в точках x_0, x_N, зі скалярним добутком

$$(y, v) = \sum_{i=0}^{N} h y_i v_i = \sum_{x \in \omega_h} h y(x)\, v(x),$$

де оператор A_h задається таким чином:

$$(A_h v)_i = \begin{cases} 0, & i = 0,\ N, \\ -(a v_{\bar{x}})_{x,i} + q_i v_i, & i = \overline{1, N-1}, \end{cases} \tag{31}$$

$$0 < c_1 \leqslant a_i \leqslant c_2, \quad 0 \leqslant q_i \leqslant c_3.$$

Розв'язання. За допомогою формул підсумовування частинами неважко дістати

$$(A_h v, v) = -\sum_{i=1}^{N-1} h\,(a v_{\bar{x}})_{x,i} v_i + \sum_{i=1}^{N-1} h q_i v_i^2 = \sum_{i=1}^{N} h a_i v_{\bar{x},i}^2 + \sum_{i=1}^{N-1} h q_i v_i^2.$$

Враховуючи, що для функції $v \in \overset{\circ}{H}(\bar{\omega}_h)$

$$v(x) = \sum_{t=h}^{x} h v_{\bar{x}}(t),$$

$$|v(x)| \leqslant \sum_{t=h}^{x} h\,|v_{\bar{x}}(t)| \leqslant \sqrt{x}\sqrt{\sum_{t=h}^{x} h v_{\bar{x}}^2(t)} \leqslant \sqrt{\sum_{t=h}^{1} h v_{\bar{x}}^2(t)},$$

маємо

$$\frac{c_1}{2}\|v\|_1^2 \equiv \frac{c_1}{2}\sum_{i=1}^{N-1} hv_i^2 + \frac{c_1}{2}\sum_{i=1}^{N} hv_{\overline{x},i}^2 \leqslant c_1 \sum_{i=1}^{N} hv_{\overline{x},i}^2 \leqslant \sum_{i=1}^{N} ha_i v_{\overline{x},i}^2 \leqslant$$

$$\leqslant \sum_{i=1}^{N} ha_i v_{\overline{x},i}^2 + \sum_{i=1}^{N-1} hq_i v_i^2 = (A_h v,\ v) \leqslant c_2 \sum_{i=1}^{N} hv_{\overline{x},i}^2 + c_3 \sum_{i=1}^{N-1} hv_i^2 \leqslant$$

$$\leqslant \max(c_2,\ c_3)\|v\|_1^2,$$

тобто енергетична норма оператора A_h еквівалентна нормі $\|\ \|_1$ простору $\overset{\circ}{\mathcal{U}}_1^2(\omega_h)$. Помножимо рівняння $A_h y = f$ скалярно на y і застосуємо у правій частині нерівність Коші — Буняковського, а в лівій — щойно утворену нерівність

$$\frac{c_1}{2}\|y\|_1^2 \leqslant (A_h y,\ y) = (f,\ y) \leqslant \|f\|\,\|y\|.$$

З теореми вкладення (див. п. 1.4 з ч. 1) маємо

$$\|y\| \equiv \|y\|_{L_2(\omega_h)} \leqslant M\|y\|_1,$$

де стала M не залежить від h і y. Тому

$$\frac{c_1}{2}\|y\|_1^2 \leqslant \|f\|\,M\|y\|_1,$$

звідки дістанемо шукану нерівність

$$\|y\|_1 \leqslant c\|f\|, \tag{32}$$

де стала $c = 2c_1^{-1}M$ не залежить від h, y і f.

2.2.5. Сітковий метод Фур'є і його застосування для дослідження стійкості сіткових схем. Нехай E_h — сукупність дійсних сіткових функцій $y(x)$, заданих в точках сітки $\bar{\omega}_h = \{x_i = ih \mid i = \overline{0, N},\ h = l/N\}$. Ці функції задовольняють умови

$$y(0) = y(l) = 0. \tag{33}$$

Введемо у цьому просторі звичайні операції додавання функцій та множення функції на число, а також скалярний добуток

$$(y, v) = \sum_{x\in\omega_h} hy(x)\,v(x), \tag{34}$$

де $\omega_h = \bar{\omega}_h \setminus \{0, l\}$. Такий простір із скалярним добутком розмірності $N-1$ позначимо через R_{N-1}. Якщо система функцій $\omega_k(jh)$ утворює ортонормований базис в R_{N-1}, то довільну функцію можна розкласти за цим базисом

$$v(x) = \sum_{k=1}^{N-1} c_k \omega_k(x), \tag{35}$$

де

$$c_k = (v, \omega_k) = \sum_{x\in\omega_h} hv(x)\,\omega_k(x), \quad (\omega_k, \omega_j) = \delta_{ij},$$

δ_{ij} — символ Кронекера. Послідовність c_k, $k = \overline{1, N-1}$, можна розглядати як деякий образ функції $v(x)$, $x \in \omega_h$, який називають *спектральним образом*. Якщо c_k, b_k — спектральні образи функцій v, u, то

$$(v, u) = \sum_{x\in\omega_h} hv(x)\,u(x) = \sum_{k=1}^{N-1} c_k b_k,$$

звідки

$$(v, v) = \sum_{k=1}^{N-1} c_k^2, \tag{36}$$

що є сітковим аналогом рівності Парсеваля.

Прикладом ортонормованого базису в просторі R_{N-1} є система функцій (див. п. 1.4 з ч. 1)

$$\omega_k(x) = \sqrt{\frac{2}{l}}\sin\frac{k\pi x}{l}, \quad x = jh; \quad j = \overline{1, N-1}; \quad h = \frac{l}{N}. \tag{37}$$

Довільну функцію v, яка задана на $\bar{\omega}_h$ і задовольняє умови (33), можна подати у вигляді суми Фур'є

$$v(x) = \sqrt{\frac{2}{l}}\sum_{k=1}^{N-1} c_k \sin\frac{k\pi x}{l},$$

$$c_k = \sqrt{\frac{2}{l}}\sum_{x\in\omega_h} hv(x)\sin\frac{k\pi x}{l}, \quad k = \overline{1, N-1}. \tag{38}$$

Зазначимо (див. п.1.4 з ч. 1), що система функцій (37) є системою власних функцій різницевої задачі

$$v_{\overline{x}x} = -\lambda v, \quad v(0) = v(l) = 0, \tag{39}$$

причому функції $\omega_k(x)$ відповідає власне значення

$$\lambda_k = \frac{4}{h^2}\sin^2\frac{k\pi h}{2l}, \quad k = \overline{1, N-1}. \tag{40}$$

Якщо кожну з функцій v, заданих на $\bar{\omega}_h$, причому $v(0) = v(l) = 0$, доозначити періодично для всіх точок jh, $j = 0, \pm1, \pm2, \ldots$, осі Ox, то дістанемо деяку l-періодичну сіткову функцію і сума Фур'є (38) буде аналогом нескінченного ряду Фур'є для непарної функції v.

Аналогічно на сітці $\bar{\omega}_h$ можна розглянути сукупність всіх комплекснозначних функцій v і кожну з них доозначити на всі точки сітки $x_j = jh$, $j = 0, \pm1, \pm2, \ldots$, таким чином, щоб утворилася l-періодична сіткова функція. Сукупність таких функцій позначимо через M_h і введемо в M_h операції додавання, множення на число та

скалярного добутку

$$(v, u) = \sum_{x \in \omega_h} h v(x) \overline{u}(x).$$

Прикладом системи лінійно незалежних l-періодичних ортогональних в M_h функцій є

$$\hat{\omega}_k(x) = e^{ik\frac{2\pi}{l}x}, \quad i = \sqrt{-1}, \quad x \in \omega_h, \tag{41}$$

$$k = 0, \pm 1, \pm 2, \ldots, \pm \frac{N-1}{2}.$$

Дійсно,

$$(\hat{\omega}_k, \hat{\omega}_m) = \sum_{x \in \omega_h} h e^{i\frac{2\pi}{l}(k-m)x} = \sum_{j=0}^{N-1} h e^{i\frac{2\pi}{l}(k-m)jh} = Nh\delta_{km} = l\delta_{km}.$$

Отже, довільну функцію з M_h можна розвинути в «суму Фур'є»

$$v(x) = \sum_{k=-\frac{N-1}{2}}^{\frac{N-1}{2}} A_k \hat{\omega}_k(x),$$

$$A_k = \frac{1}{l}(v, \hat{\omega}_k). \tag{42}$$

Функції $\hat{\omega}_k(x) = e^{ik\frac{2\pi}{l}x}$, $k = 0, \pm 1, \ldots, \pm \frac{N-1}{2}$, є власними функціями задачі

$$\frac{i}{2}(\hat{\omega}_{\bar{x}} + \hat{\omega}_x) = \lambda \hat{\omega} \tag{43}$$

і відповідають власним значенням

$$\lambda_k = \frac{N}{l} \sin \frac{2\pi k}{N}. \tag{44}$$

Функції $\hat{\omega}_k(x)$, $k = 0, \pm 1, \ldots, \pm \frac{N-1}{2}$, є також власними функціями операторів $(\cdot)_{\bar{x}}$ та $(\cdot)_x$ на M_h, причому

$$(\hat{\omega}_k(x))_x = \frac{e^{ik\frac{2\pi}{l}(l+1)h} - e^{ik\frac{2\pi}{l}jh}}{h} = e^{ik\frac{2\pi}{l}x} \frac{(e^{ik\frac{2\pi}{l}h} - 1)}{h} =$$

$$= \tilde{\lambda}_k(h)\, \hat{\omega}_k(x),$$

$$(\hat{\omega}_k(x))_{\bar{x}} = -\tilde{\lambda}_k(h)\, \hat{\omega}_k(x),$$

$$\tilde{\lambda}_k(h) = \frac{e^{ik\frac{2\pi}{l}h} - 1}{h}.$$

Тому функції $\hat{\omega}_k(x) = e^{ik\frac{2\pi}{l}x}$, $k = 0, \pm 1, \pm 2, \ldots, \pm \frac{N-1}{2}$, утворюють повну систему власних функцій для будь-якого оператора A_h виду

$$A_h v = \sum_{s,q} b_{sq} D^s D^{-q} v,$$

де $D^s v = v_{\underbrace{xx\ldots x}_{s}}$, $D^{-q} v = v_{\underbrace{\bar{x}\bar{x}\ldots\bar{x}}_{q}}$, b_{sq} — сталі коефіцієнти.

Це дає змогу питання дослідження стійкості сіткових схем пов'язати з вивченням спектра різницевих операторів, за власними функціями яких можна розвинути в суму Фур'є шукані розв'язки сіткової схеми.

Приклад 4. Знайти оцінку розв'язку різницевої схеми

$$y_t^j = y_{xx}^j, \quad x \in \omega_h, \quad j = 0, 1, \ldots,$$

$$y^j(0) = y^j(l) = 0, \tag{45}$$

$$y^0(x) = \gamma(x), \quad x \in \bar{\omega}_h,$$

де $\omega_h = \left\{x_i = ih : i = \overline{1, N-1}, \ h = \frac{l}{N}\right\}$, $\bar{\omega}_h = \omega_h \cup \{0, l\}$,

$$y^j(x) = y(x, t_j) = y(x, j\tau), \quad y_t^j = (y^{j+1}(x) - y^j(x))/\tau,$$

$$y_{\bar{x}x}^j(x) = (y^j(x-h) - 2y^j(x) + \overline{y}^j(x-h))/h^2$$

(ця схема апроксимує таку задачу для рівняння теплопровідності:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad u(0) = u(l) = 0, \quad y(x, 0) = \gamma(x), \quad x \in [0, l]).$$

Розв'язання. Шукаємо розв'язок задачі (45) у вигляді

$$y_i^j = \sum_{k=1}^{N-1} c_k^j \omega_k(ih), \tag{46}$$

де $\omega_k(x)$ визначаються формулою (37), причому c_k^j визначимо так, щоб (46) мало місце і для початкових умов, тобто $c_k^0 = (\gamma, \omega_k)$. Підставляючи (46) в (45) і враховуючи (39), (40), дістаємо

$$c_k^{j+1} = q_k c_k^j, \quad q_k = 1 - \tau\lambda_k, \quad \lambda_k = \frac{4}{h^2} \sin^2 \frac{\pi k h}{2l}.$$

Якщо $|q_k| \leqslant 1$, що має місце при $\frac{\tau}{h^2} \leqslant \frac{1}{2}$, то

$$\|y^{j+1}\|^2 = (y^{j+1}, y^{j+1}) = \sum_{x \in \omega_h} h (y^{j+1}(x))^2 = \sum_{k=1}^{N-1} (c_k^{j+1})^2 \leqslant \sum_{k=1}^{N-1} (c_k^j)^2 = \|y^j\|^2.$$

Остання нерівність означає, що схема (45) умовно стійка щодо початкових умов у сітковій нормі L_2 (див. також п. 1.4 з ч. 1), якщо $\frac{\tau}{h^2} \leqslant \frac{1}{2}$. Метод Фур'є можна застосувати також для дослідження стійкості двох'ярусних явних схем вигляду

$$y_t^j + Py^j = F^j, \quad j = 0, 1, \ldots, \tag{47}$$
$$y^0 = u_0,$$

якщо $P = P^*$, $P : R_{N-1} \to R_{N-1}$ і відомий відрізок $[\alpha, \beta]$, на якому лежать власні значення λ_k оператора P. Нехай $\omega_k(x)$ — ортонормована власна функція цього оператора, що відповідає λ_k. Підставляючи

$$y^j = \sum_k c_k^j \omega_k(x), \quad F^j = \sum_k b_k^j \omega_k(x),$$
$$y^0 = \sum_k c_k^0 \omega_k(x)$$

в (47), дістанемо для коефіцієнтів c_k^j, b_k^j співвідношення

$$\frac{c_k^{j+1} - c_k^j}{\tau} + \lambda_k c_k^j = b_k^j, \quad k = \overline{1, N-1}, \quad j \geqslant 0,$$

або

$$c_k^{j+1} = (1 - \tau\lambda_k)\, c_k^j + \tau b_k^j.$$

Позначимо $1 - \tau\lambda_k$ через q_k. Якщо $|q_k| = |1 - \tau\lambda_k| \leqslant 1$, то

$$\|y^{j+1}\| = \left\| \sum_k c_k^{j+1} \omega_k(x) \right\| = \left(\sum_k (c_k^{j+1})^2 \right)^{1/2} = \left(\sum_k (q_k c_k^j + \tau b_k^j)^2 \right)^{1/2} \leqslant$$
$$\leqslant \left(\sum_k (c_k^j)^2 \right)^{1/2} + \tau \left(\sum_k (b_k^j)^2 \right)^{1/2} = \|y^j\| + \tau \|F^j\|.$$

Оцінка $|q_k| \leqslant 1$ справедлива, якщо $\alpha \geqslant 0$, $0 < \tau\beta \leqslant 2$, і тому за цієї умови справедлива оцінка

$$\|y^{j+1}\| \leqslant \|y^0\| + \sum_{k=0}^{j} \tau \|F^k\|.$$

Сітковий метод Фур'є можна застосувати для дослідження стійкості не лише двох'ярусних, а й багатоярусних схем.

Оцінки стійкості багатоярусних схем дає також теорія, викладена в 1.4.5 з ч. 1.

2.2.6. **Принцип максимуму в багатовимірному випадку.** Нехай ω — скінченна множина вузлів (сітка) у деякій обмеженій області n-вимірного евклідового простору, $p \in \omega$ — точка сітки ω. Розглянемо рівняння відносно невідомої сіткової функції $y(x)$, заданої на ω:

$$A(p)\, y(p) = \sum_{Q \in Ш'(p)} B(p, Q)\, y(Q) + F(p), \quad p \in \omega, \tag{48}$$

де $A(p)$, $B(p, Q)$ (коефіцієнти рівняння), $F(p)$ (права частина рівняння) — задані сіткові функції. Множина точок сітки, які розглядаються у рівнянні (48), називається *шаблоном*. Він складається із самого вузла p і множини точок $Ш'(p)$, яка не містить p і називається *околом точки* p. Далі вважатимемо, що виконуються умови

$$A(p) > 0, \quad B(p, Q) > 0 \quad \forall p \in \omega, \quad Q \in Ш'(p), \tag{49}$$
$$D(p) = A(p) - \sum_{Q \in Ш'(p)} B(p, Q) \geqslant 0.$$

Точка p називається *граничним вузлом* сітки ω, якщо в ній задано значення функції $y(p)$, тобто

$$y(p) = \mu(p), \quad p \in \gamma, \tag{50}$$

де γ — множина граничних вузлів, μ — задана функція. Порівнюючи (50) з (48), помічаємо, що на границі треба формально покласти

$$A(p) \equiv 1, \quad B(p, Q) \equiv 0, \quad F(p) = \mu(p).$$

Вузли, в яких виконується рівняння (48) за умов (49), називатимемо *внутрішніми вузлами* сітки, ω — множина внутрішніх вузлів, а $\bar{\omega} = \omega \cup \gamma$ — множина всіх вузлів сітки. Можна також розглядати випадок, коли граничних вузлів немає, тобто немає рівнянь виду (50) і $\bar{\omega} = \omega$. Задача виду (48), (50) називається *першою крайовою (граничною) задачею*.

Вважаємо, що сітка $\bar{\omega}$ зв'язна, якщо для будь-яких заданих точок $\bar{p} \in \omega$ і $\bar{\bar{p}} \in \omega$ існує така послідовність околів $Ш'(p)$, що можна здійснити перехід від $\bar{p}$ до $\bar{\bar{p}}$, використовуючи лише вузли цих околів. Іншими словами, знайдуться такі вузли $p_1, p_2, \ldots, p_m$ сітки ω, що

$$p_1 \in Ш'(\bar{p}), \quad p_2 \in Ш'(p_1), \quad \ldots, \quad p_m \in Ш'(p_{m-1}), \quad \bar{\bar{p}} \in Ш'(p_m),$$

причому

$$B(p_i, p_{i+1}) \neq 0, \quad i = 1, 2, \ldots, m-1; \quad B(\bar{p}, p_1) \neq 0, \tag{51}$$
$$B(p_m, \bar{\bar{p}}) \neq 0.$$

З цього означення випливає, що точка $\bar{\bar{p}}$ може бути граничною, тобто зв'язність означає, що кожна точка границі належить околу $Ш'(p)$ принаймні одного внутрішнього вузла.

За допомогою позначення

$$Ly(p) = A(p)\, y(p) - \sum_{Q \in Ш'(p)} B(p, Q)\, y(Q) \tag{52}$$

рівняння (48) можна записати у вигляді

$$Ly(p) = F(p). \tag{53}$$

Іноді використовуватимемо ще й таку форму запису $Ly(p)$:

$$Ly(p) = D(p)\,y(p) + \sum_{Q \in \text{Ш}'(p)} B(p, Q)\,(y(p) - y(Q)). \qquad (54)$$

Приклад 5. Записати у вигляді (48) сіткову схему з ваговими коефіцієнтами

$$\frac{y_i^{j+1} - y_i^j}{\tau} = \Lambda\,(\sigma y_i^{j+1} + (1-\sigma)\,y_i^j) + \varphi_i^j, \quad i = \overline{1,\ N-1}, \quad j = 0,\ 1,\ \ldots,$$
$$\Lambda y = y_{\bar{x}x}, \quad y_i^0 = u_0(x_i), \quad y_0^j = \mu_1(t_j), \quad y_N^j = \mu_2(t_j), \qquad (55)$$

яка апроксимує таку задачу для рівняння теплопровідності (див. п. 6.1):

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad x \in (0, 1), \quad t > 0,$$
$$u(x, 0) = u_0(x); \quad u(0, t) = \mu_1(t), \quad u(1, t) = \mu_2(t).$$

Розв'язання. Схема (55) записана на сітці $\bar{\omega}_{h\tau} = \bar{\omega}_h \times \omega_\tau$ ($\times$ — знак декартового добутку), $\bar{\omega}_h = \{x_i = ih : i = \overline{0,\ N},\ h = 1/N\}$, $\omega_\tau = \{t_j = j\tau : j = 0, 1, \ldots\}$. Нехай $p = (x_i, t_{j+1})$. Тоді Ш′ (p) складається з вузлів $Q_1 = (x_i, t_j)$, $Q_2 = (x_{i-1}, t_{j+1})$, $Q_3 = (x_{i+1}, t_{j+1})$, $Q_4 = (x_{i-1}, t_j)$, $Q_5 = (x_{i+1}, t_j)$, а границя складається з вузлів $(x_i, 0)$ та $(0, t_j)$, $(1, t_j)$, $i = \overline{0,\ N}$; $j = 0, 1, \ldots$. Фіксуючи деяке t_j, перепишемо (55) у вигляді

$$\left(\frac{1}{\tau} + \frac{2\sigma}{h^2}\right) y_i^{j+1} = \frac{\sigma}{h^2}\,(y_{i-1}^{j+1} + y_{i+1}^{j+1}) + \left(\frac{1}{\tau} + \frac{2(\sigma-1)}{h^2}\right) y_i^j +$$
$$+ \frac{(1-\sigma)}{h^2}\,(y_{i-1}^j + y_{i+1}^j) + \varphi_i^j.$$

Звідси випливає, що всі коефіцієнти $B(p, Q) \geqslant 0$ лише при $\tau \leqslant \dfrac{h^2}{2(1-\sigma)}$ та $\sigma \in [0, 1]$. Крім того, $D(p) = 0$.

Доведемо таке основне твердження.

Теорема 9 (принцип максимуму). *Нехай $y(p) \not\equiv \text{const}$ — сіткова функція, визначена на зв'язній сітці $\bar{\omega}$, і виконуються умови (49), (51). Тоді, якщо $Ly(p) \leqslant 0$ ($Ly(p) \geqslant 0$) на ω, то $y(p)$ не може набувати найбільшого додатного (найменшого від'ємного) значення у внутрішніх вузлах $p \in \omega$.*

Доведення. Нехай для визначеності $Ly(p) \leqslant 0$ в усіх внутрішніх вузлах $p \in \omega$. Припустимо від супротивного, що $y(p)$ набуває найбільшого додатного значення у внутрішньому вузлі $\tilde{p} \in \omega$, тобто

$$y(\tilde{p}) = \max_{\bar{\omega}} y(p) = M_0 > 0.$$

Теорему буде доведено, якщо покажемо, що існує внутрішня точка $\bar{p}$, для якої $Ly(\bar{p}) > 0$, що суперечить умові $Ly(p) \leqslant 0$. Оскільки $y(\tilde{p}) \geqslant y(Q)\ \forall\, Q \in \text{Ш}(\tilde{p})$, то

$$Ly(\tilde{p}) = D(\tilde{p})\,y(\tilde{p}) + \sum_{Q \in \text{Ш}(p)} B(\tilde{p}, Q)\,(y(\tilde{p}) - y(Q)) \geqslant D(\tilde{p})\,y(\tilde{p}) \geqslant 0,$$

бо $D(\tilde{p}) \geqslant 0$ і $y(\tilde{p}) > 0$. Зрозуміло, що слід розглянути лише випадок $Ly(\tilde{p}) = 0$. Як випливає з (54), це можливо лише при $D(\tilde{p}) = 0$, $y(Q) = y(\tilde{p})$ для всіх $Q \in \text{Ш}'(\tilde{p})$. Нехай $p_1 \in \text{Ш}'(\tilde{p})$ — вузол з околу $\tilde{p}$, в якому $y(p_1) = y(p) = M_0$. Повторимо ті самі міркування. Оскільки $y(p) \not\equiv \text{const}$ на ω і сітка зв'язна, то існує така послідовність вузлів $p_1, p_2, \ldots, p_m, \bar{\tilde{p}}$ (для яких справедливе рівняння (51)), що

$$y(p_m) = y(\bar{\tilde{p}}) = M_0,$$

а $y(\bar{\tilde{p}}) < M_0$, $\bar{\tilde{p}} \in \text{Ш}'(p_m)$. Тоді

$$Ly(p_m) \geqslant D(p_m)\,y(p_m) + B(p_m, \bar{\tilde{p}})\,(y(p_m) - y(\bar{\tilde{p}})) \geqslant$$
$$\geqslant B(p_m, \bar{\tilde{p}})\,(y(\bar{\tilde{p}}) - y(\bar{\tilde{p}})) > 0,$$

тобто $\tilde{p} = p_m$ і перше твердження теореми доведено. Друге твердження зводиться до першого, якщо $y(p)$ замінити на $-y(p)$.

Наслідок 1. Нехай виконуються умови (49), (51) і сіткова функція $y(p)$ визначена на $\omega \cup \gamma$, невід'ємна на границі ($y(p) \geqslant 0\ \forall\, p \in \gamma$) і $Ly(p) \geqslant 0$ на ω. Тоді $y(p)$ невід'ємна на $\omega \cup \gamma$, тобто $y(p) \geqslant 0\ \forall\, p \in \omega \cup \gamma$. Якщо $y(p) \leqslant 0$ на γ, $Ly(p) \leqslant 0$ на ω, то $y(p) \leqslant 0$ на $\omega \cup \gamma$.

Доведення. Нехай $Ly(p) \geqslant 0$ на ω, $y(p) \geqslant 0$ на γ. Припустимо від супротивного, що $y(p_0) < 0$ хоча б в одному внутрішньому вузлі $p_0 \in \omega$. Тоді $y(p)$ набуватиме найменшого від'ємного значення в ω, що неможливо в силу теореми 9, бо $y(p) \not\equiv \text{const}$ на ω ($y(p_0) < 0$, $y|_\gamma \geqslant 0$). Друга частина наслідку 1 доводиться аналогічно.

Наслідок 2. Задача

$$Ly(p) = 0, \quad p \in \omega,$$
$$y(p) = 0, \quad p \in \gamma \qquad (56)$$

за умов (49), (51) має лише тривіальний розв'язок $y(p) \equiv 0$.

Доведення. Неважко помітити, що $y(p) \equiv 0$ є розв'язком задачі (56). Нехай існує розв'язок такий, що $y(p) \not\equiv 0$ хоча б в одній точці. Тоді в силу наслідку 1 мають одночасно виконуватися нерівності $y(p) \geqslant 0$ та $y(p) \leqslant 0$, що можливо лише при $y(p) \equiv 0$.

Безпосередньо з наслідку 2 випливає наступне твердження.

Наслідок 3. За умовами (49), (51) існує єдиний розв'язок задачі (48), (50).

Теорема 10 *(теорема порівняння). Нехай $y(p)$ — розв'язок задачі (48) — (51), а $Y(p)$ — розв'язок задачі*

$$\begin{aligned} LY(p) &= \overline{F}(p), \quad p \in \omega, \\ Y(p) &= \bar{\mu}(p), \quad p \in \gamma, \end{aligned} \tag{57}$$

причому

$$\begin{aligned} |F(p)| &\leqslant \overline{F}(p), \quad p \in \omega, \\ |\mu(p)| &\leqslant \bar{\mu}(p), \quad p \in \gamma. \end{aligned} \tag{58}$$

Тоді

$$|y(p)| \leqslant Y(p), \quad p \in \omega \cup \gamma. \tag{59}$$

Доведення. З наслідку 1 маємо $Y(p) \geqslant 0$, $p \in \omega \cup \gamma$. Функції $u(p) = Y(p) + y(p)$ та $v(p) = Y(p) - y(p)$ задовольняють рівняння

$$Lu = F_u \equiv \overline{F} + F \geqslant 0, \quad Lv = F_v \equiv \overline{F} - F \geqslant 0$$

та граничні умови

$$u|_\gamma = (Y + y)|_\gamma = \bar{\mu} + \mu \geqslant 0, \quad v|_\gamma = (Y - y)_\gamma = \bar{\mu} - \mu \geqslant 0.$$

З наслідку 1 дістаємо $u \geqslant 0$, $v \geqslant 0$ або $y \geqslant -Y$, $y \leqslant Y$. Це і означає, що $|y(p)| \leqslant Y(p)$ на $\omega \cup \gamma$. Теорему доведено.

Функція $Y(p)$ називається *мажорантою* для розв'язку задачі (48), (50). Якщо мажоранта відома, то з (59) дістанемо оцінку для $\|y\|_C$.

Наслідок 4. Для розв'язку задачі

$$\begin{aligned} Ly(p) &= 0, \quad p \in \omega, \\ y(p) &= \mu(p), \quad p \in \gamma, \end{aligned} \tag{60}$$

справедлива оцінка

$$\max_{p \in \omega \cup \gamma} |y(p)| = \|y\|_{C(\bar{\omega})} \leqslant \|\mu\|_{C(\gamma)}, \tag{61}$$

де

$$\|\mu\|_{C(\gamma)} = \max_{p \in \gamma} |\mu(p)|.$$

Доведення. Визначимо мажоранту $Y(p)$ за умов

$$\begin{aligned} LY(p) &= 0, \quad p \in \omega, \\ Y(p) &= \|\mu\|_{C(\gamma)}, \quad p \in \gamma. \end{aligned}$$

Тоді функція $Y(p) \geqslant 0$ на $\bar{\omega}$ і набуває найбільшого значення у деякому вузлі сітки $\bar{\omega}$. Якщо $Y(p) \neq \text{const}$, то цей вузол не може бути

внутрішнім, отже, $\|Y\|_{C(\bar{\omega})} = \|Y\|_{C(\gamma)} = \|\mu\|_{C(\gamma)}$. Якщо $Y(p) = \text{const}$, то $Y(p) = \|\mu\|_{C(\gamma)}$. В обох випадках $\|Y\|_{C(\bar{\omega})} = \|\mu\|_{C(\gamma)}$. Звідси і з нерівності $\|y\|_{C(\bar{\omega})} \leqslant \|Y\|_{C(\bar{\omega})}$ і випливає (61).

Розглянемо оцінки для розв'язку неоднорідної задачі (48), (50). Запишемо його у вигляді

$$y = \bar{y} + v,$$

тоді $\bar{y}(p)$ — розв'язок задачі

$$L\bar{y}(p) = 0, \quad p \in \omega; \quad \bar{y}(p) = \mu(p), \quad p \in \gamma, \tag{62}$$

а $v(p)$ — розв'язок задачі

$$Lv(p) = F(p), \quad p \in \omega, \; v(p) = 0, \; p \in \gamma. \tag{63}$$

Для $\bar{y}(p)$ ми вже маємо оцінку (61). Оцінимо $v(p)$.

Теорема 11. *Якщо $D(p) > 0$, $p \in \omega$, то для розв'язку задачі (63) справедлива оцінка*

$$\|v\|_{C(\omega)} \leqslant \|F/D\|_{C(\omega)}. \tag{64}$$

Доведення. Розглянемо мажоранту $Y(p)$, яка є розв'язком задачі

$$\begin{aligned} LY(p) &= |F(p)|, \quad p \in \omega, \\ Y(p) &= 0, \quad p \in \gamma. \end{aligned}$$

В силу наслідку 1 з принципу максимуму $Y(p) > 0$, $p \in \omega = \bar{\omega} \cup \gamma$. Нехай $Y(p)$ набуває найбільшого значення у вузлі $p_0 \in \omega$. Оскільки $Y(p_0) = \|Y\|_{C(\omega)}$, то з рівняння

$$D(p_0) Y(p_0) + \sum_{Q \in \text{Ш}'(p_0)} B(p_0, Q)(Y(p_0) - Y(Q)) = |F(p_0)|$$

випливає

$$D(p_0) Y(p_0) \leqslant |F(p_0)|, \quad Y(p_0) \leqslant \frac{|F(p_0)|}{D(p_0)} \leqslant \|F/D\|_{C(\omega)}.$$

Це і треба було довести.

Зауваження. Оцінка (64) справедлива для розв'язку задачі (63) і тоді, коли замість (49) виконуються умови

$$\begin{aligned} A(p) &\neq 0, \quad B(p, Q) \neq 0, \\ D(p) &= |A(p)| - \sum_{Q \in \text{Ш}'(p)} |B(p, Q)| > 0. \end{aligned}$$

Дійсно, нехай $|v(p)| \geqslant 0$ набуває найбільшого значення у вузлі p_0. Тоді

$$\begin{aligned} |A(p_0)|\,|v(p_0)| &= \left| \sum_{Q \in \text{Ш}'(p_0)} B(p_0, Q) v(Q) + F(p_0) \right| \leqslant \\ &\leqslant \sum_{Q \in \text{Ш}'(p_0)} |B(p_0, Q)|\,|v(p_0)| + |F(p_0)|, \end{aligned}$$

звідки

$$D(p_0)\,|\,v(p_0)\,|\leqslant|\,F(p_0)\,|,$$

$$\|v\|_{C(\omega)}=v(p_0)\leqslant\frac{|\,F(p_0)\,|}{D(p_0)}\leqslant\|F/D\|_{C(\omega)}.$$

Може статися, що $D(p)=0$ на підмножині $\mathring{\omega}$ сітки ω і $D(p)>0$ на $\omega^*=\omega\setminus\mathring{\omega}$. Тоді має місце наступна теорема.

Теорема 12. *Нехай $D(p)=0$, $p\in\mathring{\omega}$ і $D(p)>0$, $p\in\omega^*$, де $\mathring{\omega}$ — зв'язна сітка. Тоді для розв'язку задачі (63) з правою частиною*

$$F(p)=\begin{cases}0, & p\in\mathring{\omega},\\ F_1(p), & p\in\omega^*,\end{cases}$$

справедливою є оцінка

$$\|v\|_{C(\omega)}\leqslant\|F_1/D\|_{C(\omega^*)}=\max_{p\in\omega^*}|\,F_1(p)/D(p)\,|. \tag{65}$$

Доведення. Нехай $Y(p)$ — така мажоранта, що $LY=F(p)$, $p\in\omega$; $Y(p)=0$, $p\in\gamma$, $Y(p)\geqslant 0$, $p\in\bar{\omega}$. Функція $Y(p)$ має набувати найбільшого значення на скінченній множині ω у деякому вузлі цієї множини. Проте цей вузол не може бути ні граничним, бо $Y(p)=0$, $p\in\gamma$, ні вузлом сітки $\mathring{\omega}$, бо $\mathring{\omega}$ — зв'язна сітка і на ній виконується принцип максимуму. Отже,

$$\max_{p\in\omega}Y(p)=\max_{p\in\omega^*}Y(p)=Y(p_0),\quad p_0\in\omega^*.$$

За умовою $D(p_0)>0$. Тому, повторюючи міркування щодо доведення теореми 11, дістанемо (64). Має місце і відповідний аналог зауваження до теореми 11.

Таким чином, ми розглянули теореми, які за допомогою мажоранти дозволяють дістати апріорні оцінки в нормі $C(\omega)$ для розв'язування сіткової схеми виду (48), (50). У п. 2 гл. 5 використаємо ці теореми для обгрунтування методу сіток розв'язування задачі Діріхле для рівняння Пуассона.

2.3. Загальна теорія наближених методів

Значна кількість задач обчислювальної математики зводиться до апроксимації лінійних операторів (функціоналів) або до розв'язування операторних рівнянь у деяких нормованих просторах. Для цього застосовуються наближені методи, які грунтуються на різних ідеях. Для з'ясування питання про ефективність і обгрунтованість цих методів необхідні їхні теоретичні дослідження, у процесі яких постають такі задачі: а) встановлення працездатності (здійсненності) алгоритму та його збіжності; б) дослідження швидкості збіжності;

в) ефективна оцінка похибки. Ці задачі для багатьох методів розв'язуються окремо і мають свої труднощі, багато з яких ще треба подолати. Але багато результатів досліджень із вказаних вище трьох напрямків можна об'єднати у загальну теорію наближених методів, основи якої в 1948 р. заклав Л. В. Канторович. Вона охоплює також багато питань тих методів, які ми розглядаємо, зокрема методу сіток, проекційних методів та ін. Розглянемо деякі з цих питань.

Нехай X — нормований простір, $\tilde{X}$ — його повний підпростір, причому існує неперервний лінійний оператор P, що проектує простір X на $\tilde{X}$, тобто

$$P(X)=\tilde{X},\quad P^2=P.$$

Зрозуміло, що оператор P не змінює елементів з $\tilde{X}$. Якщо, наприклад, $X=C[a,\ b]$, а $\tilde{X}$ — сукупність поліномів, не вище $(n-1)$-го степеня, то P ставить у відповідність неперервній функції $x(t)\in C[a,\ b]$ її інтерполяційний многочлен за заданою системою вузлів $t_1,\ \ldots,\ t_n$.

Розглянемо два рівняння, які називатимемо рівняннями другого роду. Перше

$$Kx=x-\lambda Hx=y \tag{1}$$

— у просторі X, а друге

$$\tilde{K}\tilde{x}\equiv\tilde{x}-\lambda\tilde{H}\tilde{x}=Py \tag{2}$$

— у просторі $\tilde{X}$. Тут H— неперервний лінійний оператор в X, $\tilde{H}$ — неперервний лінійний оператор в $\tilde{X}$, λ — число. Рівняння (1) будемо називати точним, а рівняння (2) — наближеним; воно відповідає точному рівнянню (1). Розв'язок наближеного рівняння (2) розглядаємо як наближений розв'язок точного рівняння (1). Щоб пов'язати розв'язки точного і наближеного рівнянь, зробимо такі природні припущення.

I. *Умова близькості операторів H та $\tilde{H}$.* Для довільного $\tilde{x}\in\tilde{X}$

$$\|PH\tilde{x}-\tilde{H}\tilde{x}\|\leqslant\eta\|\tilde{x}\|,$$

що рівносильно нерівності

$$\|PK\tilde{x}-\tilde{K}\tilde{x}\|\leqslant|\,\lambda\,|\,\eta\|\tilde{x}\|\ \forall\,\tilde{x}\in\tilde{X},$$

причому стала η не залежить від $\tilde{x}$.

II. *Умова хорошої апроксимації елементів виду* Hx, $x \in X$, *елементами з* X. Для будь-яких $x \in X$ знайдеться $\tilde{x} \in \tilde{X}$ такий, що

$$\|Hx - \tilde{x}\| \leqslant \eta_1 \|x\|,$$

де стала не залежить від x, $\tilde{x}$.

III. *Умови хорошої апроксимації вільного члена точного рівняння.* Існує елемент $\tilde{y} \in \tilde{X}$ такий, що

$$\|y - \tilde{y}\| \leqslant \eta_2 \|y\|$$

(на відміну від попередніх умов тут η_2 може залежати від y).

Теорема 1. *Нехай виконуються умови I, II і оператор* K *має неперервний обернений. Якщо*

$$q = |\lambda|(\eta + \|I - P\|\eta_1)\|K^{-1}\| < 1, \tag{3}$$

то оператор $\tilde{K}$ *також має обернений, причому*

$$\|\tilde{K}^{-1}\| \leqslant \frac{\|K^{-1}\|}{1-q}. \tag{4}$$

Доведення. Припустимо спочатку, що простір X повний. Розглянемо оператор $K_1 = I - \lambda PH$, де I — одиничний оператор. Тоді згідно з умовою II

$$\|(K - K_1)x\| = |\lambda|\|Hx - PHx\| = |\lambda|\|Hx - \tilde{x} + P\tilde{x} -$$

$$- PH\tilde{x}\| = |\lambda|\|(I - P)(Hx - \tilde{x})\| \leqslant |\lambda|\|I - P\|\eta_1\|x\|,$$

а це в силу довільності x означає, що

$$\|K - K_1\| \leqslant |\lambda|\|I - P\|\eta_1.$$

Оператор K_1 можна записати у вигляді

$$K_1 = K(I - K^{-1}(K - K_1)), \tag{5}$$

причому

$$\|K^{-1}(K - K_1)\| \leqslant |\lambda|\|I - P\|\eta_1\|K^{-1}\| \leqslant q < 1.$$

За теоремою Банаха (див. п. 1.4 з ч. 1) існує обернений оператор

$$(I - K^{-1}(K - K_1))^{-1} \text{ і } \|I - K^{-1}(K - K_1)^{-1}\| \leqslant \frac{1}{1-q}.$$

Із зображення (5) випливає, що оператор K_1 має обернений $K_1^{-1} = (I - K^{-1}(K - K_1))^{-1}K^{-1}$, причому

$$\|K_1^{-1}\| \leqslant \frac{\|K^{-1}\|}{1 - |\lambda|\|I - P\|\eta_1\|K^{-1}\|}. \tag{6}$$

У просторі $\tilde{X}$ розглянемо оператор $\tilde{K}_1\tilde{x} = \tilde{x} - \lambda PH\tilde{x}$. Для довільного $\tilde{x} \in \tilde{X}$ буде $\tilde{K}_1\tilde{x} = K_1\tilde{x}$. Неважко помітити, що для довільного $\tilde{y} \in \tilde{X}$ буде $K_1^{-1}\tilde{y} \in \tilde{X}$. Дійсно, якщо $x' = K_1^{-1}\tilde{y}$, то $x' - \lambda PHx' = \tilde{y}$, $x' = \tilde{y} + \lambda PHx' \in \tilde{X}'$. Тому оператор $\tilde{K}_1$ має неперервний обернений, який на X' збігається з K_1^{-1}, причому

$$\|\tilde{K}_1^{-1}\| = \sup_{\tilde{x}\in\tilde{X},\,\tilde{x}\neq 0} \frac{\|\tilde{K}_1^{-1}\tilde{x}\|}{\|\tilde{x}\|} = \sup_{\tilde{x}\in\tilde{X},\,\tilde{x}\neq 0} \frac{\|K_1^{-1}\tilde{x}\|}{\|\tilde{x}\|} \leqslant$$

$$\leqslant \sup_{x\in X,\,x\neq 0} \frac{\|K_1^{-1}x\|}{\|x\|} = \|K_1^{-1}\|. \tag{7}$$

Оцінимо різницю $\tilde{K} - \tilde{K}_1$. За умовою I для довільного $\tilde{x} \in \tilde{X}$

$$\|\tilde{K}\tilde{x} - \tilde{K}_1\tilde{x}\| = |\lambda|\|PH\tilde{x} - \tilde{H}\tilde{x}\| \leqslant |\lambda|\eta\|\tilde{x}\|,$$

тобто

$$\|\tilde{K} - \tilde{K}_1\| \leqslant |\lambda|\eta. \tag{8}$$

Згідно з (3), (6) — (8)

$$\|\tilde{K}_1^{-1}\|\|\tilde{K} - \tilde{K}_1\| \leqslant \frac{\|K^{-1}\||\lambda|\eta}{1 - |\lambda|\|I - P\|\eta_1\|K^{-1}\|} =$$

$$= 1 - \frac{1-q}{1 - |\lambda|\|I - P\|\eta_1\|K^{-1}\|} < 1.$$

Тому, записуючи оператор $\tilde{K}$ у формі $\tilde{K} = \tilde{K}_1(I - \tilde{K}_1^{-1}(\tilde{K}_1 - \tilde{K}))$, за теоремою Банаха дістанемо, що існує $\tilde{K}^{-1}$. З урахуванням (6) і останньої нерівності маємо

$$\|\tilde{K}^{-1}\| \leqslant \frac{\|\tilde{K}_1^{-1}\|}{1 - \|\tilde{K}_1^{-1}\|\|\tilde{K} - \tilde{K}_1\|} \leqslant \frac{\|K_1^{-1}\|}{1 - \frac{\|K^{-1}\||\lambda|\eta}{1 - |\lambda|\|I - P\|\eta_1\|K^{-1}\|}} =$$

$$= \frac{\|K_1^{-1}\|(1 - |\lambda|\|I - P\|\eta_1\|K^{-1}\|}{1 - |\lambda|\|\eta + \|I - P\|\eta_1\|K^{-1}\|} =$$

$$= \frac{\|K_1^{-1}\|(1 - |\lambda|\|I - P\|\eta_1\|K^{-1}\|)}{1-q} \leqslant \frac{\|K^{-1}\|}{1-q}.$$

Отже, для повного простору X теорему доведено (повнота простору необхідна для застосування теореми Банаха).

Якщо простір X неповний, то, вводячи його поповнення, можна довести, що замикання операторів K, K^{-1} (тобто розширення їх за неперервністю на весь повний простір) є взаємно оберненими, а замикання оператора H також пов'язане з оператором $\tilde{H}$ умовами I, II з тією самою сталою η і новою сталою η_1, як завгодно близькою до старої. Тому випадок неповного простору X зводиться до розглянутого вище і теорему повністю доведено.

Теорема 2. *Якщо виконано умови I—III, існує неперервний обернений оператор $\tilde{K}^{-1}$ (це має місце, зокрема, при виконанні умов теореми 1) і рівняння* (1) *має розв'язок x^*, то справедлива оцінка*

$$\|x^* - \tilde{x}^*\| \leqslant p\|x^*\|,$$

де $\tilde{x}^*$ — розв'язок рівняння (2), *а*

$$p = 2|\lambda|\eta\|\tilde{K}^{-1}\| + (\eta_1|\lambda| + \eta_2\|K\|)(1 + \|\tilde{K}^{-1}PK\|).$$

Доведення. Доведемо, що елемент x^* можна апроксимувати елементом з $\tilde{X}$ з точністю до порядку $\eta_1 + \eta_2$, тобто існує $\tilde{x} \in \tilde{X}$ такий, що

$$\|x^* - \tilde{x}\| \leqslant \varepsilon\|x^*\|, \tag{9}$$

де

$$\varepsilon = \min[1, \eta_1|\lambda| + \eta_2\|K\|].$$

Дійсно, за умовами II, III можна вибрати $\tilde{y} \in \tilde{X}$, $\tilde{z} \in \tilde{X}$, так що

$$\|Hx^* - \tilde{z}\| \leqslant \eta_1\|x^*\|, \quad \|y - \tilde{y}\| \leqslant \eta_2\|y\| \leqslant \eta_2\|K\|\|x^*\|.$$

Покладаючи $\tilde{x} = \lambda\tilde{z} + \tilde{y}$, дістанемо

$$\|x^* - \tilde{x}\| = \|\lambda Hx^* + y - (\lambda\tilde{z} + \tilde{y})\| \leqslant (\eta_1|\lambda| + \eta_2\|K\|)\|x^*\|.$$

Крім того, (9) має місце при $\varepsilon = 1$, якщо покласти $\tilde{x} = 0$.

Позначимо $\tilde{x}_0 = \tilde{K}^{-1}PK\tilde{x}$. Тоді

$$\|x^* - \tilde{x}^*\| \leqslant \|x^* - \tilde{x}\| + \|\tilde{x} - \tilde{x}_0\| + \|\tilde{x}_0 - \tilde{x}^*\|,$$

причому перша складова оцінюється за допомогою (9). З умови 1 випливає

$$\|\tilde{x} - \tilde{x}_0\| = \|\tilde{K}^{-1}\tilde{K}\tilde{x} - \tilde{K}^{-1}PK\tilde{x}\| \leqslant \|\tilde{K}^{-1}\|\|\tilde{K}\tilde{x} - PK\tilde{x}\| \leqslant$$

$$\leqslant |\lambda|\eta\|\tilde{K}^{-1}\|\|\tilde{x}\|,$$

а з нерівності (9) —

$$\|\tilde{x}\| \leqslant \|x^*\| + \|x^* - \tilde{x}\| \leqslant (1 + \varepsilon)\|x^*\|,$$

тобто

$$\|\tilde{x} - \tilde{x}_0\| \leqslant |\lambda|\eta(1 + \varepsilon)\|\tilde{K}^{-1}\|\|x^*\|.$$

Оскільки $\tilde{x}^*$ — розв'язок рівняння (2), тобто $\tilde{x}^* = \tilde{K}^{-1}PKx^*$, то для складової $\|\tilde{x}_0 - \tilde{x}^*\|$ за допомогою (9) дістанемо оцінку

$$\|\tilde{x}_0 - \tilde{x}^*\| = \|\tilde{K}^{-1}PK\tilde{x} - \tilde{K}^{-1}PKx^*\| \leqslant \|\tilde{K}^{-1}PK\|\|\tilde{x} - x^*\| \leqslant$$

$$\leqslant \varepsilon\|\tilde{K}^{-1}PK\|\|x^*\|.$$

Таким чином,

$$\|x^* - \tilde{x}^*\| \leqslant [|\lambda|\eta(1 + \varepsilon)\|\tilde{K}^{-1}\| + \varepsilon(1 + \|\tilde{K}^{-1}PK\|)]\|x^*\|. \tag{9'}$$

Оскільки $1 + \varepsilon \leqslant 2$, $\varepsilon \leqslant \eta_1|\lambda| + \eta_2\|K\|$, дістанемо твердження теореми.

Зауваження. Якщо виходити з того, що нерівність (9) має місце, то для здобуття (9') не треба пропускати виконання умов II, III. Зазначимо, що в цьому разі йдеться про наближення конкретного елемента x^* і ε може від нього залежати, а у припущенні II величина η_1 від x не залежить. Величина ε характеризує можливість апроксимації розв'язку x^* рівняння (1) елементом $\tilde{x} \in \tilde{X}$, і якщо $\eta = 0$, то

$$\|x^* - \tilde{x}^*\| = O(\varepsilon\|\tilde{K}^{-1}PK\|).$$

Теореми 1, 2 можна використовувати для апріорних оцінок (оцінок коректності) сіткових схем і для оцінок похибки.

Якщо маємо послідовність наближених рівнянь і послідовність $\{\tilde{x}_n^*\}$ знайдених за їх допомогою наближених розв'язків, то простір $\tilde{X}$, а також оператори $\tilde{H}$ ($\tilde{K}$), P, сталі η, η_1, η_2, q, p і т. д. будуть залежати від індексу n (який для спрощення позначень опускаємо). Тоді умови збіжності послідовності $\{\tilde{x}_n^*\}$ до точного розв'язку даються наступною теоремою.

Теорема 3. *Якщо виконано умови:* 1) *оператор K має неперервний обернений;* 2) *при кожному $n = 1, 2, \ldots$ виконано умови I—III, причому*

$$\lim_{n\to\infty}\eta = 0, \quad \lim_{n\to\infty}\eta_1\|P\| = 0, \quad \lim_{n\to\infty}\eta_2\|P\| = 0, \tag{10}$$

то при досить великих n наближені рівняння мають розв'язки $\tilde{x}_n^*$,

які збігаються до точного x^*, тобто

$$\lim_{n\to\infty} \| x^* - \tilde{x}_n^* \| = 0,$$

причому

$$\| x^* - \tilde{x}_n^* \| \leqslant Q_0\eta + Q_1\eta_1 \| P \| + Q_2\eta_2 \| P \|, \qquad (11)$$

де Q_i, $i = 0, 1, 2$ — деякі сталі.

Доведення. Оскільки $\| P \| \geqslant 1$, то з (10) випливає, що $\lim_{k\to\infty} \eta_1 = 0$, тому при досить великих n в теоремі 1 матимемо $q < \frac{1}{2}$. Таким чином, для вказаних n існує неперервний оператор $\tilde{K}^{-1}$, причому

$$\| \tilde{K}^{-1} \| \leqslant \frac{\| K^{-1} \|}{1 - q} < 2 \| K^{-1} \|.$$

Звідси $\| \tilde{K}^{-1} \|$ — обмежена стала, яка не залежить від n. Враховуючи це, а також незалежність $\| K \|$ від n, з теореми 2 дістаємо (11), і отже, твердження теореми 3.

Теорема 1 дає змогу на основі існування розв'язку точного рівняння зробити висновок про існування розв'язку наближеного рівняння. З наступної теореми можна зробити обернений висновок.

Теорема 4. *Якщо оператор $\tilde{K}$ має неперервний обернений, виконуються умови I і II, а також*

$$r = |\lambda| \eta (1 + |\lambda| \eta_1) \| \tilde{K}^{-1} \| + |\lambda| \eta_1 (1 + \| \tilde{K}^{-1} P K \|) < 1,$$

то K має неперервний лівий обернений оператор, норма якого оцінюється за допомогою нерівності

$$\| K^{-1} \| \leqslant \frac{1 + \| \tilde{K}^{-1} P \| + |\lambda| \eta \| \tilde{K}^{-1} \| + \| \tilde{K}^{-1} P K \|}{1 - r}.$$

Розглянемо рівняння, які завдяки наявності в їхніх операторах головної частини простого вигляду можна звести до рівнянь другого роду.

Нехай маємо два нормованих простори X та Y і в кожному з них виділено повний підпростір: $\tilde{X} \subset X$, $\tilde{Y} \subset Y$. Крім того, існує неперервний лінійний оператор Φ, який проектує Y на $\tilde{Y}$.

Розглянемо точне рівняння

$$K_1 x \equiv Gx - \lambda Tx = y_1 \qquad (12)$$

і відповідне йому наближене

$$\tilde{K}_1 \tilde{x} \equiv G\tilde{x} - \lambda \tilde{T}\tilde{x} = \Phi y_1, \qquad (13)$$

де неперервні лінійні оператори G, T, K_1 відображають X в Y, а неперервні лінійні оператори $\tilde{T}$, $\tilde{K}_1$ — $\tilde{X}$—в $\tilde{Y}$. Припустимо також, що: 1) оператор G має неперервний обернений; 2) оператор G встановлює взаємно однозначну відповідність між $\tilde{X}$ та $\tilde{Y}$, тобто $G(\tilde{X}) = \tilde{Y}$ і, отже, $G^{-1}(\tilde{Y}) = \tilde{X}$. Нехай виконуються умови, аналогічні розглянутим вище умовам I—III.

I *a*. Для будь-якого $\tilde{x} \in X$ має місце нерівність

$$\| \Phi T\tilde{x} - \tilde{T}\tilde{x} \| \leqslant \mu \| \tilde{x} \|.$$

II *a*. Для довільного $x \in X$ знайдеться таке $\tilde{y} \in \tilde{Y}$, що

$$\| Tx - \tilde{y} \| \leqslant \mu_1 \| x \|.$$

III *a*. Існує елемент $\tilde{y}_1 \in \tilde{Y}$ такий, що

$$\| y_1 - \tilde{y}_1 \| \leqslant \mu_2 \| y_1 \|.$$

Покажемо, що за цих припущень вивчення рівнянь (12), (13) зводиться до дослідження рівнянь другого роду.

Дійсно, застосувавши до обох частин рівнянь (12), (13) оператор G^{-1}, дістанемо рівняння

$$Kx \equiv G^{-1}K_1 x \equiv x - \lambda G^{-1}Tx = G^{-1}y_1, \qquad (14)$$

$$\tilde{K}\tilde{x} \equiv G^{-1}\tilde{K}_1\tilde{x} \equiv \tilde{x} - \lambda G^{-1}\tilde{T}\tilde{x} = G^{-1}\Phi y_1, \qquad (15)$$

які мають вигляд рівнянь (1), (2), причому роль операторів H, $\tilde{H}$, P, K, $\tilde{K}$ відіграють оператори $G^{-1}T$, $G^{-1}\tilde{T}$, $G^{-1}\Phi G$, $G^{-1}K_1$, $G^{-1}\tilde{K}_1$, а роль y — елемент $G^{-1}y_1$.

Перевіримо виконання умов I—III.

I. З I *a* випливає

$$\| PH\tilde{x} - \tilde{H}\tilde{x} \| = \| G^{-1}\Phi G G^{-1}T\tilde{x} - G^{-1}\tilde{T}\tilde{x} \| \leqslant \| G^{-1} \| \mu \| \tilde{x} \|.$$

II. У відповідності з II *a* для довільного $x \in X$ знайдеться елемент $\tilde{y} \in \tilde{Y}$ такий, що $\| Tx - \tilde{y} \| \leqslant \mu_1 \| x \|$. Покладаючи $\tilde{x} = G^{-1}\tilde{y}$, матимемо

$$\| Hx - \tilde{x} \| = \| G^{-1}Tx - G^{-1}\tilde{y} \| \leqslant \| G^{-1} \| \mu_1 \| x \|.$$

III. Нехай $\tilde{y}_1$ визначається згідно з III *a*. Позначивши $\tilde{y} = G^{-1}\tilde{y}_1$, матимемо

$$\| y - \tilde{y} \| = \| G^{-1}y_1 - G^{-1}\tilde{y}_1 \| \leqslant \| G^{-1} \| \mu_2 \| y_1 \| \leqslant \| G^{-1} \| \| G \| \mu_2 \| y \|.$$

Таким чином, умови I—III виконуються і сталі η, η_1, η_2 в них виражаються через μ_1, μ_2, μ за формулами

$$\eta = \mu \| G^{-1} \|, \quad \eta_1 = \mu_1 \| G^{-1} \|, \quad \eta_2 = \mu_2 \| G^{-1} \| \| G \|.$$

Тому для рівнянь (14), (15) мають місце розглянуті вище теореми 1—4. Щоб спростити їхні формулювання, припустимо додатково, що норми у просторах X, Y узгоджені за допомогою оператора G:

$$\| x \|_X = \| Gx \|_Y, \quad \| y \|_Y = \| G^{-1} y \|_X, \quad x \in X, \quad y \in Y. \tag{16}$$

При цьому відображення G ізометричне, $\| G \| = \| G^{-1} \| = 1$, завдяки цьому

$$\| P \| = \| \Phi \|, \quad \| K \| = \| K_1 \|, \quad \| \tilde{K} \| = \| \tilde{K}_1 \|, \tag{17}$$
$$\| H \| = \| T \|, \quad \| \tilde{H} \| = \| \tilde{T} \|.$$

Сформулюємо відповідні теореми для рівнянь (12), (13).

Теорема 1а. *Якщо виконано умови Ia, IIa, існує неперервний оператор* $\tilde{K}_1^{-1}$ *і*

$$q = |\lambda| [\mu + \| I - \Phi \| \mu_1] \| \tilde{K}_1^{-1} \| < 1, \tag{18}$$

то $\tilde{K}_1$ *також має неперервний обернений оператор, причому*

$$\| \tilde{K}_1^{-1} \| \leqslant \frac{\| K_1^{-1} \|}{1 - q}. \tag{19}$$

Теорема 2а. *Якщо виконуються умови Ia—IIIa, існує неперервний оператор* $\tilde{K}_1^{-1}$ *та існує розв'язок* x^* *рівняння* (12), *то справедлива оцінка*

$$\| x^* - \tilde{x}^* \| \leqslant p \| x^* \|, \tag{20}$$

де $\tilde{x}^*$ *— розв'язок рівняння* (13), *а*

$$p = 2 |\lambda| \mu \| \tilde{K}_1^{-1} \| + (\mu_1 |\lambda| + \mu_2 \| K_1 \|) (1 + \| \tilde{K}_1^{-1} \Phi K_1 \|). \tag{21}$$

Зауваження. Якщо існує елемент $\hat{x} \in \tilde{X}$ такий, що

$$\| x^* - \hat{x} \| \leqslant \varepsilon \| x^* \|,$$

то оцінка (20) справедлива без припущень IIa, IIIa, причому в цьому випадку

$$p = 2 |\lambda| \mu \| \tilde{K}_1^{-1} \| + \varepsilon (1 + \| \tilde{K}_1^{-1} \Phi K_1 \|).$$

Це доводиться аналогічно теоремі 2.

Теорема 3а. *Якщо при кожному* $n = 1, 2, \ldots$ *виконуються умови Ia—IIIa, оператор* K_1 *має неперервний обернений і має місце співвідношення*

$$\lim_{n \to \infty} \mu = \lim_{n \to \infty} \mu_1 \| \Phi \| = \lim_{n \to \infty} \mu_2 \| \Phi \| = 0,$$

то послідовність наближених розв'язків (тобто розв'язків рівняння (13)) *збігається до розв'язку* x^* *рівняння* (12), *причому*

$$\| x^* - \tilde{x}_n^* \| \leqslant Q\mu + Q_1 \mu_1 \| \Phi \| + Q_2 \mu_2 \| \Phi \|.$$

Теорема 4а. *Якщо існує неперервний оператор* $\tilde{K}_1^{-1}$, *виконуються умови Ia та IIa і*

$$r = |\lambda| \mu (1 + |\lambda| \mu_1) \| \tilde{K}_1^{-1} \| + |\lambda| \mu_1 (1 + \| \tilde{K}_1^{-1} \Phi K_1 \|) < 1,$$

то K_1 *має неперервний лівий обернений оператор, причому*

$$\| K_1^{-1} \| \leqslant \frac{1 + \| \tilde{K}_1^{-1} \Phi \| + |\lambda| \mu \| \tilde{K}_1^{-1} \| + \| \tilde{K}_1^{-1} \Phi K_1 \|}{1 - r}.$$

Розглянемо окремий випадок, коли формулювання теорем спрощуються, тобто, коли наближене рівняння (13) дістають проектуванням точного рівняння. У цьому разі для побудови наближеного рівняння до обох частин точного рівняння (12) застосовується оператор Φ:

$$\tilde{K}_1 \tilde{x} \equiv \Phi K_1 \tilde{x} \equiv G\tilde{x} - \lambda \Phi T \tilde{x} = \Phi y_1,$$

тобто $\tilde{T} = \Phi T$. Оскільки

$$\| \tilde{T}\tilde{x} - \Phi T \tilde{x} \| = \| \Phi T \tilde{x} - \Phi T \tilde{x} \| = 0,$$

то умова Ia виконується при $\mu = 0$, тому скрізь у формулюваннях теорем треба покласти $\mu = 0$.

ГЛАВА 3

ЗАДАЧА КОШІ ДЛЯ ЗВИЧАЙНИХ ДИФЕРЕНЦІАЛЬНИХ РІВНЯНЬ

3.1. Постановка задачі. Класифікація методів розв'язування задачі Коші

Нехай на відрізку $t_0 \leqslant t \leqslant T$ треба знайти розв'язок диференціального рівняння n-го порядку

$$u^{(n)} = f(t, u, u', \ldots, u^{(n-1)}), \tag{1}$$

який в точці $t = t_0$ набуває заданих початкових значень

$$u^{(i)}(t_0) = u_i^0, \quad i = \overline{0, n-1}. \tag{2}$$

Ця задача називається задачею Коші для рівняння (1). Достатньою умовою існування і єдиності її розв'язку з класу $C^n[t_0, T]$ є неперервність функції f по всіх аргументах і виконання умови Ліпшиця по змінних $u, u', \ldots, u^{(n-1)}$.

Задачу Коші для диференціального рівняння n-го порядку вигляду (1) можна звести до еквівалентної нормальної системи такого вигляду:

$$\frac{dU}{dt} = F(t, U), \quad U(t_0) = U_0, \tag{3}$$

якщо покласти

$$u'(t) = u_1(t), \quad u''(t) = u_2(t), \ \ldots, \ u^{(n-1)}(t) = u_{n-1}(t)',$$
$$u^{(n)}(x) = f(t, u(t), u_1(t), \ \ldots, \ u_{n-1}(t))$$

і позначити

$$U(x) = (u(t), u_1(t), \ \ldots, \ u_{n-1}(t))',$$
$$F(x, U) = (u_1(t), u_2(t), \ \ldots, \ u_{n-1}(t), f(x, u(t), \ \ldots, \ u_{n-1}(t)))',$$
$$U_0 = (u_0, u_0', \ \ldots, \ u_0^{(n-1)})'.$$

Задача Коші для більш загальної системи n рівнянь першого порядку

$$u_i'(t) = f_i(t, u_1(t), \ \ldots, \ u_n(t)),$$
$$u_i(t_0) = u_i^0, \quad i = \overline{1, n}, \tag{4}$$

також записується у вигляді (3), якщо ввести позначення

$$U(t) = (u_1(t), \ \ldots, \ u_n(t))',$$
$$F(t, U) = (f_1, \ \ldots, \ f_n)', \quad f_i = f_i(t, U(t)),$$
$$U_0 = (u_1^0, \ \ldots, \ u_n^0)'.$$

Оскільки більшість методів розв'язування задачі Коші для рівняння першого порядку

$$u'(t) = f(t, u),$$
$$u(t_0) = u_0 \tag{5}$$

майже без змін переносяться на системи, то з метою спрощення викладок розглядатимемо їх для задачі (5).

Розглянемо нелінійний оператор A з областю визначення

$$D(A) = \{u(t): \ u \in C^1(t_0, T], \ u(t_0) = u_0\}$$

і областю значень $R(A)$, з простору $C(t_0, T]$. Оператор A діє за правилом

$$A(u) = u'(t) - f(t, u).$$

Тоді задачу Коші (5), розв'язок якої шукається на проміжку $[t_0, T]$, можна записати у вигляді нелінійного операторного рівняння

$$A(u) = 0. \tag{5'}$$

Зазначимо, що сюди включається і випадок $T = \infty$. Вихідними даними задачі (5) або, що те саме, задачі (5') є початкове значення u_0 та права частина рівняння, тобто функція $f(t, u)$. Задача (5), (5') має бути коректно поставленою, тобто мають виконуватись умови: 1) вона має єдиний розв'язок; 2) цей розв'язок неперервно залежить від вихідних даних, тобто «малі» збурення вхідних даних викликають «малі» збурення розв'язку. Умову 2) називають також умовою стійкості. Достатньою умовою розв'язності рівняння (1) є неперервність $f(t, u)$ і виконання умови Ліпшиця по u. Виявимо умови стійкості задачі (5), (5'), ввівши попередньо наступні означення.

Означення 1. Задача (5) називається *стійкою щодо початкових умов*, якщо для функції $z(t) = \tilde{u}(t) - u(t)$, де $\tilde{u}(t)$, $u(t)$ — відповідно розв'язки задач

$$\frac{du}{dt} = f(t, u), \quad u(t_0) = u_0 \tag{6}$$

та

$$\frac{d\tilde{u}}{dt} = f(t, \tilde{u}), \quad \tilde{u}(t_0) = \tilde{u}_0, \tag{7}$$

справедлива оцінка

$$\|z\|_{(1)} \leqslant C|\tilde{u}_0 - u_0|.$$

Тут C — стала, незалежна від $\tilde{u}_0, u_0$; $\|\cdot\|_{(1)}$ — деяка норма у просторі функцій $z(t)$, заданих на проміжку $[t_0, T]$, $T \leqslant \infty$. У цьому разі кажуть також, що розв'язок задачі (5) стійкий щодо початкових умов.

Означення 2. Задача (5) називається *стійкою щодо правої частини*, якщо для функції $z(t) = \tilde{u}(t) - u(t)$, де $\tilde{u}$, u — розв'язки задач

$$\frac{du}{dt} = f(t, u), \quad u(t_0) = u_0; \tag{8}$$

$$\frac{d\tilde{u}}{dt} = \tilde{f}(t, \tilde{u}), \quad \tilde{u}(t_0) = \tilde{u}_0;$$

$$\tilde{f}(t, \tilde{u}) = f(t, \tilde{u}) + \delta f(t), \tag{9}$$

має місце оцінка

$$\|z\|_{(1)} \leqslant C\|\delta f(t)\|_{(2)}.$$

Тут C — незалежна від f, $\tilde{f}$, u, $\tilde{u}$ стала; $\|\cdot\|_{(1)}$, $\|\cdot\|_{(2)}$ — деякі норми у просторі функцій, заданих на проміжку $[t_0, T]$, $T \leqslant \infty$.

Означення 3. Задача (5) називається *стійкою*, якщо для функції $z(t) = \tilde{u}(t) - u(t)$, де $\tilde{u}$, u — розв'язки задач

$$\frac{du}{dt} = f(t, u), \quad u(t_0) = u_0; \tag{10}$$

$$\frac{d\tilde{u}}{dt} = \tilde{f}(t, \tilde{u}), \quad \tilde{u}(t_0) = \tilde{u}_0, \tag{11}$$

має місце оцінка

$$\|z\|_{(1)} \leqslant C_1 \|\delta f(t)\|_{(2)} + C_2 |\tilde{u}_0 - u_0|.$$

Тут C_1, C_2 — незалежні від u_0, $\tilde{u}_0$, f, $\tilde{f}$, u, $\tilde{u}$ — сталі, $\|\cdot\|_{(1)}$, $\|\cdot\|_{(2)}$ — деякі норми у просторі функцій, заданих на проміжку $[t_0, T]$, $T \leqslant \infty$. У цьому випадку кажуть також, що розв'язок задачі (5) стійкий щодо початкових даних та правої частини.

Не обмежуючи загальності, покладемо для зручності $t_0 = 0$. Для функції z з означення 1 дістанемо задачу

$$\frac{dz}{dt} = \alpha(t) z, \quad t \in (0, T]; \quad z(0) \equiv z_0 = \tilde{u}_0 - u_0, \tag{12}$$

де

$$\alpha(t) = [f(t, \tilde{u}) - f(t, u)]/z = f'_u(t, u + \theta z), \quad \theta \in [0, 1].$$

Розв'язком задачі (12) є вираз

$$z(t) = z_0 e^{\int_0^t \alpha(\tau) d\tau}.$$

Звідси випливає, що при $f'_u \leqslant 0$ для всіх t, u матимемо $|z(t)| \leqslant z(0)$, або $\|\tilde{u} - u\|_{C[0,T]} \leqslant |\tilde{u}_0 - u_0|$, $T \leqslant \infty$. Зазначимо, що умова $f'_u \leqslant 0$ означає незростання функції $f(t, u)$ по u або монотонність по u. Таким чином, ми довели наступне твердження.

Теорема 1. *Якщо для всіх t, u виконується нерівність*

$$f'_u(t, u) \leqslant 0, \tag{13}$$

то

$$\|\tilde{u} - u\|_{C[0,T]} \leqslant |\tilde{u}_0 - u_0|, \quad T \leqslant \infty, \tag{14}$$

тобто задача (5) *стійка щодо початкових умов.*

Нехай $z(t)$ — функція з означення 3, причому

$$\tilde{f}(t, \tilde{u}) = f(t, \tilde{u}) + \delta f(t).$$

Тоді

$$\frac{dz}{dt} = \alpha(t) z + \delta f, \quad z(0) \equiv z_0 = \tilde{u}_0 - u_0.$$

Нехай $z(t) = z_1(t) + z_2(t)$, де $z_1(t)$ і $z_2(t)$ — розв'язки таких задач:

$$\frac{dz_1}{dt} = \alpha(t) z_1, \quad z_1(0) = z_0; \quad \frac{dz_2}{dt} = \alpha(t) z_2 + \delta f, \quad z_2(0) = 0.$$

Для z_1 при $f'_u(t, u) \leqslant 0$ маємо оцінку $|z_1(t)| \leqslant |z_0| = |\tilde{u}_0 - u_0|$ для всіх $t \in [0, T]$, $T \leqslant \infty$. Рівняння для z_2 помножимо на z_2 і проінтегруємо від 0 до t:

$$\int_0^t \frac{dz_2(s)}{ds} z_2(s)\, ds = \int_0^t \alpha(s) z_2^2(s)\, ds + \int_0^t \delta f(s) z_2(s)\, ds.$$

При $f'_u \leqslant 0$ маємо

$$\frac{1}{2}\int_0^t \frac{dz_2^2(s)}{ds}\, ds = \int_0^t \alpha(s) z_2^2(s)\, ds + \int_0^t \delta f(s) z_2(s)\, ds,$$

$$\frac{1}{2} z_2^2(t) \leqslant \int_0^t \delta f(s) z_2(s)\, ds \leqslant \int_0^T |\delta f(s)|\, |z_2(s)|\, ds \leqslant$$

$$\leqslant \|\delta f(s)\|_{C[0,T]} \cdot T \cdot \|z_2\|_{C[0,T]}, \tag{15}$$

де

$$\|v\|_{C[0,T]} = \max_{x \in [0,T]} |v(x)|,$$

тобто

$$\|z_2\|_{C[0,T]} \leqslant 2T \|\delta f(s)\|_{C[0,T]}, \quad T < \infty,$$

що означає стійкість задачі (5) щодо правої частини. Для $z(t)$ маємо нерівність

$$\|z\|_{C[0,T]} \leqslant 2T \|\delta f\|_{C[0,T]} + |\tilde{u}_0 - u|, \quad T < \infty. \tag{15'}$$

Таким чином, ми довели наступну теорему.

Теорема 2. *Якщо $f'_u(t, u) \leqslant 0$ для всіх t, u, то на будь-якому скінченному відрізку $[0, T]$ для задачі* (5) *виконується нерівність* (15'), *тобто вона є стійкою.*

За умов $f'_u \leqslant 0$, $\delta f \in L_1[0, T]$ з (15) за допомогою нерівності Коші — Буняковського можна дістати таку оцінку:

$$\frac{1}{2} z_2^2(t) \leqslant \|z_2\|_{C[0,T]} \|\delta f\|_{L_1[0,T]},$$

звідки

$$\|z_2\|_{C[0,T]} \leqslant 2\|\delta f\|_{L_1[0,T]}.$$

Це означає стійкість задачі (5) щодо правої частини як на скінченному відрізку $[0, T]$, так і на нескінченному $[0, \infty)$. Для $z(t)$ маємо нерівність

$$\|z\|_{C[0,T]} \leqslant 2\|\delta f\|_{L_1[0,T]} + |\tilde{u}_0 - u_0|. \qquad (15'')$$

Отже, доведено таку теорему.

Теорема 3. *Якщо функція $f(t, u)$ неперервна в області $\Omega = \{(t, u) : t \in [0, T], u \in (-\infty, \infty), T \leqslant \infty\}$, задовольняє умову Ліпшиця по u і $f'_u(t, u) \leqslant 0$ при $(t, u) \in \Omega$, то задача (5) є коректно поставленою, причому мають місце нерівності стійкості (15'), (15'').*

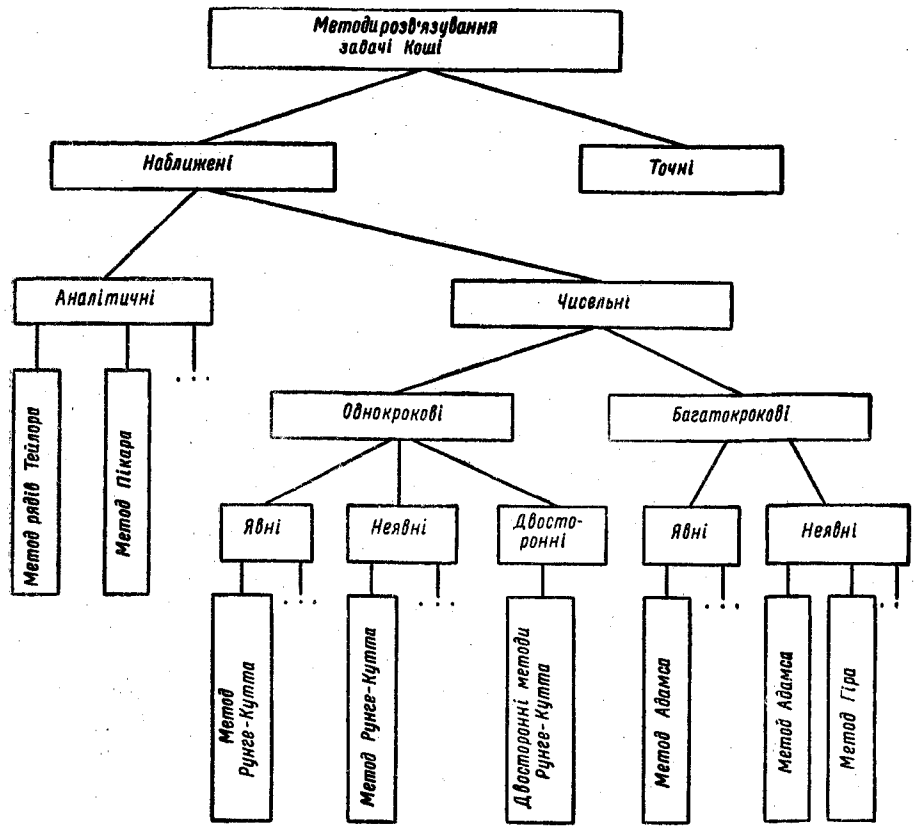


Рис. 1

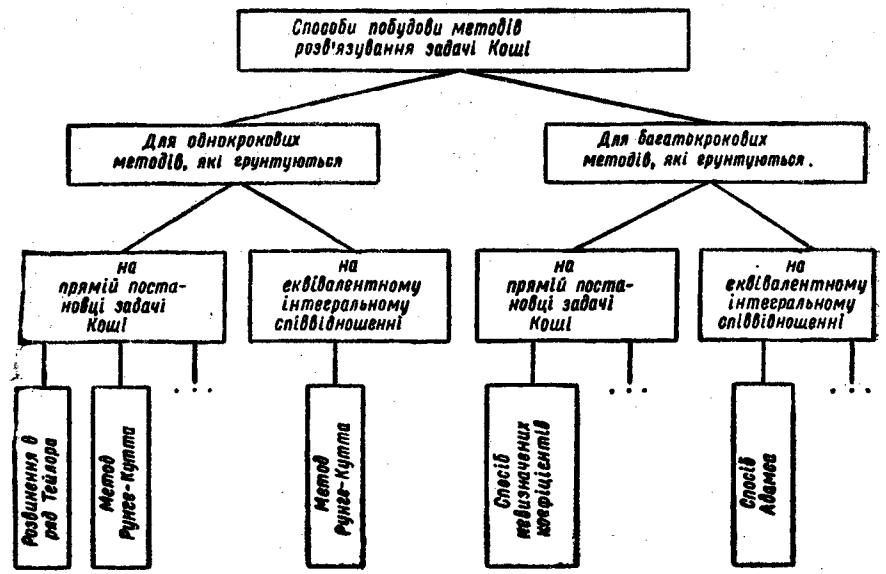


Рис. 2

З викладеного вище випливає також, що з точки зору стійкості розв'язку задача (5) аналогічна лінійній задачі

$$\frac{du}{dt} + \lambda u = 0, \quad t \in (0, T], \qquad u(0) = u_0, \qquad (16)$$

яку можна розглядати як модельну задачу при вивченні стійкості чисельних методів. Розв'язком цієї задачі є

$$u(t) = u_0 e^{-\lambda t},$$

ця функція не зростає при $\lambda \geqslant 0$, тобто

$$|u(t)| \leqslant |u_0| \quad \forall\, t \geqslant 0.$$

Умова $\lambda \geqslant 0$ відповідає умові $f'_u \leqslant 0$ для задачі (5).

Методи розв'язування задачі Коші можна (далеко не повністю) класифікувати так, як показано на рис. 1.

Можна також класифікувати способи побудови (конструювання) чисельних методів розв'язування задачі Коші (рис. 2).

3.2. Метод рядів Тейлора

Припустимо, що в задачі (5) з п. 3.1 функцію $f(t, u)$ у точці (x_0, u_0) можна розвинути в ряд вигляду

$$f(t, u) = \sum_{\alpha_0, \alpha_1} c_{\alpha_0 \alpha_1} (t - t_0)^{\alpha_0} (u - u_0)^{\alpha_1}, \tag{1}$$

$$c_{\alpha_0 \alpha_1} = \frac{1}{\alpha_0!\, \alpha_1!} \left[\frac{\partial^{\alpha_0 + \alpha_1}}{\partial t^{\alpha_0} \partial u^{\alpha_1}} f(t, u) \right]_{\substack{t = t_0 \\ u = u_0}},$$

де α_0, α_1 — цілі невід'ємні числа, $c_{\alpha_0 \alpha_1}$ — сталі коефіцієнти (іншими словами, функція $f(x, y)$ є аналітичною в точці x_0, u_0). Можна показати, що у цьому випадку розв'язок задачі (5) з п. 3.1 також є аналітичним в точці x_0 (теорема Ковалевської), і ми можемо записати наближену рівність

$$u(t) \approx \sum_{k=0}^{n} \frac{u^{(k)}(t_0)}{k!} (t - t_0)^k \tag{2}$$

для всіх t таких, що $|t - t_0| < \rho$, де ρ — радіус збіжності ряду Тейлора (в цьому випадку похибка формули (2) прямує до нуля при $n \to \infty$). Щоб визначити $u^{(k)}(t_0)$, послідовно продиференціюємо по t рівняння (5) з п. 3.1:

$$\begin{aligned} u''(t_0) &= f_t(t_0, u_0) + f_u(t_0, u_0) f(t_0, u_0), \\ u'''(t_0) &= f_{tt}(t_0, u_0) + 2 f_{tu}(t_0, u_0) f(t_0, u_0) + \\ &+ f_{uu}(t_0, y_0) [f(t_0, u_0)]^2 + f_u(t_0, u_0) f_u u''(t_0). \end{aligned} \tag{3}$$

Таким чином, метод розвинення в ряд Тейлора вимагає обчислення значення функції $f(t, u)$ і її похідних, а тому застосування цього методу можна вважати доцільним для випадку розв'язування одного і того самого диференціального рівняння вигляду (5) з п. 3.1 за різних початкових умов з використанням програм аналітичного диференціювання функцій (такі програми досить високої якості існують, наприклад програма REDUCE).

Оцінку похибки цього методу можна дістати, використовуючи оцінку залишкового члена формули Тейлора.

Приклад 1. Знайти відрізок ряду Тейлора розв'язку задачі Коші

$$u'(t) = u^2, \quad u(0) = 1$$

(точним розв'язком цієї задачі є, як неважко переконатися, $u(t) = \frac{1}{1 - t}$).

Розв'язування. Шукаємо розв'язок у вигляді

$$u(t) = u(0) + u'(0)\, t + \frac{u''(0)}{2!} t^2 + \cdots.$$

З початкової умови маємо $u(0) = 1$, а з рівняння — $u'(0) = u^2(0) = 1$. Диференціюючи послідовно рівняння $u' = u^2$, дістаємо

$$u''(t) = 2u(t) \cdot u'(t),$$

звідки з урахуванням попередніх рівностей маємо $u''(0) = 2u(0) \cdot u'(0) = 2$. Тоді

$$u'''(t) = 2[u'(t)]^2 + 2u(t)\, u''(t), \quad u'''(0) = 2 + 4 = 6,$$

$$u^{(4)}(t) = 6u'(t)\, u''(t) + 2u(t)\, u'''(t), \quad u^{(4)}(0) = 12 + 12 = 24.$$

Отже,

$$u(t) = 1 + t + t^2 + t^3 + t^4 + \cdots.$$

Неважко знайти і загальний член цього ряду, а також дослідити його збіжність до точного розв'язку. Зробіть це самостійно.

У випадку системи диференціальних рівнянь

$$\frac{du}{dt} = f(t, u), \quad u(t_0) = u_0,$$

$$u = \begin{pmatrix} u_1(t) \\ \cdots \\ u_n(t) \end{pmatrix}, \quad f(t, u) = \begin{pmatrix} f_1(t, u_1, \ldots, u_n) \\ \cdots \\ f_n(t, u_1, \ldots, u_n) \end{pmatrix}, \quad u_0 = \begin{pmatrix} u_{01} \\ \cdots \\ u_{0n} \end{pmatrix} \tag{3'}$$

можна діяти аналогічно. Якщо компоненти вектор-функції правої частини рівняння у точці $(t_0, u_{01}, \ldots, u_{0n})$ є аналітичними, тобто у деякому околі цієї точки її можна розвинути в ряди

$$f_i(t, u_1, \ldots, u_n) = \sum_{\alpha_0 \alpha_1 \ldots \alpha_n} c^{(i)}_{\alpha_0 \ldots \alpha_n} (t - t_0)^{\alpha_0} (u_1 - u_{01})^{\alpha_1} \ldots (u_n - u_{0n})^{\alpha_n},$$

$$c^{(i)}_{\alpha_0 \ldots \alpha_n} = \frac{1}{\alpha_0! \ldots \alpha_n!} \times$$

$$\times \left[\frac{\partial^{\alpha_0 + \alpha_1 + \cdots + \alpha_n}}{\partial t^{\alpha_0} \partial u_1^{\alpha_1} \ldots \partial u_n^{\alpha_n}} f_i(t, u_1, \ldots, u_n) \right]_{t = t_0}, \quad i = \overline{1, n},$$

то розв'язок задачі Коші є також аналітичною функцією, його можна подати рядами

$$u_i(t) = \sum_{k=0}^{\infty} \frac{u_i^{(k)}(t_0)}{k!} (t - t_0)^k, \quad i = \overline{1, n}, \tag{4}$$

де $u_i(t_0) = u_{0i}$, $u_i'(t_0) = f_i(t_0, u_1(t_0), \ldots, u_n(t_0))$, $i = \overline{1, n}$.

Для визначення наступних коефіцієнтів продиференціюємо рівняння (3'):

$$\frac{d^2 u}{dt^2} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial u} \frac{du}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial u} \cdot f(t, u),$$

де

$$\frac{\partial f}{\partial t} = \begin{pmatrix} \frac{\partial f_1}{\partial t} \\ \dots \\ \frac{\partial f_n}{\partial t} \end{pmatrix},$$

$$\frac{\partial f}{\partial u} = \begin{pmatrix} \frac{\partial f_1}{\partial u_1} & \frac{\partial f_1}{\partial u_2} & \dots & \frac{\partial f_1}{\partial u_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial f_n}{\partial u_1} & \frac{\partial f_n}{\partial u_2} & \dots & \frac{\partial f_n}{\partial u_n} \end{pmatrix} = f'_u.$$

Звідси

$$u^{(2)}(t_0) = f'_t(t_0, u_0) + f'_u(t_0, u_0) f(t_0, u_0).$$

Для третьої похідної

$$u^{(3)}(t) = f_{tt} + 2f_{tu}u' + f_{uu}u'u' + f_u u'',$$

де $f_{t^l u^k} = \dfrac{\partial^{l+k} f_i}{\partial t^l \partial u_{j1} \dots \partial u_{jk}}$ — тензор відповідного порядку. Аналогічно знаходять похідні вищих порядків.

Приклад 2. Знайти перші три члени розвинення в ряд Тейлора розв'язку задачі

$$\frac{du}{dt} = f(t, u),\ u(t_0) = u_0,\quad u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

де $f_1(t, u_1, u_2) = u_1 \cos t - u_2 \sin t$, $f_2(t, u_1, u_2) = u_1 \sin t + u_2 \cos t$, $t_0 = 0$, $u_1(0) = 1$, $u_2(0) = 0$.

Розв'язування. Шукаємо розв'язок у вигляді

$$u_1(t) = u_1(0) + u'_1(0)\, t + \frac{u''(0)}{2!} t^2 + \frac{u'''(0)}{3!} t^3 + \dots,$$

$$u_2(t) = u_2(0) + u'_2(0)\, t + \frac{u''_2(0)}{2!} t^2 + \frac{u'''(0)}{3!} t^3 + \dots.$$

З початкових умов дістаємо $u_1(0) = 1$, $u_2(0) = 0$, а з системи рівнянь при $t = 0$

$$u'_1(0) = 1,\quad u'_2(0) = 0.$$

Диференціюючи систему рівнянь по t, маємо

$$\begin{aligned} \frac{d^2u_1(t)}{dt^2} &= -u_1(t)\sin t - u_2(t)\cos t + \frac{du_1(t)}{dt}\cos t - \frac{du_2(t)}{dt}\sin t, \\ \frac{d^2u_2(t)}{dt^2} &= u_1(t)\cos t - u_2(t)\sin t + \frac{du_1(t)}{dt}\sin t + \frac{du_2(t)}{dt}\cos t, \end{aligned} \tag{5}$$

звідки $u''_1(0) = 1$, $u''_2(0) = 1$. Диференціюючи (5) ще раз по t, дістанемо

$$\frac{d^3u_1(t)}{dt^3} = -u_1(t)\cos t + u_2(t)\sin t + 2\left(-\frac{du_1(t)}{dt}\sin t - \frac{du_2(t)}{dt}\cos t\right) + \frac{d^2u_1(t)}{dt^2}\cos t - \frac{d^2u_2(t)}{dt^2}\sin t,$$

$$\frac{d^3u_2(t)}{dt^3} = -u_1(t)\sin t - u_2(t)\cos t + 2\left(\frac{du_1(t)}{dt}\cos t - \frac{du_2(t)}{dt}\sin t\right) + \frac{d^2u_1(t)}{dt^2}\sin t + \frac{d^2u_2(t)}{dt^2}\cos t.$$

Отже, $u'''_1(0) = -1 + 1 = 0$, $u'''_2(0) = 3$, маємо такі відрізки рядів:

$$u_1(t) = 1 + t + \frac{t^2}{2} + \dots,$$

$$u_2(t) = \frac{t^2}{2} + \frac{t^3}{2} + \dots.$$

Використовуючи описані вище методи, доцільно розбити відрізок $[t_0, T]$ точками t_j, $j = \overline{0, n}$, і шукати розв'язок за формулою (2) відповідно на відрізках $[t_0, t_1]$, $[t_1, t_2]$, ..., $[t_{n-1}, T]$, приймаючи на $[t_j, t_{j+1}]$ за початкове значення $y(t_j)$, знайдене на попередньому відрізку. Такий ускладнений метод рядів Тейлора належить до числа однокрокових методів розв'язування задачі Коші, тобто таких методів, які на кожному кроці дають змогу знайти наближений розв'язок через наближений розв'язок, знайдений на попередньому кроці. Одним з найбільш поширених однокрокових методів є метод Рунге — Кутта.

3.3. Методи Рунге—Кутта

3.3.1. Ідея методу. Нехай $u(x)$ — точний розв'язок задачі (5) з п. 3.1. При достатній гладкості функції $f(t, u)$ має місце розвинення

$$u(t_1) - u(t_0) = \sum_{k=1}^{s} \frac{\tau^k}{k!} u^{(k)}(t_0) + O(\tau^{s+1}), \tag{1}$$

де $t_1 = t_0 + \tau$, $\tau > 0$.

За співвідношенням (3) з п. 3.2 коефіцієнти при степенях τ у правій частині (1), виключаючи член $O(\tau^{s+1})$, виражаються через значення функції f та її похідних у точці (t_0, u_0).

Крім того, якщо розглянути лінійну комбінацію $L(f)$ значень функції f у точках (ξ_i, η_i), $i = \overline{1, r}$,

$$L(f) = \sum_{i=1}^{r} p_i \tau f(\xi_i, \eta_i) = \sum_{i=1}^{r} p_i k_i(\tau), \tag{2}$$

$$\xi_i = t_0 + \alpha_i \tau, \quad \alpha_1 = 0 \quad (\text{при } \alpha_1 \neq 0 \text{ матимемо неявну формулу}), \tag{3}$$

$$\eta_i = u_0 + \beta_{i1} k_1 + \beta_{i2} k_2 + \cdots + \beta_{i,i-1} k_{i-1}, \tag{4}$$

$$k_i = k_i(\tau) = \tau f(\xi_i, \eta_i), \tag{5}$$

де p_i, α_i, β_{ij}, $0 < j < i \leqslant r$ — сталі коефіцієнти, то цю лінійну комбінацію також можна записати у вигляді суми $L(f) = P(\tau) + R(\tau)$. Тут $P(\tau)$ — поліном від τ, коефіцієнти якого, як і в (1), виражаються через значення функції f, її похідні в точці (t_0, u_0) і сталі p_i, α_i, β_{ij}, а $R(\tau)$ — деякий залишковий член.

Тому можна записати рівність

$$u(t_0+\tau) - u(t_0) = \sum_{i=1}^{r} p_i k_i(\tau) + R_1(\tau), \tag{6}$$

де $R_1(\tau)$ — деякий залишковий член. Введемо на відрізку інтегрування $t_0 \leqslant t \leqslant T$ сітку $\omega_\tau = \{t_n = t_0 + n\tau,\ n = 0, 1, \ldots\}$ і позначимо через $y_n = y(t_n)$ функцію, задану на сітці ω_τ. Запишемо згідно з (6) співвідношення

$$y_1 - y_0 = \sum_{i=1}^{r} p_i k_i(\tau). \tag{7}$$

Якщо член $R_1(\tau)$ малий і $y_0 = u_0$, то y_1 наближатимемо $u(t_0 + \tau)$. Далі можна взяти y_1 за початкове значення і за тією самою формулою знайти y_2 (наближення для $u(t_2)$ і т. д.).

Розглянемо похибку $z(t_1) \equiv z(t_1, \tau) \equiv z_1(\tau) = -u(t_0 + \tau) + y(t_1) = -u(t_1) + u(t_0) + \sum_{i=1}^{r} p_i k_i(\tau)$. За формулою Тейлора при достатній гладкості функцій $f(t, u)$ і $u(t)$ маємо

$$z_1(\tau) = \sum_{i=0}^{s} \frac{\tau^i z_1^{(i)}(0)}{i!} + \tau^{s+1} \frac{z_1^{(s+1)}(\theta\tau)}{(s+1)!}, \quad |\theta| < 1. \tag{8}$$

Підберемо коефіцієнти p_i, α_i, β_{ij} таким чином, щоб для якомога більшого s виконувалися рівності

$$z_1(0) = z_1'(0) = \cdots = z_1^{(s)}(0) = 0. \tag{9}$$

Тоді величина

$$z_1(\tau) = \frac{\tau^{s+1} z_1^{(s+1)}(\theta\tau)}{(s+1)!} \tag{10}$$

(при достатній гладкості розв'язку і за умови, що $y_0 = u_0$ — точне значення) буде виражати відхилення наближеного значення y_1 від точного розв'язку $u(t)$ у точці t_1. Будемо називати її *похибкою* методу Рунге — Кутта на кроці, а s — порядком (ступенем) точності формули Рунге — Кутта (7).

Зауваження. Це означення порядку точності має зміст при достатній гладкості шуканого розв'язку. Далі покажемо, що при обмеженій гладкості порядок точності методу залежить від параметрів, які характеризують цю гладкість, а також від норми, в якій оцінюється точність. Природнішим означенням порядку точності сіткового методу є означення, наведене у гл. 2.

Нехай головну (відносно τ) частину похибки (10) формули (7) при обраних значеннях параметрів p_i, α_i, β_{ij} $(0 < j < i \leqslant r)$, які є функціями деякого параметра γ, можна подати у вигляді

$$z_1(\tau) = \gamma \tau^{s+1} \psi[f]_0 + O(\tau^{s+2}), \tag{11}$$

де $\psi[f]_0$ — деякий функціонал від функції f, обчислюваної у точці (t_0, u_0). Якщо $\gamma \neq 0$ і $\psi[f]_0 \neq 0$, τ достатньо мале, то формули вигляду (7) для двох значень параметра γ, які відрізняються лише знаком, будуть давати верхні та нижні наближення до шуканого розв'язку $u(t_1)$. Такі формули Рунге — Кутта, які відповідають значенням $\pm\gamma$, називаються *формулами двостороннього методу Рунге — Кутта.*

3.3.2. Приклад побудови формул Рунге — Кутта. Метод Ейлера. Нехай $r = 1$. Тоді згідно з (2) — (5), (7) — (9)

$$y_1 = y_0 + p_1 k_1(\tau), \quad k_1 = \tau f(t_0, y_0),$$

$$z_1(\tau) = -u(t_1) + u(t_0) + p_1 \tau f(x_0, y_0),$$

$$z_1'(0) = [-u'(t_0 + \tau) + p_1 f(x_0, y_0)]_{\tau=0} = (-1 + p_1) f(x_0, y_0),$$

$$z_1''(0) = -u''(t_0).$$

Звідси $z_1(0) = z_1'(0) = 0$ для будь-яких функцій $f(x, y)$, якщо $p_1 = 1$, а $z_1''(0)$ відмінна від нуля. Таким чином, формула

$$y_1 = y_0 + \tau f(x_0, y_0) \tag{12}$$

має перший порядок точності

$$z_1(\tau) = \frac{\tau^2}{2!} u''(t_0) + \cdots = \frac{\tau^2}{2!} z_1''(\theta\tau), \quad |\theta| < 1. \tag{13}$$

Формула (12) називається формулою Ейлера (або різницевою схемою Ейлера), і вона визначає однокроковий метод розв'язування задачі Коші (5) з п. 3.1. Метод Ейлера іноді називають методом ламаних. Визначивши y_1, можна за тією самою схемою визначити y_2

і т. д., тобто

$$y_{n+1} = y_n + \tau f(t_n, y_n), \quad n = 0, 1, \ldots, \quad y_0 = u_0, \tag{14}$$

або

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n, y_n), \quad n = 0, 1, \ldots, \quad y_0 = u_0. \tag{15}$$

Знайдемо рівняння для похибки $z_n = y_n - u_n$. Для цього підставимо $y_n = z_n + u_n$ в (15) і врахуємо, що

$$y_{n+1} - y_n = (z_{n+1} - z_n) + (u_{n+1} - u_n),$$
$$f(t_n, y_n) = f(t_n, u_n) + [f(t_n, u_n + z_n) - f(t_n, u_n)].$$

В результаті для z_n дістанемо задачу

$$\frac{z_{n+1} - z_n}{\tau} = [f(t_n, u_n + z_n) - f(t_n, u_n)] + \psi_n, \quad n = 0, 1, \ldots, \quad z_0 = 0, \tag{16}$$

де

$$\psi_n = f(t_n, u_n) - \frac{u_{n+1} - u_n}{\tau} = \frac{\partial u_n}{\partial t} - \frac{u_{n+1} - u_n}{\tau} \tag{17}$$

— нев'язка, або похибка апроксимації схеми Ейлера (15) на розв'язку $u = u(t)$ задачі (5) з п. 3.1.

3.3.3. Приклад побудови формули Рунге — Кутта другого порядку точності. Нехай $r = 2$. Шукаємо формулу Рунге — Кутта у вигляді

$$y_1 = y_0 + p_1 k_1(\tau) + p_2 k_2(\tau),$$
$$k_1(\tau) = \tau f(t_0, y_0) = \tau f_0,$$
$$k_2(\tau) = \tau f(t_0 + \alpha_2\tau, y_0 + \beta_{21} k_1) = \tau f(t^*, y^*), \tag{18}$$
$$t^* = t_0 + \alpha_2\tau, \quad y^* = y_0 + \beta_{21}\tau f(\tau_0, y_0).$$

Далі маємо

$$z_1(\tau) = y(t_1) - u(t_1) = u(t_0) - u(t_1) + p_1 k_1(\tau) + p_2 k_2(\tau); \tag{19}$$
$$z_1(0) = 0,$$
$$z_1'(0) = \{-u'(t_0 + \tau) + p_1 f_0 + p_2 f(t^*, y^*) + p_2\tau[\alpha_2 f_t'(t^*, y^*) +$$
$$+ \beta_{21} f_u'(t^*, y^*) f_0]\}_{\tau=0} = \{-f(t_1, u_1) + p_1 f_0 + p_2 f(t^*, y^*) +$$
$$+ p_2\tau[\alpha_2 f_t'(t^*, y^*) + \beta_{21} f_u'(t^*, y^*) f_0]\}_{\tau=0} = (-1 + p_1 + p_2) f_0,$$
$$z_1''(0) = \{-f_t'(t_1, u_1) - f_u'(t_1, u_1) u_t'(t_1) + p_2\alpha_2 f_t'(t^*, y^*) +$$
$$+ p_2\beta_{21} f_u'(t^*, y^*) f_0 + p_2[\alpha_2 f_t'(t^*, y^*) + \beta_{21} f_u'(t^*, y^*) f_0] +$$
$$+ p_2\tau[\alpha_2^2 f_{t^2}''(t^*, y^*) + \alpha_2\beta_{21} f_{tu}''(t^*, y^*) f_0 + \alpha_2\beta_{21} f_{ut}''(t^*, y^*) f_0 +$$
$$+ \beta_{21}^2 f_{u^2}''(t^*, y^*) f_0^2]\}_{\tau=0} = \{-f_t'(t_1, u_1) - f_u'(t_1, u_1) f(t_1, u_1) +$$
$$+ 2p_2[\alpha_2 f_t'(t^*, y^*) + \beta_{21} f_u'(t^*, y^*) f_0] + p_2\tau[\alpha_2^2 f_{t^2}''(t^*, y^*) + \tag{20}$$
$$+ 2\alpha_2\beta_{21} f_{tu}''(t^*, y^*) f_0 + \beta_{21}^2 f_{u^2}''(t^*, y^*) f_0^2]\}_{\tau=0} =$$
$$= (-1 + 2p_2\alpha_2) f_t'(t_0, u_0) + (-1 + 2p_2\beta_{21}) f_u'(t_0, u_0) f_0,$$
$$z_1'''(0) = \{-f_{t^2}''(t_1, u_1) - f_{tu}''(t_1, u_1) u_t'(t_1) - [f_{ut}''(t_1, u_1) +$$
$$+ f_{u^2}''(t_1, u_1) u_t'(t_1)] f(t_1, u_1) - f_u'(t_1, u_1)[f_t'(t_1, u_1) - f_u'(t_1, u_1) u_t'(t_1)] +$$
$$+ 2p_2[\alpha_2^2 f_{t^2}''(t^*, y^*) + \alpha_2\beta_{21} f_{tu}''(t^*, y^*) f_0 + \alpha_2\beta_{21} f_{ut}''(t^*, y^*) f_0 +$$
$$+ \beta_{21}^2 f_{u^2}''(t^*, y^*) f_0^2] + p_2[\alpha_2^2 f_{t^2}''(t^*, y^*) + 2\alpha_2\beta_{21} f_{tu}''(t^*, y^*) f_0 +$$
$$+ \beta_{21}^2 f_{u^2}''(t^*, y^*) f_0^2] + p_2\tau[\alpha_2^3 f_{t^3}'''(t^*, y^*) + \alpha_2^2\beta_{21} f_0 f_{t^2u}'''(t^*, y^*) +$$
$$+ 2\alpha_2^2\beta_{21} f_{t^2u}'''(t^*, y^*) f_0 + 2\alpha_2\beta_{21}^2 f_0^2 f_{tu^2}'''(t^*, y^*) + \alpha_2\beta_{21}^2 f_0^2 f_{u^2t}'''(t^*, y^*) +$$
$$+ \beta_{21}^3 f_0^3 f_{u^3}'''(t^*, y^*)]\}_{\tau=0} = \{-f_{t^2}''(t_1, u_1) - 2f_{tu}''(t_1, u_1) f(t_1, u_1) -$$
$$- f_{u^2}''(t_1, u_1) f^2(t_1, u_1) - f_u'(t_1, u_1) f_t'(t_1, u_1) - (f_u'(t_1, u_1))^2 f(t_1, u_1) +$$
$$+ 3p_2[\alpha_2^2 f_{t^2}''(t^*, y^*) + 2\alpha_2\beta_{21} f_{tu}''(t^*, y^*) f_0 + \beta_{21}^2 f_{u^2}''(t^*, y^*) f_0^2] +$$
$$+ p_2\tau[\alpha_2^3 f_{t^3}'''(t^*, y^*) + 3\alpha_2^2\beta_{21} f_0 f_{t^2u}'''(t^*, y^*) + 3\alpha_2\beta_{21}^2 f_0^2 f_{tu^2}'''(t^*, y^*) +$$
$$+ \beta_{21}^3 f_0^3 f_{u^3}'''(t^*, y^*)]\}_{\tau=0} = [(-1 + 3p_2\alpha_2^2) f_{t^2}'' + (-2 + 6p_2\alpha_2\beta_{21}) f_{tu}'' f +$$
$$+ (-1 + 3p_2\beta_{21}^2) f_{u^2}'' f^2 - f_u'(f_t' + f_u' f)]_0, \tag{21}$$

де $[f]_0 = f|_{t=0}$.

Із співвідношень (19) — (21) випливає, що $z_1(0) = z_1'(0) = z_1''(0) = 0$ при будь-якій функції f, коли

$$-1 + p_1 + p_2 = 0; \tag{22}$$
$$-1 + 2p_2\alpha_2 = 0; \tag{23}$$
$$-1 + 2p_2\beta_{21} = 0. \tag{24}$$

Для того щоб система (22) — (24) мала розв'язок, причому $p_2 \neq 0$, слід покласти $\alpha_2 = \beta_{21}$ (якщо $p_2 = 0$, то схема (18) збігається з відомою схемою Ейлера). Тоді (22) — (24) набуває вигляду

$$-1 + p_1 + p_2 = 0, \quad -1 + 2p_2\alpha_2 = 0$$

і ці рівності визначають однопараметричну сім'ю формул Рунге — Кутта другого порядку точності (або третього порядку на кроці). Вибираючи довільно один з параметрів, наприклад p_1, можна побудувати різні формули методу Рунге — Кутта виду (18). Наприклад, при $p_1 = 0$ маємо $p_2 = 1$, $\alpha_2 = \beta_{21} = 1/2$, і формулу Рунге — Кутта

(18) другого порядку точності записуємо у вигляді

$$y_1 = y_0 + \tau f(t_0 + \tau/2,\ y_0 + k_1/2),$$
$$k_1 = \tau f(t_0,\ y_0), \tag{25}$$
$$z_1(\tau) = \frac{\tau^3}{3!}\left\{\left[-\frac{1}{4}(f''_{t^2} + 2f''_{tu}f + f''_{u^2}f^2)\right]_0 - [f'_u(f'_t + f'_u f)]_0\right\} + O(\tau^4).$$

Цю схему можна записати інакше, вводячи проміжне значення $\bar{y}_0 = y_0 + \frac{k_1}{2}$:

$$\frac{\bar{y}_0 - y_0}{\tau/2} = f(t_0,\ y_0), \quad \frac{y_1 - y_0}{\tau} = f(t_0 + \tau/2,\ \bar{y}_0).$$

Це відома схема предиктор-коректор, або обчислення-перерахунок. Застосовуючи формулу (25) рекурентно, дістаємо таку схему Рунге — Кутта:

$$y_{n+1} = y_n + \tau f\left(t_n + \frac{\tau}{2},\ y_n + \frac{k_1}{2}\right),$$
$$k_1 = \tau f(t_n,\ y_n), \quad n = 0,\ 1,\ 2,\ \ldots,\ y_0 = u_0, \tag{25'}$$

або у вигляді схеми предиктор-коректор

$$\frac{\bar{y}_n - y_n}{\tau/2} = f(t_n,\ y_n), \quad \frac{y_{n+1} - y_n}{\tau} = f\left(t_n + \frac{\tau}{2},\ \bar{y}_n\right), \tag{25''}$$
$$n = 0,\ 1,\ \ldots,\ y_0 = u_0.$$

Схема (25′) називається також *модифікованим методом* ламаних.

Покладаючи $p_1 = p_2 = \frac{1}{2}$, $\alpha_2 = \beta_{21} = 1$, дістанемо схему Рунге — Кутта другого порядку точності

$$y_{n+1} = y_n + \frac{\tau}{2}[f(t_n,\ y_n) + f(t_{n+1},\ y_n + \tau f(t_n,\ y_n))], \quad y_0 = u_0,$$

яка називається ще *покращеним методом* ламаних.

Неважко помітити, що коли параметри $p_1,\ p_2,\ \alpha_2,\ \beta_{21}$ мають значення

$$\beta_{21} = \alpha_2 = 2/3, \quad p_2 = 3/4, \quad p_1 = 1/4, \tag{26}$$

то на класі функцій, для яких

$$f'_t f'_u + f(f'_u)^2 \equiv 0, \tag{27}$$

маємо $z'''_1(0) = 0$, і тоді $z_1(\tau) = O(\tau^4)$. Це означає, що формула Рунге — Кутта вигляду

$$y_1 = y_0 + \frac{1}{4}\tau f(t_0,\ y_0) + \frac{3}{4}\tau f\left(t_0 + \frac{2}{3}\tau,\ y_0 + \frac{2}{3}\tau f_0\right) \tag{28}$$

на класі функцій, які задовольняють умову (27), має порядок точності 3 або порядок точності 4 на кроці. До класу функцій, які задовольняють умову (27), потрапляють всі функції $f(t, u)$, що не залежать від u. Тоді

$$u_1 = \int_{t_0}^{t_0+\tau} f(x)\,dx + u_0,$$

а це означає, що формулу (28) можна використати для обчислення інтегралів.

На закінчення наведемо формули для схеми Рунге — Кутта четвертого порядку точності:

$$\frac{y_{n+1} - y_n}{\tau} = \frac{1}{6\tau}[k_1(y_n) + 2k_2(y_n) + 2k_3(y_n) + k_4(y_n)], \tag{29}$$
$$n = 0,\ 1,\ \ldots;\ y_0 = u_0,$$
$$k_1 = \tau f(t_n,\ y_n), \quad k_2 = \tau f(t_n + \tau/2,\ y_n + k_1/2),$$
$$k_3 = \tau f(t_n + \tau/2,\ y_n + k_2/2), \quad k_4 = \tau f(t_n + \tau,\ y_n + k_3).$$

3.3.4. Приклад побудови двосторонніх формул Рунге — Кутта першого порядку точності. Припустивши, що для формули вигляду (18) $z''(0) \neq 0$, де $z''(0)$ визначається формулою (20), можна дістати двосторонні формули Рунге — Кутта. Для цього покладемо

$$\alpha_2 = \beta_{21}, \quad \gamma = \frac{1}{2}(-1 + 2p_2\alpha_2) = -\frac{1}{2} + p_2\alpha_2, \tag{30}$$
$$-1 + p_1 + p_2 = 0.$$

Тоді похибка формули вигляду (18) на кроці матиме вигляд

$$z_1(\tau) = \gamma\tau^2[f'_t + f'_u f]_0 + O(\tau^3). \tag{31}$$

Для невідомих $p_1,\ p_2,\ \alpha_2,\ \gamma$ маємо систему трьох рівнянь (30). Виразимо α_2 та p_2 через p_1 і γ:

$$p_2 = 1 - p_1, \quad \alpha_2 = \frac{1 + 2\gamma}{2(1 - p_1)}, \quad \beta_{21} = \frac{1 + 2\gamma}{2(1 - p_1)} \tag{32}$$

(вважаємо, що $1 - p_1 \neq 0$). Таким чином, співвідношення (18), (32) визначають двопараметричну сім'ю двосторонніх формул Рунге — Кутта першого порядку точності (другого на кроці). Природно вимагати, щоб $0 \leqslant \alpha_2 \leqslant 1$ (α_2 — коефіцієнт при τ), тому на вибір γ будуть накладені такі обмеження:

$$0 \leqslant \frac{1 + 2\gamma}{2(1 - p_1)} \leqslant 1,$$

або

$$-1/2 \leqslant \gamma \leqslant 1/2 - p_1 \quad \text{при} \quad 1 - p_1 > 0,$$
$$1/2 - p_1 \leqslant \gamma \leqslant -1/2 \quad \text{при} \quad 1 - p_1 < 0.$$

Якщо $y_1^{(1)}$ і $y_1^{(2)}$ — наближення, знайдені для $u(t_1)$ за допомогою двосторонньої формули Рунге — Кутта (при $\pm\gamma$), то за наближений розв'язок у точці $t_1 = t_0 + h$ природно взяти

$$y_1 = \frac{y_1^{(1)} + y_1^{(2)}}{2}. \tag{33}$$

Оскільки головні (відносно h) частини похибки на кроці двосторонніх формул Рунге — Кутта рівні за абсолютною величиною і протилежні за знаком, то маємо

$$|y_1 - u(x_1)| = \left|\frac{y_1^{(1)} + y_1^{(2)}}{2} - u(x_1)\right| = O(\tau^3)$$

замість $O(\tau^2)$.

Як вже зазначалося вище, розрахунок за будь-якою з однокрокових формул починається з обчислення y_1 за заданим y_0, потім $x_1 = x_0 + \tau$ беруть за початкову точку, а y_1 — за початкове значення розв'язку і за цими значеннями обчислюють y_2 в точці $x_2 = x_1 + \tau$ і т. д. У двосторонніх формулах типу Рунге — Кутта на j-му кроці обчислюють два значення $y_j^{(1)}$ і $y_j^{(2)}$. Нехай для визначеності $y_j^{(1)} < y_j^{(2)}$. Виходячи з $y_j^{(1)}$ і виконуючи обчислення за обома формулами двостороннього методу, знайдемо $y_{j+1}^{(1,1)}$ і $y_{j+1}^{(1,2)}$. Найменше значення з $y_{j+1}^{(1,1)}$ і $y_{j+1}^{(1,2)}$ беремо за $y_{j+1}^{(1)}$ на $(j+1)$-му кроці, тобто $y_{j+1}^{(1)} = \min(y_{j+1}^{(1,1)}, y_{j+1}^{(1,2)})$. Аналогічно, виходячи з $y_j^{(2)}$, обчислюємо $y_{j+1}^{(2,1)}$, $y_{j+1}^{(2,2)}$ і найбільше з них беремо за $y_{j+1}^{(2)}$. Якщо виявиться, що $y_{j+1}^{(1)} > y_{j+1}^{(2)}$, то двосторонній метод Рунге — Кутта не можна застосовувати. За наближення для $u(t_{j+1})$ беруть величину

$$y_{j+1} = \frac{y_{j+1}^{(1)} + y_{j+1}^{(2)}}{2}, \tag{34}$$

а величини абсолютної і відносної похибок наближеного розв'язку (34) на кроці оцінюють за формулами

$$\frac{|y_{j+1}^{(1)} - y_{j+1}^{(2)}|}{2}, \quad \frac{|y_{j+1}^{(1)} - y_{j+1}^{(2)}|}{|y_{j+1}^{(1)} + y_{j+1}^{(2)}|}. \tag{35}$$

Використання двостороннього методу Рунге — Кутта призводить до збільшення обсягу обчислень, але у більшості випадків дозволяє оцінювати межі зміни розв'язку задачі Коші, якщо обчислювальна похибка не має суттєвого впливу на вірогідність цих меж.

На закінчення дістанемо двосторонні формули Рунге — Кутта другого порядку точності для задачі $u' = f(t, u)$, $u(t_0) = u_0$ і класу функцій, які задовольняють умову (27).

Для цього виберемо $\alpha_2 = \beta_{21}$, а параметр γ введемо таким чином

$$\gamma = \frac{-1 + 3p_2\alpha_2^2}{3} = -1/3 + p_2\alpha_2^2.$$

Тоді в (19), (21) випливає, що

$$z_1(0) = z_1'(0) = z_1''(0) = 0,$$

$$z_1'''(\tau) = \frac{\gamma\tau^3}{2}[f_{t^2}'' + 2f_{tu}''f + f_{u^2}''f^2]_0 + O(\tau^4)$$

ва умов

$$-1 + p_1 + p_2 = 0; \tag{36}$$

$$-1 + 2p_2\alpha_2 = 0; \tag{37}$$

$$\alpha_2 = \beta_{21}; \tag{38}$$

$$\gamma = -1/3 + p_2\alpha_2^2. \tag{39}$$

Розв'яжемо (37) і (39) відносно $p_2\alpha_2$ і $p_2\alpha_2^2$ відповідно і поділимо перше з цих співвідношень на друге:

$$p_2\alpha_2 = 1/2, \quad p_2\alpha_2^2 = \frac{3\gamma + 1}{3}, \quad \alpha_2 = \frac{2(1 + 3\gamma)}{3} \quad (p_2 \neq 0, \ \alpha_2 \neq 0).$$

Із (37) знаходимо

$$p_2 = 3/(4(1 + 3\gamma)), \quad 1 + 3\gamma \neq 0,$$

а з (36) —

$$p_1 = 1 - p_2 = \frac{1 + 12\gamma}{4(1 + 3\gamma)}.$$

3.3.5. Формули Рунге — Кутта для системи рівнянь. Розглянемо задачу Коші для системи рівнянь

$$\frac{du}{dt} = f(t, u), \quad t > 0, \quad u(t_0) = u_0, \tag{40}$$

де $u = (u_1(t), \ldots, u_n(t))^T$ — шуканий вектор; $f = (f_1, \ldots, f_n)^T$; $u_0 = (u_{01}, \ldots, u_{0n})^T$ — заданий вектор початкових значень.

Побудова формул Рунге — Кутта для системи звичайних диференціальних рівнянь (40), яка у координатній формі має вигляд

$$u_l' = f_l(t, u_1(t), \ldots, u_n(t)),$$

$$u_l(t_0) = u_l^0, \quad l = \overline{1, n},$$

вдійснюється таким чином. Для кожного l вводяться параметри

$$\alpha_{l,i}, \ \beta_{l,i,j}, \ p_{l,i}, \quad 0 < j < i \leqslant r_l, \quad l = \overline{1, n},$$

і будуються наближені рівності вигляду

$$u_l(t+\tau) - u_l(t) \approx \sum_{m=1}^{r_l} p_{l,m} k_{l,m},$$

де

$$k_{l,1} \equiv k_{l,1}(\tau;\ u) \equiv k_{l,1}(u) = \tau f_l(t,\ u_1,\ \ldots,\ u_n),\ k_{l,m} =$$
$$= \tau f_l(t + \tau\alpha_{l,m},\ u_1 + \beta_{l,m,1}k_{1,1} + \beta_{l,m,2}k_{1,2} + \cdots +$$
$$+ \beta_{l,m,m-1}k_{1,m-1},\ u_2 + \beta_{2,m,1}k_{2,1} + \cdots + \beta_{2,m,m-1}k_{2,m-1},\ \ldots,\ u_n +$$
$$+ \beta_{n,m,1}k_{n,1} + \beta_{n,m,2}k_{n,2} + \cdots + \beta_{n,m,m-1}k_{n,m-1}),$$
$$l = \overline{1,\ n},\quad m = \overline{2,\ r},$$

причому параметри вибираються за умови $u_l(t+\tau) - u_l(t) - \sum_{m=1}^{r_l} p_{l,m},\ k_{l,m} = O(\tau^s l)$ для максимально великого. Відповідний метод Рунге—Кутта має вигляд

$$y_l(t+\tau) - y_l(t) = \sum_{m=1}^{r_l} p_{l,m} k_{l,m},\quad k_{l,m} = k_{l,m}(y).$$

Як бачимо, число вільних параметрів при побудові формул методу Рунге—Кутта для системи звичайних диференціальних рівнянь значно збільшується. Отже, збільшуються можливості для знаходження формул Рунге—Кутта одного й того самого порядку точності на кроці. Якщо вибрати $\alpha_{l,i}$, $\beta_{l,i,j}$, $p_{l,i}$, r_l однаковими для всіх $l = \overline{1,\ n}$, то в цьому випадку одно- і двосторонні формули типу Рунге—Кутта, побудовані для одного рівняння, легко переносяться на випадок системи рівнянь. Наприклад, для системи двох рівнянь

$$u'(t) = f(t,\ u(t),\ v(t)),$$
$$v'(t) = g(t,\ u(t),\ v(t)),$$

позначаючи через y і z наближені значення функцій $u(x)$ і $v(x)$, маємо таку схему четвертого порядку точності:

$$y_{n+1} = y_n + 1/6\,(k_1 + 2k_2 + 2k_3 + k_4),$$
$$z_{n+1} = z_n + 1/6\,(q_1 + 2q_2 + 2q_3 + q_4),$$

де

$$k_1 = \tau f(t_n,\ y_n,\ z_n),\ q_1 = \tau g(t_n,\ y_n,\ z_n),$$
$$k_2 = \tau f(t_n + 1/2,\ y_n + 1/2k_1,\ z_n + 1/2q_1),$$
$$q_2 = \tau g(t_n + \tau/2,\ y_n + 1/2k_1,\ z_n + 1/2q_1),\ k_3 = \tau f(t_n + \tau/2,\ y_n +$$
$$+ 1/2k_2,\ z_n + \tau/2q_2),\ q_3 = \tau g(t_n + \tau/2,\ y_n + 1/2k_2,$$

$$z_n + 1/2q_2),\ k_4 = \tau f(t_n + \tau,\ y_n + k_3,\ z_n + q_3),$$
$$q_4 = \tau g(t_n + \tau,\ y_n + k_3,\ z_n + q_3).$$

Формули Рунге—Кутта можна також будувати, виходячи в такої ідеї. Нехай розв'язок задачі Коші для системи рівнянь

$$\frac{du}{dt} = f(t,\ u),\quad u(t_0) = u_0,$$

$$u = (u_1,\ \ldots,\ u_n)^T,\ f(t,\ u) = (f_1(t,\ u_1,\ \ldots,\ u_n),\ \ldots$$
$$\ldots,\ f_n(t,\ u_1,\ \ldots,\ u_n))^T,$$

у деякій точці t вже відомий і шукається цей розв'язок в точці $t + \tau$. Запишемо відому з інтегрального та диференціального числення формулу

$$u(t+\tau) = u(t) + \tau\int_0^1 u'(t + \tau h)\,dh,$$

в якій інтеграл наближено замінимо квадратурною формулою в вузлами α_i і коефіцієнтами p_i:

$$u(t+\tau) \approx u(t) + \tau\sum_{i=1}^{m} p_i u'(t + \alpha_i\tau).$$

Використовуючи систему диференціальних рівнянь, маємо (див. (6))

$$u(t+\tau) \approx u(t) + \tau\sum_{i=1}^{m} p_i f(t + \alpha_i\tau,\ u(t + \alpha_i\tau)),\qquad (6')$$

де $u(t + \alpha_i\tau)$ також поки що невідомі. Скористаємося далі формулою

$$u(t + \alpha_i\tau) = u(t) + \tau\int_0^{\alpha_i} u'(t + \tau h)\,dh$$

і знову замінимо інтеграли на відрізках $[0,\ \alpha_i]$ квадратурними формулами з тими самими вузлами α_i і коефіцієнтами β_{ij}:

$$\int_0^{\alpha_i} g(h)\,dh \approx \sum_{j=1}^{m} \beta_{ij} g(\alpha_j),\quad i = \overline{1,\ m},$$

причому коефіцієнти β_{ij} треба належним чином вибрати (вибрані вузли α_j не обов'язково всі мають лежати у проміжку $[0,\ \alpha_i]$). Оскільки остання квадратурна формула має бути точною принаймні на константі, то

$$\alpha_i = \sum_{j=1}^{m} \beta_{ij},\quad i = \overline{1,\ m}.$$

Таким чином, з (6′) для наближень $y(t) = (y_1(t), \ldots, y_n(t))^T$ до $u(t)$ дістанемо

$$y(t+\tau) = y(t) + \tau \sum_{i=1}^{m} p_i k_i, \qquad (7')$$

$$k_i = f\left(t + \alpha_i \tau,\ y(t) + \tau \sum_{j=1}^{m} \beta_{ij} k_j\right), \quad i = \overline{1, m}, \qquad (7'')$$

де k_i — наближення для $u'(t + \alpha_i \tau)$.

Взагалі кажучи, (7″) є системою m нелінійних рівнянь для m шуканих n-компонентних векторів $k_1, \ldots, k_m$, яка, крім того, має розв'язуватися на кожному кроці. Формули (7′), (7″) визначають клас m-ступеневих формул Рунге — Кутта. Неважко помітити, що при $m = 1$, $\alpha_1 = 0$, $\beta_{11} = 0$, $p_1 = 1$ ми дістаємо метод Ейлера. Якщо $\alpha_1 = 0$ і $\beta_{ij} = 0$ для $j \geqslant i$, то із (7″) k_2 можна явно виразити через $k_1 = f(t, y(t))$, k_3 — через k_1 та k_2 і т. д. В цьому випадку дістанемо явні формули Рунге — Кутта, які широко застосовують на практиці. Коефіцієнти цих формул зручно записувати за допомогою таблиці:

α_1	β_{11}	β_{12}	$\ldots$	β_{1m}
α_2	β_{21}	β_{22}	$\ldots$	β_{2m}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
α_m	β_{m1}	β_{m2}	$\ldots$	β_{mm}
	p_1	p_2	$\ldots$	p_m

Раніше ми вже розглядали класичний метод четвертого порядку з такими коефіцієнтами:

0	0	0	0	0
1/2	1/2	0	0	0
1/2	0	1/2	0	0
1	0	0	1	0
	1/6	1/3	1/3	1/6

У літературі відомі також методи цього класу, які називаються методами Хойна:

0	0	0	0
1/2	1/2	0	0
1	−1	2	0
	1/6	2/3	1/6

0	0	0	0
1/3	1/3	0	0
2/3	0	2/3	0
	1/4	0	3/4

Останній на рівномірній сітці $t_i = t_0 + i\tau$ має вигляд

$$k_1^{(i)} = f(t_i, y_i), \quad k_2^{(i)} = f\left(t_i + \frac{\tau}{3},\ y_i + \frac{\tau}{3} k_1^{(i)}\right),$$

$$k_3^{(i)} = f\left(t_i + \frac{2}{3}\tau,\ y_i + \frac{2}{3}\tau k_2^{(i)}\right),$$

$$y_{i+1} = y_i + \tau\left(\frac{1}{4} k_1^{(i)} + \frac{3}{4} k_3^{(i)}\right).$$

Якщо вибрати α_i як вузли квадратної формули Гаусса для вагової функції $\rho(x) = 1$ і для відрізку $[0, 1]$, p_i — як відповідні коефіцієнти формули Гаусса, а коефіцієнти β_{ij} вибрати так, щоб формула

$$\sum_{j=1}^{m} \beta_{ij} g(\alpha_j) = \int_0^{\alpha_i} g(x)\, dx, \quad i = \overline{1, m}$$

була точною для всіх многочленів до $(m-1)$-го степеня включно (це буде для будь-якої формули інтерполяційного типу), то дістанемо m-ступеневий неявний метод Гаусса — Рунге — Кутта. Наприклад, для $m = 2$ маємо такі значення його коефіцієнтів:

$(3-\sqrt{3})/6$	1/4	$(3-2\sqrt{3})/12$
$(3+\sqrt{3})/6$	$(3+2\sqrt{3})/12$	1/4
	1/2	1/2

Цей метод має четвертий порядок апроксимації.

3.4. Апостеріорні оцінки похибки розв'язку задачі Коші

Одну з можливостей апостеріорної оцінки похибки чисельного розв'язку задачі Коші надають двосторонні методи, наприклад двосторонні методи Рунге — Кутта. Якщо

$$y^{(1)}(t) \leqslant u(t) \leqslant y^{(2)}(t),$$

де $u(t)$ — точний розв'язок задачі Коші, $y^{(1)}(t)$, $y^{(2)}(t)$ — наближені значення, обчислені за допомогою яких-небудь чисельних методів, то абсолютна і відносна похибки наближеного значення розв'язку

$$y_{\text{набл.}}(t) = \frac{y^{(1)}(t) + y^{(2)}(t)}{2} \qquad (1)$$

задаються відповідно формулами

$$\frac{|y^{(1)}(t) - y^{(2)}(t)|}{2} \quad \text{і} \quad \frac{|y^{(1)}(t) - y^{(2)}(t)|}{|y^{(1)}(t) + y^{(2)}(t)|}. \qquad (2)$$

Другу можливість оцінки похибки надає розглянутий нами в п. 2.10 з ч. 1 метод Рунге — Ромберга.

У випадку задачі Коші його можна трактувати таким чином. Нехай $y^{(2\tau)}(t)$ — наближений розв'язок задачі Коші в точці $t = t_n + 2\tau$, обчислений за допомогою деякого однокрокового методу s-го порядку з кроком 2τ, а $y^{(\tau)}(t)$ — наближений розв'язок у точці $t = t_n + 2\tau$, знайдений за допомогою двократного застосування тієї самої формули з кроком τ. Тоді

$$y^{(2\tau)}(t) - u(t) \approx \psi(t)(2\tau)^{s+1},$$
$$y^{(\tau)}(t) - u(t) \approx 2\psi(t)\tau^{s+1}$$

(наприклад, для одностороннього методу Рунге — Кутта s-го порядку, або $(s+1)$-го порядку на кроці $\psi = \frac{z_2^{(s+1)}(0)}{(s+1)!}$).

Звідси

$$y^{(2\tau)}(t) - y^{(\tau)}(t) \approx 2\psi(t)\tau^{s+1}(2^s - 1),$$

а абсолютна похибка на кроці оцінюється величиною

$$2\psi(t)\tau^{s+1} \approx \frac{y^{(2\tau)}(t) - y^{(\tau)}(t)}{2^s - 1}. \qquad (3)$$

Використовуючи повторний розрахунок з половинним кроком, можна також здійснити уточнення розв'язку за методом Рунге (див. п. 2.10 з ч. 1)

$$u(t) = y^{(\tau)}(t) + \frac{y^{(\tau)}(t) - y^{(2\tau)}(t)}{2^s - 1} + O(\tau^{s+2}) \equiv y^*(t) + O(\tau^{s+2}). \qquad (4)$$

Іноді таке уточнення називають також уточненням за Річардсоном. Значення $y^*(t)$ має на кроці похибку $(s+2)$-го порядку, а не $s+1$, як $y^{(\tau)}(t)$.

Ще одну можливість апостеріорної оцінки похибки дають методи різного порядку точності. Нехай, наприклад, $y^{(1)}(t)$ — наближений розв'язок, обчислений за якою-небудь однокроковою формулою s-го порядку точності, тобто (у припущенні, що $y(t-\tau) \equiv u(t-\tau)$ — точне значення)

$$u(t) - y^{(1)}(t) = O(\tau^{s+1}),$$

а $y^{(2)}(t)$ — розв'язок, обчислений за однокроковою формулою $(s+1)$-го порядку, тобто

$$u(t) - y^{(2)}(t) = O(\tau^{s+2}).$$

Тоді похибка наближеного розв'язку $y^{(1)}(t)$ на кроці при малих τ може бути досить точно наближена величиною

$$z_s(t, \tau) \equiv u(t) - y^{(1)}(t) \approx y^{(2)}(t) - y^{(1)}(t), \qquad (5)$$

тобто формула (5) дає апостеріорну оцінку похибки на кроці.

При розрахунку розв'язку $y^{(2)}(t)$ (за методом $(s+1)$-го порядку точності) доцільно використовувати такий метод, який потребує додатково до розрахунків значення $y^{(1)}(t)$ (за методом s-го порядку точності) мінімуму обсягу обчислень. Наприклад, для методів Рунге — Кутта основні витрати пов'язані з обчисленням величин $k_i(\tau)$, тобто правих частин задачі Коші в різних точках. Тому для апостеріорної оцінки похибки за третім способом необхідно використовувати так звані вкладені методи Рунге — Кутта, які для підвищення порядку точності на одиницю вимагають додаткового обчислення мінімальної кількості величин $k_i(\tau)$. Наприклад, вкладені методи Рунге — Кутта четвертого і п'ятого порядків точності відповідно мають вигляд

$$y_{n+1}^{(4)} = y_n^{(4)} + \sum_{i=1}^{5} p_i^* k_i,$$

$$y_{n+1}^{(5)} = y_n^{(5)} + \sum_{i=1}^{6} p_i k_i,$$

де коефіцієнти α_i, β_{ij}, p_i, p_i^*, знайдені Фельбергом, мають такі значення:

0						
$\frac{1}{4}$	$\frac{1}{4}$					
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$				
$\frac{12}{13}$	$\frac{1932}{2197}$	$\frac{7296}{2197}$				
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$		
$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$	
p_i	$\frac{16}{135}$	0	$\frac{6656}{12\,825}$	$\frac{28\,561}{56\,430}$	$-\frac{9}{50}$	$\frac{2}{55}$
p_i^*	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	0

Добрі результати дістаємо і за допомогою формул Дорманда та Прінса

$$y_{n+1}^{(4)} = y_n^{(4)} + \sum_{i=1}^{7} p_i^* k_i \text{ (четвертий порядок)},$$

$$y_{n+1}^{(5)} = y_n^{(5)} + \sum_{i=1}^{7} p_i k_i \text{ (п'ятий порядок)},$$

$$k_j = f\Big(x + \alpha_j\tau,\ y_j + \tau\sum_{s=1}^{j-1}\beta_{js}k_s\Big),$$

що задаються такою таблицею:

0						
$\frac{1}{5}$	$\frac{1}{5}$					
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$				
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$			
$\frac{8}{9}$	$\frac{19\,372}{6561}$	$-\frac{25\,360}{2187}$	$\frac{64\,448}{6561}$	$-\frac{212}{729}$		
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46\,732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18\,656}$	
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$
p_i	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$ 0
p_i^*	$\frac{5179}{57\,600}$	0	$\frac{7571}{16\,695}$	$\frac{393}{640}$	$-\frac{92\,097}{339\,200}$	$\frac{187}{2100}$ $\frac{1}{40}$

Цей метод має ту перевагу, що сталі множники у похибці формули п'ятого порядку мінімальні (серед аналогічних методів). Зазначимо, що кут k_7 дорівнює k_1 для наступного кроку, тобто фактично на кожному кроці треба виконати лише шість обчислень функції $f(t, u)$.

Величини похибок, які обчислюються за формулами (2), (3), (5), використовують. Якщо ця похибка виявиться більшою, ніж деяка задана величина, що характеризує бажану точність, то зменшують крок, наприклад удвоє, і знову виконують обчислення. Зменшення кроку проводиться доти, поки апостеріорна похибка не стане меншою за задану величину. Цей прийом автоматичного вибору кроку для досягнення заданої точності широко використовується у стандартних програмах розв'язування задачі Коші.

Щоб зменшити обсяг обчислень і найшвидше та найкраще «вгадати» наступний крок інтегрування в однокрокових методах, здійснюють більш досконалі процедури «прогнозування кроку». Оскільки обсяг обчислень пропорціональний кількості кроків, то у найбільш поширених програмах розв'язування задачі Коші застосовують змінний крок, а при переході від точки t_i до точки $t_{i+1} = t_i + \tau_i$ крок τ_i вибирається якомога більшим, але таким, щоб похибка в точці t_{i+1} залишалася не більшою, ніж допустима. Задачу вибору кроку можна сформулювати таким чином: виходячи з відомих t_0 і u_0, вибрати для заданого однокрокового методу такий максимально можливий крок τ, для якого

$$|z(t_0 + \tau)| \approx \varepsilon, \tag{6}$$

де $z(t_0 + \tau) = y(t_0 + \tau) - u(t_0 + \tau)$ — похибка наближеного розв'язку $y(t_0 + \tau)$ по відношенню до точного розв'язку задачі Коші $u(t_0 + \tau)$ в точці $t_0 + \tau$; ε — наперед задана величина похибки. Як правило, $\varepsilon \geqslant k\,\mathrm{eps}$, де $k \approx \max\{|u(t)| : t \in [t_0, t_0 + \tau]\}$, eps — відносна точність подання чисел в ЕОМ. При такому ε вибір τ з умови (6) гарантує, що $y(t_0 + \tau)$ збігається з $u(t_0 + \tau)$ у рамках машинної точності. Іноді замість (6) використовують умову $|z(t_0 + \tau)| \approx \eta\,|y(t_0 + \tau)| + \varepsilon$, де η і ε характеризують відповідно відносну та абсолютну похибки. Для вибору τ, яке задовольняє умову (6), існує декілька методів, приклади яких розглянемо нижче.

1. Нехай для деякого однокрокового методу порядку p в кожній точці t інтервалу, на якому шукається розв'язок, має місце апріорна оцінка похибки

$$z(t;\ \tau) = e_p(t)\,\tau^p + o(\tau^p) \approx e_p(t)\,\tau^p. \tag{7}$$

Оскільки $e_p(t_0) = 0$, то

$$e_p(t) = e_p(t) - e_p(t_0) \approx (t - t_0)\,e_p'(t_0), \tag{8}$$

$$|z(t_0 + \tau;\ \tau)| \approx \varepsilon,$$

якщо

$$\varepsilon \approx |e_p(t_0 + \tau)\,\tau^p| \approx |\tau^{p+1} e_p'(t_0)|. \tag{9}$$

Звідси можна знайти τ, якщо відоме $e_p'(t_0)$. Апостеріорну оцінку для $e_p'(t_0)$ можна знайти методом Рунге. Для цього за допомогою одного і того самого однокрокового методу (з різними кроками T та $T/2$) знаходять наближені значення $y(t_0 + T;\ T)$ (за один крок T) та $y(t_0 + T;\ T/2)$ (за два кроки $T/2$).

Тоді метод Рунге дає таку апостеріорну оцінку:

$$z\left(t_0 + T;\ \frac{T}{2}\right) \approx \frac{y(t_0 + T;\ T) - y\left(t_0 + T;\ \frac{T}{2}\right)}{2^p - 1}. \tag{10}$$

Крім того, з (7), (8) дістаємо

$$z(t_0 + T;\ T/2) \approx e_p(t_0 + T)\left(\frac{T}{2}\right)^p \approx e_p'(t_0)\,T\left(\frac{T}{2}\right)^p.$$

Звідси і з (10) маємо

$$e_p'(t_0) \approx \frac{1}{T^{p+1}} \frac{2^p}{2^p - 1}\,[y(t_0 + T;\ T) - y(t_0 + T;\ T/2)].$$

Із співвідношення (9) дістанемо формулу

$$\frac{T}{\tau} \approx \left(\frac{2^p}{2^p - 1} \frac{y(t_0 + T;\ T) - y(t_0 + T;\ T/2)}{\varepsilon}\right)^{\frac{1}{p+1}}. \tag{11}$$

Ця формула є основою наступного алгоритму вибору кроку: а) вибирається $T \leqslant T_{\max}$, де $T_{\max}$ — максимальний допустимий крок, і обчислюється $y(t_0 + T;\ T)$, $y(t_0 + T;\ T/2)$ і τ за формулою (11); б) якщо $T/\tau \gg 2$, то з (11) та (10) випливає, що похибка $z(t_0 + T;\ T/2)$ значно більша, ніж допустима похибка ε, а тому крок T має бути зменшений; T замінюється на 2τ, з новим значенням T обчислюються $y(t_0 + T;\ T)$ та $y(t_0 + T;\ T/2)$ і з (11) — відповідне τ. Ця процедура повторюється доти, доки не стане $T/\tau \leqslant 2$; в) при виконанні нерівності $T/\tau \leqslant 2$ величину $y(t_0 + T;\ T/2)$ беруть за наближення для $u(t_0 + T)$ і відбувається перехід до нового кроку заміною t_0, u_0 на $t_0 + T$, $y(t_0 + T;\ T/2)$ та T на 2τ.

2. Більш елегантний спосіб, який потребує меншої кількості обчислень правої частини, був запропонований Фельбергом і базується на вкладених методах p-го та $(p+1)$-го порядків. Розглянемо його на прикладі методів Рунге—Кутта другого та третього порядків

$$\begin{aligned} \bar{y}_{i+1} &= \bar{y}_i + \tau\Phi_{\mathrm{I}}(t_i,\ \bar{y}_i;\ \tau), \\ \hat{y}_{i+1} &= \hat{y}_i + \tau\Phi_{\mathrm{II}}(t_i,\ \hat{y}_i;\ \tau), \end{aligned} \tag{12}$$

де

$$\begin{aligned} \Phi_{\mathrm{I}}(t,\ y;\ \tau) &= \sum_{k=0}^{2} p_k f_k(t;\ y;\ \tau), \\ \Phi_{\mathrm{II}}(t,\ y;\ \tau) &= \sum_{k=0}^{3} p_k^* f_k(t,\ y;\ \tau), \end{aligned} \tag{13}$$

$$f_k = f_k(t,\ y;\ \tau) = f\left(t + \alpha_k\tau,\ y + \tau\sum_{l=0}^{k-1}\beta_{kl}f_l\right), \quad k = \overline{0,\ 3},$$

і коефіцієнти p_k, p_k^*, α_k, β_{kl} задаються таблицею:

0	0			
$\frac{1}{4}$	$\frac{1}{4}$			
$\frac{27}{40}$	$-\frac{189}{800}$	$\frac{729}{800}$		
1	$\frac{214}{891}$	$\frac{1}{33}$	$\frac{650}{891}$	
p_k	$\frac{214}{891}$	$\frac{1}{33}$	$\frac{650}{891}$	
p_k^*	$\frac{533}{2106}$	0	$\frac{800}{1053}$	$-\frac{1}{78}$

Ще однією перевагою цього методу є те, що величина f_3, знайдена на i-му кроці, дорівнює f_0 для $(i+1)$-го кроку, тобто фактично кожен крок вимагає лише двох обчислень правої частини. Якщо $\bar{y}_i = u(t_i)$, $\hat{y}_i = u(t_i)$, то з означення однокрокового методу s-го порядку випливає, що

$$\bar{y}_{i+1} - u_{i+1} = C_{\mathrm{I}}(t_i)\tau^3 + \text{члени більш високого порядку},$$

$$\hat{y}_{i+1} - u_{i+1} = C_{\mathrm{II}}(t_i)\tau^4 + \text{члени більш високого порядку},$$

звідки

$$\bar{y}_{i+1} - \hat{y}_{i+1} = C_{\mathrm{I}}(t_i)\tau^3 + \text{члени більш високого порядку}. \tag{14}$$

Вважатимемо, що перехід від точки t_i до t_{i+1} відбувся успішно, тобто

$$|\bar{y}_{i+1} - \hat{y}_{i+1}| \leqslant \varepsilon. \tag{15}$$

Звідси

$$|C_{\mathrm{I}}(t_i)\tau^3| \leqslant \varepsilon.$$

Для нового кроку $\tau_H = t_{i+2} - t_{i+1}$

$$|C_{\mathrm{I}}(t_{i+1})\tau_H^3| \leqslant \varepsilon. \tag{16}$$

Якщо функція $C_{\mathrm{I}}(t)$ досить гладка, то з точністю до $O(\tau)$ можна покласти

$$C_{\mathrm{I}}(t_{i+1}) \approx C_{\mathrm{I}}(t_i). \tag{17}$$

З рівняння (14) маємо

$$|C_{\mathrm{I}}(t_{i+1})| \approx |\bar{y}_{i+1} - \hat{y}_{i+1}|/\tau^3,$$

тому з (16), (17) дістаємо

$$\frac{|\bar{y}_{i+1} - \hat{y}_{i+1}|}{\tau^3}\tau_H^3 \leqslant \varepsilon,$$

звідки

$$\tau_H \approx \tau\sqrt[3]{\frac{\varepsilon}{|\bar{y}_{i+1} - \hat{y}_{i+1}|}} \tag{18}$$

(для методів p-го та $(p+1)$-го порядків $\tau_H \approx \tau\sqrt[p+1]{\frac{\varepsilon}{|\bar{y}_{i+1} - \hat{y}_{i+1}|}}$).

Деякі автори на основі чисельних експериментів пропонують таку формулу:

$$\tau_H \approx \alpha\tau\sqrt[p]{\frac{\varepsilon}{|\bar{y}_{i+1} - \hat{y}_{i+1}|/\tau}}, \quad \alpha \approx 0{,}9.$$

3.5. Збіжність і точність однокрокових методів на проміжку інтегрування

Будемо досліджувати збіжність і точність однокрокових методів на прикладі методу Рунге — Кутта другого порядку (покращеного методу ламаних), який запишемо у вигляді

$$(A_\tau y)_n \equiv \frac{y_{n+1}-y_n}{\tau} - f\left(t_n + \frac{\tau}{2},\ y_n + \frac{\tau}{2} f(t_n,\ y_n)\right) = 0, \qquad (1)$$

$$n = 0,\ 1,\ 2,\ \ldots,\quad y_0 = u_0.$$

Відповідну задачу Коші запишемо у вигляді

$$Au \equiv \frac{du}{dt} - f(t,\ u) = 0,\quad t \in (0,\ T],\quad T \leqslant \infty, \qquad (2)$$

$$u(0) = u_0$$

і вважатимемо, що для неї виконується достатня умова стійкості

$$f'_u(t,\ u) \leqslant 0 \quad \forall t \in (0,\ T],\quad u \in (-\infty,\ \infty), \qquad (3)$$

та умова

$$|f'_u(t,\ u)| \leqslant L \ \forall t \in (0,\ T],\quad u \in (-\infty,\ \infty), \qquad (4)$$

яка забезпечує для неперервної по обох аргументах функції $f(t,\ u)$ виконання умови Ліпшиця по u і, отже, єдиність розв'язку задачі (2). Нагадаємо, що існування єдиного розв'язку та стійкість означають коректність задачі Коші.

Якщо задача (2) розглядається на скінченному проміжку $[0,\ T]$, $T < \infty$, то $y_n = y(t_n) \equiv y(t_n;\ \tau)$ в (1) є функцією дискретного аргументу, яка задана на сітці

$$\bar{\omega}_\tau = \{t_i = i\tau;\quad i = \overline{0,\ N};\quad \tau = T/N\}.$$

Якщо задача (2) розглядається на нескінченному проміжку $[0,\ \infty)$, який позначатимемо також $[0,\ T]$, $T = \infty$, то функція $y_n = y(t_n) \equiv y(t_n,\ \tau)$ задається на сітці

$$\bar{\omega}_\tau = \{t_i = i\tau : i = 0,\ 1,\ \ldots\}$$

з деяким кроком $\tau > 0$. Нехай $\omega_\tau^- = \bar{\omega}_\tau \setminus \{T\}$ для скінченного проміжку і $\omega_\tau^- = \bar{\omega}_\tau$ — для нескінченного.

Загальні ідеї дослідження збіжності сіткових схем, в тому числі і схем розв'язування задачі Коші, можна описати таким чином. Через $\tilde{B}_{1\tau}$ позначимо лінійний простір сіткових функцій, заданих на сітці $\bar{\omega}_\tau$, які при $t = t_0 = 0$ набувають заданого значення u_0, а через $\tilde{B}_{2\tau}$ — лінійний простір сіткових функцій, заданих на ω_τ^-. Тоді

схему (1) можна записати в операторному вигляді

$$A_\tau y = \varphi, \qquad (1')$$

де $\varphi = 0$ — елемент простору $\tilde{B}_{2\tau}$, $y = (y_0,\ y_1,\ \ldots,\ y_N)$ при $T < \infty$; $y = (y_0,\ y_1,\ \ldots)$ при $T = \infty$ — шуканий елемент простору $\tilde{B}_{1\tau}$; A_τ — нелінійний оператор, який задається формулою (1), $A_\tau : \tilde{B}_{1\tau} \to \tilde{B}_{2\tau}$.

Якщо через $\mathring{\tilde{B}}_{1\tau}$ позначити підпростір простору $\tilde{B}_{1\tau}$, який складається з функцій, що дорівнюють нулю при $t = t_0 = 0$, то задачу для похибки $z = (z_0,\ \ldots,\ z_N)$ для скінченного інтервалу $[0,\ T]$ і $z = (z_0,\ z_1,\ \ldots)$ для $T = \infty$ ($z_n = y_n - u_n \equiv y(t_n) - u(t_n)$) можна записати у вигляді

$$A_\tau y - A_\tau u \equiv A'_\tau z = \psi, \qquad (5)$$

де $\psi = (\psi_0,\ \ldots,\ \psi_{N-1})$, $y = (y_0,\ \ldots,\ y_N)$, $u = (u_0,\ u_1,\ \ldots,\ u_N)$ при $T < \infty$ і $\psi = (\psi_0,\ \psi_1,\ \ldots)$, $y = (y_0,\ y_1,\ \ldots)$, $u = (u_0,\ u_1,\ \ldots)$ при $T = \infty$; ψ — похибка апроксимації схеми (1′) на точному розв'язку задачі (2), $\psi = \varphi - A_\tau u$, тобто нев'язка, яку дістанемо, якщо в сіткову схему (1′) замість y підставимо точний розв'язок u (A'_τ — лінійний оператор, який називається *похідною Гато оператора* A_τ). Далі можна користуватися загальними поняттями методу сіток, які ми розглянули в гл. 2. Зокрема, при дослідженні збіжності методу при $\tau \to 0$, її швидкості можна користуватися такою універсальною схемою: 1) записати задачу для похибки вигляду (5) за допомогою сіткових операторів A_τ або A'_τ і похибки апроксимації; 2) дослідити похибку апроксимації як функцію від τ; 3) у просторах $\tilde{B}_{1\tau}$, $\tilde{B}_{2\tau}$ ввести норми $\|\cdot\|_{1\tau}$ та $\|\cdot\|_{2\tau}$ відповідно, в результаті чого ці лінійні простори перетворюються в нормовані простори $B_{1\tau}$ і $B_{2\tau}$, і знайти апріорну оцінку виду $\|z\|_{1\tau} \leqslant M \|\psi\|_{2\tau}$ з незалежною від τ сталою M, яка означає стійкість сіткової схеми (1′); 4) на основі 2), 3) дослідити збіжність і швидкість збіжності розв'язку схеми (1′), (1) до точного розв'язку задачі (2) при $\tau \to 0$. Нагадаємо, що схема вигляду (1′) називається *стійкою*, якщо для розв'язків задач

$$A_\tau y = \varphi,\quad A_\tau \tilde{y} = \tilde{\varphi}$$

справедлива оцінка

$$\|\tilde{y} - \tilde{y}\|_{1\tau} \leqslant M \|\tilde{\varphi} - \tilde{\varphi}\|_{2\tau},$$

де M — незалежна від τ, $\tilde{y}$, $\tilde{y}$, $\tilde{\varphi}$, $\tilde{\varphi}$ стала. Якщо ця оцінка має місце за деякої додаткової умови для кроку сітки τ, то схема є умовно стійкою, і безумовно стійкою — в іншому разі.

Повертаючись до схеми Рунге — Кутта (1), зазначимо так: з побудови цієї схеми випливає, що при достатній гладкості шуканого розв'язку на кожному кроці схема має похибку порядку $O(\tau^3)$, через те що попереднє значення знайдене точно. Виникають питання: 1) який порядок точності має ця сама схема, якщо розв'язок шукається на проміжку $[0, T]$ і на кожному кроці допускається деяка похибка? 2) як залежить точність наближеного розв'язку від гладкості шуканого розв'язку?

Виходячи з наведених вище загальних ідей, дамо відповідь на ці питання.

Маємо

$$(A_\tau y)_n - (A_\tau u)_n \equiv \frac{z_{n+1}-z_n}{\tau} - \left[f\left(t_n+\frac{\tau}{2},\ y_n+\frac{\tau}{2}f(t_n,\ y_n)\right) - f\left(t_n+\frac{\tau}{2},\ u_n+\frac{\tau}{2}f(t_n,\ u_n)\right)\right] =$$

$$= -\frac{u_{n+1}-u_n}{\tau} + f\left(t_n+\frac{\tau}{2},\ u_n+\frac{\tau}{2}f(t_n,\ u_n)\right) \equiv 0 - (A_\tau u)_n \equiv \psi_n.$$

Користуючись двічі формулою Лагранжа, дістаємо

$$A'_\tau z \equiv \frac{z_{n+1}-z_n}{\tau} - \beta_n\left[1+\frac{\tau\gamma_n}{2}\right]z_n = \psi_n, \quad n=0,\ 1,\ 2,\ \ldots;\ z_0=0. \tag{6}$$

де

$$\beta_n = f'_u\left(t_n+\frac{\tau}{2},\ u_n+\frac{\tau}{2}f(t_n,\ u_n)+\theta_1\left(y_n+\frac{\tau}{2}f(t_n,\ y_n)-u_n-\frac{\tau}{2}f(t_n,\ u_n)\right)\right),\quad \theta_1\in(0,\ 1),$$

$$\gamma_n = f'_u(t_n,\ u_n+\theta_2 z_n),\quad \theta_2\in(0,\ 1).$$

Щоб перетворити похибку апроксимації ψ_n, додамо і віднімемо від неї величину

$$f\left(t_n+\frac{\tau}{2},\ u_{n+1/2}\right) \equiv u'_{n+1/2},\quad u_{n+1/2} = u\left(t_n+\frac{\tau}{2}\right),$$

і скористаємося формулою Лагранжа

$$\psi_n = -\frac{u_{n+1}-u_n}{\tau} + f\left(t_n+\frac{\tau}{2},\ u_n+\frac{\tau}{2}f(t_n,\ u_n)\right) - f\left(t_n+\frac{\tau}{2},\ u_{n+1/2}\right) + u'_{n+1/2} =$$

$$= -\frac{u_{n+1}-u_n}{\tau} + u'_{n+1/2} + \delta_n\left|u_n - u_{n+1/2} + \frac{\tau}{2}f(t_n,\ u_n)\right| \equiv$$

$$\equiv \eta_{1,n} + \delta_n\eta_{2,n}, \tag{7}$$

де

$$\delta_n = f'_u\left(t_n+\frac{\tau}{2},\ u_{n+1/2}+\theta_3\left(u_n+\frac{\tau}{2}f(t_n,\ u_n)-u_{n+1/2}\right)\right),\quad \theta_3\in(0,\ 1),$$

$$\eta_{1,n} = -\frac{u_{n+1}-u_n}{\tau} + u'_{n+1/2},$$

$$\eta_{2,n} = u_n - u_{n+1/2} + \frac{\tau}{2}u'_n.$$

Неважко помітити, що ψ, η_1, η_2 є функціоналами, які залежать від точки сітки t_n, шуканого розв'язку u та від τ, тобто $\psi = \psi(t_n,\ u;\ \tau) = \psi_n(u,\ \tau)$, $\eta_i = \eta_i(t_n,\ u;\ \tau) = \eta_{i,n}(u,\ \tau)$, $i=1,\ 2$ (там, де це не викликатиме непорозумінь, деякі з аргументів опускатимемо). Нас цікавить залежність цих функціоналів від τ та від гладкості u, про що йдеться у наступних двох твердженнях.

Лема 1. *Мають місце співвідношення*

$$\eta_{1,n} = \begin{cases} -\frac{\tau^2}{24}u^{(3)}(\xi),\ \xi\in[t_n,\ t_{n+1}], & \text{якщо } u\in C^3[0,\ T], \\ -\frac{\tau}{8}[u''(\xi)-u''(\eta)],\ \xi\in[t_{n+1/2},\ t_{n+1}],\ \eta\in[t_n,\ t_{n+1/2}], & \\ & \text{якщо } u\in C^2[0,\ T]. \end{cases} \tag{8}$$

$$\eta_{2,n} = \begin{cases} -\frac{\tau^2}{8}u''(\xi),\ \xi\in[t_n,\ t_{n+1/2}], & \text{якщо } u\in C^2[0,\ T], \\ -\frac{\tau}{2}[u'(\xi)-u'(t_n)],\ \xi\in[t_n,\ t_{n+1/2}], & \text{якщо } u\in C^1[0,\ T] \end{cases} \tag{9}$$

$|\psi_n| \leqslant M\tau^k$, *якщо* $u\in C^{k+1}[0,\ T]$, $k=1,\ 2$, *де стала* M *не залежить від* τ, $n(t_n)$.

Доведення. Нехай $u\in C^3[0,\ T]$. Розвинемо $u_n = u(t_n)$, $u_{n+1} = u(t_{n+1})$ за формулою Тейлора у точці $t_{n+1/2} = t_n + \frac{\tau}{2}$:

$$u_{n+1} = u_{n+1/2} + \frac{\tau}{2}u'_{n+1/2} + \left(\frac{\tau}{2}\right)^2\frac{1}{2!}u''_{n+1/2} + \left(\frac{\tau}{2}\right)^3\frac{1}{3!}u'''(\xi_1),$$

$$\xi_1\in[t_{n+1/2},\ t_{n+1}],$$

$$u_n = u_{n+1/2} - \frac{\tau}{2}u'_{n+1/2} + \left(\frac{\tau}{2}\right)^2\frac{1}{2!}u''_{n+1/2} - \left(\frac{\tau}{2}\right)^3\frac{1}{3!}u'''(\xi_2),$$

$$\xi_2\in[t_n,\ t_{n+1/2}].$$

Підставляючи ці вирази в формулу для $\eta_{1,n}$, дістаємо

$$\eta_{1,n} = u_{n+1/2} - \frac{\tau^2}{48}[u'''(\xi_1) + u'''(\xi_2)] - u_{n+1/2} =$$

$$= -\frac{\tau^2}{48}[u'''(\xi_1) + u'''(\xi_2)].$$

За лемою 2 з п. 5.5 ч. 1

$$\eta_{1,n} = -\frac{\tau^2}{24}u^{(3)}(\xi), \quad \xi \in [t_n,\ t_{n+1}].$$

Користуючись розвиненням

$$u_{n+1/2} = u_n + \frac{\tau}{2}u_n' + \left(\frac{\tau}{2}\right)^2\frac{1}{2!}u''(\xi), \quad \xi \in [t_n,\ t_{n+1/2}],$$

для $\eta_{2,n}$ маємо

$$\eta_{2,n} = u_n - u_n - \frac{\tau}{2}u_n' - \frac{\tau^2}{8}u''(\xi) + \frac{\tau}{2}u_n' = -\frac{\tau^2}{8}u''(\xi).$$

Аналогічно доводяться оцінки для $\eta_{1,n}$, коли $u \in C^2[0,\ T]$ та для $\eta_{2,n}$, коли $u \in C^1[0,\ T]$.

Оцінки для ψ_n дістаємо з оцінок для $\eta_{1,n}$ та $\eta_{2,n}$. Наприклад, коли $u \in C^3[0,\ T]$, то

$$|\psi_n| \leqslant |\eta_{1,n}| + |\delta_n|\,|\eta_{2,n}| \leqslant$$

$$\leqslant \frac{\tau^2}{24}\|u^{(3)}\|_{C[0,T]} + L\frac{\tau^2}{8}\|u''\|_{C[0,T]} = M\tau^2,$$

$$M = \frac{1}{24}\|u^{(3)}\|_{C[0,T]} + \frac{L}{8}\|u''\|_{C[0,T]}.$$

Лему доведено.

Лема 2. *Мають місце оцінки*

$$|\eta_{1,n}| \leqslant M\tau^{k-0,5}|u|_{W_2^{k+1}(e_n)}, \quad \text{якщо } u \in W_2^{k+1}(0,\ T), \quad k = 1,\ 2, \tag{10}$$

$$|\eta_{2,n}| \leqslant M\tau^{k-0,5}|u|_{W_2^{k}(e_n)}, \quad \text{якщо } u \in W_2^{k}(0,\ T), \quad k = 1,\ 2, \tag{11}$$

$$|\psi_n| \leqslant M\tau^{k-0,5}(|u|_{W_2^{k+1}(e_n)} + |u|_{W_2^{k}(e_n)}), \tag{12}$$

$$\text{якщо } u \in W_2^{k+1}(0,\ T), \quad k = 1,\ 2,$$

$$\|\psi\|_{L_2(\omega^-)} \equiv \left(\sum_{n=0}^{N-1}\tau\psi_n^2\right)^{1/2} \leqslant \sqrt{2}\,M\tau^k(|u|_{W_2^{k+1}(0,T)} + |u|_{W_2^{k}(0,T)}), \tag{12'}$$

$$k = 1,\ 2; \quad N = \infty \text{ при } T = \infty,$$

де через M позначено різні сталі, які не залежать від τ, u, n, $e_n = (t_n,\ t_{n+1})$.

Доведення. За допомогою лінійної заміни $t = \theta\tau + t_n$ відобразимо інтервал $e_n = (t_n,\ t_{n+1})$ на інтервал $E = (0,\ 1)$ ($t \in e_n$, $\theta \in E$). Тоді

$$\eta_{1,n} = \tau^{-1}\tilde{u}'(0,5) - \tau^{-1}(\tilde{u}(0) - \tilde{u}(1)),$$

де $\tilde{u}(\theta) = u(t, (\theta)) = u(t_n + \theta\tau)$. Застосовуючи теореми вкладення (див. п. 1.5 з ч. 1), дістаємо

$$|\eta_{1,n}| \leqslant \tau^{-1}\|\tilde{u}\|_{C[0,1]} \leqslant M\tau^{-1}\|\tilde{u}\|_{W_2^{k+1}(0,1)}, \quad k \geqslant 1, \tag{13}$$

де стала M не залежить від τ і $\tilde{u}$. Зазначимо, що сталі множники в теоремах вкладення залежать від довжини відрізка. Тому відображення $e_n \to E$ було зроблено з метою позбавлення залежності сталої M від e_n і тим самим від τ. Такий прийом називається масштабуванням. Нерівність (13) означає обмеженість лінійного функціонала (від $\tilde{u}$) у просторі $W_2^{k+1}(0,\ 1)$, $k \geqslant 1$. Неважко перевірити, що функціонал $\eta_{1,n} \equiv \eta_1(t_n,\ \tilde{u};\ \tau) \equiv \eta_1(\tilde{u})$ перетворюється в нуль на многочленах до другого степеня включно.

Дійсно,

$$\eta_1(1) = \tau^{-1}[(1)'|_{\theta=0,5} - (1 - 1)] = 0,$$

$$\eta_1(\theta) = \tau^{-1}[\theta'|_{\theta=0,5} - (1 - 0)] = 0,$$

$$\eta_1(\theta^2) = \tau^{-1}\left[\frac{d\theta^2}{d\theta}\bigg|_{\theta=0,5} - (1^2 - 0^2)\right] = 0,$$

$$\eta_1(\theta^3) = \tau^{-1}\left[\frac{d\theta^3}{d\theta}\bigg|_{\theta=0,5} - (1^3 - 0^3)\right] \neq 0.$$

Тому, застосовуючи лему Брембла — Гільберта, дістаємо

$$|\eta_{1,n}| = |\eta_1(\tilde{u})| \leqslant \bar{M}M\tau^{-1}|\tilde{u}|_{W_2^{k+1}(0,1)}, \quad k = 1,\ 2,$$

де сталі $\bar{M}$, M не залежать від τ і $\tilde{u}$. Переходячи до змінної t, матимемо

$$|\tilde{u}|_{W_2^{k+1}(0,1)} = \left(\int_0^1(\tilde{u}^{(k+1)}(\theta))^2 d\theta\right)^{1/2} =$$

$$= \left(\tau^{-1}\tau^{2(k+1)}\int_{t_n}^{t_{n+1}}(u^{(k+1)}(t))^2\,dt\right)^{1/2} = \tau^{k+1/2}|u|_{W_2^{k+1}(e_n)},$$

і тому

$$|\eta_{1,n}| \leqslant M\tau^{k-0,5}|u|_{W_2^{k+1}(e_n)},$$

де стала M не залежить від τ, u, n.

Аналогічно доводиться і нерівність від $\eta_{2,n}$, а нерівність для ψ_n дістаємо з (10), (11). Враховуючи нерівність $(a+b)^2 \leqslant 2(a^2+b^2)$, з (7) маємо

$$|\psi_n|^2 \leqslant 2M^2\tau^{2k-1}\left(|u|^2_{W_2^{k+1}(e_n)} + |u|^2_{W_2^k(e_n)}\right),$$

і далі

$$\|\psi\|_{L_2(\omega^-)} = \sqrt{2}\,M\tau^k\left(\sum_{n=0}^{N-1}|u|^2_{W_2^{k+1}(e_n)} + \sum_{k=0}^{N-1}|u|^2_{W_2^k(e_n)}\right)^{1/2} =$$

$$= \sqrt{2}\,M\tau^k\left(|u|^2_{W_2^{k+1}(0,T)} + |u|^2_{W_2^k(0,T)}\right)^{1/2}.$$

З нерівності $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$, $a \geqslant 0$, $b \geqslant 0$, дістанемо

$$\|\psi\|_{L_2(\omega^-)} \leqslant \sqrt{2}\,M\tau^k\left(|u|_{W_2^{k+1}(0,T)} + |u|_{W_2^k(0,T)}\right).$$

Лему доведено.

Наступним кроком при дослідженні збіжності є визначення апріорної оцінки, яка виражає стійкість схеми (1), (1′). Якість таких оцінок залежить від вибору норм у просторах $\tilde{B}_{1\tau}$ та $\tilde{B}_{2\tau}$.

Лема 3. *Нехай задача (2) розглядається на проміжку* $[0, T]$, $T \leqslant \infty$, *і виконуються умови (3), (4). Якщо* $T < \infty$,

$$\tau L \leqslant 2, \tag{14}$$

то для задачі (6) справедливі оцінки

$$\|z\|_{C(\bar{\omega})} \leqslant T\|\psi\|_{C(\tilde{\omega})}, \tag{15}$$

$$\|z\|_{C(\bar{\omega})} \leqslant T^{1/2}\|\psi\|_{L_2(\tilde{\omega})}, \tag{16}$$

де

$$\|z\|_{C(\bar{\omega})} = \max_{n=\overline{0,N}}|z_n| = \max_{n=\overline{1,N}}|z_n|,$$

$$\|\psi\|_{C(\omega^-)} = \max_{n=\overline{0,N-1}}|\psi_n|,$$

$$\|\psi\|_{L_2(\omega^-)} = \left(\sum_{k=0}^{N-1}\tau\psi_k^2\right)^{1/2}.$$

На нескінченному проміжку $(T = \infty)$ *за умов*

$$0 < \lambda_0 \leqslant |f_u'| \leqslant L,\ \lambda_0 < L/2,\ \tau \leqslant \tau_0,\ \tau_0 L < 2, \tag{17}$$

мають місце оцінки

$$\|z\|_{C(\bar{\omega})} \leqslant c_0^{-1}\|\psi\|_{C(\omega^-)}, \tag{18}$$

$$\|z\|_{C(\bar{\omega})} \leqslant c_0^{-1/2}\|\psi\|_{L_2(\omega^-)}, \tag{19}$$

де стала

$$c_0 = \begin{cases} \tau_0\lambda_0 L/2, & \text{якщо } \lambda_0(1-\tau_0 L/2) > \tau_0\lambda_0 L/2, \\ \lambda_0(1-\tau_0 L/2), & \text{якщо } \lambda_0(1-\tau_0 L/2) \leqslant \tau_0\lambda_0 L/2, \end{cases} \tag{20}$$

не залежить від τ.

Доведення. Перепишемо рівняння (6) у вигляді

$$z_{n+1} = q_n z_n + \tau\psi_n, \qquad q_n = \tau\beta_n\left(1 + \frac{\tau\gamma_n}{2}\right) + 1. \tag{21}$$

Розглянемо спочатку випадок $T < \infty$. Тоді $\bar{\omega} = \{t_i = i\tau : i = \overline{0, N},\ \tau = T/N\}$, $n = 0, 1, \ldots, N-1$. За умов $|q_n| \leqslant 1$ матимемо

$$|z_{n+1}| \leqslant |z_n| + \tau|\psi_n| \leqslant |z_{n-1}| + \tau|\psi_{n-1}| + \tau|\psi_n| \leqslant \cdots \leqslant$$

$$\leqslant |z_0| + \sum_{k=0}^{n}\tau|\psi_k| = \sum_{k=0}^{n}\tau|\psi_k| \leqslant \sum_{k=0}^{N-1}\tau|\psi_k|.$$

Звідси

$$\|z\|_{C(\bar{\omega})} = \max_{n=1,N}|z_n| \leqslant \max_{n=\overline{0,N-1}}|\psi_n|\sum_{k=0}^{N-1}\tau = T\|\psi\|_{C(\omega^-)}.$$

Застосовуючи у правій частині попередньої нерівності нерівність Коші — Буняковського, дістаємо

$$\|z\|_{C(\bar{\omega})} \leqslant \left(\sum_{k=0}^{N-1}\tau\right)^{1/2}\left(\sum_{k=0}^{N-1}\tau\psi_k^2\right)^{1/2} = T^{1/2}\|\psi\|_{L_2(\omega^-)}.$$

Умова $|q_n| \leqslant 1$ буде виконана, якщо

$$\begin{cases} 2 + \tau\beta_n + \dfrac{\tau^2}{2}\beta_n\gamma_n \geqslant 0, \\ \tau\beta_n + \dfrac{\tau^2}{2}\beta_n\gamma_n \leqslant 0. \end{cases}$$

Оскільки в силу нерівності (3) $\beta_n \leqslant 0$, $\gamma_n \leqslant 0$, то цю систему нерівностей можна переписати у вигляді

$$\begin{cases} 2 - \tau|\beta_n| + \dfrac{\tau^2}{2}|\beta_n||\gamma_n| \geqslant 0, \\ -\tau|\beta_n| + \dfrac{\tau^2}{2}|\beta_n||\gamma_n| \leqslant 0. \end{cases}$$

Друга з цих нерівностей буде виконана при $\tau|\gamma_n| \leqslant 2$. Неважко помітити, що тоді перша нерівність виконується при $\tau|\beta_n| \leqslant 2$, а нерівність $|q_n| \leqslant 1$ тим більше буде виконана при $\tau L \leqslant 2$, оскільки $|\beta_n| \leqslant L$, $|\gamma_n| \leqslant L$. Тим самим нерівності (5), (6) доведені.

Розглянемо випадок $T = \infty$. Тоді сітка $\bar{\omega} = \{t_i = i\tau : i = 0, 1, \ldots\}$ нескінченна і в (6) $n = 0, 1, \ldots$ За умов $|q_n| \leqslant q < 1$ матимемо

$$|z_{n+1}| \leqslant q|z_n| + \tau|\psi_n| \leqslant q^2|z_{n-1}| + q\tau|\psi_{n-1}| + \tau|\psi_n| \leqslant \cdots \leqslant$$

$$\leqslant q^{n+1}|z_0| + \sum_{k=0}^{n} \tau|\psi_k| q^{n-k} = \sum_{k=0}^{n} \tau|\psi_k| q^{n-k}. \qquad (22)$$

За допомогою нерівності Коші — Буняковського маємо

$$|z_{n+1}| \leqslant \left(\sum_{k=0}^{n} \tau q^{2n-2k}\right)^{1/2} \left(\sum_{k=0}^{n} \tau|\psi_k|^2\right)^{1/2} =$$

$$= \tau^{1/2} \left(\sum_{k=0}^{n} q^{2k}\right)^{1/2} \left(\sum_{k=0}^{n} \tau|\psi_k|^2\right)^{1/2}. \qquad (23)$$

З виразу (22) випливають нерівності:

$$|z_{n+1}| \leqslant \tau \max_{k=0,\infty} |\psi_k| \sum_{k=0}^{\infty} q^k = \frac{\tau}{1-q} \|\psi\|_{C(\omega^-)},$$

$$\|z\|_{C(\bar{\omega})} \leqslant \frac{\tau}{1-q} \|\psi\|_{C(\bar{\omega})}. \qquad (24)$$

З виразу (23) маємо

$$|z_{n+1}| \leqslant \tau^{1/2} \left(\frac{1}{1-q^2}\right)^{1/2} \left(\sum_{k=0}^{\infty} \tau\psi_k^2\right)^{1/2},$$

$$\|z\|_{C(\bar{\omega})} \leqslant \left(\frac{\tau}{1-q^2}\right)^{1/2} \|\psi\|_{L_2(\omega^-)}. \qquad (25)$$

Неважко помітити, що при $\tau \leqslant \tau_0$, $\tau_0 L < 2$, $0 < \lambda_0 \leqslant |f'_u| \leqslant L$, $\lambda_0 < \frac{L}{2}$, виконуються нерівності

$$0 < 1 - \frac{\tau_0 L}{2} \leqslant 1 - \frac{\tau|\gamma_n|}{2} \leqslant 1 - \frac{\tau\lambda_0}{2},$$

$$-1 + \frac{\tau_0\lambda_0 L\tau}{2} < 1 - \tau_0 L\left(1 - \frac{\tau\lambda_0}{2}\right) \leqslant q_n = 1 + \tau\beta_n\left(1 + \frac{\tau\gamma_n}{2}\right) =$$

$$= 1 - \tau|\beta_n|\left(1 - \frac{\tau|\gamma_n|}{2}\right) \leqslant 1 - \tau\lambda_0\left(1 - \frac{\tau_0 L}{2}\right),$$

звідки

$$|q_n| \leqslant q = 1 - c\tau < 1, \qquad (26)$$

де

$$c = \begin{cases} \tau_0\lambda_0 L/2, & \text{якщо } \lambda_0(1 - \tau_0 L/2) > \tau_0\lambda_0 L/2, \\ \lambda_0(1 - \tau_0 L/2), & \text{якщо } \lambda_0(1 - \tau_0 L/2) \leqslant \tau_0\lambda_0 L/2, \end{cases}$$

$$0 < c < \lambda_0, \quad c\tau \leqslant c\tau_0 < \lambda_0\tau_0 < \frac{L\tau_0}{2} < 1.$$

Враховуючи (26), з (24), (25) дістанемо (18), (19). Лему доведено.

Зауваження. У випадку $T < \infty$ величини N і τ пов'язані співвідношенням $\tau = T/N$. Тому за умови $|f'_u| \leqslant L$ маємо $|q_n| \leqslant 1 + \tau K$, $K = L\left(1 + \frac{TL}{2}\right)$. Тому

$$|z_{n+1}| \leqslant (1 + \tau K)|z_n| + \tau|\psi_n| \leqslant \cdots \leqslant (1 + \tau K)^{n+1}|z_0| +$$

$$+ \sum_{j=0}^{n} \tau(1 + \tau K)^{n-j}|\psi_j| \leqslant \max_{j=\overline{0,N-1}} |\psi_j| \sum_{j=0}^{N-1} \tau(1 + \tau K)^j =$$

$$= \|\psi\|_{C(\omega^-)} \tau \frac{(1 + \tau K)^N - 1}{\tau K} = \|\psi\|_{C(\omega^-)} \frac{[(1 + \tau K)^{1/\tau K}]^{TK} - 1}{K} \leqslant$$

$$\leqslant \frac{e^{TK} - 1}{K} \|\psi\|_{C(\omega^-)},$$

звідки

$$\|z\|_{C(\bar{\omega})} \leqslant \frac{e^{TK} - 1}{K} \|\psi\|_{C(\omega^-)}.$$

Таким чином, у цьому разі ми маємо апріорну оцінку без обмеження (14) на τ (звичайно, крім природного обмеження $\tau \leqslant T$). Тоді і збіжність матиме місце без обмеження на τ.

З лем 1—3 випливають такі твердження.

Теорема 1. *Нехай для задачі Коші (2), яка розглядається на скінченному проміжку $[0, T]$, і для схеми Рунге—Кутта (1) виконуються умови (3), (4), (14), а розв'язок задачі (2) належить до класу $C^{k+1}[0, T]$, $k = 1, 2$. Тоді має місце оцінка*

$$\|y - u\|_{C(\bar{\omega})} \leqslant M\tau^k \|u\|_{C^{k+1}[0,T]}, \quad k = 1, 2,$$

де y — розв'язок схеми (1), u — точний розв'язок задачі (2), M — незалежна від τ і u стала.

Теорема 2. *Нехай для задачі Коші (2), яка розглядається на нескінченному проміжку $[0, \infty)$, і для схеми Рунге — Кутта (1) виконуються умови (3), (17), а розв'язок задачі (2) належить до класу $C^{k+1}[0, \infty)$, $k = 1, 2$, тобто норми $\|u\|_{C^k[0,\infty)}$, $k = 1, 2$, обмежені. Тоді має місце оцінка*

$$\|y - u\|_{C(\bar{\omega})} \leqslant M\tau^k \|u\|_{C^{k+1}[0,\infty)}, \quad k = 1, 2,$$

де y — розв'язок схеми (1), u — точний розв'язок задачі (2), M — незалежна від τ і u стала.

Теорема 3. *Нехай для задачі Коші (2), яка розглядається на скінченному проміжку $[0, T]$, і для схеми Рунге — Кутта (1) виконуються умови (3), (4), (14), а розв'язок задачі (2) належить до класу $W_2^{k+1}(0, T)$, $k = 1, 2$. Тоді справедлива оцінка*

$$\|y - u\|_{C(\bar{\omega})} \leqslant M\tau^k \|u\|_{W_2^{k+1}(0,T)}, \quad k = 1, 2,$$

де y — розв'язок схеми (1), u — точний розв'язок задачі (2), M — незалежна від τ і u стала.

Теорема 4. *Нехай для задачі Коші (2), яка розглядається на нескінченному проміжку $[0, \infty)$, і для схеми Рунге — Кутта (1) виконуються умови (3), (17), а розв'язок u задачі (2) належить до класу $W_2^{k+1}(0, \infty)$. Тоді має місце оцінка*

$$\|y - u\|_{C(\bar{\omega})} \leqslant M\tau^k \|u\|_{W_2^{k+1}(0,\infty)}, \quad k = 1, 2,$$

де y — розв'язок схеми (1), M — незалежна від τ і u стала.

Теореми 1—4 за вказаних в них умов стверджують збіжність наближеного розв'язку y до точного u при $\tau \to 0$. Крім того, вони показують, що в практичних розрахунках крок τ слід вибирати так, щоб задовольнялись умови (14) або (17), які забезпечують виконання відповідних оцінок близькості y до u.

Вправа. Дослідити на збіжність і швидкість збіжності метод Ейлера (15) з п. 3.3, використовуючи задачу для похибки (16) і вираз для нев'язки (похибки апроксимації) (17).

3.6. Про підвищення точності однокрокових методів на проміжку інтегрування

Зазначимо, що з результатів попереднього параграфа (див. теореми 1—4) випливає, що порядок точності методу Рунге — Кутта на всьому проміжку інтегрування на одиницю менший за її порядок на кроці (при досить великій гладкості точного розв'язку, а саме $u \in C^3[0, T]$ або $u \in W_2^3(0, T)$).

Уважно проаналізувавши доведення теорем 1—4, можна зазначити, що порядок точності схеми Рунге — Кутта визначається функціоналом $\eta_{1,n}$, порядок якого (по τ) на одиницю нижчий, ніж порядок функціонала $\eta_{2,n}$. Крім того, це є наслідком того, що функціонал $\eta_{1,n}$ містить звичайну і різницеву похідні від функції u, а це потребує ділення на τ. Зазначимо, що така структура функціонала $\eta_{1,n}$ обумовлена вибором ψ_n у вигляді (5) з п.3.5, точніше першим доданком $f(t_n + \tau/2)$.

Розглянемо схеми, відмінні від схем Рунге — Кутта, порядок точності яких на всьому проміжку інтегрування на одиницю вищий. Це досягається за допомогою усереднення правої частини диференціального рівняння.

Розглянемо лінійну стійку задачу Коші

$$\frac{du}{dt} + \lambda u = f(t), \quad t > 0, \quad \lambda \geqslant 0;$$
$$u(0) = u_0, \tag{1}$$

розв'язок якої треба знайти на проміжку $[0, T]$. Нехай сітка $\bar{\omega} = \{t_i = i\tau : i = \overline{0, N}, \ \tau = T/N\}$ покриває відрізок $[0, T]$, $\omega^- = \bar{\omega} \setminus \{T\}$ і $t \in \omega^-$ — точка сітки, а ξ — змінюється на проміжку $[0, T]$. Проінтегрувавши рівняння (1) на проміжку $[t, t + \tau]$ і розділивши результат на τ, дістанемо

$$u_t(t) + \frac{\lambda}{\tau}\int_t^{t+\tau} u(\xi)\, d\xi = \frac{1}{\tau}\int_t^{t+\tau} f(\xi)\, d\xi, \quad t \in \omega^-. \tag{2}$$

Замінюючи в (2) $\frac{1}{\tau}\int_t^{t+\tau} u(\xi)\, d\xi$ наближено на $u(t)$, запишемо схему розв'язування задачі (1):

$$y_t + \lambda y = \bar{f}(t), \quad t \in \omega^-; \quad y(0) = u_0, \tag{3}$$

де

$$\bar{f}(t) = \frac{1}{\tau}\int_t^{t+\tau} f(\xi)\, d\xi.$$

Дослідимо на стійкість і збіжність схему (3) на проміжку $[0, T]$. Для похибки $z = y - u$ за допомогою (2) дістанемо задачу

$$z_t + \lambda z = y_t + \lambda y - u_t - \lambda u = \tau\int_t^{t+\tau} f(\xi)\, d\xi - u_t - \lambda u =$$
$$= u_t + \frac{\lambda}{\tau}\int_t^{t+\tau} u(\xi)\, d\xi - u_t - \lambda u = \lambda\left[\frac{1}{\tau}\int_t^{t+\tau} u(\xi)\, d\xi - u(t)\right] =$$
$$= \psi(t;\ u), \quad z(0) = 0, \tag{4}$$

де $\psi(t) \equiv \psi(t;\ u) \equiv \psi(u)$ — похибка апроксимації схеми (3) на розв'язку $u(t)$ задачі (1). Введемо у просторі сіткових функцій $y(t)$, $v(t)$, заданих на сітці ω^-, скалярний добуток за формулою

$$(y, v) = \sum_{t \in \omega^-} \tau y(t) v(t).$$

Помножимо (4) скалярно на z_t і врахуємо, що

$$(z^2)_t = z_t(t)\,[z(t+\tau) + z(t)] = 2z_t(t)\,z(t) + \tau z_t^2(t),$$

$$z_t \cdot z(t) = 0{,}5\,[z^2(t)]_t - 0{,}5\tau\, z_t^2(t), \quad t \in \omega^-,$$

$$(z_t, z) = (z_t \cdot \dot{z}, 1) = 0{,}5\,((z^2)_t, 1) - 0{,}5\tau\,(z_t^2, 1) =$$

$$= 0{,}5 \sum_{t\in\omega^-} \tau\,[z^2(t)]_t - 0{,}5\tau \| z_t \|^2_{L_2(\omega^-)} =$$

$$= 0{,}5\,[z^2(T) - z^2(0)] - 0{,}5\tau\,|z|^2_{W_2^1(\omega^-)} =$$

$$= 0{,}5 z^2(T) - 0{,}5\tau\,|z|^2_{W_2^1(\omega^-)},$$

де

$$\|y\|_{L_2(\omega^-)} = (y, y)^{1/2}, \quad |z|_{W_2^1(\omega^-)} = \|z_t\|_{L_2(\omega^-)}.$$

В результаті

$$|z|^2_{W_2^1(\omega^-)} + \lambda\,[0{,}5z^2(T) - 0{,}5\tau\,|z|^2_{W_2^1(\omega^-)}] = (\psi, z_t).$$

Далі за допомогою нерівності Коші — Буняковського дістаємо

$$(1 - 0{,}5\tau\lambda)\,|z|_{W_2^1(\omega^-)} \leqslant \|\psi\|_{L_2(\omega^-)}.$$

При $\tau \leqslant \tau^* < 2/\lambda$ маємо нерівність

$$|z|_{W_2^1(\omega^-)} \leqslant (1 - 0{,}5\tau^*\lambda)^{-1} \|\psi\|_{L_2(\omega^-)}. \tag{5}$$

Враховуючи нерівності $\|z\|_{C(\omega^-)} \leqslant \sqrt{T}\,|z|_{W_2^1(\omega^-)}$, $\|z\|_{L_2(\omega^-)} \leqslant \sqrt{T}\,\|z\|_{C(\bar{\omega})}$, перша з яких в силу того, що $z(0) = 0$, має місце і для z, з (5) дістаємо апріорні оцінки

$$\|z\|_{L_2(\omega^-)} \leqslant \sqrt{T}\,\|z\|_{C(\bar{\omega})} \leqslant T\,|z|_{W_2^1(\omega^-)} \leqslant$$

$$\leqslant T\,(1 - 0{,}5\,\tau^*\lambda)^{-1} \|\psi\|_{L_2(\omega^-)}. \tag{6}$$

Оцінки (5), (6) виражають умовну стійкість схеми (3) у відповідних нормах за умов $\tau \leqslant \tau^* < 2/\lambda$, де τ^* — фіксоване число.

Оцінимо величину $\|\psi\|_{L_2(\omega^-)}$, припускаючи, що $u \in W_2^1(0, T)$. Неважко помітити, що $\psi(t) \equiv \psi(t; u) \equiv \psi(u) = \lambda \left[\frac{1}{\tau} \int_t^{t+\tau} u(\xi)\,d\xi - u(t) \right]$, $t \in \omega^-$ при фіксованому t є лінійним функціоналом від u. За

допомогою лінійної заміни $s = \frac{\xi - t}{\tau}$ відобразимо відрізок $e^+ = e^+(t) = [t, t+\tau]$ на відрізок $[0, 1]$ і позначимо $u(\xi) = u(\xi(s)) = \tilde{u}(s)$. Тоді в силу теореми вкладення

$$|\psi(t; u)| = |\psi(t; \tilde{u})| \equiv \lambda \left| \int_0^1 \tilde{u}(s)\,ds - \tilde{u}(0) \right| \leqslant$$

$$\leqslant 2\lambda \|\tilde{u}\|_{C[0,1]} \leqslant M \|\tilde{u}\|_{W_2^1[0,1]},$$

де стала не залежить від $\tilde{u}$ і від τ. Ця нерівність означає, що лінійний функціонал $\psi(t; \tilde{u}) = \psi(t; u)$ обмежений у просторі $W_2^1[0, 1]$. Він перетворюється в нуль, якщо $\tilde{u}(s)$ — многочлен нульового степеня. Тому в силу леми Брембла — Гільберта

$$|\psi(t; u)| \leqslant M\bar{M}\,|\tilde{u}|_{W_2^1[0,1]} \leqslant M\bar{M}\tau^{1/2}\,|u|_{W_2^1(e^+)}.$$

Звідси

$$\|\psi\|_{L_2(\omega^-)} = \left(\sum_{t\in\omega^-} \tau\psi^2(t; u) \right)^{1/2} \leqslant M\tau\,|u|_{W_2^1[0,T]}, \tag{7}$$

де стала M не залежить від τ і u.

З оцінок (6), (7) випливає наступна теорема.

Теорема 1. *Нехай рівняння (1) стійке, тобто $(\lambda \geqslant 0)$ і розв'язок задачі (1) належить простору $W_2^1[0, T]$ (тобто $f(t) \in L_2[0, T]$). Тоді різницева схема (3) стійка при $\tau \leqslant \tau^* < \frac{2}{\lambda}$ і мають місце оцінки*

$$\|y - u\|_{L_2(\omega^-)} \leqslant M\tau\,|u|_{W_2^1[0,T]},$$

$$\|y - u\|_{C(\bar{\omega})} \leqslant M\tau\,|u|_{W_2^1[0,T]}, \tag{8}$$

$$|y - u|_{W_2^1(\omega^-)} \leqslant M\tau\,|u|_{W_2^1[0,T]},$$

де сталі M не залежать від τ і u. Ці оцінки означають також збіжність наближеного розв'язку y до точного u у відповідних нормах при $\tau \to 0$.

Аналогічні результати справедливі і для нелінійної задачі

$$\frac{du}{dt} = f(t, u), \quad t \in (0, T],$$

$$u(0) = u_0, \tag{9}$$

для якої виконуються умови:

$$f'_u \leqslant 0 \tag{10}$$

(умова стійкості рівняння (9), або умова монотонності функції $f(t, u)$ по u);

$$|f'_u| \leqslant L; \tag{11}$$

$$u(t) \in W_2^1[0, T]. \tag{12}$$

Інтегруючи (9) від $t \in \omega^-$ до $t + \tau$ і ділячи на τ, дістанемо

$$u_t - \tau^{-1} \int_t^{t+\tau} f(\xi, u(\xi))\, d\xi = 0. \tag{13}$$

Замінюючи наближено $\int_t^{t+\tau} f(\xi, u(\xi))\, d\xi$ на $\int_t^{t+\tau} f(\xi, u(t))\, d\xi$, маємо наступну схему для задачі (9):

$$y_t - \tau^{-1} \int_t^{t+\tau} f(\xi, y(t))\, d\xi = 0, \quad t \in \omega^-, \quad y_0 = u_0. \tag{14}$$

Для похибки $z = y - u$ маємо задачу

$$z_t - \tau^{-1} \int_t^{t+\tau} [f(\xi, y(t)) - f(\xi, u(\xi))]\, d\xi = 0, \tag{15}$$

$$t \in \omega^-, \quad z(0) = 0.$$

Додаючи і віднімаючи під інтегралом величину $f(\xi, u(t))$ і застосовуючи формулу Лагранжа, дістаємо

$$z_t + q(t) z(t) = \psi(t), \quad t \in \omega^-, \tag{16}$$

$$z(0) = 0,$$

де

$$q(t) = -\tau^{-1} \int_t^{t+\tau} f'_u d\xi, \quad \psi(t) = \tau^{-1} \int_t^{t+\tau} f'_u [u(t) - u(\xi)]\, d\xi. \tag{17}$$

Для лінійного обмеженого функціонала (при фіксованих t, ξ) $\eta(t; u) = u(t) - u(\xi)$ за допомогою леми Брембла — Гільберта аналогічно попередньому маємо

$$|\eta(u, t)| \leqslant M\tau^{1/2} |u|_{W_2^1(e+)} \tag{18}$$

і з урахуванням умов (10) — (12) з (17), (18) знаходимо

$$\|\psi\|_{L_2(\omega^-)} \leqslant M\tau |u|_{W_2^1[0,T]}, \tag{19}$$

де M не залежить від τ і u. Щоб визначити апріорну оцінку для z, помножимо (16) на τz, просумуємо по t від 0 до $t^* - \tau \in \bar{\omega}$ і врахуємо (10) — (12), (4′). В результаті маємо

$$0{,}5z^2(t^*) - 0{,}5\tau \sum_{t=0}^{t^*-\tau} \tau z_t^2(t) + \sum_{t=0}^{t^*-\tau} \tau q(t) z^2(t) = \sum_{t=0}^{t^*-\tau} \tau\psi(t) z(t),$$

звідки

$$z^2(t^*) \leqslant \tau \sum_{t=0}^{t^*-\tau} \tau z_t^2(t) + 2 \sum_{t=0}^{t^*-\tau} \tau\psi(t) z(t) \leqslant$$

$$\leqslant \tau \sum_{t=0}^{T-\tau} \tau z_t^2(t) + 2 \sum_{t=0}^{T-\tau} \tau |\psi(t)| |z(t)| =$$

$$= \tau \|z_t\|^2_{L_2(\omega^-)} + 2\|\psi\|_{L_2(\omega^-)} \|z\|_{L_2(\bar{\omega})} \leqslant$$

$$\leqslant \tau \|z_t\|^2_{L_2(\omega^-)} + 2T\|\psi\|_{L_2(\omega^-)} \|z_t\|_{L_2(\omega^-)}, \tag{20}$$

$$\|z\|^2_{L_2(\bar{\omega})} = \sum_{t^* \in \bar{\omega}} \tau z^2(t^*) \leqslant \tau T \|z_t\|^2_{L_2(\omega^-)} + 2T^2 \|\psi\|_{L_2(\omega^-)} \|z_t\|_{L_2(\omega^-)}.$$

Крім того, з (16), (17), (20) і нерівності $(a + b)^2 \leqslant 2(a^2 + b^2)$ (а також теорем вкладення п. 1.5 з ч. 1) маємо

$$\|z_t\|^2_{L_2(\omega^-)} \leqslant 2L^2 \|z\|^2_{L_2(\bar{\omega})} + 2\|\psi\|^2_{L_2(\omega^-)} \leqslant$$

$$\leqslant 2L^2\tau T \|z_t\|^2_{L_2(\omega^-)} + 4L^2T^2 \|\psi\|_{L_2(\omega^-)} \|z_t\|_{L_2(\omega^-)} + 2\|\psi\|^2_{L_2},$$

або

$$(1 - 2L^2\tau T) \|z_t\|^2_{L_2(\omega^-)} \leqslant 4L^2T^2 \|\psi\|_{L_2(\omega^-)} \times$$

$$\times \|z_t\|_{L_2(\omega^-)} + 2\|\psi\|^2_{L_2(\omega^-)}. \tag{21}$$

Ця оцінка має зміст, якщо $1 - 2L^2\tau T > 0$ або $\tau < 1/(2L^2T)$. У цьому разі, розв'язуючи (21) як квадратну нерівність відносно $\|z_t\|_{L_2(\omega^-)}$, дістаємо оцінку

$$\|z_t\|_{L_2(\omega^-)} \leqslant M \|\psi\|_{L_2(\omega^-)}, \tag{22}$$

де стала M не залежить від τ і u. Ця оцінка виражає умовну стійкість схеми (14) при $\tau < 1/(2L^2T)$.

Апріорна оцінка (22) і оцінка похибки апроксимації (19) приводять до наступного твердження.

Теорема 2. *Нехай для задачі* (9) *виконано умови* (10) — (12). *Тоді при* $\tau \to 0$ *розв'язок різницевої схеми* (14) *збігається з розв'язком задачі* (9), *причому виконуються оцінки* (8) *при* $\tau < 1/(2L^2T)$.

Зауваження 1. З викладеного випливає, що схему (14) слід застосувати у випадку, коли розв'язок задачі (9) має невисоку гладкість і інтеграли $\int_t^{t+\tau} f(\xi, y(t))\, d\xi$ обчислюються досить просто.

Зауваження 2. Проаналізувавши доведення оцінок (18), (19), зазначимо, що підвищення гладкості розв'язку $u(t)$ (наприклад, якщо $u \in W_2^2[0, T]$) не приведе до покращення оцінок (18), (19) і, отже, оцінки (8). Тому в цьому разі для підвищення порядку точності треба використовувати інші схеми.

Вправа 1. Розглянути неявну схему

$$y_t - \frac{1}{\tau}\int_t^{t+\tau} f(\xi, y(t+\tau))\,d\xi = 0, \quad t \in \omega^-, \quad y_0 = u_0,$$

і довести, що вона має перший порядок точності у випадку $u(t) \in W_2^1[0, T]$ і є абсолютно стійкою.

Нехай замість умови (12) виконується умова

$$u(t) \in W_2^2[0, T]. \tag{23}$$

Розглянемо наступну неявну однокрокову схему:

$$y_t(t) - \frac{1}{t}\int_t^{t+\tau} f(\xi, y(t) + (\xi - t)\,y_t(t))\,d\xi = 0, \tag{24}$$

$$t \in \omega^-, \quad y_0 = u_0.$$

Для похибки $z = y - u$ маємо задачу

$$z_t - \frac{1}{\tau}\int_t^{t+\tau} [f(\xi, y(t) + (\xi - t)\,y_t(t)) - f(\xi, u(\xi))]\,d\xi = 0,$$

$$t \in \omega^-, \quad z(0) = 0.$$

Додаючи і віднімаючи у квадратних дужках величину $f(\xi, u(t) + (\xi - t)\,u_t(t))$ і застосовуючи формулу Лагранжа, дістанемо

$$z_t + q(t)\,z(t) + q_1(t)\,z_t(t) = \psi(t), \quad t \in \omega^-, \tag{25}$$
$$z(0) = 0,$$

де

$$q(t) = -\tau^{-1}\int_t^{t+\tau} f'_u d\xi, \quad q_1(t) = -\tau^{-1}\int_t^{t+\tau} (\xi - t)\,f'_u d\xi,$$

$$\psi(t) = \tau^{-1}\int_t^{t+\tau} [f(\xi, u(t) + (\xi - t)\,u_t(t)) - f(\xi, u(\xi))]\,d\xi =$$

$$= \tau^{-1}\int_t^{t+\tau} f'_u \eta(t, \xi; u)\,d\xi, \quad \eta(t, \xi; u) = u(t) + (\xi - t)\,u_t(t) - u(\xi).$$

Зрозуміло, що $\eta(t, \xi; u)$ є лінійним функціоналом від u (при фіксованих t, ξ). Виконавши масштабування, аналогічно попередньому маємо $\eta(t, \xi; u) = \eta(\tilde{u})$, $\tilde{u}(s) = u(\xi(s))$, причому $\tilde{u} \in W_2^2[0, 1]$, і лінійний функціонал $\eta(\tilde{u})$ обмежений у просторі $C[0, 1]$ і в силу теореми вкладення — у просторі $W_2^2[0, 1]$. Він також перетворюється в нуль на многочленах першого степеня. За допомогою леми Брембла — Гільберта дістанемо оцінки

$$|\eta(\tilde{u})| \equiv |\eta(t, \xi; u)| \leqslant M\overline{M}\,|\tilde{u}|_{W_2^2[0,1]} \leqslant$$

$$\leqslant M\tau^{1.5}\,|u|_{W_2^2(e+)}, \quad u \in W_2^2[0, T], \tag{26}$$

$$\|\psi\|_{L_2(\omega^-)} \leqslant M\tau^2\,|u|_{W_2^2[0,T]},$$

де сталі M не залежать від u і τ.

Щоб дістати апріорну оцінку розв'язку задачі (25), перепишемо її у вигляді

$$z_t(t) + Q(t)\,z(t) = \Psi(t), \quad t \in \omega^-, \quad z(0) = 0, \tag{27}$$

де

$$G(t) = q(t)/(1 + q_1(t)), \quad \Psi(t) = \psi(t)/(1 + q_1(t)).$$

Неважко помітити, що при досить малих $\tau \leqslant \tau^*$ таких, що $1 - L\tau^*/2 > 0$, де τ^* — фіксоване число, мають місце оцінки

$$|Q(t)| \leqslant |q(t)|/(1 - L\tau^*/2),$$
$$|\Psi(t)| \leqslant |\psi(t)|/(1 - L\tau^*/2),$$

тому при досить малих τ для розв'язку задачі (25) має місце та сама апріорна оцінка (22), що і для задачі (16), а в силу (26)

$$\|\Psi\|_{L_2(\omega^-)} \leqslant M\tau^2\,|u|_{W_2^2[0,T]}.$$

Тому справедливе таке твердження.

Теорема 3. *Нехай для задачі* (9) *виконуються умови* (10), (11), (23). *Тоді при* $\tau \to 0$ *розв'язок задачі* (24) *збігається з розв'язком задачі* (9) *і при* $\tau \leqslant \tau^*$, $\tau^* < \frac{2}{L}$, *виконуються оцінки*

$$\|y - u\|_{L_2(\bar{\omega})} \leqslant M\tau^2\,|u|_{W_2^2[0,T]},$$

$$\|y - u\|_{C(\bar{\omega})} \leqslant M\tau^2\,|u|_{W_2^2[0,T]}, \tag{28}$$

$$|y - u]_{W_2^1(\omega^-)} \leqslant M\tau^2\,|u|_{W_2^2[0,T]},$$

де сталі M не залежать від τ і u.

Вправа 2. Довести теорему 3 для схеми

$$y_t(t) - \tau^{-1} \int_t^{t+\tau} f(\xi,\ y(t+\tau) + (\xi - t - \tau)\, y_t(t))\, d\xi = 0, \qquad (29)$$

$$t \in \omega^-, \quad y(0) = u_0.$$

Вправа 3. Довести, що схема (29) безумовно стійка.

Вправа 4. Довести, що схема (24) умовно стійка при $\tau\lambda \leqslant 1$.

3.7. Багатокрокові схеми розв'язування задачі Коші

В п. 3.3 розглянуто методи Рунге — Кутта для чисельного розв'язування задачі Коші

$$\frac{du}{dt} = f(t, u), \quad 0 < t \leqslant T, \quad u(0) = u_0. \qquad (1)$$

Ці методи, як було зазначено, називаються *однокроковими*, бо при визначенні нового значення y_{n+1} використовується лише значення y_n. З викладеного зрозуміло, що для підвищення порядку однокрокових методів (наприклад, Рунге — Кутта) треба більше обчислень значень функції $f(t, u)$ на кожному кроці, що призводить до збільшення обчислювальної роботи, особливо у випадку розв'язування систем звичайних диференціальних рівнянь високого порядку. Можна довести (див., наприклад, Butcher J. C. The nonexistence of ten stage eight order caplicit Runge — Kutta methods. BIT 25, 521—540 (1985)), що для m-ступеневого явного методу Рунге — Кутта має місце такий взаємозв'язок між m та можливим максимальним порядком точності p^*.

m	1	2	3	4	5	6	7	8	9	>9
p^*	1	2	3	4	4	5	6	6	7	$\leqslant m-3$

Щоб зменшити кількість обчислень правої частини $f(t, u)$ на кожному кроці і при цьому забезпечити високий порядок точності, застосовуються *багатокрокові методи*. В загальному випадку для визначення наближеного розв'язку можна розглянути лінійні відносно m-крокові схеми (методи), $m \geqslant 1$, тобто схеми вигляду

$$\sum_{k=0}^{m} \frac{a_k}{\tau} y_{n-k} = \sum_{k=0}^{m} b_k f_{n-k}, \quad n = m, m+1, \ldots, \qquad (2)$$

де a_k, b_k — числові коефіцієнти, $f_{n-k} = f(t_{n-k},\ y_{n-k})$, $a_0 \neq 0$, $b_m \neq 0$, y_j — наближення для $u(t_j)$.

Схема (2) називається *явною (екстраполяційною)*, якщо $b_0 = 0$ і y_n визначається через попередні значення y_{n-1}, y_{n-2}, ..., y_{n-m} за явною формулою

$$y_n = \frac{1}{a_0} \sum_{k=1}^{m} (b_k \tau f_{n-k} - a_k y_{n-k}) = \frac{1}{a_0} F(y_{n-1}, y_{n-2}, \ldots, y_{n-m}). \qquad (3)$$

Обчислення тут починаються з y_m, тобто з $n = m$. Щоб знайти y_m, треба задати m початкових значень $y_0, y_1, \ldots, y_{m-1}$, а їх можна визначити, наприклад, за методом Рунге — Кутта.

Якщо $b_0 \neq 0$, то схема (2) називається неявною (інтерполяційною) і для відшукання y_n при кожному n треба розв'язати нелінійне рівняння

$$a_0 y_n - \tau b_0 f(t_n, y_n) = F(y_{n-1}, y_{n-2}, \ldots, y_{n-m}). \qquad (4)$$

Похибка апроксимації (нев'язка) схеми (2) на розв'язку $u = u(t)$ рівняння (1) визначається за формулою

$$\psi(t_j) \equiv \psi(t_j;\ u) \equiv \psi(t_j;\ u;\ \tau) =$$

$$= \sum_{k=0}^{m} b_k f(t_{j-k},\ u_{j-k}) - \frac{1}{\tau} \sum_{k=0}^{m} a_k u_{j-k}. \qquad (5)$$

Схема (2) має s-й порядок апроксимації у деякій сітковій нормі $\|\cdot\|$, якщо

$$\|\psi\| = O(\tau^s) \text{ або } \|\psi\| \leqslant M\tau^s, \quad s > 0, \qquad (6)$$

де M — стала, яка не залежить від τ. Схема (2) має s-й порядок апроксимації в точці t_j, якщо $\psi(t_j) = O(\tau^s)$.

Коефіцієнти a_k, b_k підбирають, виходячи з вимог апроксимації і стійкості. Оскільки коефіцієнти рівняння (2) визначаються з точністю до сталого множника, то без обмеження загальності вважаємо, що

$$\sum_{k=0}^{m} b_k = 1. \qquad (7)$$

3.7.1. Метод невизначених коефіцієнтів побудови багатокрокових схем. Виберемо коефіцієнти a_k, b_k таким чином, щоб виконувались умови

$$\lim_{\tau \to 0} \sum_{k=0}^{m} \frac{a_k u(t_j - k\tau)}{\tau} = u'(t_j), \quad 0 \leqslant t_j \leqslant T, \qquad (8)$$

$$\lim_{\tau \to 0} \sum_{k=0}^{m} b_k f(t_j - k\tau, \quad u(t_j - k\tau)) = f(t_j, u(t_j)),$$

$$0 \leqslant t_j \leqslant T, \qquad (9)$$

$$|\psi(t_j;\ u;\ \tau)| \equiv |\psi_j(\tau)| = O(\tau^s). \qquad (10)$$

Для цього значення $u(t_j - k\tau)$ і $f(t_j - k\tau, u(t_j - k\tau))$ подамо за формулою Тейлора

$$u(t_j - k\tau) = \sum_{l=0}^{s} \frac{(-k\tau)^l u^l(t_j)}{l!} + r_j^k, \tag{11}$$

$$u'(t_j - k\tau) = \sum_{l=0}^{s-1} \frac{(-k\tau)^l u^{(l+1)}(t_j)}{l!} + p_j^k, \tag{12}$$

де r_j^k, p_j^k — залишкові члени рядів Тейлора, причому

$$|r_j^k| \leqslant \frac{c_{s+1}(k\tau)^{s+1}}{(s+1)!}, \quad |p_j^k| \leqslant \frac{c_{s+1}(k\tau)^s}{s!}, \tag{13}$$

$$|u^{(s+1)}(t)| \leqslant c_{s+1} < \infty, \quad 0 \leqslant t \leqslant T.$$

Підставляючи (11), (12) в (5), маємо

$$\psi_j(\tau) = -\sum_{k=0}^{m} \frac{a_k}{\tau} u(t_j) - \sum_{k=0}^{m} (-ka_k - b_k) u'(t_j) -$$

$$- \sum_{k=0}^{m} \left[\frac{k^2 a_k}{2!} + \frac{kb_k}{1!} \right] \tau u''(t_j) - \cdots - \sum_{k=0}^{m} \left[\frac{(-k)^s a_k}{s!} - \right.$$

$$\left. - \frac{(-k)^{s-1} b_k}{(s-1)!} \right] \tau^{s-1} u^{(s)}(t_j) - \sum_{k=0}^{m} \frac{a_k r_j^k}{\tau} + \sum_{k=0}^{m} b_k p_j^k. \tag{14}$$

Введемо позначення

$$\sum_{k=0}^{m} a_k = \varphi_0,$$

$$\sum_{k=0}^{m} \left[\frac{-k}{1!} a_k - b_k \right] = \varphi_1, \tag{15}$$

$$\sum_{k=0}^{m} \left[\frac{(-k)^l}{l!} a_k - \frac{(-k)^{l-1}}{(l-1)!} b_k \right] = \varphi_l, \quad l > 1.$$

Якщо

$$\varphi_0 = \varphi_1 = \cdots = \varphi_s = 0, \tag{16}$$

то

$$|\psi_j(\tau)| \leqslant c_{s+1} \tau^s \sum_{k=0}^{m} \left(|a_k| \frac{k^{s+1}}{(s+1)!} + |b_k| \frac{k^s}{s!} \right). \tag{17}$$

Розглянемо поліноми

$$\rho(q) = \sum_{k=0}^{m} a_k q^{m-k}, \quad \sigma(q) = \sum_{k=0}^{m} b_k q^{m-k}.$$

Неважко помітити, що в термінах цих поліномів умови (16), за яких схема (2) має s-й порядок апроксимації, можна записати також у вигляді

$$\rho_1(1) = 0,$$

$$\rho_{j+1}(1) = j\sigma_j(1), \quad j = \overline{1, s},$$

де поліноми ρ_j та σ_j визначаються рекурентними формулами

$$\rho_1(q) \equiv \rho(q), \quad \sigma_1(q) \equiv \sigma(q),$$

$$\rho_{j+1}(q) = q\rho_j'(q), \quad \sigma_{j+1}(q) = q\sigma_j'(q).$$

Співвідношення (16) утворюють систему лінійних однорідних рівнянь з $2m + 2$ невідомими a_k, b_k, $k = \overline{0, m}$. Ця система матиме ненульовий розв'язок при $2m + 2 > s + 1$, а для $\psi_j(\tau)$ справедлива оцінка $\psi_j(\tau) = O(\tau^s)$.

Умови (8), (9) при виконанні рівностей (7), (16) мають місце. Дійсно, користуючись розвиненням в ряд Тейлора, дістаємо

$$\lim_{\tau \to 0} \sum_{k=0}^{m} \frac{a_k u(t_j - k\tau)}{\tau} = \lim_{\tau \to 0} \left(\sum_{k=0}^{m} \frac{a_k}{\tau} u(t_j) - \right.$$

$$\left. - \sum_{k=0}^{m} k a_k u'(t_j) + O(\tau) \right) = -u'(t_j) \sum_{k=0}^{m} k a_k,$$

$$\lim_{\tau \to 0} \sum_{k=0}^{m} b_k f(t_j - k\tau, u(t_j - k\tau)) = \lim_{\tau \to 0} \left(\sum_{k=0}^{m} b_k f(t_j, u(t_j)) + O(\tau) \right) =$$

$$= f(t_j, u(t_j)).$$

Проте в силу (15) при $l = 1$ маємо $-\sum_{k=0}^{m} k a_k - \sum_{k=0}^{m} b_k = 0$, а в силу (7) — $\sum_{k=0}^{m} k a_k = 1$. Отже,

$$\lim_{\tau \to 0} \sum_{k=0}^{m} \frac{a_k u(t_j - \delta\tau)}{\tau} = u'(t_j)$$

і умови (8) — (10) виконані. Якщо заздалегідь зафіксувати значення деяких параметрів, то порядок апроксимації зменшуватиметься.

Приклад 1. Нехай $m = 2$, $s = 4$. Система для визначення коефіцієнтів a_k, b_k ($k = \overline{0, 2}$) має вигляд

$$\begin{gathered} a_0 + a_1 + a_2 = 0 \quad (\varphi_0 = 0), \\ a_1 + 2a_2 + b_0 + b_1 + b_2 = 0 \quad (\varphi_1 = 0), \\ a_1/2 + 2a_2 + b_1 + 2b_2 = 0 \quad (\varphi_2 = 0), \\ -a_1/6 - (4/3)\, a_2 - b_1/2 - 2b_2 = 0 \quad (\varphi_3 = 0), \\ a_1/24 + (2/3)\, a_2 + b_1/6 + (4/3)\, b_2 = 0 \quad (\varphi_4 = 0), \\ b_0 + b_1 + b_2 = 1. \end{gathered} \tag{18}$$

Звідси

$$a_1 = 0, \quad a_2 = -a_0, \quad b_0 = b_2 = \frac{a_0}{3}, \quad b_1 = \frac{4a_0}{3}. \tag{19}$$

З умови (7) знаходимо $a_0 = 1/2$. Таким чином, схема четвертого порядку апроксимації має вигляд

$$y_j = y_{j-2} + \frac{\tau}{3}\,(f_j + 4f_{j-1} + f_{j-2}). \tag{20}$$

Приклад 2. Нехай $m = 2$, $b_0 = 0$, $a_2 = 0$, тобто розглянемо систему вигляду

$$b_0 = 0, \ a_2 = 0, \ a_0 + a_1 = 0 \ (\varphi_0 = 0), \quad a_1 + b_1 + b_2 = 0 \ (\varphi_1 = 0),$$
$$a_1/2 + b_1 + 2b_2 = 0 \quad (\varphi_2 = 0).$$

Розв'язком цієї системи, який задовольняє також умову (7), є

$$a_0 = 1, \ a_1 = -1, \ a_2 = 0, \ b_0 = 0, \ b_1 = 3/2, \ b_2 = -1/2. \tag{21}$$

Таким чином, маємо схему другого порядку апроксимації

$$y_j = y_{j-1} + \tau/2\,(3f_{j-1} - f_{j-2}). \tag{22}$$

Вона називається *екстраполяційною формулою Адамса.*

Приклад 3. Нехай $m = 2$ і $a_2 = 0$. Тоді система $a_2 = \varphi_0 = \varphi_1 = \varphi_2 = \varphi_3 = 0$, $\sum_{k=0}^{2} b_k = 1$ має розв'язок

$$a_0 = 1, \ a_1 = -1, \ b_0 = 5/12, \ b_1 = 2/3, \ b_2 = -1/12$$

і схема третього порядку апроксимації має вигляд

$$y_j = y_{j-1} + \tau\left(\frac{5}{12} f_j + \frac{2}{3} f_{j-1} - \frac{1}{12} f_{j-2}\right), \tag{23}$$

або

$$y_j = y_{j-1} + \tau\left(f_j - \frac{1}{2}\Delta f_j - \frac{1}{12}\Delta^2 f_{j-2}\right). \tag{24}$$

Це інтерполяційна формула Адамса.

Вправа 1. Методом невизначених коефіцієнтів побудувати однокрокову формулу другого порядку апроксимації.

Відповідь: $y_j = y_{j-1} + \frac{\tau}{2}\,(f_j + f_{j-1})$.

Вправа 2. Методом невизначених коефіцієнтів побудувати явну двокрокову схему третього порядку апроксимації.

Відповідь: $y_j = -4y_{j-1} + 5y_{j-2} + \tau\,(4f_{j-1} + 2f_{j-2})$.

3.7.2. Метод Адамса побудови багатокрокових схем. Розглянемо праву частину $f(x, u)$ рівняння (1) не на всій площині зміни її аргументів, а лише на кривій $u(t)$, що відповідає шуканому розв'язку. Тоді $f(t, u(t)) = \tilde{F}(t)$ є функцією лише одного аргументу t. Нехай вже відомий наближений розв'язок у деяких точках сітки y_n, y_{n-1}, ..., y_{n-m}. Тоді у цих точках відомі також $F(t_k) = f(t_k, y_k) \approx \approx f(t_k, u_k)$. В околі цих вузлів можна наближено замінити $\tilde{F}(t)$ інтерполяційним многочленом, який запишемо у формі Ньютона:

$$\begin{aligned} \tilde{F}(t) \approx p(t;\, F) = F(t_n) + (t - t_n)\, F(t_n, t_{n-1}) + \\ + (t - t_n)(t - t_{n-1})\, F(t_n, t_{n-1}, t_{n-2}) + \\ + (t - t_n)(t - t_{n-1})(t - t_{n-2})\, F(t_n, t_{n-1}, t_{n-2}, t_{n-3}) + \ldots. \end{aligned} \tag{25}$$

Обмежимося лише записаними членами, які забезпечують четвертий порядок апроксимації. Для обчислення розв'язку в наступній точці запишемо диференціальне рівняння в інтегральній формі

$$u_{n+1} = u_n + \int_{t_n}^{t_{n+1}} f(t, u(t))\, dt = u_n + \int_{t_n}^{t_{n+1}} \tilde{F}(t)\, dt \tag{26}$$

і підставимо в нього інтерполяційний многочлен (25). Дістанемо екстраполяційну формулу Адамса зі змінним кроком

$$\begin{aligned} y_{n+1} = y_n + \tau_n F(t_n) + \frac{1}{2}\tau_n^2 F(t_n, t_{n-1}) + \\ + \frac{1}{6}\tau_n^2 (2\tau_n + 3\tau_{n-1})\, F(t_n, t_{n-1}, t_{n-2}) + \\ + \frac{1}{12}\tau_n^2 (3\tau_n^2 + 8\tau_n\tau_{n-1} + 4\tau_n\tau_{n-2} + 6\tau_n^2 + \\ + 6\tau_{n-1}\tau_{n-2})\, F(t_n, t_{n-2}, t_{n-3}), \end{aligned} \tag{27}$$

де $\tau_n = t_{n+1} - t_n$. Ця формула має четвертий порядок апроксимації. Якщо відкинути останній доданок, то дістанемо формулу третього порядку апроксимації і т. д. Формула першого порядку збігається зі схемою Ейлера.

Частіше користуються формулою (27) для сталого кроку τ. Записуючи замість розділених різниць скінченні різниці за формулою $\Delta^p F_n = p!\, F(t_n, t_{n-1}, \ldots, t_{n-p})$, маємо

$$y_{n+1} = y_n + \tau F_n + \frac{\tau^2}{2}\Delta F_n + \frac{5}{12}\tau^3\Delta^2 F_n + \frac{3}{8}\tau^4\Delta^3 F_n. \tag{28}$$

Залишковий член цієї формули дорівнює $(251/750)\,\tau^5 F^{(4)}(\xi)$ (залишковий член формули Адамса виражається через залишковий член інтерполяційного многочлена).

Відрізок інтегрування вихідного рівняння та точки інтерполяції можна вибирати і в більш загальному вигляді.

Інтегруючи диференціальне рівняння $u'(t) = f(t, u(t))$, маємо

$$u(t_{i+1}) = u(t_{i-j}) + \int_{t_{i-j}}^{t_{i+1}} f(t, u(t))\,dt, \quad j \geqslant 0.$$

Замінимо $f(t, u(t))$ інтерполяційним многочленом p-го степеня, побудованим за точками $(t_i, f(t_i, y_i)), \ldots, (t_{i-p}, f(t_{i-p}, y_{i-p}))$, де y_j — наближення для $u(t_j)$. В результаті дістаємо явний метод

$$y_{i+1} = y_{i-j} + \sum_{k=0}^{p} f(t_{i-k}, y_{i-k}) \int_{t_{i-j}}^{t_{i+1}} \prod_{\substack{l=0 \\ l\neq k}}^{p} \frac{t - t_{i-l}}{t_{i-k} - t_{i-l}}\,dt,$$

в якому інтеграли

$$\int_{t_{i-j}}^{t_{i+1}} \prod_{\substack{l=0 \\ l\neq k}}^{p} \frac{t - t_{i-l}}{t_{i-k} - t_{i-l}}\,dt \equiv \beta_{i,j,p,k}$$

можна обчислити точно. При $p = j = 0$ маємо відомий вже метод Ейлера — Коші. Для довільного p та $j = 0$ дістаємо методи, які називаються методами Адамса — Башфорта. Для рівномірної сітки $t_{i+1} - t_i \equiv \tau$ коефіцієнти $\tau^{-1}\beta_{i,0,p,k}$ методу Адамса — Башфорта визначаються за такою таблицею:

p \ k	0	1	2	3	Порядок апроксимації
0	1				1
1	3/2	—1/2			2
2	23/12	—16/12	5/12		3
3	55/24	—59/24	37/24	—9/24	4

При $j = 1$ дістаємо методи Ністрьома, найпростіший з яких має вигляд

$$y_{i+1} = y_{i-1} + 2\tau f(t_i, y_i)$$

і називається *явним методом (правилом) середньої точки* (порядку 2).

Якщо при інтерполяції $f(t, u(t))$ використати додатково ще вузол $(t_{i+1}, f(t_{i+1}, y_{i+1}))$, то дістанемо *неявні методи* вигляду

$$y_{i+1} = y_{i-j} + \sum_{k=-1}^{p} f(t_{i-k}, y_{i-k}) \int_{t_{i-j}}^{t_{i+1}} \prod_{\substack{l=-1 \\ l\neq k}}^{p} \frac{t - t_{i-l}}{t_{i-k} - t_{i-l}}\,dt,$$

в яких, як неважко помітити, y_{i+1} треба знаходити як розв'язок нелінійного (якщо $f(t, u)$ нелінійна по u) рівняння.

При $j = 0$ дістанемо методи Адамса — Мултона, коефіцієнти

$$\overline{\beta}_{i,j,p,k} = \int_{t_{i-j}}^{t_{i+1}} \prod_{\substack{l=-1 \\ l\neq k}}^{p} \frac{t - t_{i-l}}{t_{i-k} - t_{i-l}}\,dt$$

яких для рівномірної сітки подано у наступній таблиці:

p \ k	—1	0	1	2	Порядок апроксимації
—1	1				1
0	1/2	1/2			2
1	5/12	8/12	—1/12		3
2	9/24	19/24	—5/24	1/24	4

При $p = -1$ маємо неявний метод Ейлера

$$y_{i+1} = y_i + \tau f(t_{i+1}, y_{i+1}),$$

при $p = 0$ — метод трапецій

$$y_{i+1} = y_i + \frac{\tau}{2}\left(f(t_i, y_i) + f(t_{i+1}, y_{i+1})\right).$$

Неявні методи самі по собі використовуються лише в спеціальних випадках, оскільки для відшукання y_{i+1} треба розв'язувати нелінійне рівняння. Якщо в формулах Адамса — Мултона замінити величину $f(t_{i+1}, y_{i+1})$ наближеною величиною $f(t_{i+1}, \tilde{y}_{i+1})$, де $\tilde{y}_{i+1}$ здобуте за допомогою деякого явного методу (як правило, методу Адамса — Башфорта), то дістанемо методи предиктор-коректор, які часто використовуються на практиці. Наприклад, комбінація трикрокового методу Адамса — Мултона четвертого порядку (коректор) та трикрокового методу Адамса — Башфорта третього порядку (предиктор) дає явний трикроковий метод четвертого порядку, що вимагає лише двох обчислень функції f на кожному кроці

$$\tilde{y}_{i+1} = y_i + \frac{\tau}{12}\left(33f(t_i, y_i) - 16f(t_{i-1}, y_{i-1}) + 5f(t_{i-2}, y_{i-2})\right),$$

$$y_{i+1} = y_i + \frac{\tau}{24}(9f(t_{i+1}, \tilde{y}_{i+1}) + 19f(t_i, y_i) - 5f(t_{i-1}, y_{i-1}) + f(t_{i-2}, y_{i-2}))$$

(для порівняння: однокроковий класичний метод Рунге — Кутта четвертого порядку потребує чотирьох обчислень функції f). Комбінація p-крокового методу Адамса — Мултона та $(p+1)$-крокового методу Адамса — Башфорта (предиктор) дає явний $(p+1)$-кроковий метод

$$\tilde{y}_{i+1} = y_i + \sum_{k=0}^{p} f(t_{i-k}, y_{i-k}) \int_{x_i}^{x_{i+1}} \prod_{\substack{l=0 \\ l \neq k}}^{p} \frac{t - t_{i-l}}{t_{i-k} - t_{i-l}}\, dt, \quad i \geqslant p,$$

$$y_{i+1} = y_i + f(t_{i+1}, \tilde{y}_{i+1}) \int_{t_i}^{t_{i+1}} \prod_{l=1}^{p} \frac{t - t_{i+1-l}}{t_{i+1} - t_{i+1-l}}\, dt +$$

$$+ \sum_{k=1}^{p} f(t_{i+1-k}, y_{i+1-k}) \int_{t_i}^{t_{i+1}} \prod_{\substack{l=0 \\ l \neq k}}^{p} \frac{t - t_{i+1-l}}{t_{i+1-k} - t_{i+1-l}}\, dt.$$

Цей метод має порядок $p+1$ відносно $\tau = \max\limits_j \tau_j$ для довільної послідовності кроків $\{\tau_j\}$.

Зазначимо, що багатокрокові методи предиктор-коректор мають вигляд, відмінний від (2), і є нелінійними відносно значень функції f_i.

Приклад 1. Знайти екстраполяційну формулу Адамса другого порядку апроксимації шляхом заміни функції $f(t, u)$ інтерполяційним многочленом, побудованим за точками t_{n-1}, t_{n-2} (методом Адамса).

Розв'язання. Інтегруючи диференціальне рівняння, маємо

$$u_n = u_{n-1} + \int_{t_{n-1}}^{t_n} f(t, u(t))\, dt = u_n + \int_{t_{n-1}}^{t_n} \tilde{F}(t)\, dt. \qquad (29)$$

Побудуємо інтерполяційний многочлен для функції $\tilde{F}(t)$ за точками t_{n-1}, t_{n-2} (вони розміщені зовні відрізка інтегрування $[t_{n-1}, t_n]$, тому можна говорити про екстраполяцію)

$$p(t; \tilde{F}) = \tilde{F}[t_{n-1}] + (t - t_{n-1})\tilde{F}[t_{n-1}, t_{n-2}].$$

Далі маємо

$$\tilde{F}(t_{n-1}) = f(t_{n-1}, u(t_{n-1})), \quad \tilde{F}[t_{n-1}, t_{n-2}] = \frac{\tilde{F}(t_{n-2}) - \tilde{F}(t_{n-1})}{t_{n-2} - t_{n-1}} =$$

$$= \frac{f(t_{n-2}, u(t_{n-2})) - f(t_{n-1}, u(t_{n-1}))}{-\tau}.$$

Нам невідомі точні значення $u(t_{n-1})$, $u(t_{n-2})$, тому підставимо замість них наближені: y_{n-1} і y_{n-2}. Тоді

$$F(t_{n-1}) = f(t_{n-1}, y_{n-1}) \equiv f_{n-1},$$

$$F[t_{n-1}, t_{n-2}] = \frac{f(t_{n-2}, y_{n-2}) - f(t_{n-1}, y_{n-1})}{-\tau} \equiv \frac{f_{n-1} - f_{n-2}}{\tau}, \qquad (30)$$

$$p(t; F) = f_{n-1} + (t - t_{n-1})\frac{f_{n-2} - f_{n-1}}{-\tau}.$$

Підставляючи $p(t; F)$ замість $\tilde{F}(t)$ в (29), дістанемо наближену формулу

$$y_n = y_{n-1} + \int_{t_{n-1}}^{t_n} \left[f_{n-1} + (t - t_{n-1})\frac{f_{n-1} - f_{n-2}}{\tau} \right] dt.$$

Виконуючи інтегрування, маємо екстраполяційну формулу Адамса другого порядку апроксимації

$$y_{n+1} = y_n + \frac{\tau}{2}(3f_{n-1} - f_{n-2}),$$

яка збігається з формулою (22), знайденою за методом невизначених коефіцієнтів.

Приклад 2. Знайти інтерполяційну формулу Адамса третього порядку апроксимації шляхом заміни функції $f(t, u)$ інтерполяційним многочленом, який побудовано за точками (тобто методом Адамса) t_n, t_{n-1}, t_{n-2}.

Розв'язування. Побудуємо інтерполяційний многочлен для функції $F(t)$ за точками t_n, t_{n-1}, t_{n-2}:

$$p(t; \tilde{F}) = \tilde{F}(t_n) + (t - t_n)\tilde{F}[t_n, t_{n-1}] + (t - t_n)(t - t_{n-1})\tilde{F}[t_n, t_{n-1}, t_{n-2}].$$

Введемо заміни

$$\tilde{F}(t_n) = f(t_n, u(t_n)) \approx f(t_n, y_n) = f_n \equiv F(t_n),$$

$$\tilde{F}[t_n, t_{n-1}] = \frac{f(t_{n-1}, u(t_{n-1})) - f(t_n, u(t_n))}{t_{n-1} - t_n} \approx$$

$$\approx \frac{f_n - f_{n-1}}{\tau} \equiv F[t_n, t_{n-1}],$$

$$\tilde{F}[t_n, t_{n-1}, t_{n-2}] = \frac{\tilde{F}[t_{n-1}, t_{n-2}] - \tilde{F}[t_n, t_{n-1}]}{t_{n-2} - t_n} \approx$$

$$\approx \frac{(f_n - f_{n-1})/\tau - (f_{n-1} - f_{n-2})/\tau}{2\tau} = \frac{f_n - 2f_{n-1} + f_{n-2}}{2\tau^2} \equiv F[t_n, t_{n-1}, t_{n-2}];$$

$$p(t; F) = f_n + (t - t_n)\frac{f_n - f_{n-1}}{\tau} + (t - t_n)(t - t_{n-1})\frac{f_n - 2f_{n-1} + f_{n-2}}{2\tau^2}. \qquad (31)$$

Підставляючи замість $\tilde{F}(t)$ в (29) наближене значення її інтерполяційного многочлена, яке обчислимо за формулою (31), і інтегруючи, дістаємо наступну інтерполяційну

формулу Адамса, знайдену методом невизначених коефіцієнтів (див. (23)):

$$y_n = y_{n-1} + \tau\left(\frac{5}{12} f_n + \frac{2}{3} f_{n-1} - \frac{1}{12} f_{n-2}\right).$$

Ця формула має третій порядок апроксимації.

3.7.3. Методи побудови багатокрокових схем, що грунтуються на формулах чисельного диференціювання. В цих методах виходять з рівності

$$u'(t_{i+1}) = f(t_{i+1}, u(x_{i+1})) \tag{32}$$

і похідну в лівій частині заміняють похідною інтерполяційного многочлена, обчисленого в точці t_{i+1}. Побудуємо, наприклад, інтерполяційний многочлен $(k+1)$-го степеня $p(t)$ за точками (t_{i+1}, y_{i+1}), ..., (t_{i-k}, y_{i-k}) і визначимо невідомі y_{i+1}, ..., y_{i-k} з рівняння, яке є аналогом (32):

$$\frac{dp(t)}{dt}\Big|_{t=t_{i+1}} = f(t_{i+1}, y_{i+1}). \tag{33}$$

Многочлен $p(t)$ можна, наприклад, записати в формі Ньютона

$$p(t) = y_{i+1} + \frac{y_i - y_{i+1}}{t_i - t_{i+1}}(t - t_{i+1}) + y[i+1, i, i-1](t - t_{i+1}) \times$$

$$\times (t - t_i) + \cdots + y[i+1, \ldots, i-k](t - t_{i+1})(t - t_i) \times \cdots$$

$$\cdots \times (t - t_{i+1-k}),$$

де $y[i+1, \ldots, i-j]$—поділена різниця $j+1$-го порядку. Співвідношення (33) можна записати у вигляді

$$y_{i+1} = \sum_{j=0}^{k} \tilde{\alpha}_{k,j} y_{i-j} + \tilde{\beta} f(x_{i+1}, y_{i+1})$$

з відповідними сталими $\tilde{\alpha}_{k,j}$, $\tilde{\beta}$, а у випадку рівновіддалених точок x_i $(x_{i+1} - x_i = \tau \ \forall\, i)$ у вигляді

$$y_{i+1} = \sum_{j=0}^{k} \alpha_{k,j} y_{i-j} + \tau\beta f(x_{i+1}, y_{i+1}). \tag{34}$$

Коефіцієнти $\alpha_{k,j}$, β неважко обчислити, при різних k вони визначаються за такою таблицею:

k	β	$\alpha_{k,0}$	$\alpha_{k,1}$	$\alpha_{k,2}$	$\alpha_{k,3}$	$\alpha_{k,4}$	$\alpha_{k,5}$
0	1	1					
1	2/3	4/3	—1/3				
2	6/11	18/11	—9/11	2/11			
3	12/25	48/25	—36/25	16/25	—3/25		
4	60/137	300/137	—300/137	200/137	—75/137	12/137	
5	60/147	360/147	—450/147	400/147	—225/147	72/147	—10/147

(Неважко помітити, що при $k = 0$ маємо неявну схему Ейлера. Формули (34) відомі також як формули Гіра, або метод BDF (backward differentiation formula). Для $k > 5$ ці формули не використовуються, бо для них порушується умова коренів (див. п. 8).

3.8. Збіжність і точність багатокрокових методів

3.8.1. Умова коренів. Стійкість лінійних різницевих рівнянь. Розглянемо умови збіжності і похибку розв'язування задачі Коші багатокроковим методом (див. п. 3.7). Застосуємо цей метод до задачі

$$\frac{du}{dt} = 0, \quad t > 0; \quad u(0) = u_0. \tag{1}$$

Дістанемо рекурентне співвідношення (різницеве рівняння)

$$a_0 y_n + a_1 y_{n-1} + \cdots + a_m y_{n-m} = 0, \tag{2}$$

$$n = m, m+1, \ldots.$$

Щоб виконувати розрахунки за формулою (2), треба мати початкові значення y_0, y_1, ..., y_{m-1}, які вважаємо відомими. Необхідно, щоб розв'язок рівняння (2) відображав поведінку точного розв'язку задачі (1), який є сталим: $u(t) \equiv u_0$, Тому природно вважати y_n обмеженими при $n \to \infty$. Оскільки розв'язок рівняння (2) зі сталими коефіцієнтами виражається через корені характеристичного рівняння

$$p(q) \equiv a_0 q^m + a_1 q^{m-1} + \cdots + a_m = 0, \tag{3}$$

то ці корені мають бути розміщені в одиничному крузі $|q| \leqslant 1$ комплексної площини, причому на границі круга немає кратних коренів. Ця умова на корені рівняння (3) називається *умовою коренів*. Виявляється, що умова коренів забезпечує не лише обмеженість розв'язку рівняння (2) при $n \to \infty$, але має суттєве значення для збіжності розв'язку методу (2) з п. 3.7 до точного розв'язку задачі (1) з п. 3.7 при $\tau \to 0$. Розглянемо далі питання збіжності багатокрокових методів.

Означення 1. Рівняння (2) стійке щодо початкових умов, якщо існує незалежна від n стала M така, що за будь-яких початкових даних y_0, y_1, ..., y_{m-1} виконується оцінка

$$|y_n| \leqslant M \max_{j=\overline{0,m-1}} |y_j|, \quad n = m, m+1, \ldots. \tag{3'}$$

Отже, стійкість щодо початкових даних означає рівномірну по n обмеженість розв'язку рівняння (2).

Рівняння (2) можна записати у вигляді еквівалентної системи рівнянь

$$y_{n-m+1} = y_{n-m+1},\ \ldots,\ y_{n-1} = y_{n-1},$$

$$y_n = -\frac{a_m}{a_0} y_{n-m} - \frac{a_{m-1}}{a_0} y_{n-m+1} - \cdots - \frac{a_1}{a_0} y_{n-1}$$

або у векторній формі

$$Y_n = SY_{n-1},\quad n = m,\quad m+1,\ \ldots, \tag{4}$$

де

$$Y_n = (y_{n-m+1},\ y_{n-m+2},\ \ldots,\ y_n),$$

$$S = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \ldots & 1 \\ -\frac{a_m}{a_0} & -\frac{a_{m-1}}{a_0} & -\frac{a_{m-2}}{a_0} & \ldots & -\frac{a_1}{a_0} \end{bmatrix}. \tag{5}$$

Початковий вектор $Y_{m-1} = (y_0,\ y_1,\ \ldots,\ y_{m-1})$ є заданим.

Множина власних чисел матриці S збігається з множиною коренів характеристичного рівняння (3).

Лема 1. *Якщо виконується умова коренів, то існує векторна норма* $\|\cdot\|_*$ *така, що для узгодженої норми матриці* S *буде справедлива нерівність* $\|S\|_* \leqslant 1$.

Доведення. Перш за все нагадаємо, що для будь-якої квадратної матриці S порядку m і, зокрема, для матриці (5), існує невироджена матриця P така, що матриця $\tilde{S} = P^{-1}SP$ має жорданову канонічну форму

$$\tilde{S} = \begin{bmatrix} \tilde{S}_1 & 0 & \ldots & 0 \\ 0 & \tilde{S}_2 & \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \ldots & \tilde{S}_l \end{bmatrix},$$

де $\tilde{S}_k$ — або власне значення матриці S, або жорданова клітина, тобто матриця вигляду

$$\tilde{S}_k = \begin{bmatrix} s_k & 1 & 0 & 0 & \ldots & 0 \\ 0 & s_k & 1 & 0 & \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 0 & 0 & \ldots & s_k \end{bmatrix},$$

а s_k — власне значення матриці S.

Матрицю S можна звести і до модифікованої жорданової форми. Для цього до матриці $\tilde{S}$ застосовується перетворення подібності $D^{-1}\tilde{S}D$ з діагональною матрицею $D = \operatorname{diag}[1,\ \varepsilon,\ \ldots,\ \varepsilon^{m-1}]$, де ε — будь-яке додатне число. Неважко переконатися, що матриця

$$\hat{S} = D^{-1}\tilde{S}D$$

має таку саму блочно-діагональну структуру, що й матриця $\tilde{S}$, але з жордановими клітинами вигляду

$$\hat{S}_k = \begin{bmatrix} s_k & \varepsilon & 0 & 0 & \ldots & 0 \\ 0 & s_k & \varepsilon & 0 & \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \ldots & \varepsilon \\ 0 & 0 & 0 & 0 & \ldots & s_k \end{bmatrix}.$$

Матриці S та $\hat{S}$ пов'язані рівністю

$$\hat{S} = Q^{-1}SQ,\quad Q = PD. \tag{6}$$

Позначаючи Q через $\tilde{Q}^{-1}$, рівність (6) можна записати також у вигляді

$$\hat{S} = \tilde{Q}S\tilde{Q}^{-1}. \tag{6'}$$

Оцінимо норму матриці $\hat{S}$ для матриці S вигляду (5):

$$\|\hat{S}\|_\infty = \max_{i=\overline{1,m}} \sum_{j=1}^{m} |\hat{s}_{ij}| = \max_{i=\overline{1,m}} (|\hat{s}_{ii}| + |\hat{s}_{i,i+1}|),$$

де через $\hat{s}_{ij}$ позначено елементи матриці $\hat{S}$, причому в кожному рядку лише елементи $\hat{s}_{ii}$ та $\hat{s}_{i,i+1}$ відмінні від нуля, $\hat{s}_{ii}$ збігається з одним із власних чисел матриці S (тобто коренем характеристичного рівняння (3)),

$$\hat{s}_{i,i+1} = \begin{cases} 0 \text{ для простого кореня } \hat{s}_{ii}, \\ \varepsilon \text{ для кратного кореня } \hat{s}_{ii}. \end{cases}$$

Якщо $\hat{s}_{i,i+1} = 0$ для деякого i, то за умовою леми $|\hat{s}_{ii}| \leqslant 1$. Якщо $\hat{s}_{i,i+1} = \varepsilon$, то $\hat{s}_{ii}$ — кратний корінь і за умовою $|\hat{s}_{ii}| < 1$. Тоді при досить малому ε

$$|\hat{s}_{ii}| + |\hat{s}_{i,i+1}| < 1.$$

Таким чином, вибираючи ε досить малим, маємо

$$\|\hat{S}\|_\infty \leqslant 1.$$

Введемо норму вектора за формулою

$$\|y\|_* = \|\tilde{Q}y\|_\infty.$$

Тоді

$$\|S\|_* = \sup_{y\neq 0} \frac{\|Sy\|_*}{\|y\|_*} = \sup_{y\neq 0} \frac{\|\tilde{Q}Sy\|_\infty}{\|\tilde{Q}y\|_\infty} = \sup_{x\neq 0} \frac{\|\tilde{Q}S\tilde{Q}^{-1}x\|_\infty}{\|x\|_\infty} = \|\hat{S}\|_\infty \leqslant 1,$$

що і треба було довести.

Теорема 1. *Щоб рівняння (2) було стійким щодо початкових даних, необхідно і достатньо, щоб виконувалась умова коренів.*

Доведення. *Необхідність.* Нехай рівняння (2) є стійким щодо початкових умов. Доведемо, що виконується умова коренів. Припустимо від супротивного, що рівняння (3) має простий корінь q, для якого $|q| > 1$, або кратний корінь q_1 такий, що $|q_1| \geqslant 1$. Якщо $|q| > 1$, то, вибираючи початкові дані у вигляді $y_j = q^j$, $j = \overline{0, m-1}$, дістанемо розв'язок $y_n = q^n$, $n \geqslant m$, який необмежено зростає при $n \to \infty$. Це означає, що не може виконуватись нерівність (3′). А це суперечить припущенню про стійкість рівняння (2). Якщо q_l є r_l-кратним коренем рівняння $\rho(q) = 0$, $i = \overline{1, l}$, $\sum_{i=1}^{l} r_i = m$, то загальний розв'язок рівняння (2) має вигляд

$$y_n = c_0^{(1)}q_1^n + c_1^{(1)}nq_1^n + \cdots + c_{r_1-1}^{(1)}n^{r_1-1}q_1^n + \cdots + c_0^{(l)}q_l^n + \\ + c_1^{(l)}nq_l^n + \cdots + c_{r_l-1}n^{r_l-1}q_l^n,$$

де $c_l^{(l)}$ — довільні сталі. Нехай, наприклад, $r_1 > 1$, $|q_1| = 1$, тобто корінь q_1 має кратність, більшу від одиниці.

Початкові дані $y_0, y_1, \ldots, y_{m-1}$ можна вибрати так, що

$$y_n = nq_1^n$$

(для цього досить взяти $y_j = jq_1^j$, $j = \overline{0, m-1}$) і $|y_n| \to \infty$ при $n \to \infty$, що також суперечить припущенню про стійкість рівняння (2). Ці суперечності доводять необхідність умови коренів.

Достатність. Нехай виконується умова коренів. Доведемо, що рівняння (2) стійке за початковими умовами.

Із рівняння (4), враховуючи лему 1, дістаємо

$$\|Y_n\|_* \leqslant \|S\|_* \|Y_{n-1}\|_* \leqslant \|Y_{n-1}\|_*,$$

тобто для всіх $n = m, m+1, \ldots,$

$$\|Y_n\|_* \leqslant \|Y_{m-1}\|_*. \qquad (7)$$

За означенням норми $\|\cdot\|_*$ маємо

$$\|Y\|_* = \|\tilde{Q}Y\|_\infty \leqslant \|\tilde{Q}\|_\infty \|Y\|_\infty.$$

Крім того,

$$\|Y\|_\infty = \|\tilde{Q}^{-1}\tilde{Q}Y\|_\infty \leqslant \|\tilde{Q}^{-1}\|_\infty \|\tilde{Q}Y\|_\infty = \|\tilde{Q}^{-1}\|_\infty \|Y\|_*.$$

Таким чином,

$$(\|\tilde{Q}^{-1}\|_\infty)^{-1}\|Y\|_\infty \leqslant \|Y\|_* \leqslant \|\tilde{Q}\|_\infty \|Y\|_\infty. \qquad (8)$$

З (7) і (8) дістаємо

$$\|Y_n\|_\infty \leqslant M_1 \|Y_{m-1}\|_\infty, \qquad (9)$$

де $M_1 = \|\tilde{Q}^{-1}\|_\infty \|\tilde{Q}\|_\infty = \operatorname{cond}_\infty \tilde{Q}$ — стала, яка залежить від m, але не від n. З (9) випливає нерівність

$$|y_n| \leqslant M_1 \max_{j=\overline{0,m-1}} |y_j|,$$

яка і означає стійкість рівняння (2) щодо початкових даних. Теорему 1 доведено.

Розглянемо задачу Коші для неоднорідного різницевого рівняння

$$a_0y_n + a_1y_{n-1} + \cdots + a_my_{n-m} = \tau g_{n-m}, \\ n = m, m+1, \ldots, \qquad (10)$$

де $y_0, \ldots, y_{m-1}$ — задані величини, g_k, $k = 0, 1, \ldots,$ — задана функція дискретного аргументу. Зазначимо, що при $a_0 \neq 0$ розв'язок задачі Коші (10) існує, єдиний і може бути знайдений рекурентно за формулою

$$y_n = -\frac{a_m}{a_0}y_{n-m} - \frac{a_{m-1}}{a_0}y_{n-m+1} - \cdots - \frac{a_1}{a_0}y_{n-1} + \\ + \frac{\tau g_{n-m}}{a_0}, \quad n = m, m+1, \ldots,$$

виходячи із заданих $y_0, \ldots, y_{m-1}$ та g_k.

Означення 2. Рівняння (10) називається *стійким,* якщо для будь-яких початкових даних $y_0, \ldots, y_{m-1}$ і для будь-якої правої частини g_k має місце оцінка

$$\|y\|_{(1)} \leqslant M_1\|y^{(0)}\|_{(2)} + M_2\|g\|_{(3)}, \qquad (11)$$

де сталі M_1, M_2 не залежать від y, $y^{(0)}$, g, $y = (y_m, y_{m+1}, \ldots)$, $y^{(0)} = (y_0, \ldots, y_{m-1})$, $g = (g_0, g_1, \ldots)$, $\|\cdot\|_{(i)}$, $i = \overline{0, 3}$ — деякі функціонали з властивостями норми.

Зазначимо, що нерівність (11) виражає стійкість щодо початкових умов і правої частини.

Теорема 2. *Якщо однорідне рівняння* (2) *є стійким щодо початкових даних, то неоднорідне рівняння* (10) *є стійким, а саме має місце оцінка*

$$|y_n| \leqslant M_1 \max_{j=0,\overline{m-1}} |y_j| + M_2 \sum_{k=0}^{n-m} \tau |g_k|. \qquad (12)$$

Доведення. За теоремою 1 стійкість щодо початкових даних означає виконання умови коренів.

Подамо рівняння (10) у векторному вигляді

$$Y_n = SY_{n-1} + \tau G_{n-1}, \quad n = m, m+1, \ldots, \qquad (13)$$

де

$$Y_n = (y_{n-m+1}, y_{n-m+2}, \ldots, y_n), \quad G_{n-1} = (0, 0, \ldots, 0, g_{n-m}/a_0),$$

матриця S має вигляд (5). За лемою $1 \| S \|_* \leqslant 1$, і тому з (13) дістаємо

$$\| Y_k \|_* \leqslant \| Y_{k-1} \|_* + \tau \| G_{k-1} \|_*, \quad k = m, m+1, \ldots,$$

Підсумовуючи ці нерівності по k від m до n, маємо

$$\| Y_n \|_* \leqslant \| Y_{m-1} \|_* + \sum_{k=m-1}^{n-1} \tau \| G_k \|_*.$$

Використовуючи еквівалентність норм $\| \cdot \|_*$ та $\| \cdot \|_\infty$ (див. (8)), маємо

$$(\| \tilde{Q}^{-1} \|_\infty)^{-1} \| Y_n \|_\infty \leqslant \| \tilde{Q} \|_\infty \left(\| Y_{m-1} \|_\infty + \sum_{k=m-1}^{n-1} \tau \| G_k \|_\infty \right). \qquad (13')$$

Неважко помітити, що $\| G_k \|_\infty = |g_{k+1-n}| / |a_0|$, а тому

$$\sum_{k=m-1}^{n-1} \tau \| G_k \|_\infty = \frac{1}{|a_0|} \sum_{k=0}^{n-m} \tau |g_k|.$$

Звідси і з (13′) дістаємо нерівність (12), де $M_1 = \| \tilde{Q}^{-1} \|_\infty \times \| \tilde{Q} \|_\infty$, $M_2 = M_1 |a_0|^{-1}$. Теорему доведено.

3.8.2. Оцінка похибки багатокрокового методу розв'язування нелінійної задачі Коші. Розглянемо похибку m-крокового методу

$$\frac{a_0 y_n + a_1 y_{n-1} + \cdots + a_m y_{n-m}}{\tau} = b_0 f(t_n, y_n) + b_1 f(t_{n-1}, y_{n-1}) + \cdots + b_m f(t_{n-m}, y_{n-m}), \quad n = m, m+1, \ldots, \qquad (14)$$

де значення $y_0, y_1, \ldots, y_{m-1}$ вважаються заданими, при розв'язуванні з його допомогою задачі Коші

$$\frac{du}{dt} = f(t, u), \; t > 0; \quad u(0) = u_0. \qquad (15)$$

Підставляючи в ліву частину (14) замість y_j вирази $u(t_j) + z_j$, $j = n, n-1, \ldots, n-m$, де $z_j = y_j - u(t_j)$ — похибка, $t_j = j\tau$, маємо

$$\frac{a_0 z_n + a_1 z_{n-1} + \cdots + a_m z_{n-m}}{\tau} = -\frac{a_0 u_n + a_1 u_{n-1} + \cdots + a_m u_{n-m}}{\tau} + b_0 f(t_n, y_n) + b_1 f(t_{n-1}, y_{n-1}) + \cdots + b_m f(t_{n-m}, y_{n-m}).$$

Додавши і віднявши в правій частині цієї рівності вираз

$$b_0 f(t_n, u_n) + b_1 f(t_{n-1}, u_{n-1}) + \cdots + b_m f(t_{n-m}, u_{n-m}),$$

дістанемо наступне рівняння для похибки:

$$\frac{a_0 z_n + a_1 z_{n-1} + \cdots + a_m z_{n-m}}{\tau} = \varphi_{n-m} + \psi_{n-m}, \quad n = m, m+1, \ldots, \qquad (16)$$

де через ψ_{n-m} позначена похибка апроксимації

$$\psi_{n-m} = -\frac{a_0 u_n + a_1 u_{n-1} + \cdots + a_m u_{n-m}}{\tau} + b_0 f(t_n, u_n) + b_1 f(t_{n-1}, u_{n-1}) + \cdots + b_m f(t_{n-m}, u_{n-m}), \qquad (17)$$

а через φ_{n-m} — вираз

$$\varphi_{n-m} = b_0 [f(t_n, y_n) - f(t_n, u_n)] + b_1 [f(t_{n-1}, y_{n-1}) - f(t_{n-1}, u_{n-1})] + \cdots + b_m [f(t_{n-m}, y_{n-m}) - f(t_{n-m}, u_{n-m})]. \qquad (18)$$

Похибка апроксимації ψ_{n-m} оцінена в п. 7.1, де було знайдено умови, за яких $\psi_{n-m} = O(\tau^p)$, $p > 0$, тобто $\psi_{n-m} \to 0$ при $\tau \to 0$. Далі вважаємо виконаною умову Ліпшиця для функції $f(t, u)$ по другому аргументу, тобто

$$|f(t, u_1) - f(t, u_2)| \leqslant L |u_1 - u_2|$$

для всіх t, u_1, u_2 з області визначення функції $f(t, u)$.

Тоді з (18) маємо

$$|\varphi_{n-m}| \leqslant bL (|z_n| + |z_{n-1}| + \cdots + |z_{n-m+1}| + |z_{n-m}|), \qquad (19)$$

де $b = \max (|b_0|, |b_1|, \ldots, |b_m|)$.

Неважко помітити, що коли функція $f(t, u)$ лінійна по u, тобто $f(t, u) = u + f_0(t)$, то рівняння для похибки (16) можна записати у вигляді

$$\frac{c_0 z_n + c_1 z_{n-1} + \cdots + c_m z_{n-m}}{\tau} = \psi_{n-m}, \quad n = m, m+1, \ldots, \qquad (20)$$

де $c_0 = a_0 - b_0\tau$, $c_1 = a_1 - b_1\tau$, $\ldots$, $c_m = a_m - b_m\tau$. Користуючись далі теоремою 2, можна одразу дістати оцінку похибки z через похибку апроксимації ψ і збіжність $z_n \to 0$ при $\tau \to 0$.

Дістанемо оцінку для розв'язку задачі (16) у загальному випадку нелінійної по u функції $f(t, u)$.

Теорема 3. *Нехай задача* (15) *розглядається на скінченному проміжку* $[0, T]$, *причому* $|f'_u(t, u)| \leqslant L$ *для всіх* $t \in [0, T]$, $u \in (-\infty, \infty)$. *Тоді якщо виконується умова коренів для рівняння* (2), *а рівняння* (16) *розглядається на сітці* $\bar{\omega}_\tau = \{t_i = i\tau : i = \overline{0, M};\ \tau = T/M\}$, *яка покриває відрізок* $[0, T]$, *причому*

$$M \geqslant \frac{T}{\tau_0},\quad \tau_0 = \frac{|a_0|}{(1+\varepsilon)|b_0|L}$$

$$(M \geqslant 0,\ \tau_0 = \infty,\ \text{при}\ b_0 = 0), \tag{21}$$

де $\varepsilon > 0$ — *довільне фіксоване число, то для розв'язку рівняння* (16) *має місце оцінка*

$$|z_n| \leqslant M_1 e^{c_1 T} \max_{j=\overline{0,m-1}} |z_j| + M_2 \sum_{k=0}^{n-m} \tau |\psi_k|,\quad n = m, m+1, \ldots, M, \tag{22}$$

де

$$M_1 = \|\tilde{Q}\|_\infty \|\tilde{Q}^{-1}\|_\infty,\quad M_2 = (1+\varepsilon) M_1/(|a_0|\varepsilon),$$

$$c_1 = M_1 L v,\ v = (1+\varepsilon)\sum_{k=1}^{m} (|a_0||b_k| + |a_k||b_0|)/(a_0^2\varepsilon).$$

Доведення. За формулою Лагранжа про скінченні прирости маємо

$$f(t_{n-k}, y_{n-k}) - f(t_{n-k}, u_{n-k}) = l_{n-k} z_{n-k},$$

де

$$l_{n-k} = f'_u(t_{n-k}, \tilde{u}_{n-k}),\ \tilde{u}_{n-k} = y_{n-k} - \theta z_{n-k},\quad 0 \leqslant \theta \leqslant 1,$$

$h = \overline{0, m}$. Тому функцію φ_{n-m} з (18) можна подати у вигляді

$$\varphi_{n-m} = b_0 l_n z_n + \sum_{k=1}^{m} b_k l_{n-k} z_{n-k}.$$

Підставляючи цей вираз у рівняння (16), дістаємо

$$(a_0 - b_0 l_n \tau) z_n = \sum_{k=1}^{m} (-a_k + \tau b_k l_{n-k}) z_{n-k} + \tau\psi_{n-m}. \tag{23}$$

З умови (21) маємо $\tau \leqslant \tau_0$, звідки випливає нерівність

$$|a_0 - \tau b_0 l_n| \geqslant \frac{|a_0|\varepsilon}{1+\varepsilon} > 0, \tag{23'}$$

а тому рівняння (23) можна розв'язати відносно z_n:

$$z_n = \sum_{k=1}^{m} \frac{-a_k + \tau b_k l_{n-k}}{a_0 - \tau b_0 l_n} z_{n-k} + \tau \frac{\psi_{n-m}}{a_0 - \tau b_0 l_n}. \tag{24}$$

Додамо і віднімемо в правій частині (24) вираз

$$\sum_{k=1}^{m} \frac{a_k}{a_0} z_{n-k}.$$

Тоді (24) можна записати у вигляді

$$z_n = -\sum_{k=1}^{m} \frac{a_k}{a_0} z_{n-k} + \tau \sum_{k=1}^{m} v_{nk} z_{n-k} + \frac{\tau\psi_{n-m}}{a_0 - \tau b_0 l_n}, \tag{25}$$

де

$$v_{nk} = \frac{a_0 b_k l_{n-k} - a_k b_0 l_n}{a_0 (a_0 - \tau b_0 l_n)}. \tag{26}$$

Запишемо (25) у векторній формі

$$Z_n = S Z_{n-1} + \tau V_{n-1} Z_{n-1} + \tau \Psi_{n-1}, \tag{27}$$

де

$$Z_n = (z_{n-m+1}, z_{n-m+2}, \ldots, z_n)^T,$$

$$\Psi_{n-1} = (0, 0, \ldots, 0, \psi_{n-m}/(a_0 - \tau b_0 l_n))^T,$$

$$V_{n-1} = \begin{pmatrix} 0 & \ldots & 0 \\ \cdot & \ldots & \cdot \\ 0 & \ldots & 0 \\ v_{nm} & \ldots & v_{n1} \end{pmatrix},\quad S = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ 0 & 0 & 0 & \ldots & 1 \\ \frac{-a_m}{a_0} & \frac{-a_{m-1}}{a_0} & \frac{-a_{m-2}}{a_0} & \ldots & \frac{-a_1}{a_0} \end{bmatrix}.$$

В силу леми 1 маємо $\|S\|_* \leqslant 1$. Тому з (27) дістаємо

$$\|Z_n\|_* \leqslant \|Z_{n-1}\|_* + \tau \|V_{n-1}\|_* \|Z_{n-1}\|_* + \tau \|\Psi_{n-1}\|_*. \tag{28}$$

Доведемо, що $\|V_{n-1}\|_*$ обмежена числом, яке не залежить від n і τ. Дійсно, згідно з (23'), (26)

$$|v_{nk}| \leqslant \frac{1+\varepsilon}{\varepsilon} \frac{(|a_0||b_k| + |a_k||b_0|)\cdot L}{a_0^2},\quad k = \overline{1, m}.$$

Звідси

$$\|V_{n-1}\|_\infty = \sum_{k=1}^{m} |v_{nk}| \leqslant vL,$$

де

$$v = \frac{1+\varepsilon}{\varepsilon a_0^2} \sum_{k=1}^{m} (|a_0||b_k| + |a_k||b_0|).$$

Для будь-якого вектора x справедливі нерівності

$$\|V_{n-1}x\|_* = \|\tilde{Q}V_{n-1}x\|_\infty = \|(\tilde{Q}V_{n-1}\tilde{Q}^{-1})(\tilde{Q}x)\|_\infty \leqslant$$

$$\leqslant \|\tilde{Q}V_{n-1}\tilde{Q}^{-1}\|_\infty \|\tilde{Q}x\|_\infty \leqslant M_1 \|V_{n-1}\|_\infty \|x\|_*,$$

де $M_1 = \| \tilde{Q} \|_\infty \| \tilde{Q}^{-1} \|_\infty$. Отже,

$$\| V_{n-1} \|_* = \sup_{x \neq 0} \frac{\| V_{n-1} x \|_*}{\| x \|_*} \leqslant M_1 \| V_{n-1} \|_\infty = M_1 v L \equiv c_1.$$

Підставляючи цю оцінку в (28), дістаємо

$$\| Z_n \|_* \leqslant (1 + \tau c_1) \| Z_{n-1} \|_* + \tau \| \Psi_{n-1} \|_* \leqslant \cdots \leqslant (1 + \tau c_1)^{n-m} \| Z_{m-1} \|_* +$$

$$+ \tau \sum_{k=m}^{n} \| \Psi_{k-1} \|_* \leqslant e^{c_1 t_{n-m}} \| Z_{m-1} \|_* + \tau \sum_{k=m}^{n} \| \Psi_{k-1} \|_*, \qquad (29)$$

де $t_{n-m} = \tau (n - m) \leqslant T$. Враховуючи далі (8) та (23), маємо

$$\| Z_n \|_* \geqslant (\| \tilde{Q}^{-1} \|_\infty)^{-1} \| Z_n \|_\infty =$$

$$= (\| \tilde{Q}^{-1} \|_\infty)^{-1} \max (| z_{n-m+1} |, | z_{n-m+2} |, \ldots, | z_n |) \geqslant$$

$$\geqslant (\| \tilde{Q}^{-1} \|_\infty)^{-1} | z_n |, \quad \| Z_{m-1} \|_* \leqslant \| \tilde{Q} \|_\infty \| Z_{m-1} \|_\infty =$$

$$= \| \tilde{Q} \|_\infty \max_{j=\overline{0,m-1}} | z_j |, \quad \| \Psi_{k-1} \|_* \leqslant \| \tilde{Q} \|_\infty \| \Psi_{k-1} \|_\infty =$$

$$= \frac{\| \tilde{Q} \|_\infty \cdot | \psi_{k-m} |}{| a_0 - \tau b_0 l_n |} \leqslant \frac{(1 + \varepsilon) \| \tilde{Q} \|_\infty}{\varepsilon | a_0 |} | \psi_{k-m} |.$$

Підставляючи ці нерівності в (29), здобудемо оцінку (22), що і треба було довести.

Наслідок. За умов теореми 3 має місце нерівність

$$\| z \|_+ \leqslant M_1 e^{c_1 T} \| z \|_- + M_2 \| \psi \|_{L_1},$$

де

$$\| z \|_+ = \max_{j=\overline{m,M}} | z_j |, \quad \| z \|_- = \max_{j=\overline{0,m-1}} | z_j |,$$

$$\| \psi \|_{L_1} = \sum_{k=0}^{M-m} \tau | \psi_k |.$$

Теорема 4. *Нехай виконуються умови теореми 3, m-кроковий метод* (14) *має p-й порядок апроксимації і* $\| z \|_- = O(\tau^q)$. *Тоді при* $\tau \to 0$ *розв'язок задачі* (14) *збігається до точного розв'язку задачі* (15), *причому має місце оцінка швидкості збіжності*

$$\| z \|_+ \leqslant c \tau^r,$$

де $r = \min (p, q)$, c — *незалежна від* τ *стала.*

Теорема 4 випливає з попереднього наслідку та означення багатокрокового методу p-го порядку апроксимації.

З результатів цього пункту випливає, що багатокрокові методи, для яких не виконується умова коренів, повинні бути «забраковані».

Отже, схема третього порядку апроксимації із вправи 2, п. 7 не придатна для практичного застосування, бо квадратний тричлен

$$p(q) = q^2 + 4q - 5$$

має корені $z_1 = 1$, $|z_2| = 5 > 1$. Доведено, що при $p > m + 2$ (p — порядок апроксимації m-крокового методу) у випадку явної схеми і при $p > m$ у випадку неявної схеми серед коренів характеристичного рівняння є корінь, який за модулем більший за одиницю. Точніше, для неявного методу має місце таке твердження.

Теорема 5 (Дальквіста). *Максимальним порядком апроксимації неявного m-крокового методу, для якого виконується умова коренів, є*

$$p = \begin{cases} m + 1, & m \text{ — непарне,} \\ m + 2, & m \text{ — парне.} \end{cases}$$

Ця теорема справедлива не лише для лінійних багатокрокових методів, але й для методів загального вигляду

$$\sum_{k=0}^{m} \frac{a_k y_{n-k}}{\tau} = F(t_m; y_0, \ldots, y_{n-m}; f)$$

з функцією F, яка задовольняє умову Ліпшиця по y_i.

Порядок $m + 1$ має m-кроковий метод предиктор-коректор, побудований за формулами Адамса — Башфорта — Мултона. Прикладом двокрокового методу четвертого порядку ($m = 2$) є метод Мілна — Сімпсона

$$y_{i+1} = y_{i-1} + \frac{\tau}{3} (f(t_{i+1}, y_{i+1}) + 4 f(t_i, y_i) + f(t_{i-1}, y_{i-1})),$$

який мало придатний для практичного застосування (див. нижче). Можна довести, що формули (7.33) при $k \geqslant 6$ не задовольняють умову коренів і тому не використовуються на практиці.

Проте, якщо виконуються умова коренів та інші умови збіжності методу, то це ще не означає, що він придатний для практичного застосування. Так, неважко помітити, що згаданий вище метод Мілна — Сімпсона задовольняє умову коренів. Якщо застосувати його до модельного рівняння

$$u' = \lambda u, \quad u(0) = 1, \quad \lambda < 0,$$

з точним розв'язком $u(t) = \exp(\lambda t)$, то дістаємо лінійне різницеве рівняння

$$(1 - q) y_{i+1} - 4q y_i - (1 + q) y_{i-1} = 0, \quad y_0 = 1, \quad q = \tau\lambda/3. \qquad (30)$$

Характеристичне рівняння

$$(1 - q) \xi^2 - 4q\xi - (1 + q) = 0$$

має корені

$$\xi_1 = \frac{2q + \sqrt{1+3q^2}}{1-q} = \frac{-2|q| + \sqrt{1+3q^2}}{1+|q|},$$

$$\xi_2 = \frac{2q - \sqrt{1+3q^2}}{1-q} = \frac{-2|q| - \sqrt{1+3q^2}}{1+|q|},$$

тому задача (30) має однопараметричну сім'ю розв'язків

$$y_j = c\xi_1^j + (1-c)\,\xi_2^j. \tag{31}$$

Вибравши y_1 так, що $y_1 - \exp(\lambda\tau) = O(\tau^4)$, матимемо

$$c\xi_1 + (1-c)\,\xi_2 = y_1,$$

звідки

$$c = (y_1 - \xi_2)/(\xi_1 - \xi_2) = \frac{(y_1 - \xi_2)(1+|q|)}{2\sqrt{1+3q^2}} =$$

$$= \frac{(y_1 - e^{3q})(1+|q|) + (1+|q|)\,e^{3q} + \sqrt{1+3q^2} + 2|q|}{2\sqrt{1+3q^2}}.$$

Розвинувши e^{3q} за формулою Тейлора, дістаємо

$$e^{3q} = e^{-3|q|} = 1 - 3|q| + \frac{9}{2}|q|^2 - \frac{9}{2}|q|^3 + \frac{27}{8}|q|^4 + O(|q|^5),$$

$$(1+|q|)\,e^{3q} = 1 - 2|q| + \frac{3}{2}|q|^2 - \frac{9}{8}|q|^4 + O(|q|^5),$$

$$\sqrt{1+3q^2} = 1 + \frac{3}{2}|q|^2 - \frac{9}{8}|q|^4 + O(|q|^6),$$

$$(1+|q|)\,e^{3q} + 2|q| = \sqrt{1+3q^2} + O(|q|^4), \quad \frac{1}{1+|q|} = 1 - |q| + |q|^2 + \cdots.$$

Звідси

$$c = 1 + O(|q|^4) = 1 + O(\tau^4),$$

$$e^{3q} = \xi_1 + O(\tau^5),$$

$$\xi_1 = 1 + 3q + O(\tau^2) = 1 - \tau|\lambda| + O(\tau^2),$$

$$\xi_2 = \frac{-\sqrt{1+3q^2} - 2|q|}{1+|q|} = \frac{-1 - \frac{3}{2}|q|^2 + O(\tau^4) - 2|q|}{1+|q|} =$$

$$= \frac{-1 - |q| - |q|\left(1 + \frac{3}{2}|q| + O(\tau^3)\right)}{1+|q|} = \tag{32}$$

$$= -1 - |q|\left(1 + \frac{3}{2}|q| + O(\tau^3)\right)\left(1 - |q| + |q|^2 + O(|q|^3)\right) =$$

$$= -1 - \frac{\tau|\lambda|}{3} + O(\tau^2).$$

Якщо y_1 вибрано так, що $c \neq 1$ (а при розрахунках на ЕОМ внаслідок похибок заокруглення можна завжди вважати $c \neq 1$), то наближений розв'язок містить осцилюючий експоненціально зростаючий зі зростанням $t = j\tau$ член $(1-c)\,\xi_2^j$. Порівняємо точний розв'язок $u(t)$ та наближений $y_n = y(t_n; \tau) = y(t; \tau)$ при фіксованому t, $t = n\tau$, $\tau = t/n$, $n = t/\tau$, $\tau \to 0$, $n \to \infty$. Маємо

$$\xi_1^n = \xi_1^{\frac{t}{\tau}} = e^{\frac{t}{\tau}\ln\xi_1} = e^{\frac{t}{\tau}\ln(e^{3q} - O(\tau^5))}.$$

Оскільки $3q = \tau\lambda$, $e^{3q} = O(1)$ при $\tau \to 0$, то

$$\xi_1^n = e^{\frac{t}{\tau}\ln[e^{3q}(1-O(\tau^5))]} = e^{\frac{t}{\tau}[3q + \ln(1-O(\tau^5))]} =$$

$$= e^{\lambda t} e^{\frac{t}{\tau}\ln(1-O(\tau^5))} = e^{\lambda t}[1 + O(\tau^4)],$$

$$\xi_2^n = \xi_2^{\frac{t}{\tau}} = (-1)^{\frac{t}{\tau}} e^{\frac{t}{\tau}\ln|\xi_2|} = (-1)^{\frac{t}{\tau}} e^{\frac{t}{\tau}\ln\left(1 + \frac{\tau|\lambda|}{3} + O(\tau^2)\right)} =$$

$$= (-1)^{\frac{t}{\tau}} e^{\frac{t}{\tau}\left(\frac{|\tau\lambda|}{3} + O(\tau^2)\right)} = (-1)^{\frac{t}{\tau}} e^{\frac{t|\lambda|}{3} - O(\tau)}, \tag{33}$$

$$1 - c = O(\tau^4),$$

і, таким чином,

$$y(t) - u(t) = y_n - u(t) = c\xi_1^n + (1-c)\,\xi_2^n - e^{\lambda t} = O(\tau^4)$$

при досить малих τ, тобто метод має четвертий порядок швидкості збіжності. Проте при практичних розрахунках на ЕОМ τ не може бути вибрано довільно малим і вплив осцилюючого члена $(1-c)\,\xi_2^n$ стає порівняно з експоненціально спадним точним розв'язком дуже помітним. Наприклад, при $\tau = 5 \cdot 10^{-2}$, $y_1 = e^{\lambda\tau}$, $\lambda = -4$ зростаючі за амплітудою коливання наближеного розв'язку починаються з моменту $t = 7{,}1$, хоча при $\lambda = -3$ вони не спостерігаються до $t = 10$.

Описаний ефект зростаючих паразитичних осциляцій проявляється і для інших лінійних (по f) m-крокових методів порядку $m + 2$. Він має місце і для правила середньої точки (другого порядку апроксимації)

$$y_{i+1} = y_{i-1} + 2\tau f(t_i, y_i).$$

Вправа. Довести, що для наближеного розв'язку задачі Коші

$$\frac{du}{dt} = -u, \quad u(0) = 1, \quad u(t) = e^{-t},$$

знайденого за правилом середньої точки

$$y_{i+1} + 2\tau y_i - y_{i-1} = 0,$$

для $t_0 = 0$, $y_0 = 1$, $y_1 = 1 - \tau$ і фіксованого $t = n\tau$ має місце формула

$$y_n = e^{-t} + \tau^2 \sum_{k=1}^{\infty} (e^{-t} v_k(t) + (-1)^n e^t u_k(t)) \tau^{2k-2},$$

де u_k, v_k — аналітичні функції. Таким чином, метод середньої точки внаслідок наявності осцилюючого і експоненціально зростаючого члена не забезпечує добру апроксимацію для експоненціально спадних розв'язків.

Для правила середньої точки у вигляді

$$y_{i+1} = y_{i-1} + 2\tau f(t_i, y_i), \quad i \geqslant 1,$$
$$y_0 = u_0, \quad y_1 = u_0 + \tau f(t_0, u_0)$$

і для фіксованого $t = n\tau - t_0$ можна довести таке зображення

$$y(t; \tau) = u(t) + \sum_{k=1}^{m} \tau^{2k} [v_k(t) + (-1)^{(t-t_0)/\tau} u_k(t)] + O(\tau^{2m+2}), \qquad (34)$$

де $u(t)$ — точний розв'язок задачі Коші $\frac{du}{dt} = f(t, u)$, $u(t_0) = u_0$, $f \in C^{2m+1}([t_0, t] \times R^n)$. Для функції

$$S(t; \tau) = \frac{1}{2}(y(t; \tau) + y(t - \tau; \tau) + \tau f(t, y(t, \tau)))$$

має місце нове розвинення, в якому осцилююча складова має менший вплив:

$$S(t; \tau) = u(t) + \tau^2 \left(v_1(t) + \frac{1}{2} u''(t)\right) + \sum_{k=2}^{m} \tau^{2k} (v_k^*(t) +$$
$$+ (-1)^{\frac{t-t_0}{\tau}} u_k^*(t)) + O(\tau^{2m+2}).$$

На основі цього будується досить ефективний метод GBS Грега — Булірша — Штоера (Gragg, Bulirsch, Stoer). Виходячи з x_0 і вибравши деяке невелике T, покладають $t = t_0 + T$. Для кроків $\tau_i = T/n_i$, $i = 0, 1, \ldots$; n_i — парні, наприклад, $n_i \in \{2, 4, 6, 8, 12, 16, \ldots\}$ $n_i = 2n_{i-2}$, $i \geqslant 3$, обчислюють величини

$$S_i = S(t; \tau_i), \quad i = 0, \ldots, M, \quad M \leqslant m$$

і будують інтерполяційний многочлен за точками (τ_i^2, S_i), $i = \overline{0, M}$.

Його значення в точці O беруть за наближення для $u(t)$, тобто

$$y(t; T) = \sum_{i=0}^{M} S_i \prod_{\substack{k=0 \\ k \neq i}}^{M} \frac{-\tau_k^2}{\tau_i^2 - \tau_k^2},$$

причому можна довести, що $y(t; T) - u(t) = O(T^{2M+2})$. Далі покладають $t_0 = t$ і процес повторюють знову доти, доки не дістають розв'язок на деякому великому інтервалі (порівняйте з методом Рунге — Ромберга). Перевагою цього методу є й те, що за допомогою подільних різниць можна дістати апостеріорну оцінку похибки і за її допомогою автоматично вибирати T та M. Цей метод застосовують для нежорстких диференціальних рівнянь при високих вимогах точності. Асимптотичні розвинення вигляду (34) мають місце і для загальних лінійних багатокрокових методів. Зокрема, оцінка (22) показує, що коли L, T великі, то за рахунок великих сталих множників (не залежних від τ) похибка може бути значною, а крок τ для виконання (21) мізерний.

Тому, крім збіжності, метод розв'язування задачі Коші має задовольняти ще деякі умови, які б забезпечували правильність розрахунків при реальних величинах τ і реальному часі розрахунків. Про це піде мова в наступному параграфі.

3.9. Деякі особливості побудови та використання методів чисельного інтегрування задачі Коші. Жорсткі задачі

Після побудови методу перед його програмуванням для ЕОМ доцільно визначити, як сітковий розв'язок передає основні властивості точного розв'язку деяких простих модельних задач. Якщо для таких задач на реальних сітках (тобто при розумних величинах τ) метод дає незадовільний результат, то від його застосування доцільно відмовитися. Такий підхід частіше використовується на практиці, ніж детальний загальний теоретичний аналіз, бо дає великий виграш у витратах часу.

Наприклад, для розв'язку модельної задачі (п. 1)

$$\frac{du}{dt} + \lambda u = 0, \quad \lambda \geqslant 0, \ t > 0,$$
$$u(0) = u_0 \qquad (1)$$

має місце нерівність

$$|u(t)| \leqslant |u_0| \forall t \geqslant 0. \qquad (2)$$

Тому треба, щоб наближений розв'язок y_n задачі (1), який знайдено за допомогою деякого чисельного методу (схеми), задовольняв умову

$$|y_n| \leqslant |y_0|, \quad n = 1, 2, \ldots, \qquad (2')$$

Такий метод іноді називають *асимптотично стійким*, причому якщо (2′) виконується за додаткових умов на крок сітки τ, то він називається умовно стійким і безумовно стійким в іншому разі.

Розглянемо приклади дослідження на стійкість.

Приклад 1. Дослідити на асимптотичну стійкість явну схему Ейлера

$$\frac{y_{n+1}-y_n}{\tau}=f(t_n,\ y_n), \tag{3}$$

y_0 задано.

Розв'язання. Запишемо схему Ейлера для модельного рівняння

$$\frac{y_{n+1}-y_n}{\tau}+\lambda y_n=0,\quad y_{n+1}=(1-\tau\lambda)\,y_n.$$

Звідси випливає, що умова стійкості $|y_n|\leqslant|y_0|$ виконується при $|1-\tau\lambda|\leqslant 1$, або $-1\leqslant 1-\tau\lambda\leqslant 1$, тобто при

$$\tau\lambda\leqslant 2. \tag{4}$$

Таким чином, схема Ейлера умовно стійка при $\tau\leqslant 2/\lambda$, $\lambda>0$.

Приклад 2. Дослідити на асимптотичну стійкість схему Рунге — Кутта (25) з п. 3.3, тобто

$$y_{n+1}=y_n+\tau f\left(t_n+\frac{\tau}{2},\ y_n+\frac{\tau k_1}{2}\right),\quad k_1=f(t_n,\ y_n). \tag{5}$$

Розв'язання. Підставляючи в (5) $f=-\lambda y$, дістаємо

$$y_{n+1}=qy_n,\quad q=1-\tau\lambda+\frac{1}{2}\,\tau^2\lambda^2.$$

Схема стійка, якщо $|q|=\left|1-\tau\lambda+\frac{1}{2}\,\tau^2\lambda^2\right|\leqslant 1$, що виконується при

$$\tau\lambda\leqslant 2, \tag{6}$$

тобто схема Рунге — Кутта умовно стійка за тієї самої умови, що й схема Ейлера.

Приклад 3. Дослідити на асимптотичну стійкість схему Рунге — Кутта четвертого порядку (в означенні п. 3)

$$\frac{y_{n+1}-y_n}{\tau}=\frac{1}{6}\,[k_1(y_n)+2k_2(y_n)+2k_3(y_n)+k_4(y_n)],$$

$$k_1=f(t_n,\ y_n),\ k_2=f\left(t_n+\frac{\tau}{2},\ y_n+\frac{\tau k_1}{2}\right), \tag{7}$$

$$k_3=f\left(t_n+\frac{\tau}{2},\ y_n+\frac{\tau k_2}{2}\right),\quad k_4=f(t_n+\tau,\ y_n+\tau k_3).$$

Розв'язання. Підставляючи в (7) $f=-\lambda y$, дістаємо

$$y_{n+1}=qy_n,\quad q=1-\tau\lambda+\frac{1}{2}\,\tau^2\lambda^2-\frac{1}{6}\,\tau^3\lambda^3+\frac{1}{24}\,\tau^4\lambda^4.$$

Нерівність $|q|\leqslant 1$ виконується при

$$\tau\lambda\leqslant 2{,}78, \tag{8}$$

тобто умова стійкості схеми (7) дещо слабкіша, ніж умова (6) для схеми (5).

Ці приклади показують, що явні однокрокові схеми умовно стійкі. Якщо $\lambda>0$ і велике, то крок τ в силу (4), (6), (8) треба вибирати достатньо малим, що ускладнює практичну реалізацію методу. Однак є і безумовно стійкі схеми чисельного розв'язування задачі Коші, приклади яких наведено у наступних вправах.

Вправа 1. Довести, що неявна схема Ейлера

$$\frac{y_{n+1}-y_n}{\tau}=f(t_{n+1},\ y_{n+1})$$

асимптотично безумовно стійка.

Вправа 2. Довести, що схема з ваговими коефіцієнтами

$$\frac{y_{n+1}-y_n}{\tau}=\sigma f(t_{n+1},\ y_{n+1})+(1-\sigma)\,f(t_n,\ y_n)$$

асимптотично безумовно стійка (тобто стійка для всіх τ) при $\sigma\geqslant 1/2$ і умовно стійка при $\sigma<1/2$, якщо $\tau\leqslant 1/((1/2-\sigma)\,\lambda)$.

Для багатокрокових методів, як і для однокрокових, має виконуватись умова асимптотичної стійкості (2′). Розглянемо стійкість конкретних багатокрокових методів.

Перепишемо явну двокрокову схему Адамса (див. (22) з п. 3.7) у вигляді

$$\frac{y_j-y_{j-1}}{\tau}=\frac{3}{2}\,f_{j-1}-\frac{1}{2}\,f_{j-2}.$$

Застосовуючи цю схему до модельного рівняння $u'(t)+\lambda u(t)=0$, маємо

$$\frac{y_j-y_{j-1}}{\tau}+\lambda\left(\frac{3}{2}\,y_{j-1}-\frac{1}{2}\,y_{j-2}\right)=0.$$

Підставляючи сюди $y_j=q^j$, для q дістаємо характеристичне рівняння

$$q^2-\left(1-\frac{3}{2}\,\mu\right)q-\frac{1}{2}\,\mu=0,\quad \mu=\lambda\tau. \tag{7′}$$

Дискримінант цього квадратного рівняння $D=1-\mu+\frac{9}{4}\,\mu^2>0$ $\forall\,\mu$, тому корені q_1 і q_2 дійсні і різні. Стійкість означає, що $|y_j|\leqslant\leqslant\text{const}$, а це буде тоді, коли $|q_1|\leqslant 1$, $|q_2|\leqslant 1$. Безпосередньо можна перевірити, що корені квадратного рівняння $q^2+bq+c=0$ не перевищують за модулем одиницю, якщо $|b|\leqslant 1+c$, $c\leqslant 1$. Для рівняння (7′) за цих умов маємо

$$\left|1-\frac{3}{2}\,\mu\right|\leqslant 1-\frac{\mu}{2},\quad 1-\frac{\mu}{2}\leqslant 1.$$

Неважко перевірити, що обидві нерівності виконуються при $\mu\leqslant\leqslant 1$ або $\tau\lambda\leqslant 1$, тобто схема (22) з п. 3.7 умовно стійка, причому крок τ у неї має бути у два рази меншим допустимого кроку в методі Ейлера.

Неявну двокрокову схему Адамса (див. (23) з п. 3.7) перепишемо у вигляді

$$\frac{y_j - y_{j-1}}{\tau} = \frac{1}{12}(5f_j + 8f_{j-1} - f_{j-2}).$$

Для модельної задачі вона має вигляд

$$\frac{y_j - y_{j-1}}{\tau} + \frac{\lambda}{12}(5y_j + 8y_{j-1} - y_{j-2}) = 0,$$

а її характеристичне рівняння запишемо таким чином:

$$aq^2 + bq + c = 0,\ a = 1 + \frac{5}{12}\tau\lambda,\ \ b = \frac{8}{12}\tau\lambda - 1,\ c = -\frac{1}{12}\lambda\tau.$$

Аналогічно знаходимо, що схема асимптотично стійка при $\tau\lambda \leqslant 6$. Умовна стійкість є недоліком методу, оскільки при великих $\lambda > 0$ вимушує застосувати занадто малий крок τ, що призводить до збільшення обсягу обчислювальної роботи і збільшення впливу похибок заокруглень. Безумовно стійкі методи позбавлені цього недоліку, але оскільки вони неявні, то на кожному кроці доводиться розв'язувати нелінійне рівняння.

Значні труднощі виникають при чисельному розв'язуванні так званих жорстких задач Коші. Такі задачі досить різноманітні за причинами, які породжують жорсткість, і тому дати строге математичне означення цього явища дуже не просто. Суть жорсткості полягає в тому, що треба обчислити поволі змінний розв'язок, на який накладаються швидкозмінні збурення. Виявляється, що багато чисельних методів нездатні подавляти вплив цих збурень, і тому необхідно розробляти спеціальні методи. Пояснимо явище жорсткості на такому прикладі.

Розв'язком задачі

$$\frac{du}{dt} = \lambda u(t) + \frac{dF(t)}{dt} - \lambda F(t),\quad t \geqslant 0, \tag{9}$$

$$u(0) = u_0,\quad \lambda \ll 0 \tag{10}$$

є функція $u(t) = F(t) + e^{\lambda t}[u_0 - F(0)]$, де $F(t)$ — функція від t, яка поволі змінюється (тобто має малу похідну). Оскільки $\lambda \ll 0$, то зрозуміло, що вже після досить невеликого проміжку часу швидко змінна (перехідна) частина розв'язку (або його жорстка компонента) $e^{\lambda t}[u_0 - F(0)]$ майже не впливає на точний розв'язок $u(t)$, тобто на досить великому проміжку $[0, T]$ цей розв'язок повністю визначається поволі змінною (неперехідною) компонентою $F(t)$. Застосовуючи до задачі (9) явний

$$y_{n+1} = y_n + \tau f(t_n, y_n)$$

та неявний

$$y_{n+1} = y_n + \tau f(t_{n+1}, y_{n+1})$$

методи Ейлера, відповідно дістаємо

$$y_{n+1} = (1 + \tau\lambda)[y_n - F(t_n)] + F(t_n) + \tau F'(t_n),$$

$$y_{n+1} = (1 - \tau\lambda)^{-1}[y_n - F(t_n)] + (1 - \tau\lambda)^{-1}[F(t_n) + \tau F'(t_{n+1}) -$$

$$- \tau\lambda F(t_{n+1})] = (1 - \tau\lambda)^{-1}[y_n - F(t_n)] + (1 - \tau\lambda)^{-1}[(1 - \tau\lambda)F(t_n) +$$

$$+ \tau F'(t_{n+1}) - \tau\lambda F'(\xi)],$$

$$\xi \in [t_n,\ t_{n+1}].$$

Порівнюючи ці формули з точним розв'язком

$$u(t_{n+1}) = e^{\lambda t_{n+1}}[u_0 - F(0)] + F(t_{n+1}) = e^{\lambda\tau}[u(t_n) - F(t_n)] + F(t_{n+1}),$$

помічаємо, що неявний метод Ейлера досить задовільно відбиває поведінку точного розв'язку і «зменшує» різницю $y_n - F(t_n)$, яку можна розглядати як збурення, що завжди присутнє в реальних обчисленнях внаслідок заокруглень. В явному методі Ейлера різниця $|y_n - F(t_n)|$ «зменшує» лише при $-2 < \tau\lambda < 0$. Ця умова чисельної стійкості накладає серйозне обмеження на крок τ при $\lambda \ll 0$ навіть коли величина $y_n - F(t_n)$ дуже мала. Така ситуація є типовою при застосуванні явних методів до жорстких рівнянь.

Додаткові труднощі можуть виникати при розв'язуванні задачі Коші для системи рівнянь, хоча багато з розглянутих вище методів переносяться без змін на системи рівнянь.

Розглянемо задачу Коші для лінійної системи звичайних диференціальних рівнянь

$$\frac{du_i(t)}{dt} + \sum_{j=1}^{N} a_{ij}(t)\,u_j(t) = f_j(t),\quad t > t_0,\ u_i(t_0) = u_i^0,\quad i = \overline{1, N},$$

або у векторній формі

$$\frac{du}{dt} + Au = f(t),\quad A = (a_{ij})_{i,j=\overline{1,N}}. \tag{11}$$

На практиці часто зустрічаються системи рівнянь, які називають жорсткими. Чисельне розв'язування таких систем породжує значні труднощі. Нехай λ_k — власні числа матриці A (якщо A несиметрична, то λ_k можуть бути комплексними). Систему рівнянь (11) називають *жорсткою*, якщо $\operatorname{Re}\lambda_k > 0$, $k = \overline{1, N}$, і відношення $\xi = \max\limits_k |\operatorname{Re}\lambda_k| / \min\limits_k |\operatorname{Re}\lambda_k|$ велике. Якщо матриця A симетрична, то всі її власні значення дійсні і жорсткість системи (3) означає, що матриця A додатно означена і погано обумовлена, тобто

$$\xi = \frac{\max\limits_k \lambda_k}{\min\limits_k \lambda_k} \gg 1.$$

Зазначимо, що поняття жорсткості є умовним, бо в означенні не вказано чітко, що означає «велике ξ». Вплив величини ξ залежить від типу ЕОМ, програм розв'язування задачі Коші і т. д., тобто одна й та сама задача Коші може бути жорсткою, наприклад, при розв'язуванні її на одній ЕОМ і нежорсткою на іншій. У випадку нелінійної системи (40) з п. 3.5 запишемо її в околі розв'язку u^* у вигляді

$$u' = f(t, u^*) + f'_u(t, u^*)(u - u^*) + o(\|u - u^*\|). \qquad (12)$$

Це зображення показує, що в околі розв'язку система рівнянь (40) з п. 3.3.5 добре апроксимується лінійною системою

$$\frac{du}{dt} = f'_u(t, u^*)\, u + \tilde{f}(t), \qquad (13)$$

де $f'_u(t, u) = \left\{\frac{\partial f_i(t, u)}{\partial u_j}\right\}_{i,j=\overline{1,N}}$ — матриця Якобі правої частини, $\tilde{f}(t) = f(t, u^*) - f_u(t, u^*)\, u^*$. Якщо матриця $f'_u(t, u^*)$ має широкий діапазон зміни від'ємних дійсних частин власних значень, то система (40) з п. 3.3.5 також називається жорсткою. Щоб висвітлити поняття жорсткості ще з одного боку, розглянемо систему (11) зі сталою матрицею A, що має різні власні значення λ_i, $\mathrm{Re}\lambda_i > 0$, які відповідають лінійно незалежним власним векторам μ_i, $i = \overline{1, N}$, $\mu_i = (\mu_{1i}, \ldots, \mu_{Ni})^T$. Тоді матрицю A можна подати у вигляді $A = M\Lambda M^{-1}$, де $M = (\mu_1, \ldots, \mu_N)$ — матриця, складена із власних векторів, M^{-1} — її обернена матриця, $\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_N)$ — діагональна матриця власних чисел. Систему (11) запишемо у вигляді

$$\frac{du}{dt} + M\Lambda M^{-1}u = f(t).$$

Її загальний розв'язок буде сумою загального розв'язку однорідної системи

$$\frac{du_0}{dt} + M\Lambda M^{-1}u_0 = 0$$

та частинного розв'язку неоднорідної системи. Помножимо останнє рівняння зліва на M^{-1}, справа на M, і покладемо $M^{-1}u_0M = \tilde{u}(t)$. В результаті дістаємо систему

$$\frac{d\tilde{u}}{dt} + \Lambda\tilde{u} = 0,$$

яка розпадається на N рівнянь $\tilde{u}'_i(t) + \lambda_i\tilde{u}_i(t) = 0$ з розв'язками

$$\tilde{u}_i(t) = c_i e^{-\lambda_i t},$$

де c_i — деякі сталі. Виконавши зворотне перетворення, дістанемо

$$u_0(t) = M\,\mathrm{diag}\,(c_i e^{-\lambda_i t})_{i=\overline{1,N}}\, M^{-1}.$$

Великий «розмах» у величин дійсних частин власних значень означає, що серед розв'язків системи є дуже поволі і дуже швидко змінні компоненти, а це, як і у випадку розглянутого вище одного рівняння, означає жорсткість системи. Явні методи виявляються непридатними для чисельного розв'язування жорстких систем, бо вони призводять до значних обмежень на крок через вимоги стійкості, а малий крок з урахуванням похибок заокруглення спричиняє зменшення точності і недопустимо великий час розрахунків. Покажемо це на прикладі системи

$$\frac{du}{dt} + Au = 0, \qquad (14)$$

де $u = (u_1; u_2)^T$, $A = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}$, $a_2 \gg a_1 > 0$. Неважко помітити, що ця система розпадається на два незалежних рівняння, а розв'язком її є вектор

$$u(t) = (u_1(0)\, e^{-a_1 t},\ u_2(0)\, e^{-a_2 t})^T,$$

компоненти якого спадають зі зростанням t, причому $|u_2(t)| \ll |u_1(t)|$ при досить великих t.

Розглянемо явну схему Ейлера

$$\frac{y_1^{n+1} - y_1^n}{\tau} + a_1 y_1^n = 0, \quad \frac{y_2^{n+1} - y_2^n}{\tau} + a_2 y_2^n = 0, \quad n = 0, 1, , \ldots,$$

$y_i^n = y_i(t_n)$, $i = \overline{1, 2}$. Ці два рівняння можна розв'язати окремо, але вони пов'язані вибором спільного кроку τ. Схема стійка, якщо одночасно $a_1\tau \leqslant 2$ і $a_2\tau \leqslant 2$. Оскільки $a_2 \gg a_1$, то обидві умови виконуються, якщо $\tau \leqslant \frac{2}{a_2}$, тобто припустимий крок визначається максимальним власним числом матриці A або тією компонентою $u_2(t)$ розв'язку, яка швидше змінюється (спадає).

Вправа 2. Показати, що неявна схема

$$\frac{y_1^{n+1} - y_1^n}{\tau} + a_1 y_1^{n+1} = 0, \quad \frac{y_2^{n+1} - y_2^n}{\tau} + a_2 y_2^{n+1} = 0$$

стійка при будь-яких τ і $a_1 \geqslant 0$, $a_2 \geqslant 0$.

Аналогічні ускладнення виникають і при розв'язуванні будь-якої системи звичайних диференціальних рівнянь. Це неважко показати на прикладі системи

$$\frac{du}{dt} + Au = 0, \qquad (15)$$

де $u = (u_1(t), \ldots, u_N(t))^T$, A — матриця розмірності $N \times N$, яку перетворенням подібності $M^{-1}AM = \Lambda$ можна звести до діагонального вигляду. Тоді заміна $u = Mv$ перетворює систему (15) на систему незалежних рівнянь

$$\frac{dv}{dt} + \Lambda v = 0, \tag{16}$$

матриця Λ якої має ті самі власні значення, що й матриця A. Оскільки, взагалі кажучи, матриця може мати комплексні власні значення, то чисельний метод, орієнтований на розв'язування систем рівнянь, доцільно випробувати на рівнянні

$$\frac{du}{dt} + \lambda u = 0, \tag{17}$$

де λ — довільне комплексне число.

Якщо багатокроковий метод (2) з п. 3.7 застосувати до рівняння (17), то

$$\sum_{k=0}^{m} (a_k - \mu b_k) y_{n-k} = 0, \quad n = m, m+1, \ldots, \tag{18}$$

де $\mu = \tau\lambda$ — комплексний параметр. Будемо шукати розв'язок рівняння (18) у вигляді $y_n = q^n$, де q — деяка стала. Підставляючи цей вираз у (18), для відшукання q дістаємо характеристичне рівняння

$$\sum_{k=0}^{m} (a_k - \mu b_k) q^{m-k} = 0. \tag{19}$$

Якщо $\lambda = \lambda_R + i\lambda_I$, де $i = \sqrt{-1}$, то розв'язок рівняння (17) можна подати у вигляді

$$u = e^{-\lambda t} = e^{-\lambda_R t} \cdot e^{-\lambda_I t i} = e^{-\lambda_R t}(\cos \lambda_I t - i \sin \lambda_I t).$$

Отже, при $\lambda_R = \operatorname{Re} \lambda \geqslant 0$ розв'язок рівняння (17) залишається обмеженим при $t \to \infty$ або є асимптотично стійким. В цьому разі кажуть, що рівняння (17) асимптотично стійке. Доцільно вимагати, щоб і різницевий розв'язок відбивав властивість асимптотичної стійкості точного розв'язку. Це буде тоді, коли всі корені рівняння (19) не перевищують за модулем одиницю, причому немає кратних коренів з модулем, що дорівнює одиниці. *Областю стійкості* методу (2) з п. 3.7 називається множина всіх точок комплексної площини $\mu = \tau\lambda$, для яких виконується така умова.

Розглянемо, наприклад, явний метод Ейлера

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n, y_n). \tag{20}$$

Застосувавши його до рівняння (17), дістанемо

$$y_{n+1} = (1 + \mu) y_n, \quad \mu = \tau\lambda.$$

Отже, умова стійкості $|1 + \mu| \leqslant 1$ для комплексного $\mu = \mu_R + i\mu_I$ означає, що $(\mu_R + 1)^2 + \mu_I^2 \leqslant 1$. Таким чином, областю стійкості методу (20) є круг одиничного радіуса з центром у точці $(-1, 0)$.

Вправа 3. Довести, що для неявного методу Ейлера

$$\frac{y_{n+1} - y_n}{\tau} = f(t_{n+1}, y_{n+1})$$

областю стійкості є зовнішність круга одиничного радіуса з центром у точці $(1, 0)$.

Часто вигляд області стійкості беруть за основу класифікації чисельних методів.

Чисельний метод розв'язування задачі Коші називають A-*стійким*, якщо область його стійкості містить ліву напівплощину $\operatorname{Re} \mu < 0$. Метод називають A (α)-*стійким*, якщо область його стійкості містить кут

$$|\arg(-\mu)| < \alpha, \quad \mu = \tau\lambda.$$

Зокрема, A $\left(\frac{\pi}{2}\right)$-стійкість збігається з A-стійкістю.

Неважко помітити, що явний метод Ейлера не є A-стійким, а неявний метод Ейлера є A-стійким.

Оскільки при $\operatorname{Re} \lambda < 0$ рівняння (17) є асимптотично стійким, то суть поняття A-стійкості методу полягає в тому, що він є абсолютно (безумовно) асимптотично стійким, якщо рівняння (17) є асимптотично стійким.

Вправа 4. Довести, що однокроковий метод

$$\frac{y_{n+1} - y_n}{\tau} = 0{,}5\,(f(t_{n+1}, y_{n+1}) + f(t_n, y_n)) \tag{21}$$

є A-стійким.

Оскільки A-стійкі методи не накладають обмеження на крок τ, то їх доцільно застосовувати для розв'язування жорстких систем рівнянь. Проте клас A-стійких методів вигляду (2) з п. 3.7 (тобто лінійних відносно f_j) виявився досить вузьким, про що свідчать наступні два твердження Дальквіста: а) серед методів вигляду (2) з п. 3.7 не існує явних A-стійких методів; б) серед неявних лінійних багатокрокових методів немає A-стійких методів, порядок точності яких більший ніж два. Прикладом A-стійкого методу другого порядку точності є симетрична схема (21). Можна довести, що A-стійким є метод Гаусса — Рунге — Кутта. Клас $A(\alpha)$-стійких методів є дещо ширшим. Можна довести, що для жодного α не існує явного A (α)-

стійкого лінійного багатокрокового методу вигляду (2) з п. 3.7. Побудовані неявні A (α)-стійкі методи третього і четвертого порядків точності. До них, зокрема, належать так звані чисто неявні методи вигляду (2) з п. 3.7 при $b_0 = 1$, $b_1 = b_2 = \ldots = b_m = 0$, тобто

$$\sum_{k=0}^{m} a_k y_{n-k} = \tau f(t_n, y_n). \qquad (22)$$

Наприклад, схема

$$\frac{25y_n - 48y_{n-1} + 36y_{n-2} - 16y_{n-3} + 3y_{n-4}}{12\tau} = f(t_n, y_n)$$

має четвертий порядок точності і є A (α)-стійкою, де $\alpha = \operatorname{arctg} \sqrt{6} \approx 68°$.

Метод Гіра (2) з п. 3.7 є A (α)-стійким для кутів α та кількості кроків k, які даються таблицею:

k	1	2	3	4	5	6
α	90°	90°	88°02′	73°21′	51°50′	17°50′

Області стійкості k-крокових методів Адамса — Мултона ($k \geqslant 2$), Адамса — Башфорта ($k \geqslant 1$), методів предиктор-коректор і явних методів Рунге — Кутта є досить вузькими. Дійсні інтервали стійкості $(\gamma, 0)$ для явних m-ступеневих методів Рунге — Кутта m-го порядку (зокрема, для класичного методу Рунге — Кутта четвертого порядку) даються таблицею:

m	1	2	3	4
γ	−2	−2	−2,51	−2,78

Для k-крокових методів Адамса — Мултона та Адамса — Башфорта вони даються відповідно таблицями:

k	1	2	3	4	5
γ	−∞	−6	−3	−90/49	−45/38
k	1	2	3	4	5
γ	−2	−1	−6/11	−3/10	−90/551

Через вузьку область стійкості методи Адамса — Башфорта в чистому вигляді застосовуються рідко.

Для жорстких систем диференціальних рівнянь останнім часом найчастіше застосовують метод трапецій (двокроковий метод Адамса — Мултона)

$$y_{i+1} = y_i + \frac{\tau}{2}(f_i + f_{i+1}),$$

неявний метод середньої точки

$$y_{i+1} = y_i + \tau_i f\left(t_i + \frac{\tau_i}{2}, \frac{y_i + y_{i+1}}{2}\right)$$

або методи Гіра для $k \leqslant 6$. Добрі результати для жорстких систем дають також методи Розенброка, зокрема A- та A (α)-стійкі методи вищих порядків, для яких є також вдалі програми. Ефективними для жорстких систем є також так звані дробово-раціональні методи.

Зазначимо, що теорія стійкості, викладена вище, яка орієнтується на модельну задачу $\frac{du}{dt} + \lambda u = 0$, не завжди дає правильні висновки стосовно чисельних методів для рівнянь зі змінними коефіцієнтами, а також для нелінійних та жорстких рівнянь. Як приклад можна розглянути задачу Коші

$$\frac{du}{dt} = \lambda(t)\, u, \quad u(0) = u_0$$

з функцією $\lambda(t) < 0$, $\lambda(t) \in C^1[0, \infty]$, і методи трапецій, Ейлера (неявний) та неявний метод середньої точки. Неважко переконатися, що метод Ейлера

$$y_{i+1} = \frac{1}{1 - \tau\lambda(t_{i+1})} y_i$$

є безумовно стійким, тобто $|y_{i+1}| \leqslant |y_j|\ \forall\, i$ без обмежень на τ.

Метод трапецій дає $y_{i+1} = y_i + \frac{\tau}{2}(f(t_i, y_i) + f(t_{i+1}, y_{i+1}))$

$$y_{i+1} = \frac{2 + \lambda(t_i)\tau}{2 - \lambda(t_{i+1})\tau} y_i,$$

і за умови $\sup\limits_{t \in [0,\infty)} \lambda'(t) = \beta > 0$ дістаємо нерівність стійкості $|y_{i+1}| \leqslant |y_i|$, якщо

$$\tau \leqslant 2/\sqrt{\beta}.$$

Неявний метод середньої точки

$$y_{i+1} = y_i + \tau f\left(\frac{t_i + t_{i+1}}{2}, \frac{y_{i+1} + y_i}{2}\right)$$

є еквівалентним методу трапецій для лінійних систем зі сталими коефіцієнтами, але в даному випадку дає формулу

$$y_{i+1} = \frac{2 + \tau\lambda\left(\frac{t_i + t_{i+1}}{2}\right)}{2 - \tau\lambda\left(\frac{t_i + t_{i+1}}{2}\right)} y_i$$

і на відміну від методу трапецій є безумовно стійким. Тому активно

ведуться дослідження, спрямовані на пошук відповідних означень стійкості нелінійних систем рівнянь і на конструювання відповідних стійких методів, які пристосовані до властивостей шуканого розв'язку. Розглянемо один з таких підходів для систем

$$\frac{du}{dt} = f(t,\ u),\ u(0) = u_0, \tag{23}$$

де функція $f : [0,\ \infty) \times R^n \to R^n$ задовольняє умову

$$\langle f(t,\ u) - f(t,\ v),\ u - v\rangle \leqslant m\langle u - v,\ u - v\rangle\ \forall\, t \in [0,\ \infty),\ u,\ v \in R^n. \tag{24}$$

Такі системи при $m \leqslant 0$ називаються *дисипативними*. Як бачимо, в означення дисипативності не входить матриця Якобі $J(f) = \frac{\partial f}{\partial u}$, і тому в цей клас потрапляють довільні жорсткі задачі.

Введемо таке означення.

Означення. Деякий чисельний метод для задачі (23) називається *оптимально B-збіжним* з порядком p на класі всіх дисипативних задач, якщо

$$\| u(t_i) - y_i \| \leqslant c(t_i)\,\tau^p\ \forall\, \tau < \bar{\tau},\quad \forall\, i,$$

де $c(t_i)$ залежить лише від $\max\{\| u^{(j)}(t) \| : t_0 \leqslant t \leqslant t_i,\ 1 \leqslant j \leqslant \bar{p}\}$ для деякого $\bar{p}$ та від m, $u(t)$ — розв'язок задачі (23), права частина якої задовольняє (24) з $m \leqslant 0$.

Перевагою оптимально B-збіжних методів є те, що обмеження на крок сітки, а також похибка методу не залежать від жорсткості задачі. Але, наприклад, явні методи Рунге — Кутта p-го порядку мають головний член похибки $\tau^p c(t_i)$, де $c(t_i)$ залежить від норми матриці Якобі $J(f)$. У роботі Шнайда доведено таку теорему.

Теорема. *Неязний метод Ейлера є оптимально B-збіжним з порядком 1. Метод трапецій та неявний метод середньої точки є оптимально B-збіжним з порядком 2 на класі дисипативних задач Коші.*

Можна також довести, що m-ступеневий метод Гаусса — Рунге — Кутта на класі дисипативних задач є A-стійким та оптимально B-збіжним з порядком m.

3.10. Задача Коші для рівняння другого порядку. Методи Штьормера

Розглянемо задачу Коші

$$\frac{d^2u}{dt^2} = f(t,\ u(t)),\quad t > 0,\quad u(0) = u_0,\ \frac{du}{dt}(0) = u_1. \tag{1}$$

Найбільш поширеними є методи Штьормера, які мають вигляд

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2} = \sum_{k=-1}^{m} b_k f(t_{n-k},\ y_{n-k}),\quad m \geqslant 0,$$

$$n = 1,\ 2,\ \ldots,\ y_0 = u_0,\ y_1 = \bar{u}_1 \left(\text{або } \frac{y_1 - y_0}{\tau} = \tilde{u}_1\right). \tag{2}$$

Значення $\bar{u}_1$ (або $\tilde{u}_1$) вибирається так, щоб похибка апроксимації другої з початкових умов $\nu = \frac{1}{\tau}[u(\tau) - u(0)] - \tilde{u}_1$ мала вигляд $\nu = 0(\tau^p)$, де p — порядок апроксимації. Наприклад, при $n = 2$ знаходимо

$$u(\tau) = u(0) + \tau u'(0) + \frac{1}{2}\tau^2 u''(0) + O(\tau^3),$$

$$\nu = u'(0) + \frac{1}{2}\tau u''(0) + O(\tau^2) - \tilde{u}_1 = u_1 + \frac{\tau}{2}u''(0) - \tilde{u}_1 +$$

$$+ O(\tau^2) = \frac{\tau}{2} f(0,\ u(0)) + u_1 - \tilde{u}_1 + O(\tau^2),$$

і якщо покласти

$$\tilde{u}_1 = u_1 + \frac{\tau}{2} f(0,\ u_0),\quad \bar{u}_1 = u_0 + \tau\tilde{u}_1, \tag{3}$$

то $\nu = O(\tau^2)$.

Якщо $b_{-1} = 0$, то схема (2) буде явною, бо до правої частини будуть входити лише відомі значення y_n, y_{n-1}, ..., y_{n-m}. Якщо $b_{-1} \neq 0$, то схема (2) є неявною і для визначення y_{n+1} треба розв'язувати рівняння вигляду

$$y_{n+1} - b_{-1} f(t_{n+1},\ y_{n+1}) = F(y_n,\ y_{n-1},\ \ldots,\ y_{n-m}).$$

Для відшукання коефіцієнтів різницевої схеми (2) можна застосувати метод невизначених коефіцієнтів або метод типу Адамса. Можна також скористатися таким способом. Помножимо рівняння (1) на функцію $v(t)$, яку називатимемо пробною функцією, і проінтегруємо від t_{n-1} до t_{n+1}:

$$\int_{t_{n-1}}^{t_{n+1}} u''(t)\,v(t)\,dt = \int_{t_{n-1}}^{t_{n+1}} f(t,\ u(t))\,v(t)\,dt. \tag{4}$$

Виберемо функцію $v(t)$ у вигляді (рис. 3)

$$v(t) = \begin{cases} 1 - \tau^{-1}|t - t_n|, & t \in [t_{n-1},\ t_{n+1}], \\ 0, & t \notin [t_{n-1},\ t_{n+1}]. \end{cases}$$

Тоді, враховуючи, що $v'(t) = \begin{cases} \frac{1}{\tau}, & t \in [t_{n-1}, t_n], \\ -\frac{1}{\tau}, & t \in [t_n, t_{n+1}], \end{cases}$ і інтегруючи частинами, дістаємо

$$\int_{t_{n-1}}^{t_{n+1}} u''(t)\, v(t)\, dt = \int_{t_{n-1}}^{t_n} u''v dt + \int_{t_n}^{t_{n+1}} u''v dt =$$

$$= u'v\Big|_{t_{n-1}}^{t_n} - \int_{t_{n-1}}^{t_n} u'v' dt + u'v\Big|_{t_n}^{t_{n+1}} - \int_{t_n}^{t_{n+1}} u'v' dt =$$

$$= u'v|_{t=t_n} - \frac{1}{\tau}\int_{t_{n-1}}^{t_n} u' dt - u'v|_{t=t_n} + \frac{1}{\tau}\int_{t_n}^{t_{n+1}} u' dt \equiv \tau u_{\bar{t}t,n}.$$

Підставляючи останній вираз в (4), знаходимо

$$u_{\bar{t}t,n} = \frac{1}{\tau}\int_{t_{n-1}}^{t_{n+1}} f(t,\ u(t))\, v(t)\, dt.$$

Похибка апроксимації схеми (2) на розв'язку $u(t)$ (або нев'язка схеми (2)) визначається формулою

$$\psi_n = \sum_{k=-1}^{m} b_k f(t_{n-k},\ u_{n-k}) - u_{\bar{t}t,n} \qquad (5)$$

і в силу (5) може бути записана у вигляді

$$\psi_n = \sum_{k=-1}^{m} b_k f(t_{n-k},\ u_{n-k}) - \frac{1}{\tau}\int_{t_{n-1}}^{t_{n+1}} f(t,\ u(t))\,[1 - \tau^{-1}\,|t - t_n|]\, dt. \qquad (6)$$

Перший доданок в (6) можна розглядати як квадратурну формулу для інтегралу від функції $F(t) \equiv f(t, u(t))$ з вагою $v(t) \geqslant 0$, і похибка апроксимації схеми (2) визначається похибкою цієї квадратурної формули. Застосуємо, наприклад, формулу інтерполяційного типу, замінивши $F(t)$ інтерполяційним многочленом нульового степеня, поклавши

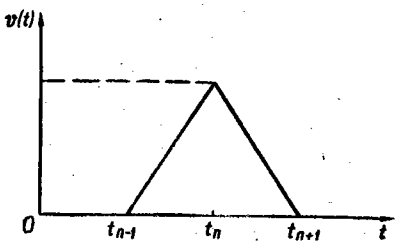


Рис. 3

$$F(t) = p_0(t;\ F) + R_0(t;\ F),$$

де

$$p_0(t;\ F) = F(t_n); \quad R_0(t;\ F) = \frac{F'(\xi)}{1!}(t - t_n),$$

$$\xi \in [y_1,\ y_2], \quad y_1 = \min(t,\ t_n), \quad y_2 = \max(t,\ t_n).$$

Тоді, враховуючи, що

$$\int_{t_{n-1}}^{t_{n+1}} v(t)\, dt = \tau,$$

матимемо

$$\int_{t_{n-1}}^{t_{n+1}} v(t)\, F(t)\, dt = \tau F(t_n) + R(F), \qquad (7)$$

де

$$R(F) = \int_{t_{n-1}}^{t_{n+1}} v(t)\, F'(\xi(t))\,(t - t_n)\, dt.$$

Неважко помітити, що при обмеженості похідних від функції $f(t, u)$ по t і $u(t)$ буде $R(F) = O(\tau^2)$. Тому, якщо покласти $b_{-1} = 0$, $m = 0$ і

$$\sum_{k=-1}^{m} b_k f(t_{n-k},\ u_{n-k}) \equiv F(t_n) \equiv f(t_n,\ u_n),$$

то з (7) знаходимо

$$\psi_n \equiv -\tau^{-1}R(F) = f(t_n,\ u_n) - \tau^{-1}\int_{t_{n-1}}^{t_{n+1}} v(t)\, F(t)\, dt = O(\tau).$$

Таким чином, схема

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2} = f(t_n,\ y_n)$$

має перший порядок апроксимації.

Щоб дослідити цю схему на стійкість, запишемо її для модельної задачі $u''(t) + \lambda u = 0$, $t > 0$; $u(0) = 0$, $u'(0) = u_1$:

$$y_{\bar{t}t,n} + \lambda y_n = 0.$$

Підставляючи сюди $y_n = q^n$, знаходимо $q^2 - 2(1 - \tau^2\lambda/2)\, q + 1 = 0$. Неважко показати, що $|q| < 1$, якщо $\tau\sqrt{\lambda} < 2$, тобто за цієї умови схема буде стійкою в тому розумінні, що $y_n \to 0$ при $n \to \infty$.

3.11. Загальні зауваження про методи розв'язування задачі Коші

Вибір чисельного методу залежить від багатьох обставин: обсягу обчислень, необхідного об'єму оперативної пам'яті ЕОМ, порядку точності, стійкості до похибок заокруглення і т. п. Який же з розглянутих вище методів розв'язування задачі Коші слід вибрати? Щоб відповісти на це запитання, треба брати до уваги і особливості задачі, яка розв'язується, і особливості кожного з методів, але навіть з урахуванням цих обставин дістати точну відповідь часто не вдається. Тоді критерієм істини має стати практика. Зробимо кілька зауважень з приводу розглянутих методів.

1. Достоїнствами багатьох методів Рунге — Кутта є такі: а) вони явні, тобто значення y_{n+1} обчислюється за раніше знайденим значенням скінченою кількістю дій за відомими формулами; б) всі схеми допускають розрахунок зі змінним кроком, і тому неважко зменшити крок там, де розв'язок швидко змінюється, і збільшити там, де він змінюється поволі; в) для початку розрахунків досить вибрати сітку і задати значення y_0 (воно відоме з постановки задачі), а далі обчислення здійснюються циклічно за одними і тими самими формулами; г) методи Рунге — Кутта мають «високу» точність. Останнє пояснимо більш детально. Вище встановлено лише асимптотичну поведінку залишкових членів конкретних формул Рунге — Кутта і показано, що вони є досить громіздкими і складно залежать від максимумів модулів відповідних похідних від правої частини $f(x, u)$. Наочне уявлення про величину цих залишкових членів можна дістати в окремому випадку $f(t, u) \equiv f(t)$. При цьому розв'язання диференціального рівняння зводиться до квадратури, а всі схеми чисельного розв'язування задачі Коші переходять у квадратурні формули. Наприклад, схема четвертого порядку точності набуває вигляду

$$y_{n+1} = y_n + \frac{\tau}{6}\left[f(t_n) + 4f\left(t_n + \frac{\tau}{2}\right) + f(t_{n+1})\right],$$

тобто збігається зі схемою, яка утворюється, якщо в рівності $u_{n+1} = u_n + \int_{t_n}^{t_{n+1}} f(t)\,dt$ застосувати для обчислення інтеграла формулу Сімпсона. Залишковий член в цьому разі має вигляд (див. залишковий член квадратурної формули Сімпсона) $R = \frac{t_{n+1} - t_n}{2880}\tau^4 \max |f^{(4)}|$, тобто числовий коефіцієнт у залишковому члені досить малий, і це є однією з причин високої точності схем Рунге — Кутта.

Таким чином, якщо права частина диференціального рівняння неперервна і обмежена разом зі своїми четвертими похідними (які не дуже великі), то формули Рунге — Кутта четвертого порядку точності по-

винні давати хороші результати, і точність їх при зменшенні кроку швидко зростає. Якщо права частина не має вказаних властивостей, то граничний порядок точності цих схем не може бути досягнутий. Тоді не гірші (але і не кращі) результати можна дістати за допомогою більш простих схем меншого порядку.

2. Досягти необхідної точності можна за рахунок зменшення кроку в формулах Рунге — Кутта. Оскільки вирази для залишкових членів занадто громіздкі, то апріорними оцінками точності для вибору кроку в практичних розрахунках не користуються. Зручніше робити розрахунки на послідовностях сіток, які ущільнюються, і давати апостеріорну оцінку точності, наприклад за методом Рунге. Ряд ефективних програм використовують для цього викладені методи Рунге — Кутта.

3. У деяких задачах функції є досить гладкими, але швидко змінюються. Тоді схеми Рунге — Кутта як низького, так і високого порядку точності вимагають занадто малого кроку для знаходження доброго результату. Такі задачі є жорсткими і вимагають розробки спеціальних методів, які орієнтовані на даний вузький клас задач. Наприклад, розглянемо задачу

$$\frac{du}{dt} = \alpha(t)\,u, \quad u(0) = u_0,$$

де $\alpha(t)$ — сильно змінюється, набираючи як додатних, так і від'ємних значень (такі задачі зустрічаються в хімічній кінетиці). Розв'язок

$$u(t) = u_0 \exp\left(\int_0^t \alpha(s)\,ds\right)$$

є невід'ємною функцією при будь-яких t. Якщо $\alpha(t) \geqslant 0$, то можна користуватися схемою Ейлера $y_{n+1} = y_n + \tau\alpha_n y_n = (1 + \tau\alpha_n)\,y_n$. Якщо $\alpha(t) < 0$, то може виявитися, що $1 + \tau\alpha_{n_1} < 0$ при деякому $n = n_1$, і тоді чисельний розв'язок втрачає зміст. У цьому разі можна користуватися неявною схемою $y_{n+1} = y_n + \tau\alpha_n y_{n+1}$, або $y_{n+1} = y_n/(1 - \tau\lambda_n)$, $1 - \tau\alpha_n > 1$, яка є стійкою при будь-якому τ. Звідси випливає, що для розв'язування поставленої задачі можна застосувати комбінований метод: в тих вузлах, де $\alpha(t) \geqslant 0$, треба використовувати явну схему, а в вузлах, де $\alpha(t) < 0$,— неявну.

4. Щоб почати розрахунки багатокроковими методами, наприклад методом Адамса p-го порядку точності, треба знати розв'язок в p точках, який обчислюється яким-небудь іншим методом. Ця обставина збільшує об'єм програми. Крім того, формули методу Адамса для змінного кроку досить громіздкі, а більш прості формули для сталого кроку потребують нестандартних дій при зміні кроку. Це робить метод Адамса незручним для розрахунків на ЕОМ. Достоїнством його є те,

що, наприклад, в методі четвертого порядку точності за один крок доводиться лише один раз обчислювати $f(x, u)$, в той час, як у формулі Рунге — Кутта того самого порядку точності $f(x, u)$ обчислюється на кожному кроці чотири рази. Проте коефіцієнт у залишковому члені схеми Рунге — Кутта четвертого порядку точності ($R = \tau^5 \max |f^{(4)}|/2880$) при $f(t, u) \equiv f(t)$ в 960 разів менший, ніж у відповідній формулі Адамса зі сталим кроком ($R = (251/750)\,\tau^5 \max |f^{(4)}|$).

Отже, при однаковій точності схема Рунге — Кутта дозволяє брати крок у $\sqrt[4]{960} \approx 5{,}7$ разів більший. Це приводить до того, що фактичне число обчислень $f(x, u)$ у методі Рунге — Кутта може виявитися навіть меншим, ніж в методі Адамса. Тому метод Адамса та інші аналогічні методи застосовуються рідше, ніж метод Рунге — Кутта. Часто використовуються програми з комбінацією методів Рунге — Кутта і Адамса з автоматичним вибором кроку. За допомогою методу Рунге — Кутта обчислюються початкові значення для відповідного багатокрокового методу.

Більш елегантним є застосування методів зростаючого порядку точності. Так, за допомогою методу Ейлера — Коші першого порядку обчислюють y_1, потім, маючи y_0, y_1, за допомогою явного двокрокового методу дістають y_2 і т. д. Як правило, при обчисленні початкових (стартових) значень крок τ вибирається значно меншим, ніж при наступному обчисленні за багатокроковими формулами.

5. Щоб мати можливість автоматичного вибору кроку при використанні багатокрокових методів, треба мати апостеріорні оцінки похибки на кожному кроці. Для цього можна використати, наприклад, такі дві можливості.

Перша з них грунтується на тому, що локальна похибка методів Гіра (*BDF*), методу Адамса — Мултона, методу предиктор-коректор, що складається з p-крокової формули Адамса — Мултона (коректор) і $(p+1)$-крокової формули Адамса — Башфорта (предиктор), має вигляд

$$\tau^{p+1} c_{p+1} u^{(p+2)}(t) + O(\tau^{p+2}),$$

і похідна $u^{(p+2)}(t)$ в головному члені може бути оцінена через відповідну поділену різницю величин $f(t_i, y_i)$, обчислених на попередніх кроках.

Друга можливість грунтується на тому, що значення

$$\tilde{y}_{i+1} - y_{i+1},$$

де $\tilde{y}_{i+1}$ обчислюється за формулою Адамса — Башфорта (предиктор), а y_{i+1} — за формулою Адамса — Мултона (коректор), дає апостеріорну оцінку відповідного порядку точності.

Вказані вище апостеріорні оцінки дістаємо (на відміну від вкладених формул Рунге — Кутта) без додаткового обчислення правої частини $f(t, u)$.

6. Якщо застосовується чисто неявний метод (неявний метод Ейлера, неявний метод середньої точки, методи Гіра, метод трапецій

$$y_{i+1} = y_i + \frac{\tau_i}{2}(f(t_i, y_i) + f(t_{i+1}, y_{i+1})),$$

то на кожному кроці треба розв'язувати це нелінійне відносно y_{i+1} рівняння. Для цього, здавалося б, можна застосувати ітераційний процес

$$y_{i+1}^{(k+1)} = y_i + \frac{\tau_i}{2}(f(t_i, y_i) + f(t_{i+1}, y_{i+1}^{(k)})), \quad k = 0, 1, \ldots,$$

$$y_{i+1}^{(0)} = y_i + \tau_i f(t_i, y_i),$$

для збіжності якого необхідне виконання умови $\frac{\tau_i}{2} L < 1$, де L — стала Ліпшиця для функції $f(t, u)$ по u. Якщо навіть не враховувати необхідну при цьому велику кількість обчислень функції $f(t, u)$, то цей процес незручний і з тієї точки зору, що L може бути великим (для жорстких рівнянь $L \sim 10^6 - 10^{12}$), і ми маємо дуже сильне обмеження на τ. Тому часто застосовують спрощений метод Ньютона

$$\left(I - \frac{\tau_i}{2} J(f(t_i, y_i))\right)(y_{i+1}^{(k+1)} - y_i) = -y_{i+1}^{(k)} + y_i + \frac{\tau_i}{2}(f(t_i, y_i) + f(t_{i+1}, y_{i+1}^{(k)})),$$

де через $J(f)$ позначено матрицю Якобі вектор-функції $f(\frac{\partial f}{\partial u}$ в скалярному випадку). Якщо навіть частинні похідні в $J(f)$ замінити відповідними скінченними різницями, то досить незначної кількості ітерацій (*як правило, 3—5*), щоб дістати розв'язок із бажаною точністю.

7. Оскільки алгоритми з використанням багатокрокових методів у порівнянні з тими, що використовують однокрокові методи, є складнішими, то їх доцільно застосувати тоді, коли треба мати розв'язок з високою точністю. Зазначимо, що багатокрокові методи є чутливішими до сильних змін розв'язку (бо в них використовують значення функції f на більш великому інтервалі зміни t), тому їх доцільно застосувати, коли функція $f(t, u)$ в усій області інтегрування є досить гладкою.

8. Теорія багатокрокових методів зі змінним кроком, зокрема теорія збіжності, управління кроком і т. п., ще далека до завершення. Деякі результати містяться у наведеній в кінці книги літературі.

ГЛАВА 4

ЧИСЕЛЬНЕ РОЗВ'ЯЗУВАННЯ КРАЙОВИХ ЗАДАЧ ДЛЯ ЗВИЧАЙНИХ ДИФЕРЕНЦІАЛЬНИХ РІВНЯНЬ

4.1. Постановка крайових задач для звичайних диференціальних рівнянь. Класифікація методів розв'язування крайових задач для звичайних диференціальних рівнянь

Як відомо з теорії диференціальних рівнянь, загальний розв'язок системи диференціальних рівнянь першого порядку

$$Y'(x) = F(x,\ Y(x)),$$

$$Y(x) = (y_1(x)\ \ \ldots,\ \ y_n(x))^T,\ F(x,\ Y(x)) = (f_1(x,\ y_1(x),\ \ldots,\ y_n(x)),$$

$$f_2(x\ ,y_1(x),\ \ldots,\ y_n(x)),\ \ldots,\ f_n(x,y_1(x),\ \ldots,\ y_n(x))^T \qquad (1)$$

залежить від n довільних сталих. Отже, щоб виділити якийсь один розв'язок, треба задати n додаткових умов. У задачі Коші ці n умов були задані в деякій одній точці.

Крайова задача — це задача відшукання єдиного розв'язку системи (1) на відрізку $a \leqslant x \leqslant b$, в якій додаткові умови накладаються на значення функцій $y_k(x)$ більш ніж в одній точці цього відрізка. Очевидно, що крайові задачі мають зміст для систем не нижче другого порядку.

Приклад 1. Знайти статичний прогин $u(x)$ навантаженої струни із закріпленими кінцями, якщо

$$u''(x) = -f(x),\quad a < x < b,\quad u(a) = u(b) = 0,$$

де $f(x)$ — зовнішнє вигинаюче навантаження на одиницю довжини струни, поділене на пружність струни. Ця задача належить до класу крайових задач, в яких одна частина додаткових умов задана на одному кінці відрізка ($x = a$), а інша — на другому (при $x = b$). Звідси і назва — крайові (граничні) задачі.

В термінах (1) дану задачу можна записати у вигляді

$$\begin{vmatrix} y_1'(x) \\ y_2'(x) \end{vmatrix} = \begin{bmatrix} y_2(x) \\ -f(x) \end{bmatrix},\quad x \in (a,\ b),\quad y_1(a) = y_1(b) = 0,$$

де $y_1(x) = u(x)$, $y_2(x) = u'(x)$.

Приклад 2. Статичний прогин навантаженого пружного бруска задовольняє рівняння четвертого порядку

$$u^{(4)}(x) = f(x),\quad a \leqslant x \leqslant b.$$

Якщо цей брусок лежить в точках x_i, $i = \overline{1,\ 4}$, на опорах, то додаткові (крайові) умови мають вигляд

$$u(x_i) = 0,\quad i = \overline{1,\ 4},\quad a \leqslant x_1 < x_2 < x_3 < x_4 \leqslant b.$$

Дану задачу можна звести до вигляду (1) з відповідними крайовими умовами. Ця крайова задача називається *багатоточковою*. Наведемо загальний вигляд багатоточкових крайових умов:

$$\varphi(Y(x_1),\ Y(x_2),\ \ldots,\ Y(x_k)) = d, \qquad (2)$$

де

$$\varphi(Y(x_1),\ \ldots,\ Y(x_k)) = (\varphi_i(x_1,\ y_1(x_1),\ \ldots,\ y_n(x_1),\ \ldots x_k,$$

$y_1(x_k),\ \ldots y_n(x_k)))_{i=\overline{1,n}}$, $d = (d_i)_{i=\overline{1,n}}$ — заданий вектор.

З багатоточкових задач вигляду (1), (2) окремо можна виділити групу двоточкових лінійних задач:

$$Y' - A(x)\,Y = F(x), \qquad (3)$$

$$B_1 Y(a) + B_2 Y(b) = d, \qquad (4)$$

де Y, F, d — n-вимірні вектори; $A(x) = (a_{ij}(x))_{i,j=\overline{1,n}}$, $B_1 = (b_{ij}^{(1)})_{i,j=\overline{1,n}}$; $B_2 = (b_{ij}^{(2)})_{i,j=\overline{1,n}}$ — матриці розмірності $n \times n$.

Будь-яку двоточкову крайову задачу для лінійного диференціального рівняння n-го порядку

$$y^{(n)}(x) = \sum_{i=1}^{n} p_i(x)\,y^{(n-i)}(x) + f(x),$$

$$s_i = \gamma_i,\ i = \overline{1,\ n}, \qquad (5)$$

де $s_i = \sum_{j=1}^{n} (\alpha_{ij} y^{(j-1)}(a) + \beta_{ij} y^{(j-1)}(b))$, можна записати у вигляді (3), (4). Це стає очевидним, якщо ввести позначення

$$Y(x) = (y(x),\ y'(x),\ \ldots,\ y^{(n-1)}(x))^T,$$

$$F(x) = (0,\ \ldots,\ 0,\ f(x)),\ d = (\gamma_i)_{i=\overline{1,n}},$$

$$A(x) = \begin{Vmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \\ p_n & p_{n-1} & p_{n-2} & \ldots & p_1 \end{Vmatrix},$$

$$B_1 = (\alpha_{ij})_{i,j=\overline{1,n}},\quad B_2 = (B_{ij})_{i,j=\overline{1,n}}.$$

Крайову задачу (3), (4) можна звести до однорідної лінійної системи диференціальних рівнянь з неоднорідними крайовими умовами вигляду

$$Z'(x) = M(x)\,Z(x), \qquad (6)$$

$$\hat{B}_1 Z(a) + \hat{B}_2 Z(b) = \hat{d}, \qquad (7)$$

якщо ввести позначення: $Z(x) = (Y(x), 1)^T$, $\hat{d} = (d, 1)^T$ — $(n+1)$-вимірні вектори; $M(x) = \begin{pmatrix} A(x) & F(x) \\ 0 & 0 \end{pmatrix}$, $\hat{B}_1 = \begin{pmatrix} B_1 & 0 \\ 0 & 0 \end{pmatrix}$, $\hat{B}_2 = \begin{pmatrix} B_2 & 0 \\ 0 & 0 \end{pmatrix}$ — матриці розмірності $(n+1) \times (n+1)$.

Існують задачі з додатковими умовами нелокального типу (тобто заданими за допомогою деяких інших функціоналів від розв'язку), наприклад:

$$\int_a^b y_k^2(x)\,dx = 1, \quad k = \overline{1, n}.$$

Такі умови нормування зустрічаються в квантовій механіці.

Питання, пов'язані з існуванням розв'язку, і побудова наближених методів для крайових задач, як правило, складніші, ніж відповідні питання для задачі Коші. Методи розв'язування крайових задач (рис. 4) можна класифікувати так: методи зведення крайової задачі до задачі Коші, до яких належать метод стрільби, метод ортогоналізації, метод спряжених рівнянь, метод лінеаризації з наступним зведенням до задачі Коші, метод продовження розв'язку за параметром; сіткові методи, до яких належать метод скінченних елементів, метод скінченних різниць та проекційно-варіаційні методи, розглянуті нами раніше.

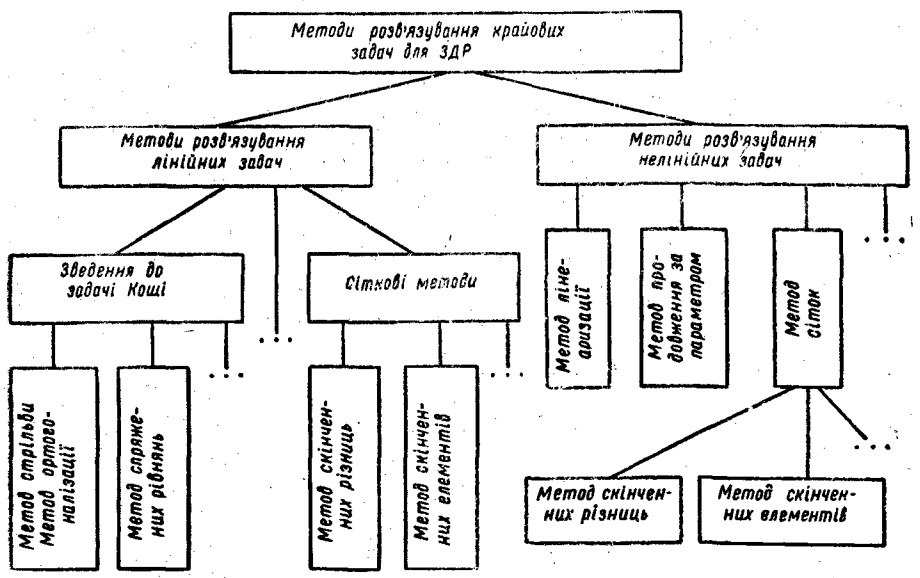


Рис. 4

Зазначимо, що в наведених вище постановках крайових задач вимагається, щоб невідома функція мала похідні, які входять до рівняння. Ці диференціальні рівняння на практиці часто утворюються з деяких інтегральних законів, до яких входять похідні від невідомої функції нижчих порядків. При цьому часто буває так, що вищі похідні від розв'язку можуть не існувати. Тоді крайова задача розглядається в узагальненій постановці. Розглянемо, наприклад, першу крайову задачу для звичайного диференціального рівняння другого порядку

$$Lu = -\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) + q(x)u = f(x), \quad 0 < x < 1, \tag{8}$$

$$u(0) = \mu_1, \; u(1) = \mu_2, \quad k(x) \geqslant k_0 > 0, \; q(x) \geqslant 0. \tag{9}$$

Така задача описує стаціонарний (тобто такий, що не змінюється з часом) розподіл температури у стержні довжиною 1 або концентрації речовини у деякому розчині. У першому випадку $u = u(x)$ — температура, $w(x) = -k(x)\frac{du}{dx}$ — тепловий потік, $k(x)$ — коефіцієнт теплопровідності, а саме диференціальне рівняння називається *стаціонарним рівнянням теплопровідності*. У другому випадку $u(x)$ — концентрація речовини, $w(x)$ — потік речовини, $k(x)$ — коефіцієнт дифузії, а рівняння називається *стаціонарним рівнянням дифузії*. Можливі й інші крайові умови при $x = 0$ і $x = 1$. Наприклад, $ku' = \sigma_1 u - \mu_1$ при $x = 0$ і $-ku' = \sigma_2 u - \mu_2$ при $x = 1$ ($\sigma_1 \geqslant 0$, $\sigma_2 \geqslant 0$). Якщо $\sigma_1 > 0$, то умова $ku' = \sigma_1 u - \mu_1$ називається умовою третього роду, а при $\sigma_1 = 0$ — другого роду. Можливі також комбінації різних умов при $x = 0$ і $x = 1$. Нехай $\mu_1 = \mu_2 = 0$, $D(L)$ — множина функцій з нормованого простору $W_2^2(0, 1)$, які дорівнюють нулю при $x = 0$ та $x = 1$. Цю множину позначають також $W_{2,0}^2(0, 1)$, тобто $D(L) \equiv W_{2,0}^2(0, 1)$. Нехай $k(x)$ має обмежену похідну, функція $q(x)$ обмежена, а функція $f(x)$ належить простору $R(L) \equiv L_2(0, 1)$. Тоді задачу (8), (9) запишемо у вигляді операторного рівняння

$$Lu = f, \tag{10}$$

де оператор L визначається формулою

$$Lu \equiv -\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) + q(x)u$$

і має область визначення $D(L)$ та область значень $R(L)$. Оскільки $W_{2,0}^2(0, 1) \subset \mathring{W}_2^1(0, 1) \subset L_2(0, 1)$, то оператор можна розглядати як оператор, який діє у просторі зі скалярним добутком $L_2(0, 1)$. Будь-який нормований простір або простір зі скалярним добутком можна поповнити (тобто зробити його відповідно банаховим чи гільбертовим), а властивість повноти використовується, як правило, лише при обгрунтуванні збіжності методів. Звернемо далі основну увагу на алгоритміч-

ні аспекти, відступаючи іноді в незначній мірі від математичної строгості.

Запишемо для оператора L функціонал енергії (див. п. 1.2)

$$J(v) = (Lv,\ v) - 2(f,\ v) = \int_0^1 \left[-\frac{d}{dx}\left(k(x)\frac{dv(x)}{dx}\right) + q(x)v(x)\right] v(x)\,dx - 2\int_0^1 f(x)v(x)\,dx.$$

Вправа 1. Переконатися, що $(L\ v,\ v) \geqslant \mu \| v \|^2$. Інтегруючи частинами, маємо

$$J(v) = \int_0^1 \left[k(x)\left(\frac{dv}{dx}\right)^2 + q(x)v^2(x)\right] dx - 2\int_0^1 f(x)v(x)\,dx. \qquad (11)$$

За деяких додаткових умов, на які зараз не звертаємо увагу, в п. 1.2 було доведено, що задача мінімізації функціонала (11) на $D(L)$ і задача (10) еквівалентні.

Те саме можна довести й іншим способом. Дійсно, замінюючи в (11) функцію v на $u + (v - u)$, перепишемо (11) у вигляді

$$J(v) = J(u) + \int_0^1 \left[k(x)\left(\frac{d(v-u)}{dx}\right)^2 + q(x)(v-u)^2\right] dx +$$

$$+ 2\int_0^1 \left[k(x)\frac{du}{dx}\frac{d(v-u)}{dx} + q(x)u(v-u) - f(x)(v-u)\right] dx =$$

$$= J(u) + \int_0^1 \left[k(x)\left(\frac{d(v-u)}{dx}\right)^2 + q(x)(v-u)^2\right] dx +$$

$$+ 2\int_0^1 \left[-\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) + q(x)u - f\right](v-u)\,dx =$$

$$= J(u) + \int_0^1 \left[k(x)\left(\frac{d(v-u)}{dx}\right)^2 + q(x)(v-u)^2\right] dx.$$

Тоді

$$J(v) \geqslant J(u)\ \forall\, v \in \mathring{W}_2^1,$$

тобто розв'язок задачі (8), (9) мінімізує функціонал (11). Нехай, $v(x) \in \in W_{2,0}^2(0, 1)$ мінімізує функціонал (11). Покажемо, що тоді ця функція є розв'язком задачі (8), (9). Для цього в (11) підставимо функцію $v(x) + \alpha\varphi(x)$, де α — параметр, а $\varphi(x)$ — довільна функція з $\mathring{W}_2^1(0, 1)$.

Функція $J(v + \alpha u)$ як функція від α має мінімум при $\alpha = 0$, тому

$$\frac{d}{d\alpha} J(v + \alpha\varphi)|_{\alpha=0} = 0. \qquad (12)$$

Вираз у лівій частині цієї рівності називається першою варіацією функціонала і позначається δJ. Неважко показати, що

$$\delta J = 2\int_0^1 \left[k(x)\frac{dv}{dx}\frac{d\varphi}{dx} + q(x)v(x)\varphi(x) - f(x)\varphi(x)\right] dx.$$

Із (12) дістаємо

$$\int_0^1 \left[k(x)\frac{dv}{dx}\frac{d\varphi}{dx} + q(x)v(x)\varphi(x)\right] dx =$$

$$= \int_0^1 f(x)\varphi(x)\,dx\ \forall\, \varphi(x) \in \mathring{W}_2^1(0,\ 1). \qquad (13)$$

Інтегруючи частинами, приходимо до рівності

$$\int_0^1 \left[-\frac{d}{dx}\left(k(x)\frac{dv(x)}{dx}\right) + q(x)v(x) - f(x)\right]\varphi(x)\,dx = 0,$$

і в силу довільності φ робимо висновок, що v задовольняє рівняння (8), тобто є розв'язком задачі (8), (9).

Таким чином, задача мінімізації функціонала (11) еквівалентна задачі відшукання функції $v \in \mathring{W}_2^1(0, 1)$, яка задовольняє тотожність (13) при довільній функції $\varphi(x) \in \mathring{W}_2^1(0, 1)$. Отже, всі ці три задачі еквівалентні. Зазначимо, що рівняння (13) можна дістати з рівності

$$(Lu - f,\ \varphi) = 0\ \forall\, \varphi \in \mathring{W}_2^1(0,\ 1)$$

(в силу (10)), інтегруючи частинами.

Зазначимо, що вимоги до гладкості шуканої функції і функцій k, q у задачах (13) та мінімізації функціонала (11) можуть бути більш слабкі, ніж у задачі (8), (9) (в (13), (11) досить існування перших похідних від u, а в (9), (8) треба, щоб існували другі похідні). Тому розв'язок рівняння (13), який належить простору $\mathring{W}_2^1(0, 1)$, має назву узагальненого розв'язку крайової задачі (8), (9) при $\mu_1 = \mu_2 = 0$. Якщо крайова задача (8), (9) має розв'язок в $W_{2,0}^2(0, 1)$, то він буде і узагальненим розв'язком. Від функції k, наприклад, в (13) достатньо вимагати обмеженості. Таким чином, постановка задачі у вигляді (13) дозволяє поширити поняття розв'язку крайової задачі (8), (9), яка у даному випадку є лише символічним записом більш загальної задачі (13).

Нагадаємо ще один факт. Хоча розв'язок задачі (8), (9) шукається у просторі $W^2_{2,0}(0, 1)$, при побудові наближеного розв'язку за допомогою тотожності (13) його можна шукати в більш широкому просторі $\mathring{W}^1_2$, що буває зручно на практиці. У цьому випадку говорять про зовнішню апроксимацію на відміну від внутрішньої апроксимації, коли наближення і точний розв'язок містяться в одному просторі.

Як вже зазначалося, на практиці часто зустрічаються граничні умови другого роду

$$u'(0) = 0, \quad u'(1) = 0, \tag{14}$$

або умови Неймана. Тоді задача (8), (9) називається другою граничною задачею. Узагальненим розв'язком задачі (8), (9) називатимемо функцію $u(x) \in W^1_2(0, 1)$ (на відміну від $\mathring{W}^1_2(0, 1)$), яка задовольняє тотожність

$$\int_0^1 \left[k(x)\frac{du}{dx}\frac{d\varphi}{dx} + q(x)u\varphi\right]dx = \int_0^1 f\varphi dx \;\forall\, \varphi \in W^1_2(0, 1). \tag{14'}$$

Тут, на відміну від першої граничної задачі, розв'язок шукають серед функцій, «вільних» на кінцях відрізка, тобто у просторі $W^1_2(0, 1)$, а не в $\mathring{W}^1_2(0, 1)$, такою самою є функція φ.

Вправа 2. Показати, що коли узагальнений розв'язок другої граничної задачі належить $W^2_2(0, 1)$, то він задовольняє граничні умови (14).

Граничні умови (14) називають також природними, а граничні умови (9) головними. Таким чином, природні граничні умови випливають з узагальненої постановки задачі, а головні умови ми «нав'язуємо» розв'язку.

4.2. Зведення крайової задачі до послідовності задач Коші

4.2.1. Метод стрільби та метод ортогоналізації для розв'язування лінійних систем. Найпростіше цей метод реалізується для лінійної крайової задачі вигляду (6), (7) з п. 4.1. У цьому разі розв'язок задачі Коші для рівняння (6) з п. 4.1 з *деякими початковим значенням* $Z(a)$ можна подати у вигляді

$$Z(x) = \Gamma(x)Z(a), \tag{1}$$

де $\Gamma(x)$ — фундаментальна матриця (тобто матриця фундаментальних розв'язків) системи (6) з п. 4.1. Підставляючи (1) в крайову умову (7) з п. 4.1, дістанемо таку систему лінійних алгебраїчних рівнянь відносно $Z(a)$:

$$[\hat{B}_1 + \hat{B}_2\Gamma(b)]Z(a) = \hat{d}, \tag{2}$$

визначник якої у випадку єдиності розв'язку задачі (6), (7) з п. 4.1 відмінний від нуля (припустивши протилежне, можна легко переконатися, що однорідна крайова задача має ненульовий розв'язок). Якщо $Z(a)$ — розв'язок системи (2), то вектор-функція (1) задовольнятиме систему (6) і крайові умови (7) з п. 4.1. У цьому і полягає ідея методу стрільби. Обчислювальна схема цього методу складається з таких етапів: 1) будуємо матрицю $\Gamma(x)$, чисельним інтегруванням розв'язуючи для $x \in [a, b]$ задачі Коші $z_i'(x) = M(x)z_i(x)$, $z_i(a) = e_i \equiv (\delta_{ij})_{j=\overline{1,n+1}}$, δ_{ij} — символ Кронекера; тоді $\Gamma(b) = (z_1(b), \ldots, z_{n+1}(b))$; 2) розв'язуючи систему (2), знаходимо $Z(a)$; 3) за $\Gamma(x)$ і $Z(a)$ знаходимо $Z(x)$ за формулою (1).

Метод стрільби спрощується, якщо крайові умови в (7) з п. 4.1 розділені. Розглянемо наступну крайову задачу з розділеними крайовими умовами

$$Y'(x) = A(x)Y(x) + F(x), \tag{3}$$

$$PY(a) = \omega_1, \tag{4}$$

$$QY(b) = \omega_2, \tag{5}$$

де $P = (p_{ij})^{j=\overline{1,n}}_{i=\overline{1,k}}$, $Q = (q_{ij})^{j=\overline{1,n}}_{i=\overline{1,n-k}}$ — задані прямокутні матриці відповідно порядків $k \times n$ та $(n-k) \times n$ $(k < n)$; $F(x)$ — задана вектор-функція розмірності n; ω_1, ω_2 — задані вектори розмірності k та $n-k$ відповідно; $A(x)$ — задана квадратна матриця порядку $n \times n$, причому ранг P дорівнює k, ранг Q дорівнює $n-k$. Тоді алгоритм методу стрільби набуває такого вигляду.

1. Знаходимо лінійно незалежні розв'язки однорідної системи лінійних алгебраїчних рівнянь

$$PY^{(i)} = 0. \tag{6}$$

Нехай це будуть вектори $Y^{(i)}$, $i = \overline{1, n-k}$ (їх $n-k$, бо ранг матриці P дорівнює k). Знаходимо окремий (частинний) розв'язок системи

$$PY = \omega_1, \tag{7}$$

який позначимо $Y^{(n-k+1)}$.

2. За допомогою чисельного інтегрування розв'язуємо $n-k$ задач Коші для однорідного рівняння (3)

$$Y_i'(x) = A(x)Y_i(x),$$

$$Y_i(a) = Y^{(i)}, \quad i = \overline{1, n-k}, \tag{8}$$

та задачу Коші для неоднорідного рівняння (3)

$$Y'_{n-k+1}(x) = A(x)\,Y_{n-k+1}(x) + F(x),$$

$$Y_{n-k+1}(a) = Y^{(n-k+1)}. \tag{9}$$

За допомогою $Y_i(x)$, $i = \overline{1, n-k+1}$, будь-який розв'язок системи (3), що задовольняє крайову умову (4), можна подати у вигляді

$$Y(x) = \sum_{i=1}^{n-k} c_i Y_i(x) + Y_{n-k+1}(x), \tag{10}$$

де c_i — довільні сталі. Підставляючи (10) в крайові умови (5), дістаємо систему лінійних алгебраїчних рівнянь для визначення c_i:

$$Rc = t, \tag{11}$$

де $R = QV$, $V = (Y_1(b), \ldots, Y_{n-k+1}(b))$, $c = (c_1, \ldots, c_{n-k})^T$, $t = \omega_2 - w$, $w = QY_{n-k+1}(b)$.

3. Розв'язуємо систему (11) відносно c_i, $i = \overline{1, n-k}$, і за формулою (10) знаходимо розв'язок крайової задачі (3) — (5).

Як зазначалося вище у припущенні, що задача (3) — (5) має єдиний розв'язок, визначник системи (11) відмінний від нуля. Якщо зберігати всі проміжні значення $Y_i(x)$, $i = \overline{1, n-k+1}$, утворені на кроці 2, неможливо через обмеженість пам'яті ЕОМ, то можна діяти таким чином. Внаслідок виконання кроку 2 знайти $Y_i(b)$, $i = \overline{1, n-k+1}$ («забуваючи» проміжні значення), потім розв'язати систему (11) і за формулою (10) знайти розв'язок $Y(a)$ задачі (3) — (5) в точці a.

Після цього треба ще раз розв'язати задачу Коші для рівняння (3) з початковим значенням $Y(a)$.

Вправа 1. Записати розрахункові формули розв'язування задачі

$$u^{(4)}(x) = a^4 u(x) + b,$$

$$u(0) = u'(0) = 0,\ u(l) = u'(l) = 0$$

методом стрільби.

У деяких випадках розглянутий алгоритм методу стрільби непридатний для практичного використання, бо в ньому може дуже швидко зростати обчислювальна похибка. Таке трапляється, наприклад, якщо серед розв'язків системи $Y'(x) = A(x)\,Y$ є такі, що швидко змінюються зі зростанням x. Схематично це можна пояснити таким чином. Нехай A — стала матриця, що має дійсні різні власні значення λ_j, $j = \overline{1, n}$, які відповідають власним векторам e_j, $0 < \lambda_1 < \ldots < \lambda_n$, причому λ_n значно більше за λ_{n-1}. Нехай вектори $Y^{(i)}$, знайдені на кро-

ці 1, мають такі розвинення:

$$Y^{(i)} = \sum_{k=1}^{n} \alpha_{ik} e_k.$$

Тоді

$$Y_i(b) = \sum_{k=1}^{n} \alpha_{ik} e_k \exp(\lambda_k b),$$

причому, оскільки $\exp(\lambda_n b)$ значно більше, ніж $\exp(\lambda_{n-1} b)$, то для всіх $i = \overline{1, k}$

$$Y_i(b) \sim \alpha_{in} e_n \exp(\lambda_n b).$$

Це означає, що всі рядки визначника системи (11) приблизно пропорціональні вектору e_n, і тому розв'язок системи (11) можна знайти з великою обчислювальною похибкою.

Щоб розв'язати систему (11), треба в лінійному многовиді $L(x)$, який містить вектори вигляду

$$Y_{n-k+1}(x) + \sum_{i=1}^{n-k} c_i Y_i(x) \tag{12}$$

при $x = b$, знайти вектор, що задовольняє граничну умову (5). Тут $Y_{n-k+1}(x)$ — розв'язок неоднорідного рівняння (3), $Y_i(x)$, $i = \overline{1, n-k}$ — лінійно незалежні розв'язки однорідного рівняння (3). Розв'язання цієї задачі методом стрільби може призвести до незадовільних результатів, бо базис з векторів $Y_i(x)$ під час просування до точки b «сплющується». Тому виникає ідея: по мірі просування до точки b періодично переходити до іншого базису, тобто до іншого способу задання многовиду (12). На цій ідеї грунтується побудова методу ортогоналізації, до розгляду якого ми і переходимо.

Многовид $L(x)$ вигляду (12) можна задавати будь-яким набором базисних векторів, наприклад, набором $Y_i(x)$, $i = \overline{1, n-k+1}$, або набором векторів $g_i(x)$, $i = \overline{1, n-k+1}$, здобутих ортогоналізацією $Y_i(x)$, $i = \overline{1, n-k+1}$. Ці набори утворюють прямокутні матриці $G_Y(x)$ та $G_Y^{\perp}(x)$ відповідно. Іншими словами, многовид $L(x)$ можна задавати матрицею $G_Y(x)$, або $G_Y^{\perp}(x)$, або будь-якою іншою матрицею, яка складена з лінійно незалежних векторів цього многовиду. Перетворення, яке переводить $G_Y(x)$ в $G_Y^{\perp}(x)$, позначимо через U, тобто $G_Y^{\perp}(x) = U(G_Y(x))$.

Розіб'ємо відрізок $[a, b]$ на частини точками $a = X_1 < X_2 < \ldots < X_m = b$ і введемо наступне перетворення M_s^X для деякої матриці $G_p(X_s) = (p_1(X_s), \ldots, p_{n-k}(X_s), p_{n-k+1}(X_s))$ розмірності $n \times (n - k + 1)$. Якщо через $p_j^s(x)$, $j = \overline{1, n-k}$, позначити розв'язок

системи $Y'(x) = A(x)\,Y(x)$ за початкових умов $Y(X_s) = p_j(X_s)$, а через $p^s_{n-k+1}(x)$ — розв'язок системи $Y'(x) = A(x)\,Y(x) + f(x)$ за початкових умов $Y(X_s) = p_{n-k+1}(X_s)$, $x \in [X_s,\ X_{s+1}]$, то, за означенням,

$$M^X_s(G_Y(X_s)) = (p^s_1(x),\ \ldots,\ p^s_{n-k}(x),\ p^s_{n-k+1}(x)),$$

або

$$M_s(p_j(X_s)) = p^s_j(X_{s+1}),\quad j = \overline{1,\ n-k+1}.$$

Покладемо $M_s \equiv M^{X_s+1}_s$. Якщо метод стрільби застосовується так, що за допомогою чисельного інтегрування спочатку за $Y^{(i)} \equiv Y_i(a)$, $i = \overline{1,\ n-k+1}$, визначається $Y_i(X_2)$, $i = \overline{1,\ n-k+1}$, потім за $Y_i(X_2)$ визначається $Y_i(X_3)$, $i = \overline{1,\ n-k+1}$, і т. д. доти, доки не дістанемо $Y_i(b)$, $i = \overline{1,\ n-k+1}$, то в термінах перетворення M_s цей процес («прямий хід» методу стрільби) можна записати так:

$$G_Y(X_2) = M_1(G_Y(X_1)),\quad G_Y(X_3)) = M_2(G_Y(X_2)),\ \ldots,$$
$$G_Y(X_m) = M_{m-1}(G_Y(X_{m-1})),$$

тобто послідовно визначаємо многовиди $L(X_2),\ \ldots,\ L(X_m) = L(b)$, причому $L(X_s)$ є многовидом всіх векторів $Y(X_s)$, які визначаються розв'язками системи $Y'(x) = A(x)\,Y(x) + f(x)$, що задовольняють ліву граничну умову $PY(a) = \omega_1$. Розглянемо питання: що станеться, якщо в цьому процесі матриці $G_Y(X_i)$, $i = \overline{1,\ m-1}$, які визначають лінійні многовиди $L(x_i)$, $i = \overline{1,\ m-1}$, замінити іншими матрицями, що визначають ті самі многовиди, наприклад матрицями $G^{\perp}_Y(X_i) \equiv U(G_Y(X_i))$, $i = \overline{1,\ m-1}$, де U позначає операцію ортогоналізації стовпчиків матриці $G_Y(X_i)$ і нормування перших $n-k$ з них. Тоді процес обчислення формально можно подати так: спочатку з матрицею $G_Y(X_1) \equiv G_Y(a)$ виконуємо перетворення U, в результаті якого дістаємо матрицю $G^{\perp}_Y(X_1) \equiv G_1(X_1)$, потім до $G_1(X_1)$ застосовуємо перетворення M_1, в результаті чого маємо матрицю $G_1(X_2) = M_1(G_1(X_1))$ (стовпчики в $G_1(X_2)$ не обов'язково ортогональні). Після цього знаходимо $G_2(X_2) = U(G_1(X_2))$ і т. д., поки не знайдемо матриці $G_{m-1}(X_m) = M_{m-1}(G_{m-1}(X_{m-1}))$, $G_m(X_m) = U(G_{m-1}(X_m))$. Матриця $G_s(X_s)$ визначає деякий лінійний многовид $\tilde{L}(X_s)$.

Лема 1. *Лінійні многовиди $\tilde{L}(X_s)$ і $L(X_s)$ збігаються, тобто многовид $\tilde{L}(X_s)$ є многовидом всіх значень $Y(X_s)$ розв'язків системи $Y(x) = A(x)\,Y(x) + f(x)$, що задовольняють ліву граничну умову $PY(a) = \omega_1$.*

Доведення. При $s = 1$ це твердження очевидне. Нехай воно має місце для $s = r$. Покажемо, що тоді воно справедливе і при $s = r+1$.

Кожен вектор многовиду $L(X_r)$ можна подати у вигляді

$$Y(X_r) = \sum_{j=1}^{n-k+1} c_j g^r_j(X_r) = G_r(X_r)\cdot c, \tag{13}$$

де $c = (c_j)_{j=\overline{1,n-k+1}}$ — вектор довільних сталих,

$$(g^r_1(X_r),\ \ldots,\ g^r_{n-k+1}(X_r)) = G_r(X_r) = U\,G_{r-1}(X_r) =$$
$$= U(M_{r-1}(G_{r-1}(X_{r-1}))).$$

Здійснюючи операцію M_r, переходимо від многовиду $L(X_r)$ всіх значень $Y(X_r)$ розв'язків системи $Y'(x) = A(x)\,Y(x) + f(x)$, які задовольняють ліву граничну умову, до многовиду $L(X_{r+1})$ значень таких розв'язків в точці X_{r+1}, тобто

$$Y(X_{r+1}) = G_r(X_{r+1})\cdot c. \tag{14}$$

Запишемо $G_r(X_{r+1})$ через $G_{r+1}(X_{r+1})$, пригадавши, що

$$G_{r+1}(X_{r+1}) \equiv (q_j)_{j=\overline{1,n-k+1}} = U(G_r(X_{r+1})),$$

$$G_r(X_{r+1}) = (g^r_1(X_{r+1}),\ \ldots,\ g^r_{n-k}(X_{r+1}),\ g^r_{n-k+1}(X_{r+1})).$$

Процес ортогоналізації Шмідта з нормуванням здійснюється формулами

$$z_j = g^r_j(X_{r+1}) - \sum_{i=1}^{j-1}(g^r_j(X_{r+1}),\ q_i)\,q_i,$$

$$q_j = z_j/\sqrt{(z_j,\ z_j)},\quad j = \overline{1,\ n-k},$$

$$q_{n-k+1} = g^r_{n-k+1}(X_{r+1}) - \sum_{i=1}^{n-k}(g^r_{n-k+1}(X_{r+1}),\ q_i)\,q_i$$

(вектор q_{n-k+1} не нормується). Перепишемо ці співвідношення у вигляді

$$\sqrt{(z_j,\ z_j)}\,q_j + \sum_{i=1}^{j-1}(g^r_j(X_{r+1}),\ q_i)\,q_i = g^r_j(X_{r+1}),\quad j = \overline{1,\ n-k},$$

$$\sum_{i=1}^{n-k}(g^r_{n-k+1}(X_{r-1}),\ q_i)\,q_i + q_{n-k+1} = g^r_{n-k+1}(X_{r+1}).$$

Систему співвідношень можна записати у матричній формі

$$G_r(X_{r+1}) = G_{r+1}(X_{r+1})\,\Omega_{r+1}, \tag{15}$$

де

$$\Omega_{r+1} = \begin{bmatrix} \omega_{11}^{(r+1)} & \omega_{12}^{(r+1)} & \dots & \omega_{1,n-k}^{(r+1)} & \omega_{1,n-k+1}^{(r+1)} \\ 0 & \omega_{22}^{(r+1)} & \dots & \omega_{2,n-k}^{(r+1)} & \omega_{2,n-k+1}^{(r+1)} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \omega_{n-k,n-k}^{(r+1)} & \omega_{n-k,n-k+1}^{(r+1)} \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix},$$

$$\omega_{ij} = \begin{cases} 0, & i > j, \\ \sqrt{(z_j, z_j)} = \sqrt{\|g_j^r(X_{r+1})\|^2 - \sum_{l=1}^{i-1} \omega_{lj}^2}, & i = j, \\ (g_j^r(X_{r+1}), q_i), & i < j \leqslant n-k+1. \end{cases} \quad (16)$$

Таким чином, з (14) дістаємо

$$Y(X_{r+1}) = G_{r+1}(X_{r+1})\,\Omega_{r+1}c.$$

Покладемо $\Omega_{r+1}c = \tilde{c}$, причому в силу того, що $\omega_{n-k+1,n-k+1}^{(r+1)} = 1$, остання компонента вектора $\tilde{c}$ дорівнює 1. Перепозначаючи $\tilde{c} = c$, маємо такий самий, як і (13), запис елементів многовиду $L(X_{r+1})$, що в силу принципу математичної індукції і завершує доведення леми.

На кожному з відрізків $[X_r, X_{r+1}]$ розв'язок задачі (3) — (5) можна записати у вигляді

$$Y(x) = \sum_{j=1}^{n-k+1} c_j^{(r)} g_j^r(x),$$

де $g_j^r(x)$ — елементи матриці $G_r(x) = M_r^X(G_r(X_r))$, $c_{n-k+1} = 1$, або інакше

$$Y(x) = G_r(x)\,c^{(r)}, \quad (17)$$

$c^{(r)} = (c_1^{(r)}, \dots, c_{n-k}^{(r)}, 1)^T$. Поклавши в (17) $x = X_{q+1}$, $r = q$, дістаємо

$$Y(X_{q+1}) = G_q(X_{q+1})\,c^{(q)},$$

а при $x = X_{q+1}$, $r = q+1$ маємо (після ортогоналізації $g_i^r(X_{q+1})$, $j = \overline{1,\ n-k+1}$)

$$Y(X_{q+1}) = G_{q+1}(X_{q+1}) \cdot c^{(q+1)}.$$

З останніх двох рівностей дістаємо

$$c^{(q+1)} = [G_{q+1}(X_{q+1})]^{-1} G_q(X_{q+1}),$$

або з урахуванням (15)

$$c^{(q+1)} = \Omega_{q+1}c^{(q)}. \quad (18)$$

Значення $c^{(m-1)}$ знаходять із системи лінійних алгебраїчних рівнянь, яка утворюється підстановкою (17) при $r = m-1$ (після ортогоналізації $G_{m-1}(X_m)$) в праву граничну умову

$$QY(X_m) = QUG_{m-1}(X_m)\,c^{(m-1)} = QG_m(X_m)\,c^{(m-1)} = \omega_2.$$

Далі за $c^{(m-1)}$ з (18) послідовно визначаємо $c^{(m-2)}, \dots, c^{(1)}$. Після цього для відшукання розв'язку на відрізку $[X_r, X_{r+1}]$ можна було б скористатися формулою (7). Однак такий спосіб вимагає великої пам'яті ЕОМ в силу необхідності запам'ятовувати значення матриці $G_r(x)$ у великій кількості точок. Для економії пам'яті доцільніше знайти значення $Y(X_s)$, а потім на кожному з відрізків $[X_r, X_{r+1}]$ розв'язувати задачі Коші.

Щодо реалізації цього алгоритму на ЕОМ зробимо ще такі зауваження: 1) для економії пам'яті і часу розрахунків за початкову точку інтегрування треба вибрати той кінець проміжку $[a, b]$, на якому задана більша кількість граничних умов; 2) з тією ж метою серед точок ортогоналізації виділяються точки X_{s_p}, в яких видаються результати, тоді матриці $G_r(X_{r+1})$ зберігаються лише для $r+1 = s_p$, $X_1 \leqslant X_{s_p} < X_m$, а замість матриць Ω_r в точках $X_1 < X_{s_p} < X_m$ зберігаються матриці трикутного вигляду $\Delta = \Omega_{s'}\Omega_{s'-1} \dots \Omega_{s+1}$ і вектори $c^{(s)}$ визначаються через $c^{(s')}$, $s < s' \leqslant m$, за формулою $\Delta c^{(s)} = c^{(s')}$; 3) кількість точок ортогоналізації в зоні сильної зміни розв'язку слід збільшити.

Не аналізуючи обчислювальну стійкість алгоритму ортогоналізації, зазначимо, що він мало чутливий до похибок заокруглень, якщо відрізки $[X_r, X_{r+1}]$ не дуже великі.

4.2.2. Метод спряжених рівнянь для розв'язування лінійних систем. Цей метод застосовується при розв'язуванні крайових задач вигляду (3), (4) з п. 4.1 для різних стовпців вільних членів $F(x)$. Для цього спочатку задача (3), (4) з п. 4.1 зводиться до вигляду (6), (7) з п. 4.1. Далі припустимо, що матриця $\hat{B}_2$ є значенням деякої змінної матриці $\hat{B}_2(x)$ при $x = b$.

Розглянемо співвідношення

$$\hat{B}_1 Z(a) + \hat{B}_2(x)\,Z(x) = \hat{d}. \quad (19)$$

Визначивши яким-небудь способом $\hat{B}_2(a)$, звідси знайдемо $Z(a)$, і тим самим зведемо крайову задачу (6), (7) з п. 4.1 до задачі Коші з початковими значеннями $Z(a)$. Для відшукання $Z(a)$ треба розв'язати систему лінійних алгебраїчних рівнянь

$$[\hat{B}_1 + \hat{B}_2(a)]\,Z(a) = \hat{d}. \quad (20)$$

Продиференціювавши (19), маємо

$$\frac{d\hat{B}_2(x)}{dx} Z(x) + \hat{B}_2(x) \frac{dZ(x)}{dx} = 0.$$

Врахувавши (6) з п. 4.1, знаходимо

$$\left(\frac{d\hat{B}_2(x)}{dx} + \hat{B}_2(x) M(x)\right) Z(x) = 0.$$

Оскільки це співвідношення має виконуватися в будь-якій точці проміжку $[a, b]$, то

$$\frac{d\hat{B}_2(x)}{dx} + \hat{B}_2(x) M(x) = 0,$$

або

$$\frac{d\hat{B}_2^*(x)}{dx} = -M^*(x) \hat{B}_2^*(x), \quad M^*(x) = \begin{pmatrix} A^*(x) & 0 \\ F^*(x) & 0 \end{pmatrix}.$$

Останню рівність можна записати у вигляді

$$\frac{d\hat{B}_{2i}(x)}{dx} = -M^*(x) \hat{B}_{2i}(x), \quad i = \overline{1, n+1}, \tag{21}$$

де $\hat{B}_{2i}$ — i-й рядок матриці $\hat{B}_2(x)$. Система (21) називається *спряженою* до рівняння (6) з п. 4.1. Із порівняння (7) з п. 4.1 і (19) випливає, що

$$\hat{B}_{2i}(b) = \hat{B}_{2i}, \quad i = \overline{1, n+1}, \tag{22}$$

де $\hat{B}_{2i}$ — i-й рядок матриці $\hat{B}_2$. Отже, для визначення матриці $\hat{B}_2(a)$ треба розв'язати $n+1$ задачу Коші (21), (22). Після цього з (20) знаходимо $Z(a)$ і потім розв'язуємо задачу Коші з цим початковим значенням для системи (6) з п. 4.1.

Якщо треба розв'язати ряд задач вигляду (3), (4) з п. 4.1, які відрізняються лише $F(x)$, то для першої з них доведеться розв'язати $n+1$ задачу Коші (21), (22), а для кожної наступної — лише додатково одну задачу Коші

$$\frac{d\hat{B}_{2,n+1}(x)}{dx} = -(F^*(x), 0) \cdot \hat{B}_{2,i}(x), \quad \hat{B}_{2,n+1}(b) = (0, \ldots, 0, 1),$$

де $(F^*(x), 0)$ — $(n+1)$-вимірний вектор-рядок, $\hat{B}_{2i}(x)$ — $(n+1)$-вимірний вектор-стовпець.

4.2.3. Метод лінеаризації для розв'язування нелінійних крайових задач. Розглянемо двоточкову крайову задачу

$$Y'(x) = F(x, Y(x)); \tag{23}$$

$$\varphi(Y(a), Y(b)) = d. \tag{24}$$

Метод лінеаризації можна застосувати або до задачі (23), (24), або до системи нелінійних алгебраїчних рівнянь

$$\varphi(Y(a), \omega(b, Y(a))) = d, \tag{25}$$

де $\omega(x, Y(a))$ — розв'язок задачі Коші для рівняння (23) з початковою умовою $Y(a)$.

У першому випадку розв'язок подається у вигляді

$$Y(x) = Y^{(m)}(x) + \tilde{U}^{(m)}(x),$$

де $Y^{(m)}(x)$ деяке відоме наближення до шуканого розв'язку, а $\tilde{U}^{(m)}(x)$ — невідома поправка. Припускаючи достатню гладкість функцій F та φ і використовуючи формулу скінченних приростів Лагранжа, дістаємо

$$\frac{d(Y^{(m)}(x) + \tilde{U}^{(m)}(x))}{dx} = F(x, Y^{(m)}(x)) + \Phi_F(x, Z_1^{(m)}, \ldots, Z_n^{(m)}) \tilde{U}^{(m)}(x),$$

$$\varphi(Y^{(m)}(a) + \tilde{U}^{(m)}(a), Y^{(m)}(b) + \tilde{U}^{(m)}(b)) =$$

$$= \varphi(Y^{(m)}(a), Y^{(m)}(b)) + \Phi_0(\omega_1^{(m)}(a), \ldots, \omega_n^{(m)}(a),$$

$$\omega_1^{(m)}(b), \ldots, \omega_n^{(m)}(b)) \tilde{U}^{(m)}(a) + \Phi_1(\omega_1^{(m)}(a), \ldots,$$

$$\omega_n^{(m)}(a), \omega_1^{(m)}(b), \ldots, \omega_m^{(m)}(b)) \tilde{U}^{(m)}(b), \tag{26}$$

де $\Phi_F(x, Z_1^{(m)}, \ldots, Z_n^{(m)}) = \left(\frac{\partial F_l(x, z_{l_1}, \ldots, z_{l_n})}{\partial y_j}\right)_{l,j=\overline{1,n}}$ — матриця Якобі вектор-функції

$$F(x, Y(x)) = F(x, y_1(x), \ldots, y_n(x)),$$

$$Z_l^{(m)} = (z_{l_1}^{(m)}, \ldots, z_{l_n}^{(m)})^T, \Phi_0(\omega_1^{(m)}(a), \ldots, \omega_n^{(m)}(a), \omega_1^{(m)}(b), \ldots, \omega_n^{(m)}(b)) =$$

$$= \left(\frac{\partial \varphi_j(\omega_{j1}^{(m)}(a), \ldots, \omega_{jn}^{(m)}(a), \omega_{j1}^{(m)}(b), \ldots, \omega_{jn}^{(m)}(b))}{\partial y_k(a)}\right)_{j,k=\overline{1,n}},$$

$$\Phi_1(\omega_1^{(m)}(a), \ldots, \omega_n^{(m)}(a), \omega_1^{(m)}(b), \ldots, \omega_n^{(m)}(b)) =$$

$$= \left(\frac{\partial \varphi_j(\omega_{j1}^{(m)}(a), \ldots, \omega_{jn}^{(m)}(a), \omega_{j1}^{(m)}(b), \ldots, \omega_{jn}^{(m)}(b))}{\partial y_k(b)}\right)_{j,k=\overline{1,n}}$$

— матриці Якобі вектор-функції $\varphi(Y(a), Y(b))$, причому

$$\|Z_k^{(m)}(x) - Y^{(m)}(x)\| \leqslant \|\tilde{U}^{(m)}(x)\|,$$

$$\|\omega_k^{(m)}(a) - Y^{(m)}(a)\| \leqslant \|\tilde{U}^{(m)}(a)\|,$$

$$\|\omega_k^{(m)}(b) - Y^{(m)}(b)\| \leqslant \|\tilde{U}^{(m)}(b)\|, \quad k = \overline{1, n}.$$

Якщо припустити, що $\| \tilde{U}^{(m)}(x) \|$, $\| \tilde{U}^{(m)}(a) \|$ і $\| \tilde{U}^{(m)}(b) \|$ малі, то останні нерівності означають близькість невідомих векторів $Z_k^{(m)}(x)$, $\omega_k^{(m)}(a)$, $\omega_k^{(m)}(b)$ до відомих векторів $Y^{(m)}(x)$, $Y^{(m)}(a)$ і $Y^{(m)}(b)$ відповідно. Тому, замінивши в якобіанах з (26) невідомі вектори відомими, дістанемо лінійну крайову задачу для визначення вектора $U^{(m)}(x)$, який буде наближати $\tilde{U}^{(m)}(x)$:

$$\frac{dU^{(m)}(x)}{dx} = \Phi_F(x, Y^{(m)}(x))\, U^{(m)}(x) + F(x, Y^{(m)}(x)) - \frac{\partial Y^{(m)}(x)}{dx}; \quad (27)$$

$$\Phi_0(Y^{(m)}(a), Y^{(m)}(b))\, U^{(m)}(a) + \Phi_1(Y^{(m)}(a), Y^{(m)}(b))\, U^{(m)}(b) = d - \varphi(Y^{(m)}(a), Y^{(m)}(b)). \quad (28)$$

Розв'язок лінійної задачі (27), (28) можна знайти, наприклад, викладеними вище методами зведення її до розв'язування задачі Коші. Після цього можна знайти наступне наближення до розв'язку задачі (23) і (24) за формулою $Y^{(m+1)}(x) = Y^{(m)}(x) + U^{(m)}(x)$, потім з (27) і (28), замінивши $Y^{(m)}$ на $Y^{(m+1)}$, знайти $U^{(m+1)}$ і т. д., доти, доки не здобудемо розв'язок вихідної задачі з потрібною точністю. Питання про збіжність цього методу ми не розглядаємо.

Вправа 2. Записати обчислювальні формули методу лінеаризації для розв'язування крайової задачі

$$\frac{d^2c}{dx^2} + \frac{1}{2}\,\frac{\left(\frac{dc}{dx}\right)^2}{1 - \frac{c}{2}} = 0, \quad c(0) = c_0, \quad c(\delta) = 0,$$

яка описує зміну концентрації речовини в процесі хімічної реакції з деяким каталізатором.

Аналогічно можна лінеаризувати і систему нелінійних рівнянь (25). Нехай відоме деяке наближення $Y^{(m)}(a)$ для $Y(a)$, а $\tilde{U}^{(m)}(a)$ — невідома поправка, тобто $Y(a) = Y^{(m)}(a) + \tilde{U}^{(m)}(a)$. Тоді, якщо норма $\| \tilde{U}^{(m)}(a) \|$ досить мала, можна покласти

$$\Psi(Y(a)) \equiv \varphi(Y(a), \omega(b, Y(a))) \approx \approx \Psi(Y^{(m)}(a)) + \tilde{\Phi}_\Psi(Y^{(m)}(a))\, \tilde{U}^{(m)}(a), \quad (29)$$

де $\tilde{\Phi}_\Psi = \left(\frac{\partial \psi_k(Y^{(m)}(a))}{\partial y_j(a)}\right)_{k,j=\overline{1,n}}$ — матриця Якобі вектор-функції $\Psi(Y(a))$. У викладеному вище методі утворювалась матриця Якобі від невідомих вектор-функцій $F(x, Y(x))$ і $\varphi(Y(a), Y(b))$. А при лінеаризації

рівняння (25) елементи матриці Φ_ψ виявляються невідомими, бо невідомою є залежність $\omega(b, Y(a))$. У цьому разі можна діяти таким чином. Похідну змінюємо різницевими виразами

$$\frac{\partial \psi_k(Y^{(m)}(a))}{\partial y_j(a)} \approx \frac{\psi_k(Y^{(m)}(a) + he_j) - \psi_k(Y^{(m)}(a))}{h}, \quad (30)$$

де $e_j = (\delta_{ij})_{i=\overline{1,n}}$ — j-й координатний орт; h — досить мале число. Щоб знайти $\psi_k(Y^{(m)}(a))$ і $\psi_k(Y^{(m)}(a) + he_j)$, розв'язуємо задачі Коші

$$\frac{dY_0(x)}{dx} = F(x, Y_0(x)), \quad Y_0(a) = Y^{(m)}(a),$$

$$\frac{dY_j(x)}{dx} = F(x, Y_j(x)), \quad Y_j(a) = Y^{(m)}(a) + he_j, \quad j = \overline{1, n},$$

в результаті чого знаходимо $\omega(b, Y^{(m)}(a)) \equiv Y_0(b)$, $\omega(b, Y^{(m)}(a) + he_j) \equiv Y_j(b)$, $j = \overline{1, n}$. Тим самим визначаємо величини

$$\psi_k(Y^{(m)}(a)) = \varphi_k(Y^{(m)}(a), \omega(b, Y^{(m)}(a))), \quad k = \overline{1, n};$$

$$\psi_k(Y^{(m)}(a) + he_j) = \varphi_k(Y^{(m)}(a) + he_j, \omega(b, Y^{(m)}(a) + he_j)), \quad j, k = \overline{1, n},$$

і потім обчислюємо матрицю $\Phi_\Psi(Y^{(m)}(a))$ з елементами вигляду (30):

$$\Phi_\Psi(Y^{(m)}(a)) = \left(\frac{\psi_k(Y^{(m)}(a) + he_j) - \psi_k(Y^{(m)}(a))}{h}\right)_{k,j=\overline{1,n}}.$$

Ця матриця є наближенням для матриці $\tilde{\Phi}_\Psi(Y^{(m)}(a))$. Після цього можна визначити наближення $U^{(m)}(a)$ для $\tilde{U}^{(m)}(a)$, розв'язуючи систему рівнянь

$$\Phi_\Psi(Y^*(a))\, U^{(m)}(a) = d - \varphi(Y^{(m)}(a), Y_0(b)),$$

яка в силу (29) є «наближенням» системи (25). Сума $Y^{(m)}(a) + U^{(m)}(a) = Y^{(m+1)}(a)$ є нове наближення для $Y(a)$, після цього весь процес можна повторити до утворення задовільного наближення.

4.2.4. Метод продовження розв'язку нелінійної крайової задачі за параметром. Цей метод дозволяє звести розв'язування нелінійної крайової задачі (23), (24) до розв'язування послідовності лінійних задач Коші. Спочатку вибирають довільний вектор Y_0 і розв'язують задачу Коші для рівняння (23) з початковим значенням Y_0. Позначимо її розв'язок через $\hat{Y}(x)$. За цим розв'язком можна знайти $d_0 = \varphi(\hat{Y}(a), \hat{Y}(b)) = \varphi(Y_0, \hat{Y}(b))$. Якщо для кожного $\lambda \in [0,1]$ існує розв'язок

задачі

$$\tilde{Y}'(x, \lambda) = F(x, \tilde{Y}); \tag{31}$$

$$\varphi(\tilde{Y}(a, \lambda), \tilde{Y}(b, \lambda)) = \lambda(d - d_0) + d_0, \tag{32}$$

який неперервно залежить від λ, то $\tilde{Y}(x, 1) = Y(x)$, де $Y(x)$ — розв'язок задачі (23), (24), а $\tilde{Y}(x, 0) = \hat{Y}(x)$. Таким чином, вихідна задача (23), (24) зводиться до визначення $\tilde{Y}(x, 1)$.

Диференціюючи (31), (32) по λ, маємо

$$\frac{d}{dx}\left(\frac{d\tilde{Y}}{d\lambda}\right) = \sum_{j=1}^{n} \frac{\partial F}{\partial \tilde{y}_j} \frac{\partial \tilde{y}_j}{\partial \lambda}, \tag{33}$$

$$\sum_{j=1}^{n} \frac{\partial \varphi}{\partial \tilde{y}_j(a, \lambda)} \frac{\partial \tilde{y}_j(a, \lambda)}{\partial \lambda} + \sum_{j=1}^{n} \frac{\partial \varphi}{\partial \tilde{y}_j(b, \lambda)} \frac{\partial \tilde{y}_j(b, \lambda)}{\partial \lambda} = d - d_0. \tag{34}$$

Введемо позначення:

$$\frac{\partial \tilde{Y}(x, \lambda)}{\partial \lambda} = U(x, \lambda), \quad U(a, \lambda) = U_a(\lambda); \quad U(b, \lambda) = U_b(\lambda);$$

$$\Phi_F = \left(\frac{\partial F}{\partial \tilde{y}_1}, \ldots, \frac{\partial F}{\partial \tilde{y}_n}\right),$$

$$\Phi_a = \left(\frac{\partial \varphi}{\partial \tilde{y}_1(a, \lambda)}, \ldots, \frac{\partial \varphi}{\partial \tilde{y}_n(a, \lambda)}\right),$$

$$\Phi_b = \left(\frac{\partial \varphi}{\partial \tilde{y}_1(b, \lambda)}, \ldots, \frac{\partial \varphi}{\partial \tilde{y}_n(b, \lambda)}\right).$$

Тоді співвідношення (33), (34) запишемо таким чином:

$$\frac{dU(x, \lambda)}{dx} = \Phi_F(x, \tilde{Y}(x, \lambda)) U(x, \lambda); \tag{35}$$

$$\Phi_a(\tilde{Y}(a, \lambda), \tilde{Y}(b, \lambda)) U_a(\lambda) + \Phi_b(\tilde{Y}(a, \lambda), \tilde{Y}(b, \lambda)) U_b(\lambda) = \tag{36}$$

$$= d - d_0.$$

Задача (35), (36) для $U(x, \lambda)$ є лінійною крайовою задачею. Для визначення $\tilde{Y}(x, \lambda)$ маємо задачу Коші

$$\frac{d\tilde{Y}(x, \lambda)}{d\lambda} = U(x, \lambda), \quad \tilde{Y}(x, 0) = \hat{Y}(x). \tag{37}$$

Як правило, для апроксимації задачі (37) застосовують метод Ейлера:

$$\tilde{Y}(x, \lambda + \Delta\lambda) = \tilde{Y}(x, \lambda) + \Delta\lambda U(x, \lambda), \tag{38}$$

$$\tilde{Y}(x, 0) = \hat{Y}(x).$$

Розрахунки виконують таким чином: 1) за $\tilde{Y}(x, 0)$ з (35), (36) визначають $U(x, 0)$; 2) за $\tilde{Y}(x, 0)$, $U(x, 0)$ з явних формул (38) знаходять $\tilde{Y}(x, \Delta\lambda)$; 3) за $\tilde{Y}(x, \Delta\lambda)$ із (35), (36) визначають $U(x, \Delta\lambda)$ і т. д.

Всі розглянуті методи приводять до багаторазового розв'язування задач Коші, тому їхня точність в основному визначається точністю розв'язування задач Коші. Кількість обчислювальної роботи при застосуванні цих методів досить значна. Тому на практиці часто використовують сіткові методи, які зводять лінійну або нелінійну крайову задачу до розв'язування відповідно системи лінійних чи нелінійних алгебраїчних рівнянь, які ми розглянемо далі.

Вправа 3. Записати обчислювальні формули методу продовження розв'язку за параметром для нелінійної крайової задачі

$$1 + (u')^2 + uu'' = 0, \quad x \in (0, 1),$$

$$u(0) = u_0, \quad u^2(1) + u^2(1) = u_1$$

(u_0, u_1 — задані числа), яка зустрічається в нелінійній оптиці.

4.3. Метод сіток розв'язування крайових задач для звичайних диференціальних рівнянь

Існує багато сіткових схем, які апроксимують одну і ту саму диференціальну задачу. Серед таких схем нам треба вибрати ту, яка на порівняно грубих сітках (а не тільки при достатньо малих кроках сітки, коли розмірність задачі може бути дуже великою) гарантувала б бажану точність для знайденого наближення розв'язку. Для цього сіткова схема має відбивати основні властивості початкового операторного рівняння: самоспряженість, додатну визначеність, сіткові аналоги типових апріорних оцінок для диференціального оператора тощо. Оскільки ці властивості диференціальної задачі в математичній фізиці є наслідками певних фізичних законів, то можна сказати, що сіткова схема має бути такою, щоб для її розв'язування на сітці виконувались основні (які визначають задачу) фізичні закони.

Розглянемо методи побудови сіткових схем, які найчастіше застосовуються на практиці і апроксимують крайові задачі для звичайних диференціальних рівнянь. Оскільки ці схеми часто виражаються через різниці шуканих функцій, то традиційно їх називають *різницевими схемами*.

4.3.1. Метод формальної заміни похідних скінченнорізницевими відношеннями. У цьому методі похідні, які входять до диференціального рівняння, і граничні умови замінюються використанням формул чисельного диференціювання (див. п. 2.9 з ч. 1).

Якщо $u \in C^3[a, b]$, то в усіх внутрішніх вузлах сітки $\bar{\omega}_h = \{x_i = ih;\ i = \overline{0, N};\ h = (b-a)/N\}$ (тобто в точках сітки $\omega_h = \{x_i = ih;\ i = \overline{1, N-1};\ h = (b-a)/N\}$) першу похідну $u'(x)$ можна апроксимувати правою різницевою похідною

$$u_{x,i} = \frac{u_{i+1} - u_i}{h} = \frac{u(x_{i+1}) - u(x_i)}{h}, \quad i = \overline{1, N-1} \tag{1}$$

(у безіндексних позначеннях $u_x(x) = (u(x+h) - u(x))/h$, $x \in \omega_h$), лівою різницевою похідною

$$u_{x,i} = \frac{u_i - u_{i-1}}{h}, \quad i = \overline{1, N-1}\ (u_{\bar{x}}(x) = (u(x) - u(x-h))/h), \tag{2}$$
$$x \in \omega_h$$

або, використовуючи лінійну комбінацію (1), (2),— виразом

$$u^{(\sigma)}_{x,i} = \sigma u_{x,i} + (1-\sigma) u_{\bar{x},i}, \quad i = \overline{1, N-1}$$
$$(u^{(\sigma)}_x(x) = \sigma u_x(x) + (1-\sigma) u_{\bar{x}}(x), \quad x \in \omega_h), \tag{3}$$

де σ — будь-яке дійсне число. При $\sigma = \frac{1}{2}$ дістаємо апроксимацію, яка називається центральною різницевою похідною

$$u^{(1/2)}_{x,i} = u_{\overset{\circ}{x},i} = (u_{i+1} - u_{i-1})/(2h)\ (u_{\overset{\circ}{x}}(x) = (u(x+h) -$$
$$- u(x-h))/(2h), \quad x \in \omega_h). \tag{4}$$

Формули (27), (28) п. 2.9 з ч.1 дають також і апроксимації:

$$u_{x,i+0} = \frac{-3u_i + 4u_{i+1} - u_{i+2}}{2h}; \tag{5}$$

$$u_{x,i-0} = \frac{3u_i - 4u_{i-1} + u_{i-2}}{2h}, \tag{6}$$

причому наведені вище формули далеко не вичерпують множину можливих апроксимацій першої похідної.

Вправа 1. За допомогою розвинення функції $u(x)$ в ряд Тейлора в околі точки $x = x_i$ дістати такі формули для локальної похибки апроксимації $\psi_h(x)$:

$$\psi_h(x_i) = u^{(\sigma)}_{x,i} - u'(x_i) = \left(\sigma - \frac{1}{2}\right) hu''(x_i) + O(h^2);$$

$$\psi_h(x_i) = u_{\overset{\circ}{x},i} - u'(x_i) = \frac{h^2}{6} u^{(3)}(x_i) + O(h^4) = O(h^2);$$

$$\psi_h(x_i) = u_{x,i+0} - u'(x_i) = -\frac{h^2}{3} u^{(3)}(x_i) + O(h^3) = O(h^2); \tag{7}$$

$$\psi_h(x_i) = u_{x,i-0} - u'(x_i) = \frac{h^2}{3} u^{(3)}(x_i) + O(h^3) = O(h^2).$$

При якій гладкості функції вони мають місце?

В п. 2.9 з ч.1 показано також, як дістати сіткові апроксимації для похідних вищих порядків та наведено порядок похибки цих апроксимацій, який може бути вищим для рівномірних сіток та спеціального розміщення вузлів і нижчим для нерівномірних сіток.

Наприклад, для апроксимації другої похідної $u''_{x^2}(x)$, $u \in C^4[a, b]$, на нерівномірній сітці можна використовувати формули:

$$u_{\bar{x}\hat{x},i} = \frac{1}{\hbar_{i+1}} \left[\frac{u_{i+1} - u_i}{h_{i+1}} - \frac{u_i - u_{i-1}}{h_i} \right], \quad \hbar_{i+1} = \frac{1}{2}(h_i + h_{i+1}), \tag{8}$$

$$u_{\bar{x}\hat{x},i} = \frac{1}{h_i} \left[\frac{u_{i+1/2} - u_{i-1/2}}{\hbar_{i+1}} - \frac{u_{i-1/2} - u_{i-3/2}}{\hbar_i} \right], \tag{9}$$

де $u_{i\pm1/2} = u\left(x_i \pm \frac{h_{i+1}}{2}\right)$, $\hbar_i = x_i - x_{i-1}$.

Вправа 2. За допомогою розвинення в ряд Тейлора утворити наступні вирази для локальної похибки апроксимації:

$$\psi_h(x_i) = u_{\bar{x}\hat{x},i} - u''(x_i) = \frac{h_{i+1} - h_i}{3} u^{(3)}(x_i) + O(h^2); \tag{10}$$

$$\psi_h(x_i) = u_{\bar{x}\hat{x},i} - u''(x_i) = \frac{h_{i+1} - 2h_i + h_{i-1}}{4h_i} u''(x_i) + O(h), \tag{11}$$

де $h = \max h_i$.

Якщо сітка рівномірна і $u \in C^4[a, b]$, то за допомогою формули Тейлора неважко переконатися, що

$$u_{\bar{x}\hat{x},i} = u_{\bar{x}x,i} = (u_{i+1} - 2u_i + u_{i-1})/h^2 = u''(x_i) + \frac{h^2}{24}(u^{(4)}(\xi) +$$
$$+ u^{(4)}(\xi_1)) = u''(x_i) + \frac{h^2}{12} u^{(4)}(\eta),$$

де точки ξ, ξ_1, η розміщені на відрізку $[x_i - h, x_i + h]$. Із співвідношення (11) випливає, що локальна апроксимація в точці x_i на нерівномірній сітці відсутня, тому що

$$c_i = (h_{i+1} - 2h_i + h_{i-1})/(4h_i) = O(1).$$

Тому іноді нерівномірну сітку вибирають так, щоб c_i були невеликими. Цього можна досягти, якщо, наприклад, нерівномірну сітку будувати так, щоб $h_{i+1} = qh_i$, де q — стала, близька до 1. Тоді $c_i =$

$= (q - 2 + 1/q)/4 = (q-1)^2/(4q) \to 0$ при $q \to 1$. Виходячи з формул (1) — (11), можна будувати сіткові апроксимації і для більш складних диференціальних виразів. Розглянемо, наприклад, такий диференціальний вираз:

$$Lu = \frac{d}{dx}\left(k(x)\frac{du}{dx}\right) = k'(x)\,u'(x) + k(x)\,u''(x), \qquad (12)$$

де $u \in C^3[0, 1]$. Вибравши для наближення перших похідних відповідну формулу, на сітці $\omega_h = \{x_i = ih \mid i = \overline{1, N-1},\ h = 1/N\}$, можна розглянути таку апроксимацію:

$$L_h u_h = (a u_{\bar{x}})_{x,i} = h^{-1}\left(a_{i+1}u_{\bar{x},i+1} - a_i u_{\bar{x},i}\right) = h^{-1}\left(a_{i+1}u_{x,i} - a_i u_{\bar{x},i}\right).$$

Враховуючи (7) і (12), для локальної похибки апроксимації дістаємо

$$L_h u_h - (Lu)_h = \frac{a_{i+1}-a_i}{h}\,u'(x_i) + \frac{a_{i+1}+a_i}{2}\,u''(x_i) + O(h^2) -$$

$$- k'(x_i)\,u'(x_i) - k(x_i)\,u''(x_i) = \left(\frac{a_{i+1}-a_i}{h} - k'(\tilde{x}_i)\right)u'(x_i) +$$

$$+ \left(\frac{a_{i+1}+a_i}{2} - k(x_i)\right)u''(x_i) + \frac{a_{i+1}-a_i}{6}\,hu^{(3)}(x_i) + O(h^2),$$

де $(Lu)_h$ — значення виразу (12) в точці x_i сітки. Звідси випливає, що для утворення другого порядку апроксимації за умови $u \in C^3[0, 1]$ треба вибрати a_i так, щоб

$$\frac{a_{i+1}+a_i}{2} - k(x_i) = O(h^2),$$

$$\frac{a_{i+1}-a_i}{h} - k'(x_i) = O(h^2),$$

$$\frac{a_{i+1}-a_i}{6}\,hu^{(3)}(x_i) = O(h^2).$$

Для виконання цих співвідношень достатньо, щоб $k(x) \in C^3[0, 1]$ та $a_i = k(x_i - 0{,}5h)$, або $a_i = (k_i + k_{i-1})/2$, або $a_i = (k_i^{-1} + k_{i-1}^{-1})^{-1}$ (перевірте це за допомогою розвинення за формулою Тейлора).

Приклад 1. Методом заміни похідних різницевими відношеннями побудуємо різницеву схему для такої граничної задачі:

$$-u''(x) + g(x)\,u(x) = \varphi(x),$$
$$-u'(a) + \alpha_0 u(a) = \gamma_0, \qquad (13)$$
$$u'(b) + \alpha_1 u(b) = \gamma_1,$$

де $\alpha_0 \geqslant 0$, $\alpha_1 > 0$, $g(x) \geqslant 0\ \forall\, x \in [a, b]$. Введемо сітку $\bar{\omega}_h = \{x_i : x_i = a_i + ih,\ i = \overline{0, N},\ h = (b-a)/N\} = \omega_h \cup \{a, b\}$. Використовуючи формули (8) на рівномірній сітці для апроксимації другої похідної, (5) — для апроксимації першої похідної в граничній умові в точці a та (6) — для апроксимації першої похідної в граничній умові в точці b, дістаємо схему

$$-y_{\bar{x}x,i} + g_i y_i = \varphi_i, \quad i = \overline{1, N-1},$$
$$-\frac{-3y_0 + 4y_1 - y_2}{2h} + \alpha_0 y_0 = \gamma_0, \qquad (14)$$
$$\frac{3y_N - 4y_{N-1} + y_{N-2}}{2h} + \alpha_1 y_N = \gamma_1.$$

Для апроксимації з другим порядком граничних умов можна використати формули (7) та рівняння (13). Тоді

$$-u'(a) = -u_{x,0} + \frac{h}{2}\,u''(a) + O(h^2) = -u_{x,0} + \frac{h}{2}\,(g(a)\,y(a) - \varphi(a)) +$$

$$+ O(h^2) = -u_{x,0} + \frac{h}{2}\,(g_0 u_0 - \varphi_0) + O(h^2).$$

Аналогічно

$$u'(b) = u_{\bar{x},N} + \frac{h}{2}\,(g_N u_N - \varphi_N) + O(h^2),$$

і дістаємо схему

$$-y_{\bar{x}x,i} + g_i y_i = \varphi_i, \quad i = \overline{1, N-1},$$
$$-y_{x,0} + \frac{h}{2}\,(g_0 y_0 - \varphi_0) + \alpha_0 y_0 = \gamma_0, \qquad (15)$$
$$y_{\bar{x},N} + \frac{h}{2}\,(g_N y_N - \varphi_N) + \alpha_1 y_N = \gamma_1,$$

або в безіндексних позначеннях

$$-y_{\bar{x}x}(x) + g(x)\,y(x) = \varphi(x), \quad x \in \omega_h,$$
$$-y_x(a) + \frac{h}{2}\,(g(a)\,y(a) - \varphi(a)) + \alpha_0 y(a) = \gamma_0, \qquad (16)$$
$$y_{\bar{x}}(b) + \frac{h}{2}\,(g(b)\,y(b) - \varphi(b)) + \alpha_1 y(b) = \gamma_1.$$

Із попереднього викладу випливає, що у випадку $u \in C^4[a, b]$ схеми (14) і (16) мають другий порядок локальної апроксимації (в точці). Для апроксимації похідних, що входять до граничних умов, діють аналогічно. Вводять сітку

$$\bar{\omega}_h^{(1)} = \left\{x_i = x_0 + ih, \quad i = \overline{0, N}, \quad x_0 = a - \frac{h}{2},\right.$$
$$\left. h = (b-a)/(N-1)\right\} \equiv \omega_h^{(1)} \cup \left\{a - \frac{h}{2},\ b + \frac{h}{2}\right\},$$

граничні вузли $a - \frac{h}{2}$ та $b + \frac{h}{2}$ якої не належать відрізку $[a, b]$, на якому шукають розв'язок початкової задачі (13). Ці вузли називають

фіктивними. За допомогою формули Тейлора дістаємо, що у випадку, коли $u(x) \in C^3[a, b]$ і її можна продовжити за межі відрізка $[a, b]$ із збереженням класу гладкості, то на сітці $\bar{\omega}_h^{(1)}$

$$u'(a) = u_{x,0} - \frac{h^2}{48} u^{(3)}(\xi) = \frac{u\left(a + \frac{h}{2}\right) - u\left(a - \frac{h}{2}\right)}{h} - \frac{h^2}{48} u^{(3)}(\xi),$$

де ξ — деяка точка з відрізка $\left[a - \frac{h}{2}, a + \frac{h}{2}\right]$.

Крім того, якщо $u \in C^2\left[a - \frac{h}{2}, a + \frac{h}{2}\right]$, то

$$u(a) = \frac{u\left(a - \frac{h}{2}\right) + u\left(a + \frac{h}{2}\right)}{2} - \frac{h^2}{4} u''(\xi), \quad \xi \in \left[a - \frac{h}{2}, a + \frac{h}{2}\right].$$

Таким чином, за припущення $u(x) \in C^4\left[a - \frac{h}{2}, a + \frac{h}{2}\right]$ маємо ще одну схему другого порядку локальної апроксимації на сітці $\bar{\omega}_h^{(1)}$

$$-y_{\bar{x}x,i} + g_i y_i = \varphi_i, \quad i = \overline{1, N-1},$$

$$-y_{x,0} + \alpha_0 \frac{y_0 + y_1}{2} = \gamma_0, \tag{17}$$

$$y_{\bar{x},N} + \alpha_1 \frac{y_N + y_{N-1}}{2} = \gamma_1.$$

Різницеві схеми (14) — (17) являють собою системи лінійних алгебраїчних рівнянь порядку $N+1$ з невідомими y_i, $i = \overline{0, N}$. Їх можна записати у вигляді

$$AY = F, \tag{18}$$

де $Y = (y_0, \ldots, y_N)^T$ — невідомий вектор; F — заданий вектор; A — відома матриця. Для схеми (14) F та A мають вигляд

$$F = \left(\frac{2\gamma_0}{h}, \varphi_1, \varphi_2, \ldots, \varphi_{N-1}, \frac{2\gamma_1}{h}\right)^T, \tag{19}$$

$$A = \frac{1}{h^2} \times$$

$$\times \begin{bmatrix} 3+2\alpha_0 h & -4 & 1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2+g_1h^2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2+g_2h^2 & -1 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & -1 & 2+g_{N-1}h^2 & -1 \\ 0 & 0 & 0 & 0 & \ldots & 1 & -4 & 3+2\alpha_1 h \end{bmatrix},$$

для схеми (15)

$$F = \left(\frac{\gamma_0}{h} + \frac{\varphi_0}{2}, \varphi_1, \varphi_2, \ldots, \varphi_{N-1}, \frac{\gamma_1}{h} + \frac{\varphi_N}{2}\right), \tag{20}$$

$$A = \frac{1}{h^2} \begin{bmatrix} 1 + \frac{h^2 g_0}{2} + \alpha_0 h & -1 & 0 & 0 & \ldots & & \\ -1 & 2+g_1h^2 & -1 & 0 & \ldots & & \\ 0 & -1 & 2+g_2h^2 & -1 & \ldots & & \\ \ldots & \ldots & \ldots & \ldots & \ldots & & \\ 0 & 0 & 0 & 0 & \ldots & & \\ 0 & 0 & 0 & 0 & \ldots & & \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ -1 & 2+g_{N-1}h^2 & -1 \\ 0 & -1 & 1 + \frac{h^2 g_N}{2} + \alpha_1 h \end{bmatrix}$$

і, нарешті, для схеми (17)

$$F = \left(\frac{2\gamma_0}{h}, \varphi_1, \varphi_2, \ldots, \varphi_{N-1}, \frac{2\gamma_1}{h}\right),$$

$$A = \frac{1}{h^2} \times$$

$$\times \begin{bmatrix} 2+h\alpha_0 & -2+h\alpha_0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2+g_1h^2 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2+g_2h^2 & -1 & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & -1 & 2+g_{N-1}h^2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -2+h\alpha_1 & 2+h\alpha_1 \end{bmatrix}. \tag{21}$$

Таким чином, за допомогою методу сіток гранична задача (13) зводиться до розв'язування системи лінійних алгебраїчних рівнянь виду (18), причому для побудови матриці A і вектора F було використано метод заміни похідних різницевими відношеннями. Оскільки для апроксимації однієї й тієї самої похідної існує багато різних формул, то за допомогою цього методу можна дістати й інші схеми, які не збігаються зі схемами (14) — (17). Якій же схемі віддати перевагу?

Відповідь на це запитання не однозначна і залежить від того, яких властивостей різницевої схеми ми хочемо добитися.

Зауважимо, що оператор граничної задачі (13) у відповідно визначеному просторі зі скалярним добутком є симетричним і додатно визначеним. Дійсно, задачу (13) можна записати у вигляді операторного

рівняння

$$Lu = f,$$

у випад
ка $[a, b]$ (22)

де

$$Lu = \begin{cases} -u''(x) + g(x)\,u(x), & x \in (a, b), \\ -u'(a) + \alpha_0 u(a), & x = a, \\ u'(b) + \alpha_1 u(b), & x = b, \end{cases}$$

$$f = f(x) = \begin{cases} \varphi(x), & x \in (a, b), \\ \gamma_0, & x = a, \\ \gamma_1, & x = b, \end{cases}$$

причому областю визначення $D(L)$ оператора L є простір (
, а
область значень $R(L)$ лежить у просторі $Q^0[a, b]$ кусково не
них
на $[a, b]$ функцій. Якщо у $Q^0[a, b]$ ввести скалярний добут
формулою

$$(u, v) = \int_a^b u(x)\,v(x)\,dx + u(a)\,v(a) + u(b)\,v(b),$$

мае
а сітці

то за допомогою інтегрування частинами неважко знайти, і

$$(Lu, v) = \alpha_1 u(b)\,v(b) + \alpha_0 u(a)\,v(a) + \int_a^b u'(x)\,v'(x)\,dx$$

$$+ \int_a^b g(x)\,u(x)\,v(x)\,dx = (u, Lv),$$

х алгеб
можна
цю
сподіватися, що матриця (оператор) різницевої схеми, яка а
ує
задачу (13), також є симетричною. Із цих міркувань слід з
ися
на схемі (15).

У п. 2.2.3 доведено, що при $\alpha_0, \alpha_1 > 0$ матриця (оператор; A це-
вої схеми (15) є додатно визначеною у просторі сіткових функ
ка-
лярним добутком

$$(y, v) = (y, v)_{\bar{\omega}_h} = \sum_{i=0}^{N} h y_i v_i.$$

Оскільки матриця A тридіагональна, то для розв'язуван
ачі
(15) можна застосувати метод прогонки.

4.3.2. Метод невизначених коефіцієнтів. За цим методо
ет-
ний аналог диференціального оператора Lu у кожній точці
кої
сітки шукаємо у вигляді

$$L_h u_h|_{x=x_0} = \sum_{k=0}^{q} a_k u(x_k), \qquad (23)$$

0
0
0
. . . .
-1
$+ 2\alpha_1 h$

де a_k — поки що невизначені коефіцієнти. Множина точок x_k, $k = \overline{0, q}$ називається *шаблоном*, а точка x_0 — *центром шаблона*. Невизначені коефіцієнти знаходять з умови найвищого відносно h порядку апроксимації в точці x_0 диференціального оператора різницевим оператором (23). Це еквівалентно тому, що значення диференціального і різницевого операторів мають збігатися на многочленах до максимально високого степеня m. При цьому для коефіцієнтів a_k утворюється система рівнянь і якщо вона має розв'язок, то

$$L_h u_h|_{x=x_0} - (Lu)_h|_{x=x_0} = \sum_{k=0}^{q} a_k u(x_k) - Lu(x_0) = O(h^m).$$

Розглянемо, наприклад, диференціальний оператор n-го порядку вигляду

$$Lu(x) = \sum_{k=0}^{n} p_k(x)\,u^{(k)}(u),$$

де $p_k(x)$ — задані неперервні функції. Припустимо, що $u(x) \in C^{n+m+1}(\Omega)$ і виберемо шаблон вигляду

$$x_k = x_i + \alpha_k h, \quad k = \overline{1, q},$$

з центром у точці x_i, який складається з q різних точок. Розглянемо різницевий оператор вигляду

$$L_h u_h = \sum_{k=1}^{q} a_k u(x_i + \alpha_k h) \qquad (24)$$

і покажемо, що при $q \geqslant n + m + 1$ і різних α_k (заданих) можна знайти a_k такі, що

$$L_h u_h - (Lu)_h = \sum_{k=0}^{q} a_k u(x_i + \alpha_k h) - Lu(x_i) = O(h^{m+1}), \qquad (25)$$

а точніше

$$L_h u_h - (Lu)_h = R(u^{(n+m+1)}(x_i + \theta_k \alpha_k h), \quad a_1, a_2, \ldots, a_q, \alpha_1, \ldots, \alpha_q),$$

$$|R| \leqslant h^{m+1} \frac{\max\limits_{x \in \Omega_1} |u^{(m+n+1)}(x)|}{(m+n+1)!} |P_n(h)|, \qquad (26)$$

де Ω_1 — відрізок, який містить всі точки шаблона $x_i + \underset{k=\overline{1,q}}{\alpha_k h}$; $P_n(h)$ — многочлен від h не вище n-го степеня, що не залежить від u.

Дійсно, розвинемо $u(x_i + \alpha_k h)$ із (25) у ряд Тейлора в околі точки x_i із залишковим членом у формі Лагранжа:

$$L_h u_h(x_i) = a_1 \sum_{j=0}^{n+m} \frac{(h\alpha_1)^j}{j!} u^{(j)}(x_i) + a_1 \frac{(\alpha_1 h)^{n+m+1}}{(n+m+1)!} u^{(n+m+1)}(x_i + \theta_1 \alpha_1 h) +$$

$$+ a_2 \sum_{j=0}^{n+m} \frac{(h\alpha_2)^j}{j!} u^{(j)}(x_i) + a_2 \frac{(\alpha_2 h)^{n+m+1}}{(n+m+1)!} u^{(n+m+1)}(x_i + \theta_2 \alpha_2 h) +$$

$$+ \cdots + a_q \sum_{j=0}^{n+m} \frac{(h\alpha_q)^j}{j!} u^{(j)}(x_i) + a_q \frac{(\alpha_q h)^{(n+m+1)}}{(n+m+1)!} u^{(n+m+1)}(x_i + \theta_q \alpha_q h) =$$

$$= u(x_i) \sum_{k=1}^{q} a_k + \frac{h}{1!} u'(x_i) \sum_{k=1}^{q} \alpha_k a_k + \cdots +$$

$$+ \frac{h^{n+m}}{(n+m)!} u^{(n+m)}(x_i) \sum_{k=1}^{q} \alpha_k^{n+m} a_k + R, \quad \theta_k \in (0, 1), \quad k = \overline{1, q}; \quad (27)$$

$$R = \frac{h^{m+1}}{(n+m+1)!} \sum_{k=1}^{q} h^n \alpha_k^{n+m+1} a_k u^{(n+m+1)}(x_i + \theta_k \alpha_k h).$$

Співвідношення (25), очевидно, справедливе, якщо a_k задовольняють таку систему рівнянь:

$$\begin{cases} \dfrac{h^l}{l!} \displaystyle\sum_{k=1}^{q} \alpha_k^l a_k = p_l(x_i), & l = \overline{0, n}, \\ \displaystyle\sum_{k=1}^{q} \alpha_k^l a_k = 0, & l = \overline{n+1, n+m}. \end{cases} \quad (28)$$

Система рівнянь (28) завжди має розв'язок, бо її визначник (визначник Вандермонда) відмінний від нуля.

Метод невизначених коефіцієнтів використовується для того, щоб серед апроксимацій даного порядку знайти ту, що задовольняє певні умови. Цей метод також використовується для побудови різницевих схем підвищеного порядку точності.

4.3.8. Інтегро-інтерполяційний метод. Цей метод грунтується на інтегральних тотожностях, які випливають з тих самих фізичних законів, що і відповідне диференціальне рівняння, наприклад, законів збереження кількості теплоти, руху, маси, імпульсу, енергії. Ці інтегральні тотожності можуть визначати узагальнений розв'язок диференціального рівняння або можуть бути побудовані за допомогою інтегрування диференціального рівняння за кількома елементарними відрізками сітки і т. д.

При побудові сіткових схем інтегро-інтерполяційним методом застосовують інтерполяцію підінтегральних виразів. Тому, змінюючи вигляд інтерполяційних многочленів, можна діставати різні схеми.

Сіткові схеми, які на сітці відображають закони збереження, називають *консервативними схемами*. Наведемо приклад побудови консервативної різницевої схеми, в якому використовується закон збереження кількості теплоти.

Приклад 2. Інтегро-інтерполяційним методом побудувати різницеву схему для стаціонарної задачі теплопровідності (ця сама математична модель описує стаціонарну дифузію)

$$-\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)u = f(x), \quad x \in (0, a), \quad (29)$$

$$u(0) = u(a) = 0, \quad (30)$$

де

$$p(x) \geqslant p_0 > 0, \; q(x) \geqslant 0, \; f(x) \in Q^0[0, a].$$

Розв'язання. Виберемо нерівномірну сітку

$$\omega_h = \left\{ x_i : x_{i+1} = x_i + h_{i+1}, \; x_0 = 0, \quad i = \overline{0, n}, \; \sum_{i=1}^{n+1} h_i = a \right\}$$

і запишемо рівняння балансу теплоти на відрізку $[x_{i-1/2}, x_{i+1/2}]$, де $x_{i+1/2} = x_i + \frac{1}{2}h_{i+1}$. Якщо $p(x)$ — коефіцієнт теплопровідності, а $u(x)$ — температура, то, очевидно, через точку $x_{i-1/2}$ на відрізок $[x_{i-1/2}, x_{i+1/2}]$ надходить потік теплоти

$$Q_{i-1/2} = \left(-p\frac{du}{dx}\right)\Big|_{x=x_{i-1/2}},$$

а через точку $x_{i+1/2}$ виділяється кількість теплоти

$$-Q_{i+1/2} = \left(p\frac{du}{dx}\right)\Big|_{x=x_{i+1/2}}.$$

Якщо відрізок $[x_{i-1/2}, x_{i+1/2}]$ містить джерела виділення теплоти, які розподілені з щільністю $f(x)$, то за їхній рахунок виділиться кількість теплоти

$$\int_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx.$$

Величина

$$\int_{x_{i-1/2}}^{x_{i+1/2}} qu\,dx,$$

де $q(x)$ — коефіцієнт тепловіддачі; qu — потужність потоків теплоти, характеризує тепловіддачу з бічної поверхні. Отже, рівняння балансу теплоти на відрізку $[x_{i-1/2}, x_{i+1/2}]$ відповідно до закону збереження

$$Q_{i-1/2} + \int_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx = Q_{i+1/2} + \int_{x_{i-1/2}}^{x_{i+1/2}} qu\,dx. \quad (31)$$

Це інтегральне співвідношення і є основою для побудови різницевої схеми. Виразимо значення $Q_{i\pm1/2}$ через шуканий розв'язок задачі u і через задані функції. Для цього проінтегруємо рівність $\frac{du}{dx} = \frac{-Q}{p(x)}$ на інтервалі (x_{i-1}, x_i) і замінимо Q за допомогою найпростішої інтерполяційної формули многочленом нульового степеня. Матимемо

$$u_i - u_{i-1} = -\int_{x_{i-1}}^{x_i} \frac{Q(x)}{p(x)}\,dx \approx -Q_{i-1/2}\int_{x_{i-1}}^{x_i}\frac{dx}{p(x)},$$
$$u_{i+1} - u_i = -\int_{x_i}^{x_{i+1}} \frac{Q}{p(x)}\,dx \approx -Q_{i+1/2}\int_{x_i}^{x_{i+1}}\frac{dx}{p(x)}. \qquad (32)$$

Підставляючи вирази для $Q_{i+1/2}$, знайдені з (32), а також замінюючи $\int_{x_{i-1/2}}^{x_{i+1/2}} qu\,dx \approx u_i \int_{x_{i-1/2}}^{x_{i+1/2}} q(x)\,dx$ в (31), дістанемо таку різницеву схему:

$$L_h y = \frac{1}{h_{i+1}}\left[a^*_{i+1}\frac{y_{i+1}-y_i}{h_{i+1}} - a^*_i\frac{y_i - y_{i-1}}{h_i}\right] - b^*_i y_i + f^*_i = 0, \quad i = \overline{1, n}; \qquad (33)$$
$$y_0 = y_{n+1} = 0,$$

де y_i — наближені значення для u_i;

$$a^*_i = \left(h_i^{-1}\int_{x_{i-1}}^{x_i}\frac{dx}{p(x)}\right)^{-1}, \quad b^*_i = \hbar^{-1}_{i+1}\int_{x_{i-1/2}}^{x_{i+1/2}} q(x)\,dx,$$
$$f^*_i = \hbar^{-1}_{i+1}\int_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx, \quad \hbar_{i+1} = \frac{1}{2}(h_{i+1} + h_i). \qquad (34)$$

Щоб уникнути обчислення інтегралів при визначенні коефіцієнтів різницевої схеми (33) за формулами (34), можна знову застосувати найпростішу інтерполяцію і покласти

$$a^*_i \approx p_{i-1/2} \equiv p(x_{i-1/2}), \quad b^*_i \approx \hbar^{-1}_{i+1} q_i (x_{i+1/2} - x_{i-1/2}) = q_i,$$
$$f^*_i \approx \hbar^{-1}_{i+1} f_i (x_{i+1/2} - x_{i-1/2}) = f_i$$

(неважко впевнитися, що ці формули мають апроксимацію за умов $p(x) \in C^3[0, a]$, $q(x) \in C^2[0, a]$). Тоді схема (33) має вигляд

$$L_h y = (a y_{\bar{x}})_{\hat{x}} - q_i y_i + f_i = 0, \quad i = \overline{1, n}; \quad y_0 = y_{n+1} = 0. \qquad (35)$$

Якщо позначити $\tilde{Q}_{i-1/2} = -a_i\frac{y_i - y_{i-1}}{h_i}$ і підсумувати рівність (35) по $i = \overline{1, n}$, то неважко знайти співвідношення

$$Q_{1/2} - Q_{n+1/2} + \sum_{i=1}^{n}\hbar_{i+1} q_i y_i - \sum_{i=1}^{n}\hbar_{i+1} f_i = 0, \qquad (36)$$

яке є різницевим аналогом інтегрального закону збереження енергії.

При побудові різницевих схем для практичних задач слід завжди стежити за тим, щоб виконувались не лише різницеві аналоги основних законів збереження, але й інші співвідношення, які спричинюються фізичними законами даної задачі. Такі схеми називаються *повністю консервативними* і дають змогу вести розрахунки на грубих сітках.

Інтегральне співвідношення, на основі якого можна будувати різницеву схему, можна дістати і таким чином. Проінтегруємо рівняння (29) на інтервалі $(x_{i-1/2}, x)$:

$$Q_{i-1/2} - p\frac{du}{dx} + \int_{x_{i-1/2}}^{x}(qu - f)\,ds = 0.$$

Останнє співвідношення поділимо на $p(x)$ і проінтегруємо на інтервалі (x_{i-1}, x_i):

$$Q_{i-1/2}\int_{x_{i-1}}^{x_i}\frac{dx}{p(x)} - u_i + u_{i-1} + \int_{x_{i-1}}^{x_i}\frac{1}{p(x)}\int_{x_{i-1/2}}^{x}(qu - f)\,ds dx = 0.$$

Знайдемо звідси $Q_{i-1/2}$ через відомі функції та шуканий розв'язок:

$$Q_{i-1/2} = \left(\int_{x_{i-1}}^{x_i}\frac{dx}{p(x)}\right)^{-1}\left[u_i - u_{i-1} - \int_{x_{i-1}}^{x_i}\frac{1}{p(x)}\int_{x_{i-1/2}}^{x}(qu - f)\,ds dx\right]. \qquad (37)$$

Аналогічно

$$Q_{i+1/2} = \left(\int_{x_i}^{x_{i+1}}\frac{dx}{p(x)}\right)^{-1}\left[u_{i+1} - u_i - \int_{x_i}^{x_{i+1}}\frac{1}{p(x)}\int_{x_{i+1/2}}^{x}(qu - f)\,ds dx\right]. \qquad (38)$$

Підставивши (37), (38) в (31), дістанемо основну інтегральну тотожність

$$\tilde{a}_{i+1}(u_{i+1} - u_i) - \tilde{a}_i(u_i - u_{i-1}) - \int_{x_{i-1/2}}^{x_{i+1/2}}(qu - f)\,ds =$$

$$= \tilde{a}_{i+1} \int_{x_i}^{x_{i+1}} \int_{x_{i+1/2}}^{x} \frac{q(s)u(s) - f(s)}{p(x)}\, ds dx -$$

$$- \tilde{a}_i \int_{x_{i-1}}^{x_i} \int_{x_{i-1/2}}^{x} \frac{q(s)u(s) - f(s)}{p(x)}\, ds dx, \qquad (39)$$

де

$$\tilde{a}_i = \left(\int_{x_{i-1}}^{x_i} \frac{dx}{p(x)} \right)^{-1}.$$

Інтегральне співвідношення (39) також може бути основою для відшукання різницевої схеми. Наприклад, якщо знехтувати правою частиною (39), а в лівій застосувати квадратурну формулу середніх прямокутників (або, що те саме, замість $u(s)$ використати її інтерполяційний поліном нульового степеня $u(x_i)$), то дістанемо схему (33).

4.3.4. Метод апроксимації квадратичного функціонала. Знову розглянемо побудову схеми для задачі

$$Lu = -(pu')' + qu = f(x), \quad x \in (0, 1);$$
$$u(0) = 0, \quad u(1) = 0; \quad p(x) > 0; \quad q(x) \geqslant 0, \qquad (40)$$

яка, як ми бачили у п. 1.4.1 (див. також гл. 1, метод Рітца), еквівалентна задачі про відшукання мінімуму квадратичного функціонала

$$J(u) = \int_0^1 (p(u')^2 + qu^2 - 2fu)\, dx \qquad (41)$$

на множині функцій $u(x)$, заданих на $(0, 1)$, і таких, що $u(0) = u(1) = 0$. Введемо на відрізку $[0, 1]$ сітку $\bar{\omega}_h = \{x_i = ih: i = \overline{0, N}, h = 1/N\}$ і зобразимо

$$J(u) = \sum_{i=1}^{N} J_i(u), \quad J_i(u) = \int_{x_{i-1}}^{x_i} (p(u')^2 + qu^2 - 2fu)\, dx.$$

Далі $J_i(u)$ апроксимуємо за допомогою якої-небудь квадратурної формули, наприклад,

$$\int_{x_{i-1}}^{x_i} p(u')^2\, dx \approx a_i (u_{\bar{x},i})^2 h, \quad a_i = h^{-1} \int_{x_{i-1}}^{x_i} p(x)\, dx,$$

$$\int_{x_{i-1}}^{x_i} (qu^2 - 2fu)\, dx \approx \frac{h}{2} [(qu^2 - 2fu)_i + (qu^2 - 2fu)_{i-1}]$$

(останнє співвідношення апроксимовано за формулою трапецій). У результаті дістанемо такий дискретний аналог функціонала (41)

$$J_h(y) = \sum_{k=1}^{N} h a_k (y_{\bar{x},k})^2 + \sum_{k=1}^{N} h (q_k y_k^2 - 2 f_k y_k),$$

де $y_i = y(x_i)$ — сіткова функція, що перетворюється на нуль при $i = 0, N$.

Функціонал $J_h(y)$ можна розглядати як функцію $(N - 1)$ змінних $y_1, \ldots, y_{N-1}$. Запишемо необхідну умову мінімуму

$$\frac{dJ_h}{dy_i} = 2a_i y_{\bar{x},i} - 2a_{i+1} y_{\bar{x},i+1} + (2q_i y_i - 2f_i) h = 0. \qquad (42)$$

Неважко помітити, що виконується також достатня умова мінімуму:

$$\frac{\partial^2 J_h}{\partial y_i^2} = \frac{2a_i}{h} + \frac{2a_{i+1}}{h} + 2q_i h > 0.$$

Отже, з умови (42) маємо, що елемент $y = y(x)$, $x \in \bar{\omega}_h$, $y(0) = y(1) = 0$, який мінімізує функціонал $J_h(y)$, є розв'язком різницевої схеми

$$-(a y_{\bar{x}})_{x,i} + q_i y_i = f_i, \quad i = \overline{1, N-1}; \quad y_0 = y_N = 0. \qquad (43)$$

Фактично розглядаємо дискретний аналог методу Рітца (див. гл. 1), згідно з яким рівняння

$$Ay = \varphi \left(\text{або } \sum_{j=1}^{N} a_{ij} y_j = \varphi_i \right), \quad A = A^* > 0$$

має розв'язок, що мінімізує функціонал

$$I_A(y) = (Ay, y) - 2(\varphi, y) \equiv \sum_{i,j=1}^{N} a_{ij} y_j y_i - 2 \sum_{i=1}^{N} \varphi_i y_i.$$

Якщо ввести простір $\mathring{H}_h$ функцій дискретного аргументу, заданих на сітці $\bar{\omega}_h$ і таких, що $y_0 = y_N = 0$ зі скалярним добутком

$$(y, v) = \sum_{x \in \bar{\omega}_h} h y(x) v(x),$$

і в цьому просторі ввести оператор

$$Ay = \{-(a y_{\bar{x}})_{x,i} + q_i y_i\}_{i = \overline{1, N-1}},$$

то за допомогою формули підсумування частинами неважко дістати

$$I_A(y) \equiv (Ay, y) - 2(f, y) = \sum_{i=1}^{N-1} h \{-(a y_{\bar{x}})_{x,i} + q_i y_i\} y_i -$$

$$- 2 \sum_{i=1}^{N-1} f_i y_i h = \sum_{i=1}^{N} h \{a_i (y_{\bar{x},i})^2 + q_i y_i^2\} - 2 \sum_{i=1}^{N} f_i y_i \equiv J_h(y).$$

Тому за методом Рітца елемент, що мінімізує функціонал $J_h(y)$, є розв'язком схеми (43).

Цей метод побудови різницевих схем особливо зручно застосовувати тоді, коли вихідна задача поставлена як задача мінімізації функціонала.

4.3.5. Метод апроксимації інтегральної тотожності (метод суматорних тотожностей). Розглянемо знову задачу (40). Якщо помножити рівняння (40) на довільну диференційовну функцію $v(x)$, $v(0) = v(1) = 0$ і проінтегрувати по x від 0 до 1, то дістанемо тотожність

$$I(u, v) = \int_0^1 (pu'v' + quv - fv)\,dx = 0, \quad \forall v \in \mathring{\tilde{H}}^1(0, 1), \tag{44}$$

де

$$\mathring{\tilde{H}}^1(0, 1) = \{v(x) : v \in C^1[0, 1],\ v(0) = v(1) = 0\}.$$

З п. 4.1 зрозуміло, що (44) є узагальненням задачі (40), в якій до шуканої функції $u(x)$ ставляться слабкіші вимоги гладкості. По аналогії з п. 4.3 замінимо в (44) інтеграли і похідні деякими їхніми дискретними апроксимаціями (на сітці $\bar{\omega}_h$), наприклад, похідні замінимо лівими різницевими похідними, а інтеграли замінимо за формулою правих прямокутників. Тоді дістанемо такий аналог (44), який називатимемо суматорною тотожністю:

$$I_h(y, v) = \sum_{i=1}^{N} a_i y_{\bar{x},i} v_{\bar{x},i} h + \sum_{i=1}^{N-1} (q_i y_i - f_i) v_i h = 0,$$

$$\forall v \in \mathring{H}_h, \tag{45}$$

де

$$\mathring{H}_h = \{v(x) : x \in \bar{\omega}_h,\ v(0) = v(1) = 0\},$$

$$a_i = h^{-1} \int_{x_{i-1}}^{x_i} p(x)\,dx \quad \text{або} \quad a_i = p_i.$$

Покладаючи тепер в (45) $v_i = \delta_{i,i_0} = \begin{cases} 1, & i = i_0 \\ 0, & i \neq i_0, \end{cases}$ $0 < i_0 < N$ і враховуючи, що $v_{\bar{x},i} = 0$ при $i < i_0$ та при $i > i_0 + 1$, а $v_{\bar{x},i_0+1} = -1/h$, $v_{\bar{x},i_0} = 1/h$, дістаємо

$$-h\left(\frac{1}{h} a_{i+1} y_{x,i} - \frac{1}{h} a_i y_{\bar{x},i}\right) + (q_i y_i - f_i) h = 0, \qquad i = i_0.$$

Таким чином, маємо різницеву схему

$$-(a y_{\bar{x}})_x + qy = f(x), \quad x \in \omega_h; \quad y(0) = y(1) = 0.$$

Цей метод побудови різницевих схем особливо зручний, коли вихідна задача має вигляд інтегральної тотожності, наприклад, для узагальнених розв'язків диференціальних рівнянь. Цей метод застосовується і для нелінійних задач.

4.3.6. Методи Рітца та Бубнова — Гальоркіна (варіаційно-різницеві методи). Згадаємо, що метод Рітца (див. гл. 1) грунтується на еквівалентності задачі розв'язування операторного рівняння

$$Au = f,$$

де A — самоспряжений, додатно визначений лінійний оператор у просторі V зі скалярним добутком $(,)$, та задачі знаходження мінімуму функціонала

$$I(u) = (Au, u) - 2(f, u).$$

Далі вводимо послідовність скінченновимірних просторів V_n з базисом $\{\varphi_i^{(n)}\}_{i=\overline{1,n}}$ і шукаємо елемент

$$u_n = \sum_{j=1}^{n} y_j \varphi_j \in V_n \quad (y_i \text{ — невідомі коефіцієнти}), \tag{46}$$

який мінімізує $I(u)$ не на всьому V, а лише на V_n. Неважко помітити, що

$$I(u_n) = \sum_{i,j=1}^{n} \alpha_{ij} y_i y_j - 2 \sum_{i=1}^{n} \beta_i y_i,$$

$$\alpha_{ij} = \alpha_{ji} = (A\varphi_i, \varphi_j), \quad \beta_i = (f, \varphi_i),$$

тобто $I(u_n)$ — функція від $y_1, \ldots, y_n$. З необхідної умови мінімуму дістаємо систему n рівнянь

$$\frac{\partial I(u_n)}{\partial y_i} = 2 \sum_{j=1}^{n} \alpha_{ij} y_j - 2\beta_i = 0, \quad i = \overline{1, n}, \tag{47}$$

з якої визначаємо y_i, $i = \overline{1, n}$.

Вправа 1. Записати систему рівнянь (47) для задачі (40), вибравши

$$\varphi_i(x) = \eta\left(\frac{x - x_i}{h}\right) = \eta_i(x), \quad \eta(s) = \begin{cases} 0, & s < -1,\ s > 1, \\ 1 + s, & -1 < s < 0, \\ 1 - s, & 0 < s < 1, \end{cases} \tag{48}$$

при $x_i \in \bar{\omega}_h$.

У методі Бубнова — Гальоркіна розв'язок u_n також шукаємо у вигляді (46), але коефіцієнти y_i знаходимо з умови ортогональності нев'язки $Au_n - f$ до базисних функцій $\varphi_i(x)$:

$$(Au_n - f, \varphi_i) = 0, \quad i = \overline{1, n},$$

причому тут не потребуємо самоспряженості оператора A (див. гл. 1).

Вправа 2. Для задачі (40) побудувати систему рівнянь для визначення y_i методом Бубнова — Гальоркіна, вибравши φ_i за формулами (48) і впевнитися, що вона збігається з системою методу Рітца.

Методи Рітца і Бубнова — Гальоркіна при спеціальному виборі координатних функцій (наприклад, (48)) приводять до методу, який має спеціальну назву і в останні роки широко використовується. Це *метод скінченних елементів,* який ми розглянемо далі.

4.3.7. Метод усереднюючих операторів. Розглянемо спочатку усереднюючі оператори Стєклова, які є операторами із простору $L_p(\Omega)$ в $L_p(\Omega)$, $p \geqslant 1$, $\Omega \subset R^n$. Одновимірні оператори усереднення, які для функції багатьох змінних діють у кожному напрямі x_i, $i = \overline{1, n}$, визначаються так:

$$S_i u(x) = \frac{1}{h_i} \int\limits_{x_i - 0{,}5h_i}^{x_i + 0{,}5h_i} u(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_n)\, d\xi =$$

$$= \int\limits_{-0{,}5}^{0{,}5} u(x_1, \ldots, x_{i-1}, x_i + sh_i, x_{i+1}, \ldots, x_n)\, ds,$$

$$S_i^+ u(x) = \frac{1}{h_i} \int\limits_{x_i}^{x_i + h_i} u(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_n)\, d\xi = \qquad (49)$$

$$= \int\limits_{0}^{1} u(x_1, \ldots, x_{i-1}, x_i + sh_i, x_{i+1}, \ldots, x_n)\, ds,$$

$$S_i^- u(x) = \frac{1}{h_i} \int\limits_{x_i - h_i}^{x_i} u(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_n)\, d\xi =$$

$$= \int\limits_{-1}^{0} u(x_1, \ldots, x_{i-1}, x_i + sh_i, x_{i+1}, \ldots, x_n)\, ds.$$

При $n = 1$ ці оператори позначатимемо відповідно S, S^+, S^-. Оператор усереднення по Стєклову в R^n визначається як добуток одновимірних усереднюючих операторів S_i по всіх i, тобто $S = S_1 S_2 \ldots S_n$. Неважко переконатися, що усереднюючі оператори мають такі властивості:

$$S_i^+ \frac{\partial u(x)}{\partial x_i} = u_{x_i}(x), \quad S_i^- \frac{\partial u(x)}{\partial x_i} = u_{\bar{x}_i}(x),$$

$$S_i \frac{\partial u(x)}{\partial x_i} = \frac{1}{h_i} \left(u^{(+0,5i)}(x) - u^{(-0,5i)}(x)\right), \quad i = \overline{1, n}, \qquad (50)$$

де

$$u^{(\pm 0,5i)}(x) = u(x_1, \ldots, x_{i-1}, x_i \pm 0{,}5h_i, x_{i+1}, \ldots, x_n).$$

Якщо ввести ядро усереднення

$$\varkappa_h(|x|) = \prod_{i=1}^{n} \frac{1}{h_i} \varkappa\left(\frac{|x|}{h_i}\right), \quad \varkappa(s) = \begin{cases} 1, & |s| < 0{,}5 \\ 0, & |s| \geqslant 0{,}5, \end{cases}$$

то можна записати

$$Su(x) = \int\limits_{R^n} u(y)\, \varkappa_h(|x - y|)\, dy = \frac{1}{h_1 \ldots h_n} \int\limits_{x_1 - 0,5h_1}^{x_1 + 0,5h_1} \ldots \int\limits_{x_n - 0,5h_n}^{x_1 + 0,5h_n} u(y)\, dy,$$

і оскільки ядро усереднення належить простору $L_\infty(R^n)$, то вираз $Su(x)$ має певне значення для кожної точки $x \in R^n$, якщо $u(x)$ сумовна в кожній підобласті R^n.

Розглянемо також оператори повторного усереднення

$$T_i = S_i^2 \quad (i = \overline{1, n}). \qquad (51)$$

Неважко помітити, що $T_i = S_i^+ S_i^- = S_i^- S_i^+$. Знайдемо явний вигляд оператора $T_1 \equiv T$ в одновимірному випадку. Інтегруючи частинами, маємо

$$Tu(x) = S^+ S^- u(x) = \frac{1}{h^2} \int\limits_{x}^{x+h} \int\limits_{\xi - h}^{\xi} u(\eta)\, d\eta d\xi =$$

$$= \frac{1}{h^2} \left\{ \int\limits_{x-h}^{x} (\eta + h - x)\, u(\eta)\, d\eta + \int\limits_{x}^{x+h} (x + h - \eta)\, u(\eta)\, d\eta \right\}.$$

Тому

$$T_i u(x) = \frac{1}{h_i^2} \int\limits_{x_i - h_i}^{x_i} (\xi + h_i - x_i)\, u(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_n)\, d\xi +$$

$$+ \frac{1}{h_i^2} \int\limits_{x_i}^{x_i + h_i} (x_i + h_i - \xi)\, u(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_n)\, d\xi = \qquad (52)$$

$$= \frac{1}{h_i^2} \int\limits_{x_i - h_i}^{x_i + h_i} (h_i - |\xi - x_i|)\, u(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_n)\, d\xi =$$

$$= \int\limits_{-1}^{1} (1 - |s|)\, u(x_1, \ldots, x_{i-1}, x_i + sh_i, x_{i+1}, \ldots, x_n)\, ds, \quad i = \overline{1, n}.$$

Зазначимо таку важливу властивість операторів повторного усереднення, яка випливає з їхнього означення і властивостей (50):

$$T_i \frac{\partial^2 u(x)}{\partial x_i^2} = S_i^+ S_i^- \frac{\partial^2 u(x)}{\partial x_i^2} = u_{\bar{x}_i x_i}(x), \quad i = \overline{1, n}. \tag{53}$$

Іноді використовують добутки одновимірних операторів повторного усереднення:

$$T = \prod_{k=1}^{n} T_k, \quad \hat{T}_i = \prod_{\substack{k=1 \\ k \neq i}}^{n} T_k.$$

Неважко зрозуміти, що $T = S^2$. Якщо ввести ядро усереднення

$$\varphi_h(|x|) = \prod_{i=1}^{n} \frac{1}{h_i} \varphi\left(\frac{|x|}{h_i}\right), \quad \varphi(t) = \begin{cases} 1 - |t|, & |t| \leqslant 1, \\ 0, & |t| \geqslant 1, \end{cases}$$

то можна записати

$$Tu(x) = \int_{R^n} u(y) \varphi_h(|x - y|)\, dy.$$

Застосування операторів усереднення для побудови сіткових схем (різницевих схем) проілюструємо на задачі

$$-(ku')' + q(x)u(x) = f(x), \quad x \in (0, 1), \quad u(0) = u(1) = 0. \tag{54}$$

Покриємо відрізок $[0, 1]$ рівномірною сіткою $\bar{\omega}_h = \{x_i = ih: i = \overline{0, N}, h = 1/N\}$, $\omega_h = \bar{\omega}_h \setminus \{0, 1\}$. Припустимо, що $f(x) \in L_2(0, 1)$, $q \in L_\infty(0, 1)$, $k \in W_\infty^1(0, 1)$, $u(x) \in W_2^2(0, 1)$. Застосовуючи оператор S до рівняння (54), дістаємо

$$S[-(ku')' + q(x)u(x)] = Sf(x), \quad x \in \left|\frac{h}{2}, 1 - \frac{h}{2}\right|,$$

тобто це співвідношення виконується для всіх $x \in \omega_h$. Враховуючи далі (50), для всіх $x \in \omega_h$ дістаємо

$$-\frac{1}{h}\left[k\left(x + \frac{h}{2}\right)u'\left(x + \frac{h}{2}\right) - k\left(x - \frac{h}{2}\right)u'\left(x - \frac{h}{2}\right)\right] + \int_{x-0,5h}^{x+0,5h} q(\xi)u(\xi)\, d\xi = Sf(x), \quad x \in \omega_h. \tag{54'}$$

Останнє співвідношення аналогічно тому, як це робилося в інтерполяційному методі, можна використовувати як основу для побудови різницевих схем. Наприклад, поклавши

$$u'\left(x + \frac{h}{2}\right) \approx u_{\bar{x}}(x + h), \quad u'\left(x - \frac{h}{2}\right) \approx u_{\bar{x}}(x),$$
$$u(\xi) \approx u(x), \quad \xi \in [x - 0{,}5h,\ x + 0{,}5h],$$

з (54′) дістаємо різницеву схему для задачі (54)

$$-(ay_{\bar{x}})_x + d(x)y(x) = \varphi(x), \quad x \in \omega_h, \quad y(0) = y(1) = 0, \tag{54''}$$

де

$$a(x) = k\left(x - \frac{h}{2}\right), \quad d(x) = \frac{1}{h}\int_{x-0,5h}^{x+0,5h} q(\xi)\, d\xi \equiv Sq(x),$$

$$\varphi(x) = Sf(x) = \frac{1}{h}\int_{x-0,5h}^{x+0,5h} f(\xi)\, d\xi.$$

До рівняння (54) можна застосувати і оператор повторного усереднення T:

$$T(-(ku')' + q(x)u(x)) = Tf(x), \quad x \in [h, 1 - h].$$

Це співвідношення справджується для кожного $x \in \omega_h$. Із (53) при $k(x) \equiv 1$ дістаємо

$$-u_{\bar{x}x}(x) + T(q(x)u(x)) = Tf(x), \quad x \in \omega_h, \tag{55}$$

що також є основою для побудови різницевих схем.

Наприклад, якщо замість $u(\xi)$ взяти її інтерполяційний поліном нульового степеня $u(x_i)$, то для задачі (54) при $k(x) \equiv 1$ дістаємо різницеву схему

$$-y_{\bar{x}x} + Tq(x) \cdot y(x) = Tf(x), \quad x \in \omega_h, \quad y(0) = y(1) = 0. \tag{56}$$

Якщо ж взяти інтерполяційний многочлен по точках x_i та x_{i+1} у вигляді

$$u(x_i) + (\xi - x_i)\, u_x(x_i),$$

то дістанемо схему

$$-y_{\bar{x}x} + Tq(x) \cdot y(x) + T(q(x)(x - x_i)) \cdot y_x(x) = Tf(x), \quad x \in \omega_h, \quad y(0) = y(1) = 0. \tag{57}$$

Якщо $k(x) \equiv 1$, $q(x) \equiv 0$, то (55) набирає вигляду

$$-u_{\bar{x}x}(x) = Tf(x), \quad x \in \omega_h. \tag{58}$$

Останнє співвідношення задовольняє точний розв'язок задачі (54), а разом з крайовими умовами $u(0) = u(1) = 0$ співвідношення (58) є замкненою системою лінійних алгебраїчних рівнянь відносно невідомих значень розв'язку $u(x)$ в точках сітки. Оскільки точний розв'язок диференціальної задачі є розв'язком сіткової задачі, то схема (58) разом з крайовими умовами називається *точною схемою*. У свою чергу оператор T називають *усереднюючим оператором точної різницевої схеми*.

Якщо $\varphi(x) \in W_2^1(0, 1)$, то, помноживши (54) на $\varphi(x)$ і проінтегрувавши частинами, дістанемо

$$a(u, v) \equiv \int_0^1 (u'(x)\varphi'(x) + q(x)u(x)\varphi(x))\,dx =$$

$$= \int_0^1 f(x)\varphi(x)\,dx \equiv l(\varphi), \quad \forall \varphi \in \mathring{W}_2^1(0, 1).$$

Ця тотожність може бути означенням узагальненого розв'язку задачі (54) з простору $\mathring{W}_2^1(0, 1)$. Неважко помітити, що і вигляд лінійного функціонала $l(\varphi)$, $\varphi \in \mathring{W}_2^1(0, 1)$ можна узагальнити таким чином:

$$l(\varphi) = \int_0^1 f_0(x)\varphi(x)\,dx - \int_0^1 f_1(x)\varphi'(x)\,dx \equiv \langle f, \varphi\rangle,$$

де $f_0, f_1 \in L_2(0, 1)$. Символічно (оскільки $f_1 \notin C^1[0, 1]$) це можна записати так:

$$l(\varphi) = \langle f, \varphi\rangle,$$

де $f = f_0(x) + \frac{df_1(x)}{dx}$, $f_0, f_1 \in L_2(0, 1)$.

Цьому виразу можна надати і строгого математичного змісту, ввівши простір W_2^{-1}, але це виходить за рамки нашої книги. Візьмемо за означення узагальненого з $\mathring{W}_2^1$ розв'язку задачі (54) при $f \in W_2^{-1}$ тотожність

$$a(u, v) = \langle f, v\rangle \quad \forall v \in \mathring{W}_2^1(0, 1). \tag{59}$$

Переконавшись, що ядро усереднення $\varphi_h \in \mathring{W}_2^1(0, 1)$, і поклавши в (59) $v = \varphi_h$, дістанемо

$$-\int_0^1 u'(\xi)\varphi_h'(|x - \xi|)\,d\xi + \int_0^1 q(\xi)u(\xi)\varphi_h(|x - \xi|)\,d\xi =$$

$$= \int_0^1 f_0(\xi)\varphi_h(|x - \xi|)\,d\xi - \int_0^1 f_1(\xi)\varphi_h'(|x - \xi|)\,d\xi, \quad x \in \omega_h.$$

Враховуючи вигляд функції φ_h, маємо

$$-u_{\bar{x}x}(x) + T(q(x)u(x)) = \varphi(x), \quad x \in \omega_h,$$
$$\varphi(x) = Tf_0(x) - (S^+f_1(x))_{\bar{x}}. \tag{60}$$

Співвідношення (60) є основою для побудови різницевих схем.

Оператори усереднення використовують для побудови різницевих схем для задач з узагальненими розв'язками, тому що вони дають змогу подати похибку апроксимації в зручному для досліджень вигляді (див. п. 8).

4.3.8. Збіжність і оцінки похибки методу сіток. Згідно з теоремою Лакса — Філіпова (див. п. 2.1) збіжність і оцінку похибки сіткових схем проводитимемо за такою схемою: записавши задачу для похибки, по-перше, знаходимо апріорну оцінку і, по-друге, оцінюємо похибку апроксимації.

Приклад 3 (дослідження збіжності і оцінки похибки в нормі $L_2(\bar{\omega})$). Дослідити збіжність і встановити оцінку похибки різницевої схеми (див. п. 4.3.1)

$$-y_{x,0} + \frac{h}{2}g_0y_0 + \alpha_0y_0 = \gamma_0 + \frac{h}{2}\varphi_0,$$
$$-y_{\bar{x}x,i} + g_iy_i = \varphi_i, \quad i = \overline{1, N-1}, \tag{61}$$
$$y_{\bar{x},N} + \frac{h}{2}g_Ny_N + \alpha_1y_N = \gamma_1 + \frac{h}{2}\varphi_N$$

для задачі

$$-u''(x) + g(x)u(x) = \varphi(x), \quad x \in (a, b), \quad -u'(a) + \alpha_0u(a) = \gamma_0,$$
$$u'(b) + \alpha_1u(b) = \gamma_1, \tag{62}$$

де $\varphi(x)$, $g(x)$ — задані функції, α_0, α_1, γ_0, γ_1 — задані числа, $g(x) \geqslant 0$, $\alpha_0 > 0$, $\alpha_1 > 0$, $\bar{\omega}_h = \{x_i = a + ih : i = \overline{0, N}, h = (b - a)/N\}$.

Р о з в' я з а н н я. а) Запишемо сіткову задачу для похибки $z(x) = y(x) - u(x)$, $x \in \bar{\omega}_h$. Для цього на сіткову функцію z подіємо оператором сіткової схеми і врахуємо (61). Для точки x_0 маємо

$$-z_{x,0} + \frac{h}{2}g_0z_0 + \alpha_0z_0 = \psi_0,$$

де

$$\psi_0 = \frac{h}{2}\varphi_0 + \gamma_0 + u_{x,0} - \frac{h}{2}g_0u_0 - \alpha_0u_0.$$

Аналогічно для точки x_N:

$$z_{\bar{x},N} + \frac{h}{2}g_Nz_N + \alpha_1z_N = \psi_N,$$

де

$$\psi_N = \frac{h}{2}\varphi_N + \gamma_1 - u_{\bar{x},N} - \frac{h}{2}g_Nu_N - \alpha_1u_N.$$

Для точок сітки $\omega_h = \bar{\omega}_h \setminus \{a, b\}$ маємо

$$-z_{\bar{x}x,i} + g_iz_i = \psi_i, \quad i = \overline{1, N-1},$$

де $\psi_i = \varphi_i + u_{\bar{x}x,i} - g_iu_i$. Таким чином, задача для похибки z має вигляд

$$-z_{x,0} + \frac{h}{2}g_0z_0 + \alpha_0z_0 = \psi_0,$$

$$-z_{\bar{x}x,i} + g_i z_i = \psi_i, \quad i = \overline{1, N-1}, \tag{6}$$

$$z_{\bar{x},N} + \frac{h}{2} g_N z_N + \alpha_1 z_N = \psi_N,$$

де $\psi_i = \psi(x_i)$ — похибка апроксимації.

б) Запишемо нерівність стійкості сіткової схеми (нерівність коректності). Для сіткової норми $L_2(\bar{\omega}_h)$ ми її вже знайшли в п. 2.2.3:

$$\|z\| \equiv \|z\|_{L_2(\bar{\omega}_h)} = \left(\sum_{i=0}^{N} h z_i^2\right)^{1/2} \leqslant \varkappa \|\psi\|, \tag{64}$$

де $\varkappa = 2(b-a)\max\{\alpha_0^{-1}, \alpha_1^{-1}, 2(b-a)\}$ (зауважимо, що вибір норми в руках дослідника і нерівність коректності може бути записана і в інших нормах).

в) Залишилося дослідити порядок величини $\|\psi\|$ (норми похибки апроксимації) як функції від h.

Припустимо, що $u \in W_2^4(a, b)$.

Розглянемо похибку апроксимації $\psi(x)$. При фіксованому x $\psi(x)$ є лінійним функціоналом від u. Перетворимо $\psi(x)$ з урахуванням рівностей (62):

$$\psi_0 = \frac{h}{2}(\varphi_0 - g_0 u_0) + \gamma_0 - \alpha_0 u_0 + u_{x,0} =$$

$$= -\frac{h}{2} u''(a) - u'(a) + u_x(a),$$

$$\psi_i = \varphi_i + u_{\bar{x}x,i} - g_i u_i = -u''(x_i) + u_{\bar{x}x}(x_i), \quad i = \overline{1, N-1},$$

$$\psi_N = \frac{h}{2}(\varphi_N - g_N u_N) + \gamma_1 - \alpha_1 u_N - u_{\bar{x},N} = \frac{h}{2} u''(b) + u'(b) - u_{\bar{x}}(b).$$

Щоб сталі в нерівностях, які ми використовуватимемо далі, не залежали від h, виконаємо такі заміни: $x = a + sh$, $s \in [0, 1]$ для функціоналів ψ_i, $x = x_i + sh$, $s \in [-1, 1]$ для функціоналів ψ_i, $i = \overline{1, N-1}$ і, зрештою, $x = b - sh$, $s \in [0, 1]$ для функціонала ψ_N. При цьому використаємо позначення

$$u(x) = u(a + sh) = \tilde{u}(s), \quad x \in [a, a+h], \quad s \in [0, 1],$$

$$u(x) = u(x_i + sh) = \tilde{u}(s), \quad x \in [x_{i-1}, x_{i+1}], \quad s \in [-1, 1],$$

$$u(x) = u(b - sh) = \tilde{u}(s), \quad x \in [b-h, b], \quad s \in [0, 1].$$

У нових змінних функціонали похибки апроксимації набирають вигляду

$$\psi_0 \equiv \psi_0(a; u) \equiv \psi_0(u) = \psi_0(\tilde{u}) = h^{-1}\left|-\frac{1}{2}\tilde{u}''_{s^2}(0) - \tilde{u}'_s(0) + \tilde{u}(1) - \tilde{u}(0)\right|,$$

$$\psi_i = h^{-2}[-\tilde{u}''_{s^2}(0) + \tilde{u}(1) - 2\tilde{u}(0) + \tilde{u}(-1)], \quad i = \overline{1, N-1},$$

$$\psi_N = h^{-1}\left[\frac{1}{2}\tilde{u}''_{s^2}(1) + \tilde{u}'_s(1) - \tilde{u}(1) + \tilde{u}(1-h)\right].$$

Зауважимо, що коли $u(x) \in W_2^4(a, b)$, то і $\tilde{u}(s) \in W_2^4$ відповідно на інтервалах $(0, 1)$ чи $(-1, 1)$. Покажемо, що функціонали ψ_i, $i = \overline{0, N}$ обмежені в просторі W_2^4. На-

приклад, для ψ_0 маємо

$$|\psi_0| \leqslant 4h^{-1}\|\tilde{u}\|_{C^2[0,1]}$$

і далі згідно з теоремою вкладення $C^2 \subset W_2^3 \subset W_2^4$ запишемо

$$|\psi_0| \leqslant 4Mh^{-1}\|\tilde{u}\|_{W_2^3(0,1)} \equiv 4Mh^{-1}\|\tilde{u}\|_{3,2,(0,1)}, \tag{65}$$

де стала M не залежить від h і $\tilde{u}$. Пропонуємо читачеві впевнитися, що

$$|\psi_N| \leqslant 4Mh^{-1}\|\tilde{u}\|_{3,2,(0,1)}, \tag{66}$$

$$|\psi_i| \leqslant 5Mh^{-2}\|\tilde{u}\|_{4,2,(-1,1)}, \quad i = \overline{1, N-1}$$

(через M позначатимемо різні сталі, що не залежать від h та u).

Функціонал ψ_0 перетворюється на нуль на многочленах до другого степеня включно. Дійсно,

$$\psi_0(\text{const}) = 0,$$

$$\psi_0(s) = h^{-1}[-0 - 1 + 1 - 0] = 0,$$

$$\psi_0(s^2) = h^{-1}\left[-\frac{1}{2}\cdot 2\,|_{s=0} - 2s\,|_{s=0} + s^2\,|_{s=1} - s^2\,|_{s=0}\right] = 0,$$

$$\psi_0(s^3) = h^{-1}\left[-\frac{1}{2}\cdot 6s\,|_{s=0} - 3s^2\,|_{s=0} + s^3\,|_{s=1} - s^3\,|_{s=0}\right] = 1 \neq 0.$$

Тому за лемою Бремблa — Гільберта

$$|\psi_0| \leqslant 4\bar{M}Mh^{-1}|\tilde{u}|_{3,2,(0,1)}, \tag{67}$$

де стала M не залежить від h та від u. Далі переходимо до змінної x:

$$|\tilde{u}|_{3,2,(0,1)} \equiv |\tilde{u}|_{W_2^3(0,1)} = \left(\int_0^1 \left(\frac{d^3\tilde{u}(s)}{ds^3}\right)^2 ds\right)^{1/2} =$$

$$= h^{5/2}\left(\int_a^{a+h}\left(\frac{d^3u}{dx^3}\right)^2 dx\right)^{1/2} = h^{5/2}|u|_{W_2^3(e_0)}, \quad e_0 = (a, a+h), \tag{67'}$$

тоді

$$|\psi_0| \leqslant Mh^{3/2}|u|_{W_2^3(e_0)}. \tag{68}$$

Аналогічно знаходимо оцінку

$$|\psi_N| \leqslant Mh^{3/2}|u|_{W_2^3(e_N)}, \quad e_N = (b-h, b). \tag{69}$$

Неважко переконатися, що кожен з функціоналів ψ_i, $i = \overline{1, N-1}$ перетворюється на нуль на многочленах до третього степеня включно. Тому з виразу (66) і леми Бремблa — Гільберта випливає, що

$$|\psi_i| \leqslant 5\bar{M}Mh^{-2}|\tilde{u}|_{W_2^4(-1,1)}$$

і далі аналогічно попередньому

$$|\psi_i| \leqslant Mh^{3/2} |u|_{W_2^4(e_i)}, \tag{70}$$

$$e_i = (x_{i-1}, x_{i+1}), \quad i = \overline{1, N-1}.$$

З теореми вкладення і за умов (68), (69), (70), очевидно, випливають оцінки

$$|\psi_0| \leqslant Mh^{3/2} \|u\|_{W_2^3(e_0)} \leqslant Mh^{3/2} \|u\|_{W_2^4(e_0)},$$

$$|\psi_N| \leqslant Mh^{3/2} \|u\|_{W_2^4(e_N)}, \tag{71}$$

$$|\psi_i| \leqslant Mh^{3/2} \|u\|_{W_2^4(e_i)}, \quad i = \overline{1, N-1},$$

де M — найбільша із всіх сталих, що використовувались раніше, причому M не залежить від h та u. Враховуючи оцінки (71), для величини $\|\psi\|$ знайдемо оцінку

$$\|\psi\| = \|\psi\|_{L_2(\bar{\omega}_h)} = \left(\sum_{i=0}^{N} h\psi_i^2\right)^{1/2} \leqslant$$

$$\leqslant Mh^2 \left(\|u\|^2_{W_2^4(e_0)} + \|u\|^2_{W_2^4(e_N)} + \sum_{i=1}^{N-1} \|u\|^2_{W_2^4(e_i)} \right)^{1/2} =$$

$$= Mh^2 \left(\int_a^{a+h} \sum_{j=0}^{4} \left(\frac{d^j u(x)}{dx^j} \right)^2 dx + \int_{b-h}^{b} \sum_{j=0}^{4} \left(\frac{d^j u(x)}{dx^j} \right)^2 dx + \right.$$

$$\left. + \sum_{i=1}^{N-1} \int_{x_{i-1}}^{x_{i+1}} \sum_{j=0}^{4} \left(\frac{d^j u(x)}{dx^j} \right)^2 dx \right)^{1/2} =$$

$$= Mh^2 \left(2 \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} \sum_{j=0}^{4} \left(\frac{d^j u(x)}{dx^j} \right)^2 dx \right)^{1/2} =$$

$$= Mh^2 \sqrt{2} \left(\int_a^b \sum_{j=0}^{4} \left(\frac{d^j u(x)}{dx^j} \right)^2 dx \right)^{1/2} = \sqrt{2}\, Mh^2 \|u\|_{W_2^4(a,b)}. \tag{72}$$

Отже, з нерівності коректності (64) і оцінки (72) знаходимо оцінку похибки різницевої схеми (61)

$$\|z\| = \|y - u\| \leqslant Mh^2 \|u\|_{W_2^4(a,b)} \tag{73}$$

за умови, що розв'язок диференціальної задачі (62) належить простору $W_2^4(a, b)$. З оцінки (73) випливає, що за умови $u \in W_2^4(a, b)$ розв'язок різницевої схеми при $h \to 0$ збігається до точного розв'язку в нормі $L_2(\bar{\omega}_h)$ з швидкістю $0(h^2)$.

Сформулюємо дістаний результат у вигляді такої теореми.

Теорема 1. *Якщо розв'язок задачі* (62) *належить класу* $W_2^4(a, b)$, *то розв'язок різницевої схеми* (61) *збігається при* $h \to 0$ *до точного розв'язку задачі* (62) *в сітковій нормі* $L_2(\bar{\omega}_h)$, *причому має місце оцінка швидкості збіжності* (73).

Якщо $u(x) \in C^4[a, b]$ (тобто розв'язок диференціальної задачі має більшу гладкість), то замість (67′) ми мали б

$$|\tilde{u}|_{W_2^3(0,1)} = \left(\int_0^1 \left(\frac{d^3 \tilde{u}(s)}{ds^3} \right)^2 ds \right)^{1/2} = h^{5/2} \left(\int_a^{a+h} \left(\frac{d^3 u(x)}{dx^3} \right)^2 dx \right)^{1/2} \leqslant$$

$$\leqslant h^{5/2} \max_{x \in [a, a+h]} |u'''_{x^3}(x)| \left(\int_a^{a+h} dx \right)^{1/2} \leqslant h^3 \|u\|_{C^3[a,b]} \leqslant h^3 \|u\|_{C^4[a,b]}$$

і замість (68) —

$$|\psi_0| \leqslant Mh^2 \|u\|_{C^4[a,b]}.$$

Аналогічно

$$|\psi_N| \leqslant Mh^2 \|u\|_{C^4[a,b]}, \quad |\psi_i| \leqslant Mh^2 \|u\|_{C^4[a,b]}, \quad i = \overline{1, N-1},$$

і для величини $\|\psi\|$ мали б оцінку

$$\|\psi\| = \left(\sum_{i=0}^{N} h\psi_i^2 \right)^{1/2} \leqslant Mh^2 \|u\|_{C^4[a,b]} \left(\sum_{i=0}^{N} h \right)^{1/2} \leqslant$$

$$\leqslant M(b - a + h)^{1/2} h^2 \|u\|_{C^4[a,b]}.$$

Тоді з (64) при $h \leqslant h_0$ дістали б

$$\|z\| \leqslant M(b - a + h_0)^{1/2} h^2 \|u\|_{C^4[a,b]}. \tag{74}$$

Зауважимо, що за умови $u(x) \in C^4[a, b]$ оцінку виду (74) (тобто по порядку h) можна дістати і без використання леми Бремблa — Гільберта. Користуючись формулою Тейлора, маємо

$$\psi_0 = -\frac{h}{2} u''(a) - u'(a) + h^{-1}[u(a + h) - u(a)] =$$

$$= -\frac{h}{2} u''(a) - u'(a) + h^{-1} \left[u(a) + hu'(a) + \frac{h^2}{2} u''(a) + \right.$$

$$\left. + \frac{h^3}{6} u'''(\xi) - u(a) \right] = \frac{h^2}{6} u'''(\xi), \quad \xi \in (a, a + h).$$

Звідси

$$|\psi_0| \leqslant \frac{h^2}{6} \max_{x \in [a, a+h]} |u'''(x)| \leqslant \frac{h^2}{6} \|u\|_{C^3[a,b]}. \tag{75}$$

Аналогічно

$$|\psi_N| \leqslant \frac{h^2}{6} \|u\|_{C^3[a,b]}. \tag{76}$$

Для функціоналів ψ_i за допомогою формули Тейлора при $u(x) \in C^4[a, b]$ знаходимо

$$\psi_i = -u''(x_i) + u_{\bar{x}x}(x_i) = -u''(x_i) + h^{-2}[u(x_{i+1}) - 2u(x_i) + u(x_{i-1})] = -u''(x_i) + h^{-2}\left[\left(u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{3!}u'''(x_i) + \frac{h^4}{4!}u^{(4)}(\xi_1)\right) - 2u(x_i) + \left(u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{3!}u'''(x_i) + \frac{h^4}{4!}u^{(4)}(\xi_2)\right)\right] = \frac{h^2}{24}[u^{(4)}(\xi_1) + u^{(4)}(\xi_2)], \quad \xi_2 \in (x_{i-1}, x_i), \quad \xi_1 \in (x_i, x_{i+1}).$$

Оскільки $u^{(4)} \in C[a, b]$, то

$$m \leqslant \frac{u^{(4)}(\xi_1) + u^{(4)}(\xi_2) + \cdots + u^{(4)}(\xi_n)}{n} \leqslant M,$$

де $m = \min\limits_{x \in [a,b]} u^{(4)}(x)$, $M = \max\limits_{x \in [a,b]} u^{(4)}(x)$, і в силу неперервності $u^{(4)}(x)$ існує точка $\bar{\xi} \in [a, b]$ така, що

$$\frac{u^{(4)}(\xi_1) + \cdots + u^{(4)}(\xi_n)}{n} = u^{(4)}(\bar{\xi}).$$

Тому існує $\bar{\xi}_i \in [x_{i-1}, x_{i+1}]$ така, що

$$\psi_i = \frac{h^2}{12}u^{(4)}(\bar{\xi}_i), \quad i = \overline{1, N-1},$$

звідки

$$|\psi_i| \leqslant \frac{h^2}{12}\|u\|_{C^4[a,b]}, \quad i = \overline{1, N-1}. \tag{77}$$

Враховуючи вирази (75) — (77), для величини $\|\psi\|$ маємо (при $h \leqslant h_0$)

$$\|\psi\| = \left(\sum_{i=0}^{N} h\psi_i^2\right)^{1/2} \leqslant (b - a + h_0)^{1/2}\frac{h^2}{6}\|u\|_{C^4[a,b]}$$

і з умови (64) знаходимо оцінку точності

$$\|z\| = \|y - u\| \leqslant \varkappa (b - a + h_0)^{1/2}\frac{h^2}{6}\|u\|_{C^4[a,b]}. \tag{78}$$

Оцінки (74) і (78) однакові за порядком h, але в (78) точніший сталий множник при h^2. Такий самий порядок h має і оцінка (73), знайдена при менших вимогах до шуканого розв'язку, а саме $u \in W_2^1(a, b)$.

Приклад 4 (дослідження збіжності і оцінки похибки в нормі $C(\omega_h)$). Дослідити збіжність і встановити оцінку похибки різницевої схеми

$$(ay_{\bar{x}})_x - d(x)y = -\varphi(x), \quad x \in \omega_h, \qquad y(0) = \mu_1, \quad y(1) = \mu_2 \tag{79}$$

для задачі

$$(ku')' - q(x)u = -f(x), \quad x \in (0, 1), \qquad u(0) = \mu_1, \quad u(1) = \mu_2, \tag{80}$$

де $k(x) \geqslant C_1 > 0$, $q(x) \geqslant 0$, k, q, f — задані функції, μ_1, μ_2 — задані числа, q, $f \in C^2[0, 1]$, $u(x) \in C^4[0, 1]$, $k \in C^3[0, 1]$, $\omega_h = \{x_i = ih : i = \overline{1, N-1}, h = 1/N\}$, $a_i = k_{i-1/2}$, $d_i = q_i$, $\varphi_i = f_i$.

Розв'язання. а) Запишемо різницеву задачу для похибки $z(x) = y(x) - u(x)$, $x \in \omega_h$:

$$(az_{\bar{x}})_x - d(x)z = \psi(x), \quad x \in \omega_h, \quad z(0) = 0, \quad z(1) = 0,$$

де $\psi_i = (ku')'_i - (au_{\bar{x}})_{x,i}$.

б) Дослідимо похибку апроксимації. Подамо ψ_i у вигляді

$$\psi_i = \eta_{x,i}^{(1)} + \eta_i^{(2)},$$

де

$$\eta^{(1)} = (ku')_{i-1/2} - k_{i-1/2}u_{\bar{x},i} = k_{i-1/2}\bar{\eta}_i^{(1)},$$

$$\bar{\eta}_i^{(1)} = u'_{i-1/2} - u_{\bar{x},i}, \quad \eta_i^{(2)} = (ku')'_i - (ku')_{x,i-1/2}.$$

Величина $\bar{\eta}_i^{(1)}$ при фіксованому i, очевидно, є лінійним функціоналом від u, тобто $\bar{\eta}_i^{(1)} = \bar{\eta}_i^{(1)}(u)$. Дослідимо його стандартним шляхом за допомогою леми Брембла — Гільберта. Виконавши заміну $x = x_{i-1} + sh$, $s \in [0,1]$, $x \in [x_{i-1}, x_i]$, $u(x) = u(x_{i-1} + sh) \equiv \tilde{u}(s)$, дістаємо

$$\bar{\eta}_i^{(1)}(u) = \bar{\eta}_i^{(1)}(\tilde{u}) = h^{-1}\left(\tilde{u}'\left(\frac{1}{2}\right) - \tilde{u}(1) + \tilde{u}(0)\right),$$

звідки за теоремою вкладення (за умови $u \in W_2^3(0, 1)$)

$$|\bar{\eta}_i^{(1)}(\tilde{u})| \leqslant 3h^{-1}\|\tilde{u}\|_{C^1[0,1]} \leqslant 3h^{-1}M\|\tilde{u}\|_{W_2^2(0,1)} \leqslant 3Mh^{-1}\|\tilde{u}\|_{W_2^3(0,1)}, \tag{81}$$

де стала M не залежить від $\tilde{u}$ та h. Неважко переконатися, що

$$\bar{\eta}_i^{(1)}(1) = h^{-1}(0 - 1 + 1) = 0,$$

$$\bar{\eta}_i^{(1)}(s) = h^{-1}(1 - 1 + 0) = 0, \quad \bar{\eta}_i^{(1)}(s^2) = h^{-1}(2s|_{s=1/2} - 1 + 0) = 0,$$

$$\bar{\eta}_i^{(1)}(s^3) = h^{-1}(3s^2|_{s=1/2} - 1 + 0) \neq 0.$$

Тому, враховуючи (81), за лемою Брембла — Гільберта, маємо

$$|\bar{\eta}_i^{(1)}(\tilde{u})| \leqslant 3\overline{M}Mh^{-1}|\tilde{u}|_{W_2^3(0,1)},$$

звідки після переходу до координати x дістанемо

$$|\bar{\eta}_i^{(1)}(u)| \leqslant 3\overline{M}Mh^{3/2}|u|_{W_2^3(e_i)}, \quad e_i=(x_{i-1}, x_i), \quad i=\overline{1, N-1}.$$

Оскільки $k(x)\in W_\infty^3(0, 1)$, то

$$|\eta_i^{(1)}| \leqslant Mh^{3/2}\|k\|_{W_\infty^3(0,1)}|u|_{W_2^3(e_i)}, \tag{82}$$

де стала M не залежить від k, h, u.

Величина $\eta_i^{(2)}$ при фіксованому i є функціоналом від $v\equiv ku'$. Якщо $u\in W_2^4(0,1)$, $k\in W_\infty^3(0, 1)$, то $v\in W_2^3(0, 1)$, і аналогічно попередньому за допомогою леми Брембла — Гільберта дістаємо

$$|\eta_i^{(2)}| \leqslant Mh^{3/2}|ku'|_{W_2^3(e_i)} \leqslant Mh^{3/2}\|k\|_{W_\infty^3(0,1)}\|u\|_{W_2^4(e_i)}. \tag{83}$$

Зауважимо, що похибку апроксимації можна подати у вигляді

$$\psi_i=\mu_{x,i}, \tag{84}$$

де $\mu_i=\eta_i^{(1)}+\sum\limits_{k=1}^{i-1}h\eta_k^{(2)}$, причому остання сума дорівнює нулю при $i=1$.

в) Щоб знайти апріорну оцінку, запишемо задачу (79) в індексному вигляді і врахуємо (84):

$$A_iz_{i-1}-C_iz_i+B_iz_{i+1}=h\Delta\mu_i, \quad i=\overline{1, N-1}, \quad z_0=z_N=0,$$

де

$$\Delta\mu_i=\mu_{i+1}-\mu_i, \ A_i=a_i, \ B_i=a_{i+1}, \ C_i=d_i+a_{i+1}+a_i.$$

Неважко помітити, що $a_i\geqslant c_1>0$, $d_i\geqslant 0$, а тому $A_i>0$, $B_i>0$, $C_i-A_i-B_i\geqslant 0$. Тобто виконуються всі умови, за яких справедливий наслідок 6 з теореми 4 (п. 7.2). Тому

$$\|z\|_C\leqslant\frac{2}{c_1}\sum_{k=1}^{N}h|\mu_k|=\frac{2}{c_1}\|\mu\|_{L_1(\omega_h^+)}. \tag{85}$$

Враховуючи вигляд μ_k, а також оцінки (82), (83), далі маємо

$$\|\mu\|_{L_1(\omega_h^+)}=\sum_{k=1}^{N}h|\mu_k|\leqslant\sum_{k=1}^{N}h|\eta_k^{(1)}|+\sum_{k=1}^{N}h\left|\sum_{j=1}^{k-1}h\eta_j^{(2)}\right|\leqslant$$
$$\leqslant\left(\sum_{k=1}^{N}h(\eta_k^{(1)})^2\right)^{1/2}+\left(\sum_{j=1}^{N}h(\eta_j^{(2)})^2\right)^{1/2}\leqslant$$
$$\leqslant Mh^2\|k\|_{W_\infty^3(0,1)}(|u|_{W_2^3(0,1)}+\|u\|_{W_2^4(0,1)}). \tag{86}$$

Із (85), (86) знаходимо шукану оцінку

$$\|z\|_C=\|y-u\|_C\leqslant Mh^2\|k\|_{W_\infty^3(0,1)}(|u|_{W_2^3(0,1)}+\|u\|_{W_2^4(0,1)}). \tag{87}$$

Зауважимо, що для виконання умови $u\in W_2^4(0, 1)$, очевидно, досить, щоб $q\in W_\infty^2(0, 1)$, $f\in W_2^2(0, 1)$, $k\in W_\infty^3(0, 1)$.

Сформулюємо сказане вище у вигляді наступної теореми.

Теорема 2. *Якщо в задачі* (80) $f\in W_2^2(0,1)$, $q\in W_\infty^2(0,1)$, $k\in W_\infty^3(0, 1)$, $u\in W_2^4(0, 1)$, *то розв'язок різницевої схеми* (79) *збігається до точного розв'язку в сітковій нормі* $C(\omega_h)$, *причому має місце оцінка швидкості збіжності* (87).

Вправа 1. За допомогою формули Тейлора довести, що за умови q, $f\in C^2[0, 1]$, $k\in C^3[0, 1]$, $u\in C^4[0, 1]$ для схеми (79) має місце оцінка

$$\|y-u\|_C=O(h^2).$$

Зауважимо, що високі вимоги гладкості даних задачі (80) зумовлені структурою похибки апроксимації, куди входять відповідні похідні. Отже, щоб зменшити вимоги до гладкості, треба будувати схеми, в похибки апроксимації яких входять похідні від шуканої функції (включаючи різницеві) якомога меншого порядку і які допускають апріорні оцінки, де ці похідні певним чином «поглинаються».

Приклад 5 (дослідження збіжності і оцінка похибки в енергетичній нормі). Для задачі

$$-(ku')'+q(x)u(x)=f(x), \quad x\in(0, 1), \quad u(0)=u(1)=0 \tag{88}$$

розглянути різницеву схему (див. п. 4.3.6)

$$-(ay_{\bar{x}})_x+d(x)y(x)=\varphi(x), \quad x\in\omega_h,$$
$$y(0)=y(1)=0, \tag{89}$$

де

$$k(x)\in W_\infty^1(0, 1), \quad q(x)\in L_\infty(0, 1), \quad u(x)\in W_2^2(0, 1),$$
$$k(x)\geqslant c_1>0, \quad q(x)\geqslant 0, \quad f(x)\in L_2(0, 1), \ \omega_h=\{x_i=ih: i=\overline{1, N-1},$$
$$h=1/N\}, \quad a(x)=k\left(x-\frac{h}{2}\right), \quad d(x)=Sq(x), \quad \varphi(x)=Sf(x),$$

і дослідити її збіжність та оцінити похибку в енергетичній нормі.

Розв'язання. а) Запишемо задачу для похибки $z=y(x)-u(x)$, $x\in\omega_h$. Маємо

$$\Lambda z\equiv-(az_{\bar{x}})_x+d(x)z(x)=\psi(x), \quad x\in\omega_h, \quad z(0)=z(1)=0, \tag{90}$$

де $\psi(x)=\varphi(x)-\Lambda u=Sf(x)-\Lambda u$. У силу рівності (54′) похибку апроксимації $\psi(x)$ можна перетворити так:

$$\psi(x)=\eta_x^{(1)}(x)+\eta^{(2)}(x),$$

де

$$\eta^{(1)}(x)=au_{\bar{x}}(x)-k\left(x-\frac{h}{2}\right)u'\left(x-\frac{h}{2}\right)=k\left(x-\frac{h}{2}\right)\bar{\eta}^{(1)}(x),$$
$$\bar{\eta}^{(1)}(x)=u_{\bar{x}}(x)-u'\left(x-\frac{h}{2}\right),$$

$$\eta^{(2)}(x) = S(qu)(x) - S(q)(x)\,u(x) = h^{-1}\int\limits_{x-0,5h}^{x+0,5h} q(\xi)\,u(\xi)\,d\xi -$$

$$- u(x)\,h^{-1}\int\limits_{x-0,5h}^{x+0,5h} q(\xi)\,d\xi = h^{-1}\int\limits_{x-0,5h}^{x+0,5h} q(\xi)\,[u(\xi) - u(x)]\,d\xi.$$

б) Апріорну оцінку для схеми, що досліджується, ми вже фактично встановили в п. 2.2.4. Але оскільки похибка апроксимації має певну структуру, то доцільно цю оцінку «підправити» так, щоб «поглинути» різницеву похідну, що в ній міститься. У даному разі це можливо зробити, що, очевидно, має привести до ослаблення вимог до гладкості шуканого розв'язку. Отже, повторимо метод енергетичних оцінок для задачі (90). Введемо скалярний добуток

$$(y,\ v) = \sum_{i=1}^{N-1} h y_i v_i$$

і помножимо скалярно рівняння (90) на z:

$$-((az_{\bar{x}})_x,\ z) + (dz,\ z) = (\eta^{(1)}_x,\ z) + (\eta^{(2)},\ z).$$

Ліву частину цієї рівності перетворимо і оцінимо так само, як у п. 2.2.4, а в першому доданку справа підсумуємо за частинами і застосуємо нерівність Коші — Буняковського:

$$\frac{c_1}{2}\|z\|_1^2 \leqslant (\eta^{(1)},\ z_{\bar{x}}] + (\eta^{(2)},\ z) \leqslant \|\eta^{(1)}]\|\,\|z_{\bar{x}}]\| + \|\eta^{(2)}\|\,\|z\|.$$

Тут

$$\|z\|_1 = \|z\|_{W_2^1(\omega_h)} = (\|z\|^2 + \|z_{\bar{x}}]\|^2)^{1/2} = \left(\sum_{i=1}^{N-1} h z_i^2 + \sum_{i=1}^{N} h z_{\bar{x},i}^2\right)^{1/2},$$

$$(y,\ v] = \sum_{i=1}^{N} h y_i z_i,\quad \|z_{\bar{x}}]\| \equiv |z|_{W_2^1(\omega_h)} \equiv |z|_1.$$

Враховуючи, що $\|z\|_1 = (\|z\|^2 + |z|_1^2)^{1/2}$, далі знаходимо

$$\frac{c_1}{2}\|z\|_1^2 \leqslant (\|\eta^{(1)}]\| + \|\eta^{(2)}\|)\,\|z\|_1,$$

звідки дістаємо шукану оцінку

$$\|z\|_1 \leqslant 2c_1^{-1}(\|\eta^{(1)}]\| + \|\eta^{(2)}\|). \tag{91}$$

в) Оцінимо величини $\|\eta^{(1)}]\|$ та $\|\eta^{(2)}\|$. Розглянемо функціонал від u (при фіксованому x):

$$\overline{\eta}^{(1)}(x) \equiv \overline{\eta}^{(1)}(x;\ u) \equiv \overline{\eta}^{(1)}(u) = u_{\bar{x}}(x) - u'\left(x - \frac{h}{2}\right),\quad x \in \omega_h.$$

Виконавши заміни $\xi = x + (s-1)h$, $\xi \in [x-h,\ x]$, $s \in [0,\ 1]$, $u(\xi) = u(x + (s-1)h) = \tilde{u}(s)$, перетворимо функціонал $\overline{\eta}^{(1)}$ так:

$$\overline{\eta}^{(1)}(u) = \overline{\eta}^{(1)}(\tilde{u}) = h^{-1}[\tilde{u}(1) - \tilde{u}(0) - \tilde{u}'(1/2)].$$

Застосувавши теореми вкладення, переконаємося, що функціонал $\overline{\eta}^{(1)}$ обмежений у просторі $W_2^2(0,\ 1)$:

$$|\overline{\eta}^{(1)}(\tilde{u})| \leqslant 3h^{-1}\|\tilde{u}\|_{C^1[0,1]} \leqslant 3Mh^{-1}\|\tilde{u}\|_{W_2^2(0,1)}.$$

Цей функціонал перетворюється на нуль на многочленах до другого степеня включно:

$$\eta^{(1)}(1) = h^{-1}[1 - 1 - 0] = 0,\quad \eta^{(1)}(s) = h^{-1}[1 - 0 - 1] = 0,$$

$$\eta^{(1)}(s^2) = h^{-1}[1 - 0 - 2s|_{s=1/2}] = 0,$$

$$\overline{\eta}^{(1)}(s^3) = h^{-1}[1 - 0 - 3s^2|_{s=1/2}] \neq 0.$$

Тому за лемою Брембла — Гільберта

$$|\overline{\eta}^{(1)}(\tilde{u})| \leqslant 3\overline{M}Mh^{-1}|\tilde{u}|_{W_2^2(0,1)},$$

звідки після переходу до старих змінних маємо

$$|\overline{\eta}^{(1;}(u)| \leqslant Mh^{1/2}|u|_{W_2^2(e(x))},\quad x \in \omega_h,\quad e(x) = (x-h,\ x). \tag{92}$$

Функціонал $\eta^{(2)}(x) = \eta^{(2)}(x;\ u)$ оцінимо так:

$$|\eta^{(2)}(x)| = \left|h^{-1}\int\limits_{x-0,5h}^{x+0,5h} q(\xi)\,[u(\xi) - u(x)]\,d\xi\right| =$$

$$= \left|h^{-1}\int\limits_{x-0,5h}^{x+0,5h} q(\xi)\int\limits_x^{\xi}\frac{\partial u}{\partial\xi_1}\,d\xi_1 d\xi\right| \leqslant h^{-1}\int\limits_{x-0,5h}^{x+0,5h} q(\xi)\left|\int\limits_x^{\xi}\left|\frac{\partial u}{\partial\xi_1}\right|d\xi_1\right|d\xi \leqslant$$

$$\leqslant h^{-1}\|q\|_{L_\infty(0,1)}\int\limits_{x-0,5h}^{x+0,5h}\int\limits_{x-0,5h}^{x+0,5h}\left|\frac{\partial u}{\partial\xi_1}\right|d\xi_1 d\xi =$$

$$= \|q\|_{L_\infty(0,1)}\int\limits_{x-0,5h}^{x+0,5h}\left|\frac{\partial u}{\partial\xi}\right|d\xi \leqslant h^{1/2}\|q\|_{L_\infty(0,1)}\left(\int\limits_{x-0,5h}^{x+0,5h}\left(\frac{\partial u}{\partial\xi}\right)^2 d\xi\right)^{1/2} =$$

$$= h^{1/2}\|q\|_{L_\infty(0,1)}|u|_{W_2^1(\tilde{e}(x))},\quad \tilde{e}(x) = (x - 0,5h,\quad x + 0,5h). \tag{93}$$

Із виразів (92), (93) дістаємо

$$\|\eta^{(1)}\| = \left(\sum_{x\in\omega_h^+} h\,(\eta^{(1)}(x^2)\right)^{1/2} \leqslant Mh|u|_{W_2^2(0,1)}, \tag{94}$$

$$\|\eta^{(2)}\| = \left(\sum_{x\in\omega_h} h\,(\eta^{(2)}(x))^2\right)^{1/2} \leqslant Mh|u|_{W_2^1(0,1)}, \tag{95}$$

де $\omega_h^+ = \omega_h \cup \{1\}$; M — стала, що не залежить від u та h.

Тепер з урахуванням (91), (94), (95) можна записати (при $h \leqslant h_0$, h_0 — фіксовано)

$$\| z \|_1 = \| y - u \|_1 \leqslant Mh \left(\| u \|_{W_2^2(0,1)} + | u |_{W_2^1(0,1)} \right) \leqslant Mh \| u \|_{W_2^2(0,1)}. \qquad (96)$$

Сформулюємо доведене у прикладі 3 у вигляді теореми.

Теорема 3. *Нехай розв'язок задачі* (88) *належить простору* $W_2^2(0, 1)$, $k(x) \in W_\infty^1(0, 1)$, $q(x) \in L_\infty(0, 1)$, $k(x) \geqslant c_1 > 0$, $q(x) \geqslant 0$. *Тоді розв'язок різницевої схеми* (88) *при* $h \to 0$ *збігається до точного розв'язку задачі* (88) *в сітковій нормі* $W_2^1(\omega_h)$, *причому має місце оцінка точності* (96).

Оцінки точності різницевих схем вигляду

$$\| y - u \|_{W_2^k(\omega_h)} \leqslant Mh^{s-k} \| u \|_{W_2^s(\Omega)}, \quad s \geqslant k, \qquad (97)$$

де $y(x)$, $x \in \omega_h$ — розв'язок різницевої схеми, $u(x)$, $x \in \Omega$ — розв'язок диференціальної задачі, називаються *узгодженими з гладкістю шуканого розв'язку*. Підставою для такого означення є той факт, що в деяких задачах такі оцінки мають місце для схем методу скінченних елементів (див. п. 4.4), причому в деяких випадках доведено, що вони не можуть бути покращені за порядком h.

Приклад 6 (доведення узгодженої оцінки для норми $L_2(\omega_h)$). Дослідити збіжність і встановити оцінку точності (оцінку швидкості збіжності) в $L_2(\omega_h)$ різницевої схеми (див. п. 4.3.6)

$$\Lambda y \equiv -y_{\overline{x}x} + d(x) y(x) + d_1(x) y_x(x) = \varphi(x), \qquad (98)$$

$$x \in \omega_h, \quad y(0) = y(1) = 0$$

для задачі

$$-u''(x) + q(x) u(x) = f(x), \quad x \in (0, 1), \quad u(0) = u(1) = 0, \qquad (99)$$

де

$$q(x) \subset L_\infty(0, 1), \quad 0 \leqslant q_0 \leqslant q(x) \leqslant q_1, \quad f(x) \in L_2(0, 1),$$

$$u(x) \in W_2^2(0, 1), \quad \omega_h = \{x_i = ih : i = \overline{1, N-1}, \quad h = 1/N\},$$

$$d(x) = Tq(\xi), \quad \varphi(x) = Tf(\xi), \quad x \in \omega_h, \quad d_1(x_i) = T(q(\xi)(\xi - x_i)),$$

$$Tf(\xi) = h^{-2} \int_{x-h}^{x+h} (h - |\xi - x|) f(\xi)\, d\xi, \quad x \in \omega_h.$$

Розв'язання. а) Запишемо задачу для похибки $z = y - u$. Враховуючи рівність (55) з п. 4.3.6, маємо

$$-z_{\overline{x}x} + d(x) z(x) + d_1(x) z_x(x) = \psi(x), \quad x \in \omega_h, \qquad (100)$$

$$z(0) = z(1) = 0,$$

$$\psi(x_i) = \varphi(x_i) - \Lambda u(x_i) = Tf(\xi) - \Lambda u(x) = -u_{\overline{x}x}(x) +$$

$$+ T(qu) + u_{\overline{x}x} - Tq(\xi) u(x) - T(q(\xi)(\xi - x)) u_x(x) =$$

$$= T(qu) - T(q(\xi)(u(x) + (\xi - x) u_x(x))),$$

де ξ — змінна, по якій виконується інтегрування в операторі T.

Як бачимо, застосування усереднюючого оператора T дало змогу позбутися всяких похідних у похибці апроксимації (на $u_x(x)$ можна не зважати, бо перед ним є множник $(\xi - x)$, який має порядок $O(h)$).

Запишемо рівність у вигляді

$$\psi(x) = T(qu) - T(q(\xi)(u(x) + (\xi - x) u_x(x))) =$$

$$= h^{-2} \int_{x-h}^{x+h} (h - |\xi - x|) q(\xi) [u(\xi) - u(x) - (\xi - x) u_x(x)]\, d\xi. \qquad (101)$$

б) Встановимо апріорну оцінку в нормі $L_2(\omega_h)$. Для цього скористаємося таким прийомом (у методі скінченних елементів його аналог іноді називають *прийомом Нітше*). Розглянемо допоміжну задачу

$$\Lambda_0 w = z(x), \quad x \in \omega_h,$$

$$w(0) = w(1) = 0; \quad \Lambda_0 w = -w_{\overline{x}x} + d(x) w(x).$$

Вважатимемо Λ_0 оператором, що діє в просторі сіткових функцій, які перетворюються на нуль у точках сітки 0 та 1, зі скалярним добутком

$$(y, v) = \sum_{x \in \omega_h} hy(x) v(x).$$

Підсумовуючи частинами, дістаємо

$$(\Lambda_0 v, v) = \| v_{\overline{x}} \|^2 + (dv, v) \geqslant \| v_{\overline{x}} \|^2 + q_0 \| v \|^2, \quad d(x) \geqslant q_0.$$

Враховуючи, що $\| v_{\overline{x}} \|^2 = (-v_{\overline{x}x}, v) \geqslant \lambda_1 \| v \|^2$, де $\lambda_1 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2}$ — мінімальне власне значення оператора $Av = -v_{\overline{x}x}$, $v(0) = v(1) = 0$ (див. п. 1.4.4 з ч. 1) і оскільки $\lambda_1 \geqslant 8$, то маємо

$$\| v_{\overline{x}} \|^2 \geqslant 8 \| v \|^2,$$

тому

$$(\Lambda_0 v, v) \geqslant (8 + q_0) \| v \|^2.$$

Неважко помітити, що $(\Lambda_0 v, w) = (v, \Lambda_0 w)$, тобто $\Lambda_0 = \Lambda_0^*$. Тому оператор Λ_0^{-1} також самоспряжений і (див. п. 7.2.4) $\| \Lambda_0^{-1} \| \leqslant (8 + q_0)^{-1}$. Оцінимо величину $d_1(x)$. Маємо

$$d_1(x) = T(q(\xi)(\xi - x)) = h^{-2} \int_{x-h}^{x+h} (h - |\xi - x|) q(\xi)(\xi - x)\, d\xi \leqslant$$

$$\leqslant h \| q \|_{L_\infty(0,1)} \leqslant hq_1, \qquad (102)$$

оскільки $|h - |\xi - x|| \leqslant h$, $|\xi - x| \leqslant h$, $|q(\xi)| \leqslant \| q \|_{L_\infty(0,1)} \leqslant q_1$. Помножимо рівняння (100) скалярно на функцію w і врахуємо, що $w = \Lambda_0^{-1} z$.

Матимемо $(\Lambda_0 z, w) + (d_1 z_x, w) = (\psi, w)$ і далі

$$(\Lambda_0 z, \Lambda_0^{-1} z) + (d_1 z_x, \Lambda_0^{-1} z) = (\psi, \Lambda_0^{-1} z) \leqslant \|\psi\| \|\Lambda_0^{-1} z\| \leqslant \|\psi\| (8 + q_0)^{-1} \|z\|,$$

$$(z, \Lambda_0 \Lambda_0^{-1} z) + (d_1 z_x, \Lambda_0^{-1} z) \leqslant (8 + q_0)^{-1} \|\psi\| \|z\|,$$

$$\|z\|^2 + (d_1 z_x, \Lambda_0^{-1} z) \leqslant (8 + q_0)^{-1} \|\psi\| \|z\|. \qquad (103)$$

Другий доданок у лівій частині останньої нерівності оцінимо за допомогою нерівностей (102) та Коші — Буняковського:

$$|(d_1 z_x, \Lambda_0^{-1} z)| \leqslant q_1 h \|z_x\| \|\Lambda_0^{-1} z\| \leqslant \frac{q_1 h}{8 + q_0} \|z_x\| \|z\|.$$

Враховуючи нерівність $(a \pm b)^2 \leqslant 2(a^2 + b^2)$, маємо

$$h\|z_x\| = \left(h^2 \sum_{i=1}^{N-1} h \left(\frac{z_{i+1} - z_i}{h} \right)^2 \right)^{1/2} \leqslant 2\|z\|.$$

Тому $|(d_1 z_x, \Lambda_0^{-1} z)| \leqslant \frac{2q_1}{8 + q_0} \|z\|^2$ і якщо $q_1 < 4 + \frac{q_0}{2}$, то із (103) дістаємо

$$\left(1 - \frac{2q_1}{8 + q_0}\right) \|z\|^2 \leqslant \|\psi\| \|z\|, \quad \|z\| \leqslant \frac{8 + q_0}{8 + q_0 - 2q_1} \|\psi\|. \qquad (104)$$

в) Оцінимо величину $\|\psi\|$. Розглянемо функціонал $\eta(u) = u(\xi) - u(x) - (\xi - x) u_x(x)$ при фіксованих x, ξ. Неважко переконатися, що він обмежений у просторі $W_2^2(0, 1)$ і перетворюється на нуль у многочленах до першого степеня включно. Тому стандартним шляхом за допомогою леми Брембла — Гільберта можна знайти оцінку

$$|\eta(u)| \leqslant M h^{3/2} |u|_{W_2^2(e(x))}, \quad e(x) = (x - h, x + h),$$

а з виразу (101) матимемо $|\psi(x)| \leqslant M h^{3/2} |u|_{W_2^2(e(x))}$, де M — стала, що не залежить від h та u. Тому

$$\|\psi\| = \left(\sum_{x \in \omega_h} h \psi^2(x) \right)^{1/2} \leqslant M h^2 |u|_{W_2^2(0,1)}, \qquad (105)$$

де стала M також не залежить від h та u.

Із виразів (104), (105) випливає така теорема.

Теорема 4. *Нехай розв'язок задачі* (99) *належить простору* $W_2^2(0, 1)$, $q(x) \in L_\infty(0, 1)$, $0 \leqslant q_0 \leqslant q(x) \leqslant q_1$, $f(x) \in L_2(0, 1)$, *причому* $q_1 < 4 + \frac{q_0}{2}$. *Тоді розв'язок різницевої схеми* (98) *збігається при* $h \to 0$ *до точного розв'язку в сітковій нормі* $L_2(\omega_h)$, *причому має місце узгоджена оцінка швидкості збіжності* $\|y - u\|_{L_2(\omega_h)} \leqslant M h^2 |u|_{W_{2(0,1)}^2}$, *де стала* M *не залежить від* h *та* u.

Зазначимо, що практичне застосування схем, розглянутих в останніх двох прикладах, ускладнюється, бо для знаходження їхніх коефіцієнтів потрібно обчислити інтеграли. Це можливо здебільшого лише

за допомогою квадратурних формул. Остання обставина приводить до значного збільшення обчислювальної роботи, тому ці схеми мають переважно теоретичне значення.

Зазначимо також, що наведені приклади дослідження сіткових схем далеко не вичерпують всіх задач і методів. Часто врахування індивідуальних особливостей задачі дає змогу дістати ефективніші схеми і застосувати методи їхнього дослідження.

4.3.9. Підвищення точності методу сіток та апостеріорні оцінки похибки. Обумовленість різницевих схем. Як випливає з попередніх оцінок, підвищити точність різницевої схеми (за умови її збіжності) завжди можна зменшенням h. Проте при цьому збільшується кількість різницевих рівнянь. Такий метод не завжди відповідає вимозі економності, тобто економії часу відшукання розв'язку із заданою точністю. Тому на практиці часто застосовують також інші методи, наприклад різницеві схеми на нерівномірних сітках, схеми підвищеного порядку точності, розрахунок на послідовності сіток із застосуванням принципу Рунге (див. п. 2.10 з ч. 2) та ін. Коротко розглянемо деякі з цих методів.

Однорідні схеми на нерівномірних сітках. Використання нерівномірних сіток дає змогу емпіричного підвищення точності без збільшення кількості вузлів (а, значить, різницевих рівнянь). Схеми на нерівномірних сітках доцільно застосовувати, коли є попередня інформація про поведінку шуканого розв'язку. Так, в області різкої зміни коефіцієнтів, правої частини рівнянь і точок розриву розв'язку треба згустити сітку. Поблизу точок розриву коефіцієнтів звичайно сітку згущують за законом геометричної прогресії. Щоб дістати попередню інформацію, можна провести спочатку розрахунок на грубій сітці, а вже потім — на спеціальній нерівномірній сітці.

Розглянемо знову задачу

$$(k(x) u')' - q(x) u = f(x), \quad x \in (0, 1), \quad u(0) = \mu_1, \quad u(1) = \mu_2. \qquad (106)$$

Виберемо на відрізку $[0, 1]$ довільну нерівномірну сітку

$$\hat{\omega}_h = \{x_i : i = \overline{0, N}, \ x_0 = 0, \ x_N = 1\}.$$

Різницеву схему на цій сітці можна дістати методом безпосередньої заміни похідних різницевими співвідношеннями (див. п. 4.3.1). Проте, щоб дістати консервативну схему, яка б відображала на сітці закони збереження, застосуємо інтегро-інтерполяційний метод (див. п. 4.3.3). Запишемо рівняння балансу на відрізку $[x_{i-1/2}, x_{i+1/2}]$, $x_{i+1/2} = x_i + \frac{h_{i+1}}{2}$, $h_i = x_i - x_{i-1}$:

$$w_{i+1/2} - w_{i-1/2} - \int_{x_{i-1/2}}^{x_{i+1/2}} q u dx = - \int_{x_{i-1/2}}^{x_{i+1/2}} f(x) dx, \quad w = k u', \qquad (107)$$

і апроксимуємо в ньому інтеграли і похідні таким чином:

$$w_{i-1/2} = (ku')_{i-1/2} \sim a_i \frac{u_i - u_{i-1}}{h_i}, \quad \int_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx \sim \hbar_i \varphi_i,$$

$$\int_{x_{i-1/2}}^{x_{i+1/2}} qu\,dx \sim d_i u_i h_i, \quad \hbar_i = 1/2\,(h_i + h_{i+1}),$$

де a_i, d_i, φ_i — деякі сіткові функції. Як наслідок дістаємо різницеву схему

$$\frac{1}{h_i}\left[a_{i+1}\frac{y_{i+1}-y_i}{h_{i+1}} - a_i \frac{y_i - y_{i-1}}{h_i}\right] - d_i y_i = -\varphi_i,$$

$$i = \overline{1, N-1}, \quad y_0 = \mu_1, \quad y_N = \mu_2.$$

Можна використати, наприклад, найпростіші формули $\varphi_i = f_i$, $d_i = q_i$, а коефіцієнт a_i взяти таким, як на рівномірній сітці, тобто $a_i = k_{i-1/2} + O(h_i^2)$ при $k(x) \in C^2[0, 1]$. Запишемо різницеву схему у вигляді

$$(ay_{\bar{x}})_{\hat{x}} - dy = -\varphi(x), \quad x = x_i \in \hat{\omega}_h; \quad y_0 = \mu_1, \quad y_N = \mu_2. \tag{108}$$

Для похибки $z = y - u$ маємо рівняння

$$(az_{\bar{x}})_{\hat{x}} - dz = -\psi(x), \quad x \in \hat{\omega}_h; \quad z_0 = z_N = 0, \tag{109}$$

де

$$\psi = \Lambda u + \varphi = (au_{\bar{x}})_{\hat{x}} - du + \varphi. \tag{110}$$

Лема 1. Якщо $qu \in C^2[0, 1]$, $f \in C^2[0, 1]$, *то для похибки апроксимації справедлива формула*

$$\psi = \eta_{\hat{x}} + \psi^*, \tag{111}$$

де $\eta_i = (au_{\bar{x}})_i - (ku')_{i-1/2} + h_i^2 (qu - f)_i'/8$, $\psi_i^* = O(\hbar_i^2)$.

Доведення. Запишемо тотожність (107) у вигляді

$$0 = w_{\hat{x},i} - \frac{1}{\hbar_i}\int_{x_{i-1/2}}^{x_{i+1/2}} (qu - f)\,dx, \quad w_i = (ku')_{i-1/2}$$

і віднімемо її від (110):

$$\psi = [(au_{\bar{x}})_i - (ku')_{i-1/2}]_{\hat{x},i} - (du)_i + \varphi_i + \frac{1}{\hbar_i}\int_{x_{i-1/2}}^{x_{i+1/2}} (qu - f)\,dx. \tag{112}$$

Подамо інтеграл у вигляді суми двох інтегралів по відрізках $[x_{i-1/2}, x_i]$ та $[x_i, x_{i+1/2}]$ і розкладемо функцію $\tilde{f} = qu - f$ в околі вузла x_i за формулою Тейлора:

$$\frac{1}{\hbar_i}\int_{x_{i-1/2}}^{x_{i+1/2}} \tilde{f}(x)\,dx = \frac{1}{\hbar_i}\left\{\int_{x_{i-1/2}}^{x_i} [\tilde{f}_i + (x - x_i)\tilde{f}_i']\,dx + \right.$$

$$\left. + O(h_i^3) + \int_{x_i}^{x_{i+1/2}} [\tilde{f}_i + (x - x_i)\tilde{f}_i']\,dx + O(h_{i+1}^3)\right\} =$$

$$= \tilde{f}_i + \frac{1}{8\hbar_i}(h_{i+1}^2 - h_i^2)\tilde{f}_i' + O(\hbar_i^2)$$

(бо $h_i^3 + h_{i+1}^3 < (2\hbar_i)^3$). Виконавши заміну $h_{i+1}^2 \tilde{f}_i' = h_{i+1}^2 \tilde{f}_{i+1}' + O(\hbar_i^2)$, дістанемо

$$\frac{1}{\hbar_i}\int_{x_{i-1/2}}^{x_{i+1/2}} \tilde{f}(x)\,dx = \tilde{f}_i + \frac{1}{8}(h_i^2 \tilde{f}')_{\hat{x},i} + O(\hbar_i^2).$$

Підставивши цей вираз в (112) і враховуючи, що $\tilde{f} = qu - f$, дістанемо (111). Лему доведено.

Перш ніж оцінити η_i, розглянемо величину $\eta_i^{(1)} = (au_{\bar{x}})_i - (ku')_{i-1/2}$, припускаючи, що $k \in C^2[0, 1]$, $u \in C^3[0, 1]$. Скориставшись виразами

$$a_i = k_{i-1/2} + O(h_i^2),$$

$$u_i = u_{i-1/2} + h_i u_{i-1/2}'/2 + h_i^2 u_{i-1/2}''/8 + O(h_i^3),$$

$$u_{i-1} = u_{i-1/2} - h_i u_{i-1/2}'/2 + h_i^2 u_{i-1/2}''/8 + O(h_i^3),$$

$$u_{\bar{x},i} = (u_i - u_{i-1})/h_i = u_{i-1/2} + O(h_i^2),$$

дістаємо

$$(au_{\bar{x}})_i - (ku')_{i-1/2} = (k_{i-1/2} + O(h_i^2))(u_{i-1/2}' + O(h_i^2)) -$$

$$- (ku')_{i-1/2} = O(h_i^2),$$

тобто має місце оцінка

$$\eta_i = Q(h_i^2). \tag{113}$$

Аналогічно тому, як це робилось для різницевої схеми на рівномірній сітці (див. п. 4.3.8, приклад 3), для задачі (109) з похибкою апроксимації виду (111) можна дістати апріорну оцінку (за умов $k(x) \geqslant c_1 > 0$, $q(x) \geqslant 0$)

$$\|z\|_C \leqslant \frac{1}{c_1}\{(1, |\eta|) + (1, |\psi^*|)\}, \tag{114}$$

де $(y, v] = \sum_{i=1}^{N} y_i v_i h_i$. Враховуючи (111), (113), (114), переконуємося, що справедлива наступна теорема.

Теорема 5. *Якщо* k, q, $f \in C^2[0, 1]$, $u \in C^3[0, 1]$, $k(x) \geqslant c_1 > 0$, $q \geqslant 0$, *то різницева схема* (108) *для задачі* (106) *має другий порядок точності на довільній послідовності нерівномірних сіток.*

Якщо коефіцієнт $k(x)$ має розриви першого роду в точках $\xi_i \in [0, 1)$, $i = \overline{1, m}$, то до (106) треба додати умови

$$[u]_{\xi_i} = u(\xi_i + 0) - u(\xi_i - 0) = 0, \quad (ku')(\xi_i - 0) = (ku')(\xi_i + 0) = w_i.$$

Припустивши, що $k(x) \in Q^2[0, 1]$, q, $f \in C^2[0, 1]$ і включивши точки ξ_i, $i = \overline{1, m}$ у множину вузлів нерівномірної сітки $\hat{\omega}_h$ (k), аналогічно можна довести, що схема на нерівномірній сітці матиме другий порядок точності і в класі розривних коефіцієнтів.

Схеми підвищеного порядку точності. Зразу ж зауважимо, що такі схеми доцільно будувати насамперед для рівнянь зі сталими коефіцієнтами, оскільки змінні коефіцієнти призводять до певних технічних труднощів і часто до алгоритмів, які потребують більших затрат. Ми вже зазначали, що такі схеми можна будувати методом невизначених коефіцієнтів. Проте він може привести до розростання шаблону і, як наслідок, до погіршення обчислювальних властивостей матриці системи різницевих рівнянь. Іноді вдається добитися підвищення точності іншим шляхом. Як приклад розглянемо рівняння

$$u'' - qu = -f(x), \quad q = \text{const} > 0.$$

Найпростіша різницева схема на рівномірній сітці має вигляд

$$\Lambda y = y_{\bar{x}x} - dy = -\varphi(x).$$

Виберемо d і φ так, щоб порядок апроксимації був $O(h^4)$ (як ми знаємо, порядок апроксимації обумовлює точність). Похибка апроксимації дорівнює

$$\psi = \Lambda u + \varphi = (\Lambda u - u'') - (d - q)u + \varphi - f =$$
$$= \frac{h^2}{12} u^{(4)} - (d - q)u + \varphi - f + O(h^4).$$

З рівняння маємо $u^{(4)} = qu'' - f'' = q(qu - f) - f'' = q^2 u - qf - f''$, а тому

$$\psi = -\left(d - q - \frac{h^2}{12} q^2\right) u + \varphi - \left(f + \frac{h^2}{12} qf + \frac{h^2}{12} f''\right) + O(h^4).$$

Отже матимемо $\psi = O(h^4)$, якщо покласти $d = q + \frac{h^2}{12} q^2$, $\varphi = f + \frac{h^2}{12}(qf + f'')$. Такий самий порядок апроксимації збережеться, якщо замість f'' використати $f_{\bar{x}x}$, бо $h^2 f'' = h^2 f_{\bar{x}x} + O(h^4)$.

Точні схеми. Для рівняння (106) за різноманітних граничних умов можна побудувати точну триточкову схему (тобто її шаблон містить три точки), розв'язок якої у вузлах довільної сітки збігається з точним розв'язком крайової задачі для диференціального рівняння. Оскільки безпосередньо використовувати точну схему для обчислення не можна (чому, стане зрозуміло пізніше), то на її основі будують так звані *усічені різницеві схеми m-го рангу,* порядок точності яких залежить від m. Число m залежить від нашого вибору, тому точність таких схем можна зробити як завгодно високою. Проілюструємо можливість побудови точної схеми для задачі (106) при $q(x) \equiv 0$. Проінтегрувавши рівняння $(ku')' = -f(x)$ від x_i до x (x_i — вузол довільної нерівномірної сітки $\hat{\omega}_h$), матимемо

$$(ku')(x) - (ku')_i + \int_{x_i}^{x} f(\xi)\, d\xi = 0. \tag{115}$$

Розділимо цю рівність на $k(x)$ і проінтегруємо по x від x_i до x_{i+1}:

$$u_{i+1} - u_i - (ku')_i \int_{x_i}^{x_{i+1}} \frac{dx}{k(x)} + \int_{x_i}^{x_{i+1}} \frac{dx'}{k(x')} \int_{x_i}^{x'} f(\xi)\, d\xi = 0. \tag{116}$$

Проінтегрувавши рівняння (115), поділене на $k(x)$, від x_{i-1} до x_i, дістанемо

$$u_i - u_{i-1} - (ku')_i \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)} + \int_{x_{i-1}}^{x_i} \frac{dx'}{k(x')} \int_{x_i}^{x'} f(\xi)\, d\xi = 0. \tag{117}$$

Позначимо

$$a_i^0 = \left[\frac{1}{h_i} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)}\right]^{-1}.$$

Помножимо вираз (116) на a_{i+1}^0/h_{i+1} і віднімемо від результату рівняння (117), помножене на a_i^0/h_i:

$$\frac{1}{h_i}\left[a_{i+1}^0 \frac{u_{i+1} - u_i}{h_{i+1}} - a_i^0 \frac{u_i - u_{i-1}}{h_i}\right] + \varphi_i = 0,$$

$$\varphi_i = \frac{a_i^0}{h_i h_i} \int_{x_{i-1}}^{x_i} \frac{dx'}{k(x')} \int_{x'}^{x_i} f(t)\, dt + \frac{a_{i+1}^0}{h_{i+1} h_i} \int_{x_i}^{x_{i+1}} \frac{dx'}{k(x')} \int_{x_i}^{x'} f(t)\, dt =$$

$$= \frac{h_i a_i^0}{\hbar_i} \int_{-1}^{0} \frac{ds}{k(x_i + sh_i)} \int_{s}^{0} f(x_i + \lambda h_i)\, d\lambda +$$

$$+ \frac{h_{i+1} a_{i+1}^0}{\hbar_i} \int_{0}^{1} \frac{ds}{k(x_i + sh_{i+1})} \int_{0}^{s} f(x_i + \lambda h_{i+1})\, d\lambda.$$

Разом з граничними умовами це рівняння і являє собою точну схему, яку можна записати у вигляді

$$(a^0 u_{\bar{x}})_{\hat{x},i} + \varphi_i = 0, \quad u_0 = \mu_1, \quad u_N = \mu_2,$$

причому все сказане раніше справедливе для будь-яких кусково-неперервних функцій $k(x)$, $f(x)$.

Практично використати безпосередньо точну схему важко, бо її коефіцієнти виражаються через інтеграли від k та f і обчислення їх потребує використання квадратурних формул.

Використання принципу Рунге. Для підвищення точності і апостеріорної оцінки похибки на практиці часто використовують розрахунок за однією і тією самою схемою на послідовності сіток і потім застосовують принцип Рунге (див. п. 2.10 з ч. 1). Це дає змогу підвищити точність без суттєвого збільшення часу розрахунку.

Нехай для розв'язання різницевої задачі на будь-якій рівномірній сітці з кроком h має місце формула

$$y_i^h = u_i + \alpha(x_i) h^{k_1} + O(h^{k_2}), \quad k_2 > k_1 > 0, \tag{118}$$

де u_i — точний розв'язок, $\alpha(x_i)$ не залежить від h. Знайдемо сіткову функцію $\tilde{y}_i$, задану на сітці $\tilde{\omega}_h$, таку, що

$$\hat{\tilde{y}}_i = u_i + O(h^{k_2}). \tag{119}$$

Для цього розглянемо дві сітки ω_{h_1} і ω_{h_2} з кроками h_1 і h_2, які мають множину спільних вузлів $\tilde{\omega}_h$, і позначимо через $y_i^{h_1}$ та $y_i^{h_2}$ розв'язки різницевої задачі на сітках ω_{h_1} та ω_{h_2} відповідно. Утворимо лінійну комбінацію

$$\tilde{y}_i = \sigma y_i^{h_1} + (1 - \sigma) y_i^{h_2}, \quad x \in \tilde{\omega}_h, \quad \tilde{\omega}_h = \omega_{h_1} \cap \omega_{h_2}. \tag{120}$$

Підставивши сюди (118), дістанемо

$$\tilde{y}_i = u_i + \alpha(x_i)(\sigma h_1^{k_1} + (1 - \sigma) h_2^{k_1}) + O(h^{k_2}).$$

Прирівнюючи до нуля коефіцієнти при $\alpha(x_i)$, знаходимо

$$\sigma = h_2^{k_1}/(h_2^{k_1} - h_1^{k_1}), \tag{121}$$

причому в вузлах $\tilde{\omega}_h$ має місце рівність (119).

Таким чином, щоб підвищити точність сіткового розв'язку, на множині вузлів $\tilde{\omega}_h$, треба розв'язати задачу двічі на сітках ω_{h_1} та ω_{h_2}, і на $\omega_{h_1} \cap \omega_{h_2} = \tilde{\omega}_h$ скласти їхню лінійну комбінацію (108), де σ визначається формулою (121). Наприклад, можна взяти $h_2 = h_1/2$, $h_1 = h$ і тоді $\tilde{\omega}_h = \omega_{h_1}$. Для схеми другого порядку точності маємо $k_1 = 2$, $k_2 = 4$ і $\sigma = -1/3$, $1 - \sigma = 4/3$.

Записавши на $\tilde{\omega}_h$

$$y_i^{h_1} = u_i + \alpha(x_i) h_1^{k_1} + O(h_1^{k_2}),$$

$$y_i^{h_2} = u_i + \alpha(x_i) h_2^{k_1} + O(h_2^{k_2})$$

і вибравши $h_2 = rh_1$, знаходимо вираз для головного члена похибки

$$\alpha(x_i) h_i^{k_1} = \frac{y_i^{h_2} - y_i^{h_1}}{r^{k_1} - 1} + O(h_1^{k_2}).$$

Тобто на сітці $\tilde{\omega}_h$ маємо з точністю до $O(h_1^{k_2})$

$$y_i^{h_1} - u_i \sim \frac{y_i^{h_2} - y_i^{h_1}}{r^{k_1} - 1},$$

що дає апостеріорну оцінку похибки. Її можна використовувати в програмах як критерій досягнення наближеним розв'язком заданої точності.

Обумовленість різницевих схем. Як ми бачили, застосування методу сіток для лінійних задач у звичайних диференціальних рівняннях приводить до необхідності розв'язування системи сіткових рівнянь, яка є системою лінійних алгебраїчних рівнянь відносно наближених значень розв'язку диференціальної задачі у вузлах сітки. Із результатів п. 1.2 7 з ч. 1 випливає, що при чисельному розв'язуванні рівнянь важливу роль відіграє число обумовленості матриці цієї системи. Розглянемо число обумовленості матриці системи лінійних алгебраїчних рівнянь

$$y_{\bar{x}x,i} = -f(x_i), \quad x_i \in \omega_h = \{x_i = ih:\ i = \overline{1, N-1},\ h = 1/N\},$$
$$y(0) = y(1) = 0, \tag{122}$$

яка є різницевою схемою для крайової задачі

$$u'' = -f(x), \quad x \in (0, 1); \quad u(0) = u(1) = 0.$$

Матриця системи (110) після вилучення невідомих y_0, y_N має вигляд

(розмірність її $(N-1)\times(N-1)$)

$$A=-\frac{1}{h^2}\begin{bmatrix} -2 & 1 & 0 & 0 & \dots & 0 & 0 & 0\\ 1 & -2 & 1 & 0 & \dots & 0 & 0 & 0\\ 0 & 0 & -2 & 1 & \dots & 0 & 0 & 0\\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots\\ 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1\\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & -2 \end{bmatrix}.$$

Розглянемо число обумовленості $\operatorname{cond}_2 A=\|A\|_2\|A^{-1}\|_2$. Оскільки матриця A симетрична, то $\|A_2\|=\max\limits_{k=\overline{1,N-1}}|\lambda_k|=|\lambda_{N-1}|$, де λ_k — власні числа A. Але A є також додатно визначеною, тому $\lambda_k>0$, $k=\overline{1,N-1}$, тобто $\|A\|_2=\lambda_{N-1}$. Обернена матриця A^{-1} є також симетричною і додатно визначеною, тому $\|A^{-1}\|_2=\lambda_1^{-1}$. У п. 1.4.4 з ч. 1 визначено $\lambda_k=\frac{4}{h^2}\sin^2\frac{\pi kh}{2}$, отже, при малих h

$$8<\lambda_1=\frac{4}{h^2}\sin^2\frac{\pi h}{2}=\frac{4}{h^2}\left(\frac{\pi h}{2}-\frac{\pi^3h^3}{2^3 3!}+\dots\right)^2\sim\pi^2,$$

$$8<\lambda_{N-1}=\frac{4}{h^2}\cos^2\frac{\pi h}{2}=\frac{4}{h^2}\left(1-\frac{\pi^2h^2}{2^2 2!}+\dots\right)^2\sim\frac{4}{h^2}.$$

Тому маємо

$$\operatorname{cond}_2 A=\frac{\lambda_{N-1}}{\lambda_1}\sim\frac{4}{\pi^2h^2}\xrightarrow[h\to0]{}\infty.$$

Можна довести, що і для інших диференціальних рівнянь другого порядку триточкові різницеві схеми і схеми методу скінченних елементів мають число обумовленості порядку $O(h^{-2})$.

Вправа 3. Довести, що число обумовленості різницевої схеми

$$y_{\bar{x}x\bar{x}x}(x)=f(x),\quad x\in\omega_h=\{x_i=ih:i=\overline{1,N-1}\},$$

$$y_{\bar{x}x}|_{x=0,1}=0,\quad y|_{x=0,1}=0$$

є величиною порядку $O(h^{-4})$.

Для рівнянь вищих порядків число обумовленості матриць різницевих схем і схем методу скінченних елементів росте ще швидше зі зменшенням h, наприклад, для рівнянь четвертого порядку воно є величиною $O(h^{-4})$. Тому при практичному використанні різницевих схем і схем методу скінченних елементів (як, до речі, і в багатьох інших задачах чисельного аналізу) треба задовольнити дві суперечливі умови: 1) щоб збільшити точність різницевої схеми, треба зменшувати крок

сітки h; 2) щоб збільшити стійкість розв'язку різницевої схеми до похибок заокруглень, треба зменшити її число обумовленості, а значить, збільшити h. Проте для багатьох практичних задач досить такої величини кроку h, що ця суперечність не дається взнаки, але її треба мати на увазі.

4.4. Метод скінченних елементів для звичайних диференціальних рівнянь

Метод скінченних елементів (МСЕ) коротко можна визначити як проекційний метод зі спеціальними координатними функціями. Ми бачили, що в алгоритмічному плані проекційно-варіаційні методи зводяться до розв'язування систем лінійних алгебраїчних рівнянь, матриця яких у загальному випадку повністю заповнена. Крім того, для багатьох задач ці системи лінійних алгебраїчних рівнянь недостатньо обумовлені, що при розв'язуванні на ЕОМ призводить до великої обчислювальної похибки. Ця обставина ускладнює застосування проекційно-варіаційних методів.

У наш час інтерес до цих методів значно зріс за рахунок створення фактично нового методу — методу скінченних елементів, який можна розглядати як синтез методів Гальоркіна та методу сіток. Початок МСЕ було покладено у 1943 р. Р. Курантом, який запропонував застосувати у методі Ритца спеціальні координатні функції. Виявилось, що застосування їх до граничних задач для диференціальних рівнянь приводить до систем лінійних алгебраїчних рівнянь, які мають такі самі структури та властивості, що й системи методу скінченних різниць (велика розрідженість матриці, хороші обчислювальні властивості). До цього самого методу пізніше прийшли інженери, які розв'язували задачі будівельної механіки, і він дістав назву «метод скінченних елементів».

Специфіка координатних функцій у МСЕ полягає в тому, що кожна з них «ортогональна» майже всім іншим, крім деякого скінченного числа, яке не залежить від їхньої загальної кількості («ортогональність», розуміється у тому смислі, що $(A\varphi_k,\varphi_j)=0$ при $j\neq k$, де A — деякий додатно визначений оператор, $(,)$ — скалярний добуток; якщо така «ортогональність» має місце, то системою рівнянь у методі Ритца є система з «майже» діагональною матрицею). Із цього випливає, що матриця системи лінійних алгебраїчних рівнянь у методі Ритца, як і матриця в методі скінченних різниць, розріджена, а це дуже важливо при розв'язуванні їх на ЕОМ.

4.4.1. Метод скінченних елементів як проекційний метод із спеціальними координатними функціями. Розглянемо сітку $\bar{\omega}_h=\{x_i=ih:i=\overline{0,N},h=1/N\}$, яка покриває відрізок $[0,1]$. Кожному вузлу сітки x_i поставимо у відповідність неперервну кусково-лінійну

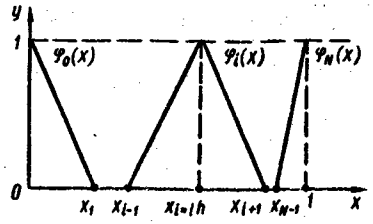


Рис. 5

функцію $\varphi_i(x)$, яка дорівнює одиниці у вузлі x_i і нулю у всіх інших вузлах (рис. 5).

Функції $\varphi_i(x)$ зручно подавати через одну стандартну функцію $\varphi(t)$, яка задана на всій дійсній осі

$$\varphi(t) = \begin{cases} 1 - |t|, & |t| \leqslant 1, \\ 0, & |t| > 1. \end{cases}$$

Легко помітити, що $\varphi_i(x) = \varphi(x/h - i)$, $i = \overline{0, N}$. Очевидно, що вектори $e_i = \{\varphi_i(x_j)\}_{j=0}^{N}$ — лінійно незалежні і утворюють базис в евклідовому просторі $\mathbb{R}^{N+1}$ розмірності $N+1$. Тому будь-яку сіткову функцію $v = \{v_i\}_{i=0}^{N}$, що розглядається як елемент $\mathbb{R}^{N+1}$, можна подати у вигляді

$$v = \sum_{i=0}^{N} v_i e_i,$$

а проекція на сітку неперервної кусково-лінійної функції

$$\tilde{v} = \sum_{i=0}^{N} v_i \varphi_i(x) \qquad (1)$$

збігається з v, причому будь-яку кусково-лінійну (лінійну на відрізках $[x_{i-1}, x_i]$) функцію можна зобразити у вигляді (1) (хвиляста лінія над літерою тут і далі означатиме кусково-лінійне поповнення).

Вправа 1. Показати, що система функцій $\varphi_i(x)$, $i = \overline{0, N}$, лінійно незалежна. Переконатися, що множина неперервних, лінійних на кожному з інтервалів $[x_i, x_{i+1}]$, $i = \overline{0, N-1}$, функцій утворює підпростір простору $W_2^1(0, 1)$.

Оскільки $\varphi_i(x)$, $i = \overline{0, N}$ лінійно незалежні, то, в силу (1), вони утворюють базис у підпросторі неперервних кусково-лінійних функцій, який позначимо через H. Неважко помітити також, що множина неперервних кусково-лінійних функцій, які дорівнюють нулю на кінцях відрізка $[0, 1]$, утворює підпростір в $\mathring{W}_2^1(0, 1)$ з базисом, який складається з функцій $\varphi_i(x)$, $i = \overline{1, N-1}$. Цей підпростір позначимо через $\mathring{H}$.

Звернемось до методу Гальоркіна для операторного рівняння (10) з п. 4. 1. Вибравши у просторі $D(L) = W_{2,0}^2(0, 1) = W_2^2(0,1) \cap \mathring{W}_2^1(0, 1)$ лінійно незалежні функції $\tilde{\varphi}_i(x)$, $i = \overline{1, N-1}$, визначимо тим самим його скінченновимірний підпростір $\mathring{H}$ і в цьому підпросторі шукаємо наближений розв'язок задачі (10) з п. 4.1 у вигляді

$$y_N = \sum_{k=1}^{N-1} c_k \tilde{\varphi}_k(x).$$

Коефіцієнти c_k відповідно до методу Гальоркіна визначаються з умови ортогональності нев'язки $Ly_N - f$ до всіх функцій підпростору $\mathring{H}$, тобто

$$(Ly_N - f, \tilde{\varphi}_i) = 0, \quad i = \overline{1, N-1}.$$

Використовуючи явний вигляд оператора L і інтегруючи частинами, перепишемо цю систему рівнянь у вигляді

$$\int_0^1 \left[k(x) \frac{dy_N}{dx} \frac{d\tilde{\varphi}_i}{dx} + q(x) y_N(x) \tilde{\varphi}_i(x) \right] dx = \int_0^1 f \tilde{\varphi}_i dx$$

або

$$\sum_{k=1}^{N-1} c_k \int_0^1 \left[k(x) \frac{d\tilde{\varphi}_k}{dx} \frac{d\tilde{\varphi}_i}{dx} + q(x) \tilde{\varphi}_k(x) \tilde{\varphi}_i(x) \right] dx = \int_0^1 f \tilde{\varphi}_i(x)\, dx, \qquad (2)$$

$$i = \overline{1, N-1}.$$

Зазначимо, що у такому вигляді система рівнянь Гальоркіна має сенс не тільки для функцій із простору $W_{2,0}^2(0, 1)$, але також і для функцій з простору $\mathring{W}_2^1(0, 1)$, що істотно використовується в методі скінченних елементів.

Вправа 2. Показати, що система рівнянь методу Рітца для задачі (10) з п. 4.1 збігається із системою (2).

Отже, шукатимемо наближений розв'язок задачі (10) з п. 4.1 у підпросторі $\mathring{H}_h$, тобто у вигляді $\tilde{v} = \sum_{i=1}^{N-1} v_i \varphi_i(x)$. Коефіцієнти v_i визначаються як розв'язок системи (2), яка в даному разі набирає вигляду

$$\sum_{i=1}^{N-1} v_i \int_0^1 \left[k \frac{d\varphi_i}{dx} \frac{d\varphi_k}{dx} + q \varphi_i \varphi_k \right] dx = \int_0^1 f \varphi_k dx, \quad k = \overline{1, N-1}. \qquad (3)$$

Неважко помітити, що $v_i = \tilde{v}(x_i)$, тобто невідомі коефіцієнти збігаються зі значенням шуканого наближеного розв'язку у вузлах сітки. Рівняння (3) являє собою систему рівнянь МСЕ.

Вправа 3. Показати, що за умов $k(x) \geqslant k_0 > 0$, $q(x) \geqslant 0$ білінійна форма

$$a(\varphi, \psi) = \int_0^1 \left[k(x) \frac{d\varphi}{dx} \frac{d\psi}{dx} + q(x) \varphi(x) \psi(x) \right] dx$$

визначає в $\mathring{W}_2^1(0, 1)$ скалярний добуток і норму $|\varphi| = \sqrt{a(\varphi, \varphi)}$, яка має назву енергетичної норми граничної задачі (10) з п. 4.1. За яких умов ця норма еквівалентна нормі простору $W_2^1(0, 1)$?

Система (3) має ту чудову властивість, що матриця її розріджена. Це є наслідком того, що система функцій $\varphi_i(x)$ у розумінні скалярного добутку $a(\varphi_i, \varphi_k)$ «майже ортогональна», а саме $a(\varphi_i, \varphi_k) = 0$ при $|i - k| > 1$ та $a(\varphi_i, \varphi_k) \neq 0$ лише при $i = k - 1, k, k + 1$ для $k = \overline{2, N-2}$, а φ_1 та φ_{N-1} неортогональні тільки до функцій φ_2 та φ_{N-2} відповідно. (Важливо, що це має місце для будь-якої кількості координатних функцій, тобто для будь-якого h.)

Вправа 4. Довести, що 1) $a(\varphi_i, \varphi_k) = 0$ при $|i - k| > 1$ та $a(\varphi_i, \varphi_k) \neq 0$ при $i = k - 1, k, k + 1$ $\forall\, k = \overline{2, N-2}$; 2) $a(\varphi_1, \varphi_2) \neq 0$, а $(\varphi_1, \varphi_k) = 0$, $k = \overline{3, N-1}$; 3) $a(\varphi_{N-1}, \varphi_{N-2}) \neq 0$, $a(\varphi_{N-1}, \varphi_k) = 0$, $k = \overline{1, N-3}$;

4) $\int\limits_{x_{k-1}}^{x_k} \varphi_k \varphi_{k-1} dx = h/6$, $\int\limits_{x_{k-1}}^{x_{k+1}} \varphi_k^2 dx = 2h/3$.

Перепишемо систему (3) у вигляді

$$v_{k-1} a(\varphi_{k-1}, \varphi_k) + v_k a(\varphi_k, \varphi_k) + v_{k+1} a(\varphi_{k+1}, \varphi_k) = (f, \varphi_k), \quad k = \overline{1, N-1} \tag{4}$$

і покладемо $v_0 = v_N = 0$. Неважко знайти, що

$$\frac{d\varphi_i}{dx} = \begin{cases} \frac{1}{h}, & x \in (x_{i-1}, x_i), \\ -\frac{1}{h}, & x \in (x_i, x_{i+1}), \\ 0, & x \notin (x_{i-1}, x_{i+1}), \end{cases} \tag{5}$$

тому система (4) має вигляд

$$v_{k-1} \int\limits_{x_{k-1}}^{x_k} [-h^{-2} k(x) + q(x) \varphi_{k-1} \varphi_k] dx + v_k \int\limits_{x_{k-1}}^{x_{k+1}} [h^{-2} k(x) + q(x) \varphi_k^2] dx + v_{k+1} \int\limits_{x_k}^{x_{k+1}} [-h^{-2} k(x) + q(x) \varphi_k \varphi_{k+1}] dx = \int\limits_{x_{k-1}}^{x_{k+1}} f \varphi_k dx, \quad k = \overline{1, N-1}, \quad v_0 = v_N = 0, \tag{4'}$$

при цьому кожне з рівнянь системи (4') природно відповідає вузлу сітки x_k.

При $k(x) = 1$, $q(x) = q \equiv \text{const}$ рівняння (4') запишуться так:

$$h^{-2}[-hv_{k-1} + 2hv_k - hv_{k+1}] + q\frac{h}{6} v_{k-1} + q\frac{2h}{3} v_k + q\frac{h}{6} v_{k+1} = f_k^h, \quad k = \overline{1, N-1},$$

де $f_k^h = \int\limits_{x_{k-1}}^{x_{k+1}} f \varphi_k dx$. Поділивши кожне з цих рівнянь на h, дістаємо

$$\left(-1 + \frac{h^2 q}{6}\right) v_{xx,k} + q v_k = h^{-1} f_k^h, \quad k = \overline{1, N-1}, \quad v_0 = v_N = 0, \tag{6}$$

що дуже близько до різницевої схеми з 9.3.

Задача (6) являє собою систему лінійних алгебраїчних рівнянь з тридіагональною матрицею і її можна розв'язати методом прогонки.

Розглянемо граничну задачу (8), (14) з п. 4.1. Оскільки тепер граничні умови природні, то наближений розв'язок шукатимемо не в $\mathring{H}_h$, а у множині функцій, які не задовольняють спеціально граничні умови, з базисом $\varphi_i(x)$, $i = \overline{0, N}$. Покладаючи $\tilde{v} = \sum\limits_{i=0}^{N} v_i \varphi_i(x)$, де φ_i — кусково-лінійні координатні функції, для визначення невідомих $v_0, v_1, \ldots, v_N$ дістаємо аналогічно попередньому рівняння (4) і, крім того, при $k = 0$, N за умови (4) знаходимо рівняння

$$h^{-2}\left[v_0 \int\limits_0^{x_1} k(x) dx - v_1 \int\limits_0^{x_1} k(x) dx\right] + v_0 \int\limits_0^{x_1} q(x) \varphi_0^2 dx + v_1 \int\limits_0^{x_1} q(x) \varphi_1 \varphi_0 dx = \int\limits_0^{x_1} f \varphi_0 dx,$$

$$h^{-2}\left[-v_{N-1} \int\limits_{x_{N-1}}^{1} k(x) dx + v_N \int\limits_{x_{N-1}}^{1} k(x) dx\right] + v_{N-1} \int\limits_{x_{N-1}}^{1} q(x) \varphi_{N-1} \varphi_N dx + v_N \int\limits_{x_{N-1}}^{1} q(x) \varphi_N^2 dx = \int\limits_{x_{N-1}}^{1} f \varphi_N dx. \tag{7}$$

Систему рівнянь (4), (7), яка має тридіагональну матрицю, після обчислення інтегралів, які визначають її елементи, можна розв'язати методом прогонки.

Вправа 5. Записати розрахункові формули методу прогонки для розв'язування системи лінійних алгебраїчних рівнянь (4), (7).

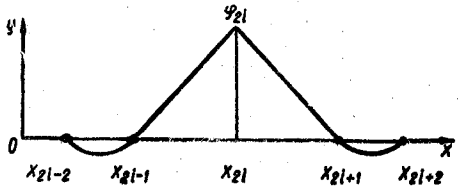


Рис. 6

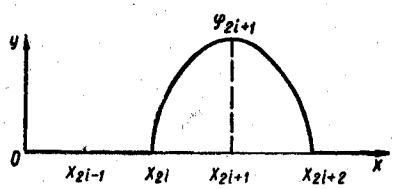


Рис. 7

Наведені вище приклади системи рівнянь МСЕ показують близькість двох підходів: МСЕ і методу скінченних різниць. Однак, вибираючи інші координатні функції, можна дістати схеми МСЕ, схожість яких зі схемами методу скінченних різниць не така очевидна.

У наведених вище прикладах для наближення розв'язку ми застосували кусково-лінійні функції. Цілком зрозуміло, що це не обов'язково і ми можемо будувати схеми МСЕ на основі наближень, які на кожному з інтервалів є поліномами вищого степеня. Від таких апроксимацій природно сподіватись і більшої точності. При цьому можна будувати апроксимації не тільки з простору W_2^1, але й більш гладкі наближення, хоча для рівнянь другого порядку вони застосовуються не настільки часто. Прикладом координатних функцій в W_2^1, відмінних від кусково-лінійних, можуть служити кусково-квадратичні функції, які задаються таким чином. Сітка на відрізку [0, 1] будується так, що $h = 1/(2N)$. Тоді координатні функції, які відповідають вузлам з парними номерами, мають вигляд (рис. 6)

$$\varphi_{2i}(x) = \begin{cases} 1 + \frac{3}{2}\frac{x - x_{2i}}{h} + \frac{1}{2}\frac{(x-x_{2i})^2}{h^2}, & x \in (x_{2i-2},\ x_{2i}), \\ 1 - \frac{3}{2}\frac{x - x_{2i}}{h} + \frac{1}{2}\frac{(x-x_{2i})^2}{h^2}, & x \in (x_{2i},\ x_{2i+2}), \\ 0, & x \notin (x_{2i-2},\ x_{2i+2}), \end{cases}$$

а вузлами з непарними номерами — вигляд (рис. 7)

$$\varphi_{2i+2}(x) = \begin{cases} 1 - \frac{(x - x_{2i-1})^2}{h^2}, & x \in (x_{2i},\ x_{2i+2}), \\ 0, & x \notin (x_{2i},\ x_{2i+2}). \end{cases}$$

Таким чином, може бути багато різних координатних функцій. Вибір їх у кожному конкретному випадку визначається гладкістю початкових даних задачі, а також тією точністю, з якою необхідно розв'язувати задачу.

4.4.2. Метод скінченних елементів як інженерний метод. МСЕ народився також як інженерний метод розрахунку різних конструкцій. Специфічні інженерні засоби та методика надали йому особливої форми, алгоритмічно відмінної від викладеної вище, де МСЕ трактувався як проекційно-варіаційний метод. Широке застосування в інженерному підході матричного формалізму виявилось дуже корисним для конструювання «зручних» алгоритмів та проведення конкретних обчислень на ЕОМ.

Щоб проілюструвати інженерний підхід, розглянемо задачу про розтягування — стиснення стержня сталого перерізу під дією власної ваги за умови, що кінці стержня закріплено. Математична модель цієї задачі має вигляд (8), (9) з п. 4.1, де $u(x)$ — переміщення (прогин), $q(x) \equiv 0$, $k(x) = E$, $f(x) = G$, E — модуль Юнга, а G — зведена об'ємна сила. Еквівалентна задача про мінімум функціонала (11) з п. 4.1 являє собою формулювання принципу мінімуму потенціальної енергії.

Для наближеного розв'язування задачі розіб'ємо стержень, один кінець якого міститься у точці 0, а другий — у точці 1, на N скінченних елементів однакової довжини, точками $x_i = ih$. Вважатимемо, що на кожному елементі $e^{(i)} = [x_{i-1}, x_i]$ наближений розв'язок зображається лінійною функцією. Тим самим для визначення наближеного розв'язку на елементі досить знати його значення у двох точках, наприклад у вузлах x_{i-1} та x_i. Позначимо наближений розв'язок через $\tilde{u}(x)$, тоді на елементі $e^{(i)}$ його можна подати так:

$$\tilde{u}(x) = \tilde{u}_{i-1}\varphi_{i-1}^{(i)}(x) + \tilde{u}_i\varphi_i^{(i)}(x), \quad x \in e^{(i)}, \tag{8}$$

де $\tilde{u}_i = \tilde{u}(x_i)$, $\varphi_{i-1}^{(i)}(x) = (x_i - x)/h$, $\varphi_i^{(i)}(x) = (x - x_{i-1})/h$, функції $\varphi_{i-1}^{(i)}(x)$, $\varphi_i^{(i)}(x)$ мають назву *функцій форми скінченного елемента* e^i.

Нехай

$$v^{(i)} = [\tilde{u}_{i-1}, \tilde{u}_i]^T, \quad \Phi^{(i)} = \Phi^{(i)}(x) = [\varphi_{i-1}^{(i)}(x), \varphi_i^{(i)}(x)],$$

де верхній індекс T означає транспортування. Тоді наближений розв'язок (8) можна записати у вигляді

$$\tilde{u}(x) = \Phi^{(i)}(x)\, v^{(i)} = \Phi^{(i)} v^{(i)}, \quad x \in e^{(i)}. \tag{9}$$

Підставимо розв'язок $\tilde{u}(x)$ у функціонал (11) з п. 4.1 (функціонал подвоєної потенціальної енергії) і подамо його у вигляді

$$J(\tilde{u}) = \sum_{i=1}^{N} J^{(i)}(\tilde{u}),$$

де

$$J^{(i)}(\tilde{u}) = \int_{e^{(i)}} E(d\tilde{u}/dx)^2\, dx - 2\int_{e^{(i)}} G\tilde{u}\, dx$$

(з урахуванням того, що $k(x) = E$, $q(x) \equiv 0$, $f(x) = G$). В силу виразів (8), (9) поздовжня деформація $\varepsilon^{(i)}(x) = d\tilde{u}/dx$ матиме вигляд

$$\varepsilon^{(i)}(x) = B^{(i)}(x)\, v^{(i)} = B^{(i)} v^{(i)},$$

де

$$B^{(i)} = B^{(i)}(x) = \frac{d}{dx}\Phi^{(i)}(x) = h^{-1}[-1,\ 1].$$

Нормальне напруження, яке визначається формулою $\sigma^{(i)} = E\varepsilon^{(i)}$, набирає вигляду $\sigma^{(i)} = EB^{(i)}v^{(i)}$, отже,

$$J_E^{(i)}(\tilde{u}) = \int\limits_{e^{(i)}} E\,(d\tilde{u}/dx)^2\,dx = \int\limits_{e^{(i)}} \sigma^{(i)}\varepsilon^{(i)}\,dx = v^{(i)T}K^{(i)}v^{(i)},$$

де

$$K^{(i)} = \int\limits_{e^{(i)}} B^{(i)T}EB^{(i)}\,dx = \int\limits_{e^{(i)}} \begin{bmatrix} -1 \\ 1 \end{bmatrix} E\,[-1,\ 1]\,h^{-2}dx = \frac{E}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

— *матриця жорсткості* елемента $e^{(i)}$. Через те що $\Phi^{(i)}v^{(i)} = v^{(i)T}\Phi^{(i)T}$, маємо

$$J_G^{(i)}(\tilde{u}) = 2\int\limits_{e^{(i)}} G\tilde{u}dx = 2v^{(i)T}F^{(i)},$$

де

$$F^{(i)} = \int\limits_{e^{(i)}} G\Phi^{(i)T}dx = Gh^{-1}\int\limits_{e^{(i)}} \begin{bmatrix} x_i - x \\ x - x_{i-1} \end{bmatrix} dx = \begin{bmatrix} 1 \\ 1 \end{bmatrix}\frac{Gh}{2},$$

і цей вектор має назву *вектора вузлових сил елемента*. Таким чином, подвоєна потенціальна енергія елемента $e^{(i)}$ на наближеному розв'язку виражається формулою

$$J^{(i)}(\tilde{u}) = J_E^{(i)}(\tilde{u}) + J_G^{(i)}(\tilde{u}) = v^{(i)T}K^{(i)}v^{(i)} - 2v^{(i)T}F^{(i)}, \qquad (10)$$

тобто є квадратичною формою вузлових значень наближеного розв'язку. Неважко помітити, що зв'язок між повним вектором невідомих $U = [\tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}^N]^T$ та вектором невідомих на елементі $e^{(i)}$ задається формулою

$$v^{(i)} = S^{(i)}U, \qquad (11)$$

де $S^{(i)}$ — прямокутна матриця розмірності $2 \times (N+1)$, яка має назву *матриці кінематичних зв'язків*. Вона має вигляд

$$S^{(i)} = \begin{bmatrix} \underbrace{0\ldots0}_{i-1} & 1 & 0 & \underbrace{0\ldots0}_{N-i} \\ \underbrace{0\ldots0}_{i-1} & 0 & 1 & \underbrace{0\ldots0}_{N-i} \end{bmatrix}.$$

З урахуванням умови (11) вираз (10) матиме вигляд

$$J^{(i)}(\tilde{u}) = U^TS^{(i)T}K^{(i)}S^{(i)}U - 2U^TS^{(i)T}F^{(i)},$$

а функціонал $J(\tilde{u})$ запишеться у вигляді

$$J(\tilde{u}) = U^TKU - 2U^TF, \qquad (12)$$

де матриця

$$K = \sum_{i=1}^{N} S^{(i)T}K^{(i)}S^{(i)} \qquad (13)$$

називається *глобальною матрицею жорсткості*, а вектор

$$F = \sum_{i=1}^{N} S^{(i)T}F^{(i)} \qquad (14)$$

є *глобальним вектором навантаження*.

З виразу (12) випливає, що значення потенціальної енергії на наближеному розв'язку є функція $(N+1)$ змінної $u_0, u_1, \ldots, u_N$, необхідною умовою мінімуму якої є перетворення на нуль її перших похідних. Виконавши диференціювання, прирівнявши похідні до нуля та врахувавши симетрію матриці K, після скорочення на множник 2 дістанемо систему рівнянь

$$KU = F. \qquad (15)$$

Для простоти та однорідності алгоритму ми вважали всі елементи однаковими, «забувши» на деякий час, що $e^{(0)}$ та $e^{(N)}$ закріплено на одному з кінців. Останнє означає, що $u_0 = 0$, $u_N = 0$, тобто якраз на ці умови треба замінити перше й останнє рівняння у системі (15). Виключивши потім u_0 та u_N із системи (15), дістанемо систему $(N-1)$-го порядку знову із симетричною матрицею. Таким чином, інженерний варіант МСЕ зводиться до такої послідовності кроків:

1) формування матриці жорсткості K за формулою (13) та глобального вектора навантаження за формулою (14);

2) розв'язання системи лінійних алгебраїчних рівнянь (15).

При цьому дуже важливо, що для формування K та E немає необхідності виконувати відповідні матричні множення у виразах (14), (15) в силу специфіки матриць $S^{(i)}$. Справді, розглянемо випадок $N = 2$, для якого послідовно дістанемо

$$K^{(1)} = \frac{E}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},\quad S^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

$$K^{(1)}S^{(1)} = \frac{E}{h}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \end{bmatrix},\quad S^{(1)T} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix},$$

$$S^{(1)T}K^{(1)}S^{(1)} = \frac{E}{h}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$K^{(2)} = \frac{E}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad S^{(2)} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$K^{(2)}S^{(2)} = \frac{E}{h}\begin{bmatrix} 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix},$$

$$S^{(2)T}K^{(2)}S^{(2)} = \frac{E}{h}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix},$$

$$K = \sum_{i=1}^{2} S^{(i)T}K^{(i)}S^{(i)} = \frac{E}{h}\begin{bmatrix} 1 & 1 & 0 \\ -1 & 1+1 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

З наведених формул випливає, що множення матриці $K^{(i)}$ на $S^{(i)}$ і $S^{(i)T}$ зводиться до відповідного оточення її нульовими рядками і стовпчиками зі збереженням симетрії. Аналогічних змін зазнає вектор $F^{(i)}$ (виконайте послідовно перетворення). Це дає змогу без виконання багатьох арифметичних операцій побудувати K і F, що і робиться при реальних розрахунках на ЕОМ.

4.4.3. Збіжність і оцінки похибки методу скінченних елементів. Збіжність методу скінченних елементів при $h \to 0$ і його точність (по h) можна досліджувати у дискретних сіткових нормах на основі загальної методики методу сіток і теореми Лакса (див. гл. 2), тобто для схеми методу скінченних елементів дістати апріорну оцінку (нерівність коректності схеми) в деякій дискретній нормі, записати задачу для похибки і оцінити похибку апроксимації, а потім повторити висновок теореми Лакса. Однак, з іншого боку, метод скінченних елементів належить до проекційно-варіаційних методів і внаслідок його застосування знаходимо розв'язок у вигляді деякої функції неперервного аргументу. Тому для методу скінченних елементів було розвинуто власну методику дослідження на основі так званих апроксимаційних нерівностей. Викладемо коротко суть цієї методики на прикладі застосування методу до задачі (10) з п. 4.1. Сіткова схема методу скінченних елементів має вигляд системи (4), наближений розв'язок якої для всіх $x \in [0, 1]$ можна записати у вигляді (1).

Наближений розв'язок $\tilde{v}(x)$ є розв'язком такої задачі мінімізації:

$$J(\tilde{v}) = \min_{\tilde{w} \in \mathring{H}} J(\tilde{w}), \quad J(\tilde{w}) = a(\tilde{w}, \tilde{w}) - 2(f, \tilde{w}),$$

яка еквівалентна задачі (4'). Хвиляста лінія означає, що відповідна функція є кусково-лінійною. Інтегруючи частинами, неважко впевнитись, що

$$J(\tilde{w}) = \int_0^1 \left[k(x) \left(\frac{d(u-\tilde{w})}{dx} \right)^2 + q(x)(u-\tilde{w})^2 \right] dx -$$

$$- \int_0^1 \left[k(x) \left(\frac{du}{dx} \right)^2 + q(x) u^2 \right] dx = a(u-\tilde{w}, u-\tilde{w}) - a(u, u),$$

де u — точний розв'язок задачі (10) з п. 4.1 з простору $\mathring{W}_2^1$. Оскільки $a(u, u)$ не залежить від $\tilde{w}$, то $\tilde{v}$ також є розв'язком задачі

$$a(u-\tilde{v}, u-\tilde{v}) = \min_{\tilde{w} \in \mathring{H}} a(u-\tilde{w}, u-\tilde{w}). \tag{16}$$

Ми дістали дуже важливе співвідношення. Воно означає, що наближений розв'язок є найкращою в розумінні енергетичної норми (тобто норми $(a(v, v))^{1/2}$, яка еквівалентна нормі простору $\mathring{W}_2^1$) апроксимацією точного розв'язку кусково-лінійними функціями. Отже, задача оцінки похибки методу скінченних елементів (а також інших проекційних методів) зводиться до класичної задачі апроксимації функцій, бо з (16) маємо

$$a(u-\tilde{v}, u-\tilde{v}) \leqslant a(u-\tilde{u}, u-\tilde{u}) \quad \forall \tilde{u} \in \mathring{H} \tag{17}$$

($\tilde{u}$ — будь-яка кусково-лінійна функція з $\mathring{H}$). Оцінимо величину $a(u-\tilde{u}, u-\tilde{u})$ за умови припущення, що $u \in W_2^2 \cap \mathring{W}_2^1$. Насамперед запишемо

$$a(u-\tilde{u}, u-\tilde{u}) = \int_0^1 \left[k(x) \left(\frac{d(u-\tilde{u})}{dx} \right)^2 + q(x)(u-\tilde{u})^2 \right] dx \leqslant$$

$$\leqslant \max_{x \in (0,1)} (k(x), q(x)) \|u-\tilde{u}\|_1^2 =$$

$$= \max_{x \in (0,1)} (k(x), q(x)) \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} \left[\left(\frac{d(u-\tilde{u})}{dx} \right)^2 + (u-\tilde{u})^2 \right] dx =$$

$$= \max \left(\|k\|_{C[0,1]}, \|q\|_{C[0,1]} \right) \sum_{i=0}^{N-1} \|u-\tilde{u}\|_{1,e_i}^2, \tag{18}$$

де

$$\|v\|_{1,e_i} = \|v\|_{W_2^1(e_i)}, \quad e_i = (x_i, x_{i+1}), \quad \|v\|_k = \|v\|_{k,(0,1)} = \|v\|_{W_2^k(0,1)}.$$

На кожному відрізку $[x_i, x_{x+1}]$ розглянемо величину $u(x) - \tilde{u}(x) \equiv F(u)$, яка, очевидно, при фіксованому x є лінійним функціоналом від u. Для того щоб сталі в наступних оцінках не залежали від h, виконаємо заміну змінних $x = x_i + sh$, $x \in [x_i, x_{i+1}]$, $s \in [0, 1]$ і позначимо $u(x) = \bar{u}(s) = u(x_i + sh)$. Тоді, оскільки

$$\tilde{u}(x) = u_i \frac{x_{i+1} - x}{h} + u_{i+1} \frac{x - x_i}{h},$$

то

$$F(u) = u(x) - \tilde{u}(x) = \bar{u}(s) - \tilde{\bar{u}}(s) = \bar{u} - \tilde{u}(x_i + sh) =$$

$$= \bar{u}(s) - \bar{u}(0)(1 - s) - \bar{u}(1)s \equiv \bar{F}(\bar{u}),$$

і далі маємо оцінку

$$|F(u)| = |\bar{F}(\bar{u})| \leqslant 3\|\bar{u}\|_{C[0,1]}.$$

За теоремою вкладення (див. п. 14.1)

$$|F(u)| = |\bar{F}(\bar{u})| \leqslant 3M\|\bar{u}\|_{2,(0,1)},$$

де стала M з теореми вкладення не залежить від $\bar{u}$ (отже і від u) та від h. Остання нерівність означає обмеженість лінійного функціонала $F(u) \equiv \bar{F}(\bar{u})$ в просторі $W_2^2(0, 1)$. Неважко помітити, що $\bar{F}(\text{const}) = \bar{F}(s) = 0$, тобто лінійний обмежений в $W_2^2(0, 1)$ функціонал $\bar{F}(u)$ перетворюється на нуль у многочленах до першого степеня включно. Тому за лемою Бремблa — Гільберта (див. п. 1.5 з ч. 1)

$$|F(u)| = |\bar{F}(\bar{u})| \leqslant 3M\bar{M}\,|\bar{u}|_{2,(0,1)},$$

де стала $\bar{M}$ не залежить від $\bar{u}$ (отже, і від u) та від h. Переходячи до змінної x і враховуючи, що $\frac{d^2\tilde{u}}{ds^2} = h^2 \frac{d^2u}{dx^2}$, $ds = h^{-1}dx$, дістаємо

$$|F(u)| \leqslant 3M\bar{M}\left(\int_0^1 \left(\frac{d^2\tilde{u}(s)}{ds^2}\right)^2 ds\right)^{1/2} =$$

$$= 3M\bar{M}h^{3/2}\left(\sum_{x_i}^{x_{i+1}} \left(\frac{d^2u}{dx^2}\right)^2 dx\right)^{1/2} = 3M\bar{M}h^{3/2}|u|_{W_2^2}(e_i). \tag{19}$$

Розглянемо аналогічно функціонал $F_1(u) \equiv \frac{du(x)}{dx} - \frac{d\tilde{u}(x)}{dx} =$

$= \frac{du(x)}{dx} - u_x(x_i)$. Після заміни $x = x_i + sh$ маємо

$$F_1(u) = \bar{F}_1(\bar{u}) = h^{-1}\left(\frac{d\bar{u}(s)}{ds} - \bar{u}(1) + \bar{u}(0)\right),$$

і далі з використанням теорем вкладення

$$|F_1(u)| = |\bar{F}_1(\bar{u})| \leqslant h^{-1}\left(\left\|\frac{d\bar{u}(s)}{ds}\right\|_{C[0,1]} + 2\|\bar{u}\|_{C[0,1]}\right) \leqslant h^{-1}M_1\|\bar{u}\|_{2,(0,1)},$$

де стала M_1 не залежить від u і від h. Неважко помітити, що $\bar{F}_1(\bar{u})$ перетворюється на нуль, якщо $\bar{u}$ є многочленом нульового або першого степеня. Тому за лемою Бремблa — Гільберта

$$|F_1(u)| = |\bar{F}_1(\bar{u})| \leqslant h^{-1}M_1\bar{M}\,|\bar{u}|_{2,(0,1)}$$

і після переходу до старих змінних

$$|F_1(u)| = |\bar{F}_1(\bar{u})| \leqslant h^{1/2}M_1\bar{M}\,|\bar{u}|^2_{2,e_i}. \tag{20}$$

Із виразів (19), (20) дістаємо

$$\int_{x_i}^{x_{i+1}} \left(\frac{du}{dx} - \frac{d\tilde{u}}{dx}\right)^2 dx \leqslant M_1\bar{M}h^2\,|u|^2_{2,e_i},$$

$$\int_{x_i}^{x_{i+1}} (u - \tilde{u})^2\,dx \leqslant 3M\bar{M}h^4\,|u|^2_{2,e_i}, \tag{21}$$

тому

$$\int_{x_i}^{x_{i+1}} \left[\left(\frac{d(u - \tilde{u})}{dx}\right)^2 + (u - \tilde{u})^2\right] dx \leqslant M_2h^2\,|u|^2_{2,e_i},$$

$$M_1\bar{M} + 3M\bar{M}h^2 \leqslant M_2,$$

M_2 не залежить від u і від h (при $h \leqslant h_0$). Зрештою, маємо

$$\|u - \tilde{u}\|_1^2 = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} \left[\left(\frac{d(u - \tilde{u})}{dx}\right)^2 + (u - \tilde{u})^2\right] dx \leqslant$$

$$\leqslant M_2h^2 \sum_{i=0}^{N-1} |u|^2_{2,e_i} = M_2h^2 \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} \left(\frac{d^2u(x)}{dx^2}\right)^2 dx =$$

$$= M_2h^2 \int_0^1 \left(\frac{d^2u(x)}{dx^2}\right)^2 dx = M_2h^2\,|u|^2_{2,(0,1)}, \tag{22}$$

де стала M_2 не залежить від h і від u. Нерівність виду (22) носить назву *апроксимаційної нерівності* і показує, наскільки добре за порядком h можна наблизити довільну функцію $u \in W_2^2(0, 1)$ інтерполяційною кусково-лінійною функцією. Тепер з (17), (18), (22) знаходимо оцінку

похибки методу скінченних елементів

$$a(u-\tilde{v},\ u-\tilde{v}) \leqslant M_2 \max(\|k\|_{C[0,1]},\ \|q\|_{C[0,1]})\, h^2 |u|^2_{2,(0,1)}, \qquad (23)$$

яка також свідчить, що при $h \to 0$ наближений розв'язок $\tilde{v}$ прямує до точного u в енергетичній нормі $\|v\|_a = [a(v,\ v)]^{1/2}$ з швидкістю $O(h)$. Якщо $k(x) \geqslant k_0 > 0$; $q \geqslant 0$, то

$$a(v,\ v) \geqslant k_0 |v|^2_1,$$

а для функцій $v(x) \in \mathring{W}^1_2(0,1)$ маємо

$$|v(x)| = \left|\int_0^x \frac{dv}{dx}\, dx\right| \leqslant \int_0^1 \left|\frac{dv}{dx}\right| \leqslant \left(\int_0^1 \left(\frac{dv}{dx}\right)^2 dx\right)^{1/2} = |v|_1,$$

$$\int_0^1 v^2(x)\, dx = |v|^2_{L_2(0,1)} \leqslant |v|^2_1,$$

$$\|v\|^2_1 \leqslant 2|v|^2_1 \equiv 2|v|^2_{\mathring{W}^1_2(0,1)}.$$

Тобто за умови (23) маємо оцінку

$$\|u-\tilde{v}\|_1 \leqslant \sqrt{2}\,|u-\tilde{v}|_1 \leqslant \sqrt{2}\,k_0^{-1/2}\,[a(u-\tilde{v},\ u-\tilde{v})]^{1/2} \leqslant$$

$$\leqslant M_3 h |u|_2, \quad M_3 = \sqrt{2}\,k_0^{-1/2} \max(\|k\|_C,\ \|q\|_C). \qquad (24)$$

Оцінки похибки виду

$$\|u-\tilde{v}\|_{W^k_2(\Omega)} \leqslant M h^{s-k} \|u\|_{W^s_2(\Omega)}, \quad s \geqslant k, \qquad (25)$$

є характерними для методу скінченних елементів і називаються *оцінками, узгодженими з гладкістю розв'язку* u. Якщо в лівій частині похибка вимірюється в сітковій нормі, то оцінкою, узгодженою з гладкістю розв'язку u, є оцінка

$$\|u-\tilde{v}\|_{W^k_2(\omega_h)} \leqslant M h^{s-k} \|u\|_{W^s_2(\Omega)}, \quad s \geqslant k. \qquad (26)$$

Для деяких схем методу скінченних елементів доведено, що оцінка (25) за порядком h не може бути покращена.

Покажемо, як можна дістати оцінку виду (25) для випадку $k = 0$ (тобто у нормі $|u-\tilde{v}\|_{L_2(0,1)} \equiv \|u-\tilde{v}\|_0$). На думку про можливість такої оцінки наводить той факт, що оцінка (24) (точніше порядок h в ній) визначалася порядком апроксимації функції $u \in W^2_2(0,1)$ кусково-лінійними функціями у нормі $W^1_2(0,1)$, а з нерівностей (21) випливає, що порядок апроксимації $u \in W^2_2(0,1)$ кусково-лінійними функціями в нормі $\|u-\tilde{u}\|_0$ має бути $\in O(h^2)$.

Згадаємо, що початкова задача (10) з п. 4.1 може бути записана у формі

$$a(u,\ v) = (f,\ v)\ \forall\, v \in \mathring{W}^1_2,$$

а наближений розв'язок $\tilde{v}(x) \in \mathring{H}_h$ задовольняє рівність (див. (3))

$$a(\tilde{v},\ \tilde{\varphi}) = (f,\ \tilde{\varphi})\ \forall\, \tilde{\varphi} \in \mathring{H}.$$

Вибравши $v = \tilde{\varphi}$, з останніх двох рівностей неважко дістати

$$a(u-\tilde{v},\ \tilde{\varphi}) = 0\ \forall\, \tilde{\varphi} \in \mathring{H}. \qquad (27)$$

Враховуючи цю рівність, маємо

$$a(u-\tilde{v},\ \Phi) = a(u-\tilde{v},\ \Phi-\tilde{\varphi})\ \forall\, \Phi \in \mathring{W}^1_2(0,\ 1),\ \tilde{\varphi} \in \mathring{H}. \qquad (28)$$

Візьмемо за Φ розв'язок задачі

$$L\Phi = u-\tilde{v}, \quad x \in (0,\ 1); \quad \Phi(0) = \Phi(1) = 0.$$

Неважко показати, що коли $f \in L_2(0,\ 1)$, то для розв'язку задачі $Lu = f$, $x \in (0,\ 1)$; $u(0) = u(1) = 0$ має місце оцінка

$$\|u\|_2 \equiv \|u\|_{W^2_2(0,1)} \leqslant c\|f\|_0,$$

де стала c не залежить від u (покажіть!). Тому

$$\|\Phi\|_2 \leqslant c\|u-\tilde{v}\|_0. \qquad (29)$$

Із узагальненої постановки задачі для Φ маємо

$$a(\Phi,\ \psi) = (u-\tilde{v},\ \psi)\ \forall\, \psi \in \mathring{W}^1_2(0,\ 1).$$

Підставляючи замість ψ у цю рівність функцію $u-\tilde{v}$ і враховуючи симетричність форми $a(u,\ v) = a(v,\ u)$, дістаємо

$$a(u-\tilde{v},\ \Phi) = (u-\tilde{v},\ u-\tilde{v}) = \|u-\tilde{v}\|^2_0.$$

Тепер (28) можна записати таким чином:

$$\|u-\tilde{v}\|^2_0 = a(u-\tilde{v},\ \Phi-\tilde{\varphi})\ \forall\, \tilde{\varphi} \in \mathring{H}.$$

Звідси, використовуючи нерівність Коші—Буняковського, знаходимо

$$\|u-\tilde{v}\|^2_0 \leqslant c\|u-\tilde{v}\|_1 \|\Phi-\tilde{\varphi}\|_1, \qquad (30)$$

де стала c залежить лише від k, q. Покладемо тепер $\tilde{\varphi} = \tilde{\Phi}$. Тоді, як

випливає з апроксимаційної нерівності (22), із (29) матимемо

$$\|\Phi - \tilde{\Phi}\|_1 \leqslant \sqrt{M_2}\, h\, |\Phi|_2 \leqslant c_2 h \|u - \tilde{v}\|_0.$$

Підставивши цю нерівність у (30), дістанемо

$$\|u - \tilde{v}\|_0^2 \leqslant c_3 h \|u - \tilde{v}\|_1 \|u - \tilde{v}\|_0.$$

Скоротивши цю нерівність на $\|u - \tilde{v}\|_0$ і застосувавши оцінку (24), остаточно знайдемо

$$\|u - \tilde{v}\|_0 \leqslant c_4 h^2 |u|_{W_{2(0,1)}^2} \leqslant c_4 h^2 \|u\|_2,$$

де стала c_4 не залежить від u та h. Як неважко помітити, ця оцінка також є узгодженою з гладкістю шуканого розв'язку $u \in W_2^2(0, 1)$.

4.5. Методи без насичення точності розв'язування крайових задач для звичайних диференціальних рівнянь

Розглянуті в попередніх двох параграфах схеми мали ту особливість, що порядок збіжності їх залишався обмеженим при зростанні гладкості шуканого розв'язку. Виникає запитання: чи можна побудувати такі сіткові схеми, порядок точності яких автоматично реагує на гладкість розв'язку і тим більший, чим більшу гладкість має шуканий розв'язок? Такі схеми називатимемо схемами без насичення точності. У зарубіжній літературі вказані методи (схеми) називають *спектральними* (див., наприклад, книгу C. Canuto, M. Y. Hussaini, A. Quarteroni, T. A. Zang. Spectral methods in fluid Dynamics. Springer — Verlag, 1988). Природно чекати, що побудувати такі схеми можна на основі апроксимації функцій, точність яких не має насичення і які можна здійснити конструктивно. Такими апроксимаціями є, наприклад, апроксимація відрізком ряду Фур'є і апроксимація інтерполяційним многочленом.

Розглянемо сіткові схеми на основі цих апроксимацій для задачі

$$-\frac{d^2u}{dx^2} + q(x)u = f(x), \quad x \in (0, 1),$$
$$u(0) = u(1) = 0. \tag{1}$$

До такого вигляду завжди можна привести довільне рівняння

$$\alpha(x)u''(x) + \beta(x)u'(x) + \gamma(x)u(x) = \varphi(x), \tag{2}$$
$$\alpha(x) \geqslant \alpha_0 > 0.$$

Дійсно, візьмемо $u(x) = \mu(x)\, v(x)$, де $v(x)$ — нова невідома функція;

$\mu(x)$ визначимо пізніше. Тоді рівність (2) набере вигляду

$$\alpha(x)\mu(x)v'' + [2\alpha(x)\mu'(x) + \beta(x)\mu(x)]v'(x) + [\gamma(x)\mu(x) + \beta(x)\mu'(x) + \alpha(x)\mu''(x)]v(x) = \varphi(x). \tag{3}$$

Знайдемо $\mu(x)$ з умови

$$2\alpha(x)\mu'(x) + \beta(x)\mu(x) = 0,$$

або

$$\frac{\mu'(x)}{\mu(x)} = -\frac{\beta(x)}{2\alpha(x)}.$$

Звідси

$$\ln \mu(x) = -\frac{1}{2}\int_0^x \frac{\beta(t)}{\alpha(t)}\,dt,$$

$$\mu(x) = \exp\left(-\frac{1}{2}\int_0^x \frac{\beta(t)}{\alpha(t)}\,dt\right).$$

Розділивши вираз (3) на $\alpha(x)\mu(x)$, дістанемо рівняння виду (1).

4.5.1. Найкращі сіткові схеми з точним спектром. Ідея побудови цих схем опирається на апроксимацію відрізком ряду Фур'є за власними функціями оператора $L_0u = -u''(x)$, $u(0) = u(1) = 0$. Неважко переконатися, що ці власні функції (тобто розв'язки задачі $L_0u = \lambda u$) мають вигляд

$$u_k(x) = \sqrt{2}\sin k\pi x \tag{4}$$

і відповідають власним значенням

$$\lambda_k = k^2\pi^2, \quad k = 1, 2, 3, \ldots. \tag{5}$$

Власні функції (4) утворюють ортонормовану систему в тому розумінні, що

$$\int_0^1 u_k(x)\,u_m(x)\,dx = \delta_{km},$$

де $\delta_{km} = \begin{cases} 0, & k \neq m \\ 1, & k = m \end{cases}$ — символ Кронекера. Позначимо через $f(x)$ функцію, знайдену непарним продовженням функції $f(x)$, $f(0) = f(1) = 0$ на відрізок $[-1, 0]$ і потім періодичним поширенням з відрізка $[-1, 1]$ на всю числову вісь. Через $S_{N-1}(x)$ позначимо частинну суму ряду Фур'є, тобто

$$S_{N-1}(x;\, f) \equiv S_{N-1}(x) = \sum_{k=1}^{N-1} a_k \sqrt{2}\sin k\pi x,$$

$$a_k = \sqrt{2}\int_0^1 f(x)\sin k\pi x\,dx.$$

Відомо, що коли $\tilde{f}(x) \in C^l(-\infty, +\infty)$, то

$$|R_{N-1}(x; f)| \equiv |f(x) - S_{N-1}(x; f)| < c \frac{\ln N}{N^l} \sup_x |f^{(l)}(x)|, \quad x \in [0, 1], \tag{6}$$

де c — стала, незалежна від n та f; $R_{N-1}(x; f) = \sum_{k=N}^{\infty} a_k \sqrt{2} \sin k\pi x$ — залишок ряду Фур'е.

Нехай сітка $\bar{\omega}_h^* = \{x_i : i = \overline{0, N},\ x_i \neq x_j$ при $i \neq j,\ x_0 = 0, x_N = 1\}$ покриває відрізок $[0, 1]$, $\omega_h^* = \bar{\omega}_h^* \setminus \{0, 1\}$. Позначимо

$$U = \sqrt{2} \begin{pmatrix} \sin \pi x_1 & \dots & \sin (N-1)\pi x_1 \\ \sin \pi x_2 & \dots & \sin (N-1)\pi x_2 \\ \dots & \dots & \dots \\ \sin \pi x_{N-1} & \dots & \sin (N-1)\pi x_{N-1} \end{pmatrix},$$

$$\Lambda = \begin{pmatrix} \pi^2 & & & \\ & 2^2\pi^2 & & 0 \\ & & \ddots & \\ & & & \\ 0 & & & (N-1)^2\pi^2 \end{pmatrix}.$$

Тоді, очевидно, матриця

$$C = U\Lambda U^{-1}$$

має власні значення $\lambda_k = k^2\pi^2$, які відповідають власним векторам $y_k = \sqrt{2}(\sin k\pi x_i)_{i=\overline{1,N-1}} = (y_k(x_i))_{i=\overline{1,N-1}}$, $k = \overline{1, N-1}$, тобто власні значення матриці C збігаються з першими $N-1$ власними значеннями оператора L_0, а її власні вектори є проекцією на сітку ω_h перших $N-1$ власних функцій оператора L_0. Це дає підставу назвати матрицю C *найкращою сітковою схемою з точним спектром* для диференціального оператора L_0 і вважати її «сітковим наближенням» для L_0.

Якщо сітка ω_h^* — рівномірна, тобто

$$\omega_h^* = \omega_h = \{x_i = ih : i = \overline{1, N-1},\quad h = 1/N\}, \tag{7}$$

то неважко переконатися, що $U = U^*$ (матриця U симетрична) і

$$(y_k, y_m) = \sum_{j=1}^{N-1} h2 \sin k\pi x_j \sin m\pi x_j = \delta_{km}, \tag{8}$$

а тому $U^{-1} = hU$ і

$$C = P\Lambda P, \quad P = \sqrt{h}\, U = \sqrt{\frac{2}{N}} \left(\sin \frac{ij\pi}{N}\right)_{i,j=\overline{1,N-1}},$$

$$P = P^* = P^{-1}. \tag{9}$$

Таким чином, у цьому випадку матриця C легко обчислюється. Введемо матрицю

$$Q = \begin{pmatrix} q(x_1) & & 0 \\ & \ddots & \\ 0 & & q(x_{N-1}) \end{pmatrix} \tag{10}$$

і вектори $y = (y_1, \dots, y_{N-1})^T$, $\varphi = (f(x_1), \dots, f(x_{N-1}))^T$. Задачі (1) на сітці ω_h поставимо у відповідність сіткову схему

$$Cy + Qy = \varphi, \tag{11}$$

де вектор y вважатимемо наближенням для функції $u(x)$ на сітці ω_h.

Дослідимо збіжність схеми (11) і порядок похибки $z = y - u$, $u = (u(x_1), \dots, u(x_N))^T$. Для z маємо задачу

$$Cz + Qz = \psi, \tag{12}$$

де похибка апроксимації ψ визначається формулами

$$\psi = (\psi_i)_{i=\overline{1,N-1}},$$

$$\psi_i = f(x_i) - [(C+Q)u]_i = -u''(x_i) + q(x_i)u(x_i) - [(C+Q)u]_i = -u''(x_i) - (Cu)_i.$$

Оцінку похибки апроксимації дає така лема.

Лема 1. *Нехай* $\tilde{u}(x) \in C^l(-\infty, \infty)$. *Тоді*

$$\|\psi\| \leqslant M \frac{\ln N}{N^{l-2}} \|u^{(l)}\|_{C[0,1]}, \tag{13}$$

де стала M *не залежить від* N *та від* u, $\|\psi\| = (\psi, \psi)^{1/2}$.

Доведення. Розвинемо $\tilde{u}(x)$ в ряд Фур'е

$$\tilde{u}(x) = S_{N-1}(x; \tilde{u}) + R_{N-1}(x, \tilde{u}), \tag{14}$$

де

$$S_{N-1}(x; \tilde{u}) = \sum_{k=1}^{N-1} a_k \sqrt{2} \sin k\pi x, \quad a_k = \sqrt{2} \int_0^1 u(x) \sin k\pi x dx.$$

Ряд Фур'е функції $\tilde{u}''(x)$ матиме вигляд

$$-\tilde{u}''(x) = -S''_{N-1}(x;\ \tilde{u}) - R''_{N-1}(x;\ \tilde{u}) =$$

$$= S_{N-1}(x;\ -\tilde{u}'') + R_{N-1}(x;\ -\tilde{u}''), \qquad (15)$$

де

$$S_{N-1}(x,\ -\tilde{u}'') = S''_{N-1}(x;\ -\tilde{u}) = \sum_{k=1}^{N-1} k^2\pi^2 a_k \sqrt{2}\sin k\pi x. \qquad (16)$$

Оскільки $-u'' \in C^{l-2}[0,\ 1]$, то в силу (6)

$$|R_{N-1}(x;\ -\tilde{u}'')| \leqslant M\frac{\ln N}{N^{l-2}}\left\|\frac{d^{l-2}(u'')}{dx^{l-2}}\right\|_{C[0,1]} =$$

$$= M\frac{\ln N}{N^{l-2}}\|u^{(l)}\|_{C[0,1]}, \qquad (17)$$

де стала M не залежить від N та u. Крім того,

$$|R_{N-1}(x;\ \tilde{u})| \leqslant M\frac{\ln N}{N^l}\|u^{(l)}\|_{C[0,1]}. \qquad (18)$$

Візьмемо проекцію (14), (15) на сітку ω_h, тоді

$$\tilde{u}(x_i) = \sum_{k=1}^{N-1} a_k y_k(x_i) + R_{N-1}(x_i;\ \tilde{u}),$$

$$-\tilde{u}''(x_i) = \sum_{k=1}^{N-1} a_k k^2\pi^2 y_k(x_i) + R_{N-1}(x_i;\ -\tilde{u}'').$$

Оскільки

$$(Cu)_i = \sum_{k=1}^{N-1} a_k k^2\pi^2 y_k(x_i) + (C\tilde{R}^{\tilde{u}}_{N-1})_i,$$

де $\tilde{R}^{\tilde{u}}_{N-1} = (R_{N-1}(x_1;\ \tilde{u}),\ \ldots,\ R_{N-1}(x_{N-1};\ \tilde{u}))^T$, то

$$\psi_i = R_{N-1}(x_i;\ -\tilde{u}'') - (C\tilde{R}^{\tilde{u}}_{N-1})_i. \qquad (19)$$

Внаслідок самоспряженості матриці C її норма, яка узгоджена з нормою вектора $\|v\|^2 = (v,\ v)$, дорівнює максимальному власному значенню, тобто

$$\|C\| = (N-1)^2\pi^2, \qquad (20)$$

тому

$$\|C\tilde{R}^{\tilde{u}}_{N-1}\| \leqslant \|C\|\|\tilde{R}^{\tilde{u}}_{N-1}\| \leqslant \|C\|\|\tilde{R}^{\tilde{u}}_{N-1}\|_{C(\tilde{\omega}_h)},$$

і в силу (18), (20)

$$\|C\tilde{R}^{\tilde{u}}_{N-1}\| \leqslant M\frac{\ln N}{N^{l-2}}\|u^{(l)}\|_{C[0,1]}. \qquad (21)$$

Із виразів (17), (19), (21) дістаємо оцінку (13). Лему доведено.

Знайдемо тепер апріорні оцінки для схеми (12).

Лема 2. *Нехай* $q(x) \in C[0,\ 1]$, $q(x) > -\pi^2$. *Тоді для схеми* (12) *має місце оцінка*

$$\|z\| \leqslant [\pi^2 + \min_{x\in[0,1]} q(x)]^{-1}\|\psi\|. \qquad (22)$$

Доведення. Помножимо (12) скалярно на z і врахуємо, що $(Cz,\ z) \geqslant \pi^2\|z\|^2$, $(Qz,\ z) \geqslant \min_i q(x_i)\|z\|^2$. Дістанемо

$$[\pi^2 + \min q(x_i)]\|z\|^2 \leqslant (Cz,\ z) + (Qz,\ z) = (\psi,\ z) \leqslant \|\psi\|\|z\|.$$

Оскільки $q(x) > -\pi^2$, то звідси знаходимо шукану оцінку.

Лема 3. *Якщо* $v(x)$ *— сіткова функція, задана на* $\tilde{\omega}_h$, *яка перетворюється в нуль на границі, то*

$$\|v\|_{C(\omega_h)} \equiv \max_i |v(x_i)| \leqslant \frac{1}{3\sqrt{5}}\|Cv\|.$$

Доведення. Розкладемо функцію $v(x)$ по системі власних векторів матриці C:

$$v(x) = \sum_{k=1}^{N-1} a_k y_k(x), \quad y_k(x) = \sqrt{2}\sin k\pi x.$$

Неважко помітити, що

$$Cv = \pi^2\sum_{k=1}^{N-1} a_k k^2 y_k(x), \quad \|Cv\|^2 = \pi^4\sum_{k=1}^{N-1} a_k^2 k^4,$$

$$|v(x_h)| \leqslant \sum_{k=1}^{N-1}|a_k|\max_l|y_k(x_l)|,$$

$$\max_l|y_k(x_l)| = \max_l\left(\sqrt{2}\left|\sin\frac{kl\pi}{N}\right|\right) \leqslant \sqrt{2}.$$

Із нерівності Коші — Буняковського дістаємо

$$|v(x_h)|^2 \leqslant 2\left(\sum_{k=1}^{N-1}|a_k|\right)^2 \leqslant 2\left(\sum_{k=1}^{N-1} a_k^2 k^4\right)\left(\sum_{k=1}^{N-1}\frac{1}{k^4}\right) =$$

$$= \frac{2}{\pi^4}\|Cv\|^2\sum_{k=1}^{N-1}\frac{1}{k^4}.$$

Оскільки

$$\sum_{k=1}^{N-1}\frac{1}{k^4} < \sum_{k=1}^{\infty}\frac{1}{k^4} = \frac{\pi^4}{90},$$

то

$$\|v\|_C^2 \leqslant \frac{1}{45}\|Cv\|^2,$$

що і треба було довести.

Лема 4. *Якщо* $\|q\|_{C[0,1]} < 3\sqrt{5}$, *то для розв'язування схеми* (12) *має місце оцінка*

$$\|z\|_{C(\omega_h)} \leqslant (3\sqrt{5} - \|q\|_{C[0,1]})^{-1}\|\psi\|.$$

Доведення. Безпосередньо з (12) маємо

$$\|Cz\| - \|q\|_{C[0,1]}\|z\| \leqslant \|Cz + Qz\| = \|\psi\|.$$

Користуючись лемою 3, далі можемо записати

$$3\sqrt{5}\,\|z\|_{C(\omega_h)} - \|q\|_{C[0,1]}\|z\|_{(C\omega_h)} \leqslant \|\psi\|,$$

звідки за умови леми 4

$$\|z\|_{C(\omega_h)} \leqslant (3\sqrt{5} - \|q\|_{C[0,1]})^{-1}\|\psi\|,$$

що і треба було довести.

Безпосередньо з лем 1, 2, 4 випливає таке твердження про збіжність і порядок точності схеми (11).

Теорема 1. *Нехай* $u(x)$ *— розв'язок диференціальної задачі* (1), *причому* $u(x) \in C^l(-\infty, \infty)$, $l \geqslant 3$. *Тоді якщо* $q(x) \in C[0, 1]$, $q(x) > -\pi^2$, *то сіткові схеми* (11) *збігаються при* $N \to \infty$ *до точного розв'язку і має місце оцінка швидкості збіжності*

$$\|y - u\| \leqslant M\frac{\ln N}{N^{l-2}}\|u^{(l)}\|_{C[0,1]},$$

де стала M *не залежить від* N *та від* u.

Якщо ж $\|q\|_{C[0,1]} < 3\sqrt{5}$, *то має місце більш сильна оцінка*

$$\|y - u\|_{C(\omega_h)} \leqslant M\frac{\ln N}{N^{l-2}}\|u^{(l)}\|_{C[0,1]}.$$

Далі постає питання про розв'язування системи лінійних алгебраїчних рівнянь (11). На відміну від різницевих методів, матриця цієї системи є заповненою. Проте її спеціальна структура дає змогу побудувати ефективний метод розв'язування.

Розглянемо спочатку випадок, коли $q(x) \equiv 0$ і система (11) має вигляд

$$Cy = \varphi \tag{23}$$

або

$$P\Lambda Py = \varphi.$$

Враховуючи ортогональність матриці P, звідси маємо

$$\Lambda\hat{y} = P\varphi, \tag{24}$$

де $\hat{y} = Py$. У координатному вигляді рівність (24) запишеться так:

$$k^2\pi^2\hat{y}_k = (P\varphi)_k \equiv \sqrt{\frac{2}{N}}\sum_{j=1}^{N-1}\sin\frac{kj\pi}{N}f(x_j),$$

звідки

$$\hat{y}_k = \sqrt{\frac{2}{N}}k^{-2}\pi^{-2}\sum_{j=1}^{N-1}\sin\frac{kj\pi}{N}f(x_j).$$

Оскільки $y = P\hat{y}$, то далі маємо

$$y_i = \frac{2}{\pi^2 N}\sum_{k=1}^{N-1}\sin\frac{ik\pi}{N}\left(k^{-2}\sum_{j=1}^{N-1}\sin\frac{kj\pi}{N}f(x_j)\right), \tag{25}$$

тобто розв'язок (23) відшукуємо за явною формулою (25). Якщо скористаємося алгоритмом швидкого перетворення Фур'є (ШПФ), то вектор $y^{(1)} = \left(\sum_{j=1}^{N-1}\sin\frac{kj\pi}{N}f(x_j)\right)_{k=\overline{1,N-1}}$ обчислимо за $O(N\ln N)$ арифметичних операцій, далі вектор $y^{(2)} = (k^{-2}y_k^{(1)})_{k=\overline{1,N-1}}$ ще за $O(N)$ операцій і, нарешті, знову за допомогою ШПФ вектор $y^{(3)} = \left(\sum_{k=1}^{N-1}\sin\frac{ik\pi}{N}y_k^{(2)}\right)$ — за $O(N\ln N)$ операцій. У цілому, таким чином, вектор $y = (y_i)$, $i = \overline{1, N-1}$ знайдемо за $O(N\ln N)$ арифметичних операцій.

Якщо ж $q(x) \neq 0$, то для розв'язування (11) застосуємо однокроковий ітераційний метод зі стійким чебишевським набором параметрів (див. п. 1.4.7 з ч. 1)

$$C\frac{\overset{k+1}{y} - \overset{k}{y}}{\tau_{k+1}} + A\overset{k}{y} = \varphi, \quad k = 0, 1, \ldots, \tag{26}$$

де $A = C + Q$, а початкове наближення $\overset{0}{y}$ можна вибрати довільно. Для того щоб до (26) можна було застосувати загальну теорію, досить встановити оцінки (див. п. 1.4.7 з ч. 1)

$$A = A^* > 0, \quad C = C^* > 0, \quad \gamma_1 C \leqslant A \leqslant \gamma_2 C, \quad \gamma_1 > 0.$$

Оскільки $C = C^* > 0$, то умова $A = A^* > 0$, очевидно, виконується, коли

$$q(x) \in C[0, 1], \quad q(x) > -\pi^2.$$

Якщо $q(x) \geqslant 0$, то в силу нерівності $(Cy, y) \geqslant \pi^2 (y, y)$ маємо

$$(Cy, y) \leqslant (Ay, y) = (Cy, y) + (Qy, y) \leqslant (Cy, y) + \\ + \|q\|_{C[0,1]} (y, y) \leqslant (1 + \pi^{-2} \|q\|_{C[0,1]}) (Cy, y),$$

тобто має місце оцінка

$$\gamma_1 C \leqslant A \leqslant \gamma_2 C,$$

$$\gamma_1 = 1, \quad \gamma_2 = 1 + \pi^{-2} \|q\|_{C[0,1]}. \tag{27}$$

Тепер можна скористатися загальною теорією для нестаціонарного неявного ітераційного методу (26) із чебишевським стійким набором параметрів, з якої випливає таке твердження.

Теорема 2. *Нехай* $q(x) \in C[0, 1]$, $q(x) \geqslant 0$. *Тоді для методу* (26), *де*

$$\tau_k = \frac{\tau_0}{1 + \rho_0 \mu_k}, \quad k = \overline{1, n}, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2},$$

$$\rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$

$\mu_k \in \mathfrak{M}_n^*$, $\mathfrak{M}_n^*$ — *стійкий чебишевський набір параметрів,* γ_1, γ_2 *визначаються формулами* (27), *досить виконати* $n(\varepsilon)$ *ітерацій:*

$$n_0(\varepsilon) \leqslant n(\varepsilon) < n_0(\varepsilon) + 1, \quad n(\varepsilon) \geqslant [n_0(\varepsilon)] + 1,$$

$$n_0(\varepsilon) = \frac{1}{2\sqrt{\varepsilon}} \ln \frac{2}{\varepsilon};$$

при цьому виконується оцінка

$$\|\overset{n}{Ay} - \varphi\|_{C^{-1}} \leqslant q_n \|\overset{\circ}{Ay} - \varphi\|_{C^{-1}} \leqslant \varepsilon \|\overset{\circ}{Ay} - \varphi\|_{C^{-1}},$$

$$q_n = \frac{2\rho_1^n}{1 + \rho_1^n}, \quad \rho_1 = \frac{1 - \sqrt{\varepsilon}}{1 + \sqrt{\varepsilon}}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

На кожній ітерації метод (26) потребує розв'язування рівняння виду (23), розв'язок якого відшукаємо за явною формулою (25) за $O(N \ln N)$ арифметичних операцій. Оскільки, як випливає з (27), величина $\xi = \frac{\gamma_1}{\gamma_2}$ не залежить від N, то і кількість ітерацій $n_0(\xi)$ не залежить від N. Це означає, що ми знайдемо розв'язок задачі (11) методом (26) за $O(N \ln N)$ арифметичних операцій, що лише в $\ln N$ раз гірше за оцінку кількості операцій у методі прогонки для системи з тридіагональною матрицею. Що стосується числа обумовленості матриці системи (23), то для нього неважко знайти оцінку

$$\operatorname{cond}_2 C = O(N^2) = O(h^{-2}),$$

яка за порядком збігається з оцінкою числа обумовленості системи

різницевих рівнянь

$$y_{\bar{x}x} = \varphi, \quad x \in \omega_h, \quad y(0) = y(1) = 0.$$

4.5.2. Інтерполяційний метод (метод коллокацій). Метод, який ми розглянемо у цьому пункті, називається ще *методом збігу* і був запропонований для наближеного розв'язування диференціальних рівнянь в 1934 р. Л. В. Канторовичем (див. також п. 1.1).

Нехай потрібно розв'язати крайову задачу

$$\frac{d^{2m}u}{dx^{2m}} - \lambda \left[p_1(x) \frac{d^{2m-1}u}{dx^{2m-1}} + \cdots + p_{2m}(x) u \right] = f(x), \tag{28}$$

$$u(a) = u'(a) = \cdots = u^{(m-1)}(a) = 0,$$

$$u(b) = u'(b) = \cdots = u^{(m-1)}(b) = 0, \tag{29}$$

де $p_1, \ldots, p_{2m}, f$ — задані функції, λ — задане число. Інтерполяційний метод наближеного розв'язування задачі (28), (29) полягає в тому, що наближений розв'язок шукаємо у вигляді

$$\tilde{u}(x) = (x - a)^m (b - x)^m \sum_{k=1}^{n} c_k x^{k-1} \tag{30}$$

(отже, крайові умови задовольняються автоматично), а коефіцієнти $c_1, \ldots, c_n$ знаходимо за умови, що рівняння (28) задовольняється в деякій системі вузлів $\omega_n = \{x_1, \ldots, x_n\}$, $x_j \in [a, b]$,

$$\left\{ \frac{d^{2m}\tilde{u}}{dx^{2m}} - \lambda \left[p_1 \frac{d^{2m-1}\tilde{u}}{dx^{2m-1}} + \cdots + p_{2m}\tilde{u} \right] \right\}_{x=x_j} = f(x_j), \quad j = \overline{1, n}. \tag{31}$$

Неважко помітити, що відносно c_j, $j = \overline{1, n}$ система (31) є системою лінійних алгебраїчних рівнянь. Щоб дослідити цей метод, застосуємо загальну теорію, викладену в п. 2.3. Для цього розглянемо рівняння (28) як операторні рівняння в просторі $X = C^{(2m)}[a, b]$ $2m$ неперервно диференційовних функцій, що задовольняють умови (29) (норму в цьому просторі означимо нижче). За $\tilde{X}$ вважаємо множину функцій вигляду (30), за Y — простір $C[a, b]$ неперервних на $[a, b]$ функцій з нормою $\|f\| = \max\limits_{x \in [a,b]} |f(x)|$ і, нарешті, за $\tilde{Y}$ візьмемо множину многочленів степеня $n - 1$. Відображення $\Phi : Y \to \tilde{Y}$ означимо, покладаючи $\tilde{f} = P_n(x; f)$, $\tilde{f} \in \tilde{Y}$, $f \in Y$, де $P_n(x; f)$ — інтерполяційний многочлен для функції f за системою вузлів ω_n, тобто $\Phi(f) \equiv P_n(x; f)$. Як ми знаємо (див. п. 2.4.7, 2.4.8 з ч. 1), якщо ω_n є системою вузлів Чебишева, тобто

$$\omega_n = \left\{ x_j = \cos \frac{2j - 1}{2n} \pi, \ j = \overline{1, n} \right\} \text{ для } a = -1, \quad b = 1$$

або

$$\omega_n = \left\{x_i = \frac{1}{2}\left[(b-a)\cos\frac{2j-1}{2n}\pi + (b+a)\right],\quad j = \overline{1,\ n}\right\}$$

для довільних a, b, то

$$\|\Phi\| \equiv \Lambda_n = \frac{2}{\pi}\ln n + 1 - \theta_n,\quad 0 \leqslant \theta_n < \frac{1}{4}. \tag{32}$$

Відомо також, що коли $a = -1$, $b = 1$, а $x_j = x_j^{(\alpha,\beta)}$, $j = \overline{1,\ n}$, де $x_j^{(\alpha,\beta)}$ — нулі многочлена Якобі $P_n^{(\alpha,\beta)}(x)$ (або при довільних a, b $x_j = \frac{1}{2}[(b-a)\,x_j^{(\alpha,\beta)} + (b+a)]$, $j = \overline{1,\ n}$), то

$$\|\Phi\| \equiv \Lambda_n \leqslant \begin{cases} cn^{\sigma+\frac{1}{2}}, & \text{якщо } \sigma = \max(\alpha,\ \beta) > -\frac{1}{2}, \\ c\ln n, & \text{якщо } \sigma \leqslant -\frac{1}{2}, \end{cases} \tag{33}$$

зокрема, для нулів многочлена Лежандра ($\alpha = 0$, $\beta = 0$), які ще називають вузлами Гаусса

$$\|\Phi\| \equiv \Lambda_n \leqslant c\sqrt{n},$$

де c — стала, що не залежить від n. Якщо ж x_j є нулями n-го ортогонального многочлена з обмеженою вагою $\rho(x) \geqslant \rho_0 \geqslant 0$, то

$$\|\Phi\| \leqslant cn. \tag{34}$$

Крайова задача еквівалентна операторному рівнянню

$$K_1 u \equiv Gu - \lambda Tu = f, \tag{35}$$

де

$$Gu = \frac{d^{2m}u}{dx^{2m}},\quad Tu = \sum_{s=1}^{2m} p_s(x)\,u^{(2m-s)}(x). \tag{36}$$

Оберненим до G є інтегральний оператор, ядром якого є функція Гріна $g(x,\ t)$ диференціального оператора $\frac{d^{2m}u}{dx^{2m}}$ за умов (29), тобто якщо $Gu = f$, то

$$u(x) = \int_a^b g(x,\ t)\,f(t)\,dt,\quad u \in X,\quad f \in Y.$$

Відомо, що функція $g(x,\ t)$ має неперервні похідні до порядку $2m-2$, а $(2m-1)$ похідна має стрибок при $x = t$. Звідси

$$u^{(k)}(x) = \int_b^a \frac{\partial^k}{\partial x^k} g(x,\ t)\,f(t)\,dt,\quad k = \overline{0,\ 2m-1},$$

а тому

$$\max_x |u^{(k)}(x)| \leqslant A_k \|f\|_C,\quad k = \overline{0,\ 2m-1}. \tag{37}$$

Якщо $A_{2m} = 1$, то (37) матиме місце і для $k = 2m$. Покладемо

$$\|u\|_X = \|Gu\|_Y = \max_x |u^{(2m)}(x)|.$$

Таким чином, оператор G здійснює ізометрію між просторами X та Y. При цьому (37) дає змогу оцінити будь-яку похідну функції u через норму елемента u в просторі X:

$$\max_x |u^{(k)}(x)| \leqslant A_r \|u\|_X,\quad k = \overline{0,\ 2m}. \tag{38}$$

Зазначимо, що оператор G переводить елементи простору $\tilde{X}$ в елементи простору Y, тобто в поліноми не вище $(n-1)$-го степеня. Оцінимо норму T як оператора з простору $C^{(2m)}[a,\ b]$ у простір $C[a,\ b]$. Із нерівності (38) дістаємо

$$\|Tu\| = \max_x \left|\sum_{s=1}^{2m} p_s(x)\,u^{(2m-s)}(x)\right| \leqslant \left[\sum_{s=1}^{2m} B_s A_{2m-s}\right]\|u\|_X,$$

де

$$B_s = \max_x |p_s(x)|,\quad s = \overline{1,\ 2m},$$

$$\|T\| \leqslant \sum_{s=1}^{2m} B_s A_{2m-s}. \tag{39}$$

Наближене рівняння, яким у даному разі є система (31), можна записати у вигляді

$$\tilde{K}_1 \tilde{u} \equiv G\tilde{u} - \lambda\Phi T\tilde{u} = \Phi f, \tag{40}$$

оскільки система (31) утворюється підстановкою в рівняння (28) виразу (30) для $\tilde{u}$ і прирівнюванням обох частин у вузлах x_j, $j = \overline{1,\ n}$, що рівносильно прирівнюванню інтерполяційних многочленів, побудованих за значеннями у цих вузлах.

Порівнюючи (40) з (13), п. 2.3, бачимо, що умова I з п. 2.3 виконується при $\mu = 0$. Для перевірки умови II (тобто перевірки можливості апроксимації елементів вигляду Tu елементами з Y) оцінимо $\frac{d}{dx}Tu$. Враховуючи (38), маємо

$$\left\|\frac{d}{dx}Tu\right\|_C = \max_x \left|\frac{d}{dx}\sum_{s=1}^{2m} p_s(x)\,u^{(2m-s)}(x)\right| =$$

$$= \max_x \left| \sum_{s=0}^{2m} (p_s'(x) + p_{s+1}(x)) u^{(2m-s)}(x) \right| \leqslant M \| u \|_x,$$

$$p_0(x) = p_{2m+1}(x) = 0.$$

Тому внаслідок теореми Джексона (див. п. 2.4.7 з ч. 1) існує поліном $\tilde{u}$ такий, що

$$\| Tu - \tilde{u} \| \leqslant \frac{M_1 M \| u \|}{n},$$

отже, умова II виконується при $\mu_1 = O\left(\frac{1}{n}\right)$.

Якщо права частина f має неперервну похідну, то, знову застосовуючи теорему Джексона, бачимо, що існує поліном $\tilde{f}$ такий, що

$$\| f - \tilde{f} \| \leqslant \frac{M \| f \|}{n},$$

тобто виконується умова III при $\mu_2 = O\left(\frac{1}{n}\right)$. Зауважимо, що коли $p_1 = 0$ в рівнянні (28), чого завжди можна досягти заміною змінних, а інші коефіцієнти p_s двічі неперервно диференційовні, то, проводячи попередні міркування щодо $\frac{d^2Tu}{dx^2}$, дістаємо $\mu_1 = O\left(\frac{1}{n^2}\right)$, а якщо $f \in C^2[a, b]$, то і $\mu_2 = O\left(\frac{1}{n^2}\right)$.

Враховуючи оцінки для μ_1, μ_2, $\| \Phi \|$, з теорем 1, 2 з п 7.3 робимо висновок, що при досить великих n система (31) має розв'язок. За теоремою з п. 2.3 випливає збіжність наближених розв'язків до точного у випадку, коли вузли є нулями многочлена Якобі $P_n^{(\alpha,\beta)}(x)$ при $\sigma = \max(\alpha, \beta) \in \left(-\frac{1}{2}, \frac{1}{2}\right)$, причому

$$\| u^* - \tilde{u}_n^* \| = O\left(\frac{1}{n^{1/2-\sigma}}\right),$$

зокрема для вузлів Гаусса

$$\| u^* - \tilde{u}_n^* \| = O\left(\frac{1}{\sqrt{n}}\right),$$

а для вузлів Чебишева

$$\| u^* - \tilde{u}_n^* \| = O\left(\frac{\ln n}{n}\right),$$

де u^* — точний розв'язок задачі (28), (29), $\tilde{u}_n^*$ — наближений розв'язок виду (30), знайдений із системи (31).

Якщо ж $p_1 = 0$, $f \in C^2[a, b]$, то система (30) має розв'язок і збіжна й тоді, коли вузли збігаються з нулями n-го ортогонального многочлена з вагою $\rho(x) \geqslant \rho_0 > 0$, причому в цьому разі

$$\| u^* - \tilde{u}_n^* \| = O\left(\frac{1}{n}\right).$$

Покажемо, що інтерполяційний метод не має насичення точності, тобто його точність тим вища, чим більшу гладкість має точний розв'язок, а гладкість точного розв'язку, очевидно, тим більша, чим більшу гладкість мають коефіцієнти p_s, $s = \overline{1, 2m}$ і права частина f. Припустимо, що p_s, $s = \overline{1, 2m}$ і f диференційовні r раз і їхні r похідні задовольняють умову Ліпшиця з показником α. Індукцією по r легко встановити, що функція $\frac{d^{2m}u^*}{dx^{2m}}$ має r похідну, яка задовольняє умову Ліпшиця з показником α, отже, і Tu^* має r похідну і вона так само задовольняє умову Ліпшиця з показником α.

Скористаємося зауваженням до теореми 2 з п. 2.3, згідно з яким швидкість збіжності визначається рівністю

$$\| u^* - \tilde{u}_n^* \| = O(\varepsilon \| \tilde{K}_1^{-1} \Phi K_1 \|), \qquad (41)$$

де ε характеризує можливість апроксимації розв'язку u^* елементом $\tilde{u} \in \tilde{X}$. За теоремою 1 з п. 2.3 норма $\tilde{K}_1^{-1}$ при досить великих n буде обмежена величиною $\| K_1^{-1} \| / (1 - q_0)$, $q \leqslant q_0 < 1$, яка не залежить від n. Оскільки u і норма $\| K_1 \|$ від n не залежать, то $\| K_1^{-1} \Phi K_1 \| \leqslant \| K_1^{-1} \| \cdot \| \Phi \| \| K_1 \| \leqslant M \| \Phi \|$, де стала M не залежить від n. Тому

$$\| u^* - \tilde{u}_n^* \| \leqslant M\varepsilon \| \Phi \|,$$

$$\| u^* - \tilde{u} \|_X = \| G(u^* - \tilde{u}) \|_Y = \left\| \frac{d^{2m}u^*}{dx^{2m}} - \tilde{y} \right\|_Y, \quad \tilde{y} = G\tilde{u} \in Y, \qquad (42)$$

і ε визначається величиною можливої апроксимації $2m$ похідної від точного розв'язку многочленом степеня $n - 1$. На основі теореми Джексона

$$\left\| \frac{d^{2m}u^*}{dx^{2m}} - \tilde{y} \right\|_Y \leqslant \frac{M}{n^{r+\alpha}},$$

тобто $\varepsilon = O\left(\frac{1}{n^{r+\alpha}}\right)$. Із рівностей (41), (42) знаходимо

$$\| u^* - \tilde{u}_n^* \| = O\left(\frac{1}{n^{r+\alpha-\sigma-0,5}}\right), \quad \sigma > -\frac{1}{2},$$

для вузлів, що збігаються з нулями многочлена Якобі $P_n^{(\alpha,\beta)}(x)$,

$\sigma = \max(\alpha, \beta)$;

$$\|u^* - \tilde{u}_n^*\| = O\left(\frac{\ln n}{n^{r+\alpha}}\right)$$

для вузлів Чебишева;

$$\|u^* - \tilde{u}_n^*\| = O\left(\frac{1}{n^{r+\alpha-1}}\right)$$

для вузлів, що збігаються з коренями ортогонального многочлена по обмеженій знизу вазі.

Більш тонкий аналіз (його провів Г. М. Вайнікко) показує, що для всіх перелічених вище систем вузлів має місце сильніша оцінка

$$\|u^* - \tilde{u}_n^*\| = O\left(E_{n-1}\left(\frac{d^{2m}u^*}{dx^{2m}}\right)\right) = O\left(\frac{1}{n^{r+\alpha}}\right),$$

де $E_{n-1}(u^{*(2m)})$ — величина найкращого рівномірного наближення функції $u^{*(2m)}$ многочленом $(n-1)$-го степеня, $u^{*(2m)}$ має похідну порядку r, яка задовольняє умову Ліпшиця з показником $\alpha \in [0, 1)$.

Зауважимо, що для рівновіддалених вузлів збіжність розглянутого методу не має місця навіть для найпростішого рівняння $u^{(2m)} = f$, що легко зрозуміти, врахувавши можливість розбіжності інтерполяційного процесу на рівновіддалених вузлах (див. п. 2.7 з ч. 1). Розглянемо питання про розв'язування системи лінійних алгебраїчних рівнянь (31) на прикладі задачі

$$-\frac{d^2u}{dx^2} + q(x)u = f(x), \quad x \in (-1, 1), \tag{43}$$

$$u(-1) = u(1) = 0. \tag{44}$$

Виберемо сітку $\omega_h = \left\{x_i = \cos\frac{2i-1}{2n}\pi,\ i = \overline{1, n}\right\}$ з нулів многочлена Чебишева $T_n(x) = \cos n \arccos x$, $x \in (-1, 1)$ і побудуємо інтерполяційний многочлен Лагранжа з вузлами x_i, який набуває в них значень y_j, $j = \overline{1, n}$. Нехай

$$t_j(x) = \frac{T_n(x)}{(x - x_j)T_n'(x_j)}, \quad u_j(x) = \frac{1 - x^2}{1 - x_j^2}t_j(x), \quad j = \overline{1, n}, \tag{45}$$

— фундаментальні многочлени інтерполяції, тобто $t_j(x_i) = \delta_{ij}$, $u_j(x_i) = \delta_{ij}$, δ_{ij} — символ Кронекера, причому многочлени $u_j(x)$ задовольняють також крайові умови (44). Інтерполяційні многочлени за системою вузлів x_i і значеннями y_i, $i = \overline{1, n}$, мають вигляд

$$t(x; y) = \sum_{j=1}^{n} y_j t_j(x), \quad v(x; y) = \sum_{j=1}^{n} y_j u_j(x),$$

причому многочлен $v(x; y)$ задовольняє крайові умови (44). Наближений розв'язок задачі (43), (44) шукатимемо інтерполяційним методом у вигляді многочлена $v(x; y)$, а невідомий вектор $y = (y_1, \ldots, y_n)$ знайдемо з умови, що рівняння (43), записане для $v(x; y)$, задовольняється в точках сітки ω_h, тобто

$$-\frac{d^2v(x_i; y)}{dx^2} + q(x_i)v(x_i; y) = f(x_i), \quad x_i \in \omega_h, \quad i = \overline{1, n}. \tag{46}$$

Якщо позначити через A, Q матриці

$$A = \left(-\frac{d^2u_k(x_i)}{dx^2}\right)_{k=\overline{1,n}\ \text{(рядок)}}^{i=\overline{1,n}\ \text{(стовпчик)}}, \tag{47}$$

$$Q = \operatorname{diag}(q(x_i))_{i=\overline{1,n}},$$

то (46) можна записати в такому матричному вигляді:

$$(A + Q)y = f, \quad f = (f(x_1), \ldots, f(x_n)). \tag{48}$$

Лема 1. *Матриця A має обернену A^{-1}, причому, якщо*

$$A^{-1}\eta = \xi, \quad \eta = (\eta_1, \ldots, \eta_n)^T, \quad \xi = (\xi_1, \ldots, \xi_n)^T,$$

то

$$\xi_j = \sum_{i=1}^{n} \eta_i \int_{-1}^{1} K(x_j, z)t_i(z)\,dz, \quad j = \overline{1, n}, \tag{49}$$

де $K(x, z)$ — функція Гріна оператора d^2/dx^2 з крайовими умовами (44).

Доведення. Неважко впевнитись, що функція Гріна має вигляд

$$K(x, z) = \begin{cases}(1 - x)(1 + z)/2, & z \in [-1, x], \\ (1 + x)(1 - z)/2, & z \in [x, 1]\end{cases} \tag{50}$$

і за її допомогою розв'язок задачі $-d^2v/dx^2 = f(x)$, $x \in (-1, 1)$, $v(-1) = v(1) = 0$ зображається так:

$$v(x) = \int_{-1}^{1} K(x, z)f(z)\,dz. \tag{51}$$

Многочлен $v(x; \xi) = \sum_{j=1}^{n} \xi_j u_j(x) \in \pi_{n+1}$, тобто є многочленом не вище $(n+1)$-го степеня, а $v''(x; y) \in \pi_{n-1}$. Тому $v''(x; y)$ однозначно визначається у вузлах x_j, $j = \overline{1, n}$. Нехай $v''(x_i, \xi) = -\eta_i$, $i = \overline{1, n}$. Ці рівняння можна записати в матричному вигляді $A\xi = \eta$.

З іншого боку, якщо $u(x; \eta) = \sum_{j=1}^{n} \eta_j t_j(x)$, то многочлен $v''(x; \xi) + u(x; \eta)$ перетворюється на нуль у точках x_j, $j = \overline{1, n}$, а оскільки

це многочлен не вище $(n-1)$-го степеня, то

$$v''(x;\ \xi) + u(x;\ \eta) \equiv 0.$$

Звідси

$$v(x;\ \xi) = \int_{-1}^{1} K(x;\ z)\, u(z;\ \eta)\, dz.$$

Покладаючи тут $x = x_j$, $j = \overline{1, n}$, дійдемо твердження леми.

Повернемося до задачі (48). Із попереднього випливає, що

$$\| y - u \|_{C(\omega_h)} \leqslant \frac{c}{n^{r+\alpha}}, \qquad (51')$$

якщо $u'(x)$ має r неперервну похідну, яка задовольняє умову Ліпшиця з показником α.

Зауважимо, що оператор A в (48) не симетричний, тому теорія ітераційних методів, яку ми розглянули раніше (див. п. 1.4.7 з ч. 1), не може бути застосована для розв'язування системи (48). Це є в якійсь мірі недоліком, тому що оператор d^2/dx^2 з умови (44), який наближає матриця A, самоспряжений. Правда, можна довести, що матриця A в певному розумінні «майже симетрична».

Запишемо (48) у вигляді

$$Gy \equiv (I + R)\, y = \varphi, \qquad (52)$$

де $G = I + R$, $R = A^{-1} Q$, $\varphi = A^{-1}\varphi$, причому в силу леми 1 ці величини обчислюються за явними формулами. Щоб застосувати до системи (52) найпростіший метод — метод простої ітерації, треба записати її у вигляді

$$y = By + C\varphi, \qquad (53)$$

де B, C — деякі матриці (це можна зробити неоднозначно). Тоді метод простої ітерації має вигляд

$$y_{k+1} = By_k + C\varphi, \quad k = 0,\ 1,\ \ldots, \qquad (54)$$

де y_0 — задане початкове наближення. Відомо, що він збігається тоді і лише тоді, коли всі власні значення $\lambda_i(B)$ матриці B за модулем менші від одиниці. Зауважимо, що $\max |\lambda_i| \leqslant \| B \|$, де $\| B \|$ — будь-яка норма матриці B.

Вправа 1. Записавши (52) у вигляді $y = -Ry + \varphi$, розглянути ітераційний процес

$$y_{k+1} = -Ry_k + \varphi, \quad k = 0,\ 1,\ \ldots, \qquad (55)$$

де y_0 — довільний вектор. Довести, що при $\| q \|_{C[-1,1]} \equiv \max\limits_{x \in [-1,1]} |q(x)|$ має місце нерівність $\| R \| \leqslant \rho < 1$, де ρ не залежить від n. Це означатиме, що ітераційний процес (55) збігається зі швидкістю геометричної прогресії, знаменник ρ якої не залежить від n.

Вправа 2. Записати систему (52) у вигляді

$$y = (1 - \alpha)\, y - \alpha R y - \alpha\varphi,$$

де α — будь-яке число, і розглянути ітераційний процес виду (54), якщо $B = (1 - \alpha)\, I - \alpha R$, $C = -\alpha$. Довести, що при $q(x) \geqslant \delta > 0$ параметр α можна вибрати так, що всі власні числа матриці B за модулем не перевищуватимуть деякого числа $\rho < 1$, причому ρ не залежить від n. Це означатиме, що метод простої ітерації (54) збігається зі швидкістю геометричної прогресії, знаменник якої ρ, причому швидкість збіжності не залежить від n.

ГЛАВА 5

МЕТОДИ РОЗВ'ЯЗУВАННЯ КРАЙОВИХ ЗАДАЧ ДЛЯ ЛІНІЙНИХ РІВНЯНЬ З ЧАСТИННИМИ ПОХІДНИМИ

Хоча багато ідей чисельних методів розв'язування крайових задач для звичайних диференціальних рівнянь переносяться на крайові задачі для рівнянь з частинними похідними, останні доцільно розглянути самостійно. Це можна пояснити такими причинами.

По-перше, рівняння з частинними похідними записуються для функцій багатьох змінних, які можуть бути задані в багатовимірній області довільної форми. Довільність форми області і її багатовимірність значно ускладнює як самі алгоритми, так і методи дослідження їх.

По-друге, множина задач для рівнянь з частинними похідними ширша, ніж множина задач для звичайних диференціальних рівнянь, що приводить до появи методів, яким немає аналогів у одновимірному випадку. Алгоритмічний аспект методів також ускладнюється.

Виникає ще цілий ряд питань, пов'язаних зі збільшенням розмірності, які ми далі розглянемо на певних класах задач.

5.1. Постановка крайових задач для рівняння Пуассона

Нехай Ω — деяка область у площині змінних x_1, x_2 з границею $\partial\Omega$, $\bar{\Omega} = \partial\Omega \cup \Omega$. Розглянемо таку задачу: знайти функцію $u(x)$, $x = (x_1, x_2)$, $u \in C^2(\Omega)$ за умови, що

$$-\Delta u \equiv -\frac{\partial^2 u}{\partial x_1^2} - \frac{\partial^2 u}{\partial x_2^2} = f(x), \quad x \in \Omega, \qquad (1)$$

$$u(x) = \mu(x), \quad x \in \partial\Omega, \qquad (2)$$

де $f(x)$, $\mu(x)$ — задані функції.

Рівняння (1) називається *рівнянням Пуассона*, яке є представником класу еліптичних рівнянь, а крайова умова (2) називається *умовою*

Діріхле, або *граничною (крайовою) умовою першого роду*. У зв'язку з цим задача (1), (2) в цілому називається *задачею Діріхле для рівняння Пуассона*, або *крайовою задачею першого роду*. Така задача виникає, наприклад, при відшуканні положення рівноваги $u(x)$ пружної однорідної мембрани, закріпленої на границі $\partial\Omega$, яка перебуває під дією зовнішньої сили, що характеризується функцією $f(x)$.

Замість граничної умови (2) можна ставити так звану *граничну умову другого роду* або *умову Неймана*

$$\frac{\partial u}{\partial \vec{n}} = \mu(x), \quad x \in \partial\Omega, \tag{3}$$

де $\frac{\partial u}{\partial \vec{n}}$ — похідна по напряму $\vec{n}$, $\vec{n} = n(x)$, $x \in \partial\Omega$ — внутрішня нормаль до границі $\partial\Omega$. Задача (1), (3) називається *задачею Неймана для рівняння Пуассона*.

Можна поставити також граничну умову

$$\frac{\partial u}{\partial \vec{s}} = \varphi(x), \quad x \in \partial\Omega, \tag{4}$$

де вектор $\vec{s} = \vec{s}(x)$ задає деякий напрям. Тоді задача (1), (4) називається *задачею із навскісною похідною*.

Гранична умова третього роду має вигляд

$$b(x)\frac{\partial u}{\partial \vec{s}} + a(x)u(x) = \mu(x), \quad x \in \partial\Omega, \tag{5}$$

а задача (1), (5) називається *задачею третього роду для рівняння Пуассона* (іноді задачею третього роду називають задачу (1), (5) при $\vec{s} = \vec{n}$).

Рівняння (1) з правою частиною $f(x) \equiv 0$ називається *рівнянням Лапласа*, а оператор $\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}$ називається *оператором Лапласа*.

Щоб спростити міркування, розглядатимемо надалі однорідні граничні умови. Насамперед розглянемо однорідну умову Діріхле

$$u(x) = 0, \quad x \in \partial\Omega. \tag{6}$$

Введемо у просторі двічі диференційовних в $\overline{\Omega}$ функцій, які задовольняють умову (6) (позначимо його $C_0^2(\overline{\Omega})$), скалярний добуток

$$(u, v) = \int_\Omega u(x)v(x)\,dx$$

і помножимо рівняння (1) скалярно на довільну функцію $v(x)$, яка також задовольняє умову (6). Проінтегрувавши частинами, матимемо

$$a(u, v) = (f, v) \;\forall\, v \in C_0^2(\overline{\Omega}), \tag{7}$$

де

$$a(u, v) = \int_\Omega \left(\frac{\partial u}{\partial x_1}\frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2}\frac{\partial v}{\partial x_2}\right)dx. \tag{8}$$

На відміну від рівняння (1) в (7) входять лише перші похідні від функції u (а також v). Тому тотожність

$$a(u, v) = (f, v) \;\forall\, v \in \mathring{W}_2^1(\Omega) \tag{9}$$

можна взяти за означення узагальненого розв'язку задач (1), (3) з простору $\mathring{W}_2^1(\Omega)$. У п. 1.2 ми бачили, що задача розв'язування операторного рівняння $Au = f$ для самоспряженого додатно визначеного оператора (а саме таким є оператор задачі (1), (2)) еквівалентна задачі мінімізації функціонала енергії

$$I(v) = (Av, v) - 2(f, v).$$

Тому узагальнений розв'язок задачі (1), (2) та (9) можна шукати також як розв'язок задачі: знайти функцію $u \in \mathring{W}_2^1$ таку, що

$$I(u) = \inf_{v \in \mathring{W}_2^1} I(v), \quad I(v) = a(v, v) - 2(f, v). \tag{10}$$

Не вдаючись до деталей, зауважимо, що гладкість розв'язку задачі (1), (2), а також (9) і (10) залежить від гладкості вихідних даних. Якщо $\mu = 0$, $f \in L_2(\Omega)$, то $u \in W_{2,0}^2(\Omega) \equiv W_2^2(\Omega) \cap \mathring{W}_2^1(\Omega)$.

5.2. Різницеві схеми. Збіжність методу сіток

5.2.1. Різницеві схеми для задачі Діріхле для рівняння Пуассона. При побудові різницевих схем для еліптичних рівнянь можна застосувати ті самі методи, що і для звичайних диференціальних рівнянь. Розглянемо спочатку метод безпосередньої заміни похідних різницевими співвідношеннями. Покриємо площину (x_1, x_2) прямокутною сіткою з кроками h_1, h_2:

$$\Pi_h = \{(x_{1i}; x_{2j}) : x_{1i} = ih_1,\ x_{2j} = jh_2,\quad i = 0, \pm 1, \ldots;\ j = 0, \pm 1, \ldots\}$$

Точки сітки Π_h називаються також *вузлами*. Тоді аналогічно одновимірному випадку з використанням формули Тейлора можна записати (за умови $u \in C^4$)

$$u_{\bar{x}_1 x_1, i,j} \equiv u_{\bar{x}_1 x_1}(x_{1i}, x_{2j}) \equiv (u_{i+1,j} - 2u_{i,j} + u_{i-1,j})/h_1^2 =$$

$$= \frac{\partial^2 u(x_{1i}, x_{2j})}{\partial x_1^2} + \frac{h_1^2}{12}\frac{\partial^4 u(\xi_1, x_{2j})}{\partial x_1^4}, \quad \xi_1 \in (x_{1,i-1}, x_{1,i+1}), \tag{1}$$

$$u_{\bar{x}_2x_2,i,j} \equiv u_{\bar{x}_2x_2}(x_{1i},x_{2j}) \equiv (u_{i,j+1} - 2u_{i,j} + u_{i,j-1})/h_2^2 =$$

$$= \frac{\partial^2 u(x_{1i}, x_{2j})}{\partial x_2^2} + \frac{h_2^2}{12}\cdot\frac{\partial^4 u(x_{1i}, \xi_2)}{\partial x_2^4}, \quad u_{ij} = u(x_{1i}, x_{2j}), \quad \xi_2 \in (x_{2,j-1}, x_{2,j+1}).$$

Оператор Лапласа на сітці Π_h можна апроксимувати так:

$$\Delta u(x) = \Lambda u(x) + \psi(x), \quad x \in \Pi_h,$$
$$\Lambda u(x) = u_{\bar{x}_1x_1}(x) + u_{\bar{x}_2x_2}(x), \tag{2}$$

$$-\psi(x) = \frac{h_1^2}{12}\frac{\partial^4 u(\xi_1, x_2)}{\partial x_1^4} + \frac{h_2^2}{12}\frac{\partial^4 u(x_1, \xi_2)}{\partial x_2^4}. \tag{3}$$

Шаблон різницевого оператора Λ для точки $x \in \Pi_h$ подібний до хреста (рис. 8), і тому апроксимацію (2) іноді називають апроксимацією «хрест». Таку саму апроксимацію можна дістати, наприклад, методом невизначених коефіцієнтів.

Якщо розглядати задачу (1), (2) з п. 5.1 в області Ω довільної форми, то користуватися апроксимацією (2) можна лише для тих вузлів Π_h, які є центром шаблонів, що належать $\bar{\Omega}$. Шаблон, в якому вузли по кожному координатному напряму рівновіддалені, називатимемо *регулярним*.

Розглянемо різницеву апроксимацію оператора Лапласа на нерегулярному шаблоні «хрест», який складається з п'яти точок:

$$(x_1, x_2),\ (x_1 + h_{1+}, x_2),\ (x_1, x_2 + h_{2+}),\ (x_1 - h_{1-}, x_2),\ (x_1, x_2 - h_{2-}),$$

де $h_{1\pm} > 0$, $h_{2\pm} > 0$, причому $h_{\alpha+} \neq h_{\alpha-}$ принаймні для одного α (рис. 9). Для зручності занумеруємо вузли цього шаблону відповідно числами 0,1, 2, 3, 4. Кожний з операторів $L_\alpha u = \frac{\partial^2 u}{\partial x_\alpha^2}$, $\alpha = 1, 2$ апроксимуємо по трьох точках: $(x_1 - h_{1-}, x_2)$, (x_1, x_2), $(x_1 + h_{1+}, x_2)$ (або точках 3, 0, 1), та $(x_1, x_2 + h_{2+})$, (x_1, x_2), $(x_1, x_2 - h_{2-})$ (або 2, 0, 4). Для цього скористаємося виразами (див. також п. 2.9 з ч. 1, п. 4.3.1).

$$L_1 u \sim \Lambda_1^* v = \frac{1}{\hbar_1}\left[\frac{v(x_1 + h_{1+}, x_2) - v(x_1, x_2)}{h_{1+}} - \right.$$
$$\left. - \frac{v(x_1, x_2) - v(x_1 - h_{1-}, x_2)}{h_{1-}}\right] \equiv v_{\bar{x}_1\hat{x}_1}, \tag{4}$$

$$L_2 u \sim \Lambda_2^* v = \frac{1}{\hbar_2}\left[\frac{v(x_1, x_2 + h_{2+}) - v(x_1, x_2)}{h_{2+}} - \right.$$
$$\left. - \frac{v(x_1, x_2) - v(x_1, x_2 - h_{2-})}{h_{2-}}\right] \equiv v_{\bar{x}_2\hat{x}_2},$$

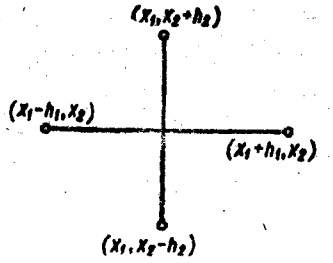


Рис. 8

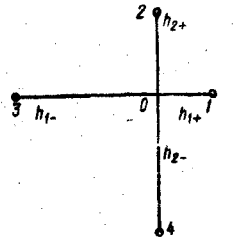


Рис. 9

де $h_\alpha = 0{,}5\,(h_{\alpha-} + h_{\alpha+})$, $\alpha = 1, 2$. Тоді апроксимація оператора Лапласа на нерегулярному шаблоні матиме вигляд

$$Lu \equiv L_1 u + L_2 u \sim \Lambda^* v \equiv \Lambda_1^* v + \Lambda_2^* v = v_{\bar{x}_1\hat{x}_1} + v_{\bar{x}_2\hat{x}_2}. \tag{5}$$

Запис аргументів сіткових функцій на нерегулярному шаблоні досить громіздкий, тому введемо ще такі позначення:

$$x^{(\pm 1_1)} = (x_1 \pm h_{1\pm}, x_2), \quad x^{(\pm 1_2)} = (x_1, x_2 \pm h_{2\pm}),$$
$$v^{(\pm 1_\alpha)} = v(x^{(\pm 1_\alpha)}), \quad v = v(x) = v(x_1, x_2), \quad \alpha = 1, 2.$$

Тоді можна записати

$$\Lambda_\alpha^* v = v_{\bar{x}_\alpha\hat{x}_\alpha} = \frac{1}{\hbar_\alpha}\left[\frac{v^{(+1_\alpha)} - v}{h_{\alpha+}} - \frac{v - v^{(-1_\alpha)}}{h_{\alpha-}}\right], \tag{6}$$
$$\hbar_\alpha = 0{,}5\,(h_{\alpha-} + h_{\alpha+}), \quad \alpha = 1, 2.$$

Розвиненням у ряд Тейлора в точці $x = (x_1, x_2)$ неважко знайти, що (див. також п. 4.3.1)

$$\Lambda_\alpha^* u - L_\alpha u = \frac{1}{3}(h_{\alpha+} - h_{\alpha-})\frac{\partial^3 u}{\partial x_\alpha^3} + O(\hbar_\alpha^2), \tag{7}$$

тобто на нерегулярному шаблоні різницевий оператор Λ^*, який визначається формулою (5), апроксимує оператор Лапласа з першим порядком.

Нам стане у нагоді ще один спосіб апроксимації оператора Лапласа на нерегулярному шаблоні, в якому замість (6) використовується вираз

$$\Lambda_\alpha^* v = \frac{1}{h_\alpha}\left[\frac{v^{(+1_\alpha)} - v}{h_{\alpha+}} - \frac{v - v^{(-1_\alpha)}}{h_{\alpha-}}\right], \tag{8}$$
$$h_\alpha = \max(h_{\alpha-}, h_{\alpha+}),$$

тобто

$$\Lambda_\alpha^* v = \frac{\hbar_\alpha}{h_\alpha} v_{\bar{x}_\alpha\hat{x}_\alpha}. \tag{9}$$

У цьому разі оператор Λ_α^* має нульовий порядок локальної апроксимації. Дійсно, враховуючи (7), маємо

$$\psi_\alpha = \Lambda_\alpha^* u - L_\alpha u = -\left(1 - \frac{h_\alpha}{h_\alpha}\right) L_\alpha u + \frac{h_\alpha}{3h_\alpha}(h_{\alpha+} - h_{\alpha-})\frac{\partial^3 u}{\partial x_\alpha^3} +$$

$$+ O(h_\alpha^2) = \pm \frac{h_{\alpha+} - h_{\alpha-}}{2h_\alpha} L_\alpha u + O(h_\alpha) = O(1). \qquad (10)$$

Пояснимо на прикладі задачі

$$u'' = -f(x), \quad x \in (0, 1); \quad u(0) = 0, \quad u(1) = 0, \qquad (11)$$

як можна використати апроксимацію (8). Виберемо сітку

$$\hat{\omega}_h = \{x_i : x_1 = h_1,\ x_{i+1} = x_i + h,\ i = \overline{1,\ N-1},\ x_{N+1} = x_N + h_2\},$$

яка є нерівномірною лише поблизу границі, причому $h_1 < h$, $h_2 < h$ $h_1 + h_2 + (N-1)h = 1$. *Вузли*, які є центрами регулярних шаблонів, називатимемо *регулярними*, а центри нерегулярних шаблонів *нерегулярними вузлами*. Замінимо у регулярних вузлах x_i, $i = \overline{2, N-1}$:

$$u_i'' \sim y_{\bar{x}x,i} = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2},$$

а в нерегулярних вузлах x_1, x_N:

$$u_1'' \sim \Lambda^* y_1 = \frac{1}{h}\left(\frac{y_2 - y_1}{h} - \frac{y_1 - y_0}{h_1}\right),$$

$$u_N'' \sim \Lambda^* y_N = \frac{1}{h}\left(\frac{y_{N+1} - y_N}{h_2} - \frac{y_N - y_{N-1}}{h}\right),$$

де $y_i = y(x_i)$ — наближення для $u_i = u(x_i)$. У результаті дістанемо різницеву схему

$$y_{\bar{x}x,i} = -f(x_i), \quad x_i = h_1 + (i-1)h, \quad i = \overline{2, N-1},$$

$$\Lambda^* y_1 = -f(x_1), \quad \Lambda^* y_N = -f(x_N); \qquad (12)$$

$$y_0 = y_{N+1} = 0.$$

Для похибки $z = y - u$ маємо задачу

$$\Lambda z = -\psi(x), \quad x = x_i, \quad i = \overline{1, N}, \quad z_0 = z_{N+1} = 0, \qquad (13)$$

де

$$\Lambda z(x) = \begin{cases} \Lambda^* z_1 & \text{при } x = x_1, \\ z_{\bar{x}x} & \text{при } x = x_i, \quad i = \overline{2, N-1}, \\ \Lambda^* z_N & \text{при } x = x_N, \end{cases}$$

$\psi_i = 0(h^2)$ при $i = \overline{2, N-1}$, $\psi_i = 0(1)$ при $i = 1$ та $i = N$. Незважаючи на те, що схема не має апроксимації в пригранічних вузлах x_1, x_N, вона має другий порядок точності в нормі $C(\hat{\omega}_h)$. Щоб це довести, перепишемо рівняння (13) при $x = x_1$, $x = x_N$ у вигляді

$$\frac{1}{h}\left(\frac{z_2 - z_1}{h} - \frac{z_1 - \bar{z}_0}{h_1}\right) = 0, \quad \frac{1}{h}\left(\frac{\bar{z}_{N+1} - z_N}{h_2} - \frac{z_N - z_{N-1}}{h}\right) = 0,$$

або

$$(z_2 - z_1) - hh_1^{-1}(z_1 - \bar{z}_0) = 0, \quad h_2^{-1}h(\bar{z}_{N+1} - z_N) -$$

$$-(z_N - z_{N-1}) = 0, \quad hh_\alpha^{-1} > 1, \quad \alpha = 1, 2,$$

де $\bar{z}_0 = hh_1\psi_1$, $\bar{z}_{N+1} = hh_2\psi_N$. Тоді задача (13) буде еквівалентна такій задачі:

$$z_{\bar{x}x,i} = -\psi_i, \quad i = \overline{2, N-1}, \quad \Lambda^* z_1 = 0, \quad \Lambda^* z_N = 0,$$

$$z_0 = hh_1\psi_1, \quad z_{N+1} = hh_N\psi_N.$$

Із наслідку 7 з п. 2.2 для z маємо оцінку

$$\|z\|_C \leqslant \max(|z_0|, |z_{N+1}|) + 2\sum_{k=1}^{N+1}\sum_{j=1}^{k-1} h^2|\psi_j| \leqslant$$

$$\leqslant hh_1|\psi_1| + hh_2|\psi_N| + 2\max_{1 \leqslant j \leqslant N}|\psi_j| \leqslant Mh^2,$$

тобто схема (12) для задачі (11) має другий порядок точності. Повернемось тепер до задачі (1), (2) з п. 5.1. У багатовимірному випадку складність цієї задачі значною мірою залежить від форми області Ω. Тому розглянемо спочатку найпростіший випадок, коли Ω є прямокутником:

$$\Omega = \{x = (x_1, x_2) : x_\alpha \in (0, l_\alpha), \quad \alpha = 1, 2\}$$

з границею $\partial\Omega$, $\bar{\Omega} = \Omega \cup \partial\Omega$. Виберемо $h_\alpha = l_\alpha/N_\alpha$, $\alpha = 1, 2$ і покриємо площину (x_1, x_2) сіткою Π_h. Вузли $x = (i_1h_1, i_2h_2)$, які лежать в прямокутнику Ω (тобто $i_1 \in (0, N_1)$, $i_2 \in (0, N_2)$), називатимемо *внутрішніми* і множину всіх внутрішніх вузлів позначимо через ω_h. Загальна кількість вузлів у ω_h є $(N_1 - 1)(N_2 - 1)$. Вузли, які лежать на границі прямокутника $\partial\Omega$ (при $i_1 = 0$, N_1, $i_2 = 0$, N_2), крім чотирьох вузлів $(0, 0)$, $(0, l_2)$, (l_1, l_2) і $(l_1, 0)$, утворюють множину γ_h *граничних вузлів* (на рис. 10 їх позначено хрестиками). Сукупність всіх внутрішніх і граничних вузлів назвемо *сіткою* $\bar{\omega}_h = \omega_h \cup \gamma_h$ в прямокутнику $\bar{\Omega}$. При $h_1 \neq h_2$ сітка називається *прямокутною*, а при $h_1 = h_2$ — *квадратною*. Для кожного внутрішнього вузла $x \in \omega_h$ можна

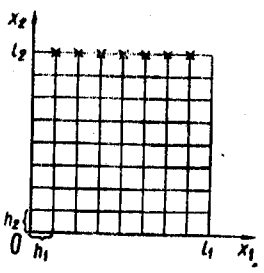


Рис. 10

побудувати п'ятиточковий регулярний шаблон «хрест», усі вузли якого належать множині $\bar{\omega}_h$. Тому в усіх внутрішніх вузлах оператор Лапласа задачі (1), (2) з п. 5.1 можна замінити за допомогою різницевих формул (2). Враховуючи, що порядок апроксимації формул (2) дорівнює 2, праву частину f доцільно замінити на φ так, щоб $\varphi(x) - f(x) = O(h^2)$, де $h = \max(h_1, h_2)$, $x \in \omega_h$. Якщо, наприклад, $f(x) \in C(\bar{\Omega})$, то можна покласти $\varphi(x) = f(x)$, $x \in \omega_h$. У результаті задачі (1), (2) з п. 5.1 поставимо у відповідність таку різницеву задачу Діріхле: знайти сіткову функцію $y(x)$, визначену на $\bar{\omega}_h$, яка у внутрішніх вузлах (на ω_h) задовольняє рівняння

$$\Lambda y = -f(x), \quad x \in \omega_h, \quad \Lambda y = y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} \tag{14}$$

і на границі γ_h набуває заданих значень

$$y(x) = \mu(x), \quad x \in \gamma_h. \tag{15}$$

Із формул (14), (15) бачимо, що значення $\mu(x)$ у вершинах прямокутника не використовуються, що і визначило вибір γ_h. Для інших задач (наприклад, для третьої крайової задачі) чи для інших схем (наприклад, схеми четвертого порядку апроксимації) границя γ_h може складатися з усіх вузлів, що лежать на $\partial\Omega$, включаючи і вершини. Підставляючи у (14), (15) $y = z + u$, де $z = y - u$ — похибка схеми (14), (15), дістанемо для z задачу

$$\Lambda z = -\psi(x), \quad x \in \omega_h, \quad z(x) = 0, \quad x \in \gamma_h, \tag{16}$$

де $\psi = \Lambda u + f$ — похибка апроксимації. Оскільки $Lu + f = 0$, то $\psi = \Lambda u + f - Lu + Lu = \Lambda u - Lu$ і з (1), (2) маємо (при $u \in C^4(\Omega)$)

$$\psi = \frac{h_1^2}{12}\frac{\partial^4 u(\xi_1, x_2)}{\partial x_1^4} + \frac{h_2^2}{12}\frac{\partial^4 u(x_1, \xi_2)}{\partial x_2^2}, \quad \xi_\alpha \in (x_\alpha - h_\alpha, x_\alpha + h_\alpha).$$

Позначивши $M_4 = \max\limits_{\substack{x\in\bar{\Omega}\\ \alpha=1,2}} \left|\frac{\partial^4 u}{\partial x_\alpha^4}\right|$, дістанемо

$$|\psi| \leqslant M_4 \frac{h^2}{6}, \quad h = \max(h_1, h_2). \tag{17}$$

У прямокутній області $\bar{\Omega}$ можна ввести і нерівномірну сітку

$$\hat{\bar{\omega}} = \{x_i = (x_1^{(i_1)}, x_2^{(i_2)}),\ i_\alpha = \overline{0, N_\alpha},\ x_\alpha^{(0)} = 0,\ x_\alpha^{(N)} = l_\alpha,\ \alpha = 1, 2\}$$

з кроками $h_1^{(i_1)} = x_1^{(i_1)} - x_1^{(i_1-1)}$, $h_2^{(i_2)} = x_2^{(i_2)} - x_2^{(i_2-1)}$. У цьому разі замість (14), (15) матимемо задачу

$$\Lambda y = -f(x), \ x \in \hat{\omega}_h, \ \Lambda y = y_{\bar{x}_1 \hat{x}_1} + y_{\bar{x}_2 \hat{x}_2}, \quad y(x) = \mu(x), \quad x \in \gamma_h. \tag{18}$$

Неважко знайти, що

$$\psi_i = (\Lambda u + f(x_i)) = \frac{1}{3}\sum_{\alpha=1}^{2}(h_\alpha^{(i_\alpha+1)} - h_\alpha^{(i_\alpha)})\frac{\partial^3 u}{\partial x^3} + O(h^2) = O(h),$$

$h = \max(h_1, h_2)$ або $h = \sqrt{h_1^2 + h_2^2}$, тобто схема на нерівномірній сітці має перший порядок локальної апроксимації.

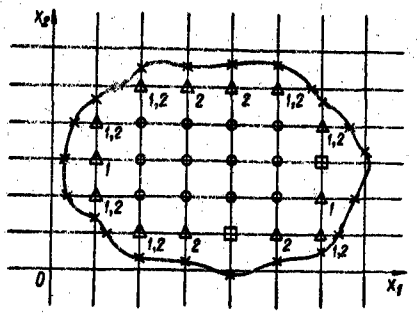


Рис. 11

Розглянемо тепер випадок довільної області $\bar{G}$ такої, що перетин G з будь-якою прямою, проведеною через внутрішню точку $x \in G$ паралельно осі координат Ox_α, $\alpha = 1, 2$, складається зі скінченного числа інтервалів. Покриємо площину (x_1, x_2) прямокутною сіткою Π_h з кроками h_1, h_2. Вузли $x_i = (i_1 h_1, i_2 h_2)$, $i = (i_1, i_2)$, які лежать всередині області G, назвемо *внутрішніми* і позначимо множину всіх таких вузлів через ω_h. Точки перетину прямих $x_\alpha^{(i_\alpha)} = i_\alpha h_\alpha$, $\alpha = 1, 2$ з границею $\partial\Omega$ області Ω назвемо *граничними* у напрямі x_α вузлами і позначимо множину цих вузлів $\gamma_{h,\alpha}$. Множина $\gamma_h = \gamma_{h,1} \cup \gamma_{h,2}$ називається *множиною граничних вузлів*, тобто це множина вузлів, які є граничними хоча б у одному з напрямів x_α. Множина всіх граничних і внутрішніх вузлів утворює сітку $\bar{\omega}_h = \omega_h \cup \gamma_h$ в області $\bar{\Omega}$ (рис. 11). Проведемо класифікацію внутрішніх вузлів. Візьмемо який-небудь внутрішній вузол $x \in \omega_h$ і проведемо через нього пряму, паралельну осі Ox_α. Її перетином з областю Ω є інтервал (або скінченна кількість інтервалів), кінці якого є граничними у напрямі Ox_α вузлами. Розглянемо вузли, що лежать на цьому інтервалі. Найближчий до кінця інтервалу вузол назвемо *приграничним* у напрямі Ox_α (або у напрямі x_α) вузлом. Якщо він віддалений від границі $\gamma_{h,\alpha}$ на величину $h_\alpha^* \neq h_\alpha$, то такий вузол є *нерегулярним* по x_α. Нехай $\omega_{h,\alpha}^*$ — множина всіх приграничних по x_α вузлів, а $\omega_{h,\alpha}^{**}$ — множина приграничних нерегулярних по x_α вузлів. Очевидно, що $\omega_{h,\alpha}^{**} \subseteq \omega_{h,\alpha}^*$. Позначимо через ω_h^* множину всіх приграничних вузлів (тобто приграничних хоча б у одному напрямі), а через ω_h^{**} — сукупність всіх нерегулярних вузлів (тобто нерегулярних хоча б у одному напрямі x_1 або x_2). Нехай $\overset{\circ}{\omega}_h$ — доповнення ω_h^* до ω_h, тобто $\omega_h = \omega_h^* \cup \overset{\circ}{\omega}_h$. Вузли з множини $\overset{\circ}{\omega}_h$ називаються *строго внутрішніми вузлами*. Введемо також позначення $\overset{\circ}{\omega}_{h,\alpha}$ для строго внутрішніх по x_α вузлів (тобто для вузла $x \in \overset{\circ}{\omega}_{h,\alpha}$ сусідні у напрямі Ox_α вузли є внутрішніми). На рис. 11 строго внутрішні

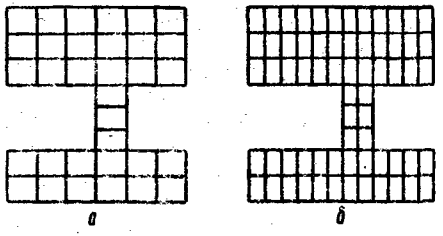


Рис. 12

вузли $\overset{\circ}{\omega}_h$ позначено кружечком, нерегулярні лише по x_α вузли — трикутником з одним індексом, нерегулярні і по x_1, і по x_2 вузли — трикутником з двома індексами, а регулярні як по x_1, так і по x_2 пригранічні вузли — квадратом.

Припустимо, що сітка $\bar{\omega}_h$ є зв'язною, тобто будь-які два внутрішніх вузли можна сполучати ламаною, ланки якої паралельні осям координат, а вершини є внутрішніми вузлами сітки. Зауважимо, що у випадку задачі Діріхле це означення зв'язності збігається з означенням у п. 2.2.6. Тоді принаймні один з чотирьох вузлів $x^{(\pm 1\alpha)}$, $\alpha = 1, 2$, п'ятиточкового шаблону $(x^{(\pm 1)}, x, x^{(\pm 2)})$ (регулярного чи нерегулярного) є внутрішнім. Вимога зв'язності сітки накладає обмеження як на вибір h_1, h_2, так і на форму області і її розміщення відносно сітки ω_h при заданих h_1, h_2. Приклади незв'язної (*а*) і зв'язної (*б*) сіток показано на рис. 12. Апроксимуємо в кожному внутрішньому вузлі $x \in \omega_h$ диференціальний оператор $L_\alpha u = \dfrac{\partial^2 u}{\partial x_\alpha^2}$, $\alpha = 1, 2$ триточковим різницевим оператором. Залежно від виду вузла можливі такі апроксимації.

Якщо вузол $x \in \omega_h$ регулярний по x_α, то різницевий оператор Λ_α записується на регулярному шаблоні $(x^{(-1\alpha)}, x, x^{(+1\alpha)})$,

$$\Lambda_\alpha y = y_{\bar{x}_\alpha x_\alpha} = \frac{y^{(+1\alpha)} - 2y + y^{(-1\alpha)}}{h^2}. \tag{19}$$

Коли вузол $x \in \omega_{h,\alpha}^{**}$, тобто нерегулярний по x_α, тоді Λ_α записується на нерегулярному шаблоні (рис. 13, *а*)

$$\Lambda_\alpha^* y = \frac{1}{h_\alpha}\left(\frac{y^{(+1\alpha)} - y}{h_\alpha} - \frac{y - y^{(-1\alpha)}}{h_\alpha^*}\right) \text{ при } x^{(-1\alpha)} \in \gamma_{h,\alpha}, \tag{20}$$

де h_α^* — відстань між вузлами x та $x^{(-1\alpha)}$ або

$$\Lambda_\alpha^* y = \frac{1}{h_\alpha}\left(\frac{y^{(+1\alpha)} - y}{h_\alpha^*} - \frac{y - y^{(-1\alpha)}}{h_\alpha}\right) \text{ при } x^{(+1\alpha)} \in \gamma_{h,\alpha}, \tag{21}$$

де h_α^* — відстань між вузлами x та $x^{(+1\alpha)}$ (рис. 13, *б*).

Якщо $x^{(-1\alpha)} \in \gamma_{h,\alpha}$ і $x^{(+1\alpha)} \in \gamma_{h,\alpha}$ (рис. 13, *в*), то маємо

$$\Lambda_\alpha^* y = \frac{1}{h_\alpha}\left(\frac{y^{(+1\alpha)} - y}{h_{\alpha+}^*} - \frac{y - y^{(-1\alpha)}}{h_{\alpha-}^*}\right) \text{ при } x^{(\pm 1\alpha)} \in \gamma_{h,\alpha}, \tag{22}$$

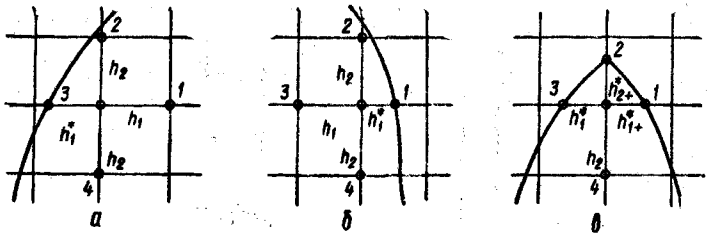


Рис. 13

де $h_{\alpha\pm}^* \neq h_\alpha$ — віддаль між x та $x^{(\pm 1\alpha)}$. Апроксимуючи оператори $L_\alpha u = \dfrac{\partial^2 u}{\partial x_\alpha^2}$ різницевими операторами за однією з формул виду (18) — (22), дістаємо для наближеного розв'язування задачі (1), (2) з п. 5.1 таку різницеву задачу Діріхле: знайти сіткову функцію $y(x)$, визначену на сітці $\bar{\omega}_h = \omega_h \cup \gamma_h$, яка у внутрішніх вузлах задовольняє рівняння

$$\Lambda y + f(x) = 0 \text{ у регулярних вузлах,} \tag{23}$$

$$\Lambda^* y + f(x) = 0 \text{ у нерегулярних вузлах} \tag{24}$$

і на границі γ_h набуває заданих значень

$$y(x) = \mu(x), \quad x \in \gamma_h, \tag{25}$$

де $\Lambda = \Lambda_1 + \Lambda_2$, $\Lambda^* = \Lambda_1^* + \Lambda_2^*$, або $\Lambda^* = \Lambda_1^* + \Lambda_2$, або $\Lambda^* = \Lambda_1 + \Lambda_2^*$.

Розглянемо похибку $z = y - u$. Підставляючи в (23) — (25) $y = z + u$, дістаємо

$$\begin{aligned} \Lambda z &= -\psi \text{ у регулярних вузлах,} \\ \Lambda^* z &= -\psi^* \text{ у нерегулярних вузлах,} \\ z(x) &= 0, \quad x \in \gamma_h, \end{aligned} \tag{26}$$

де похибка апроксимації ψ визначається формулами

$$\begin{aligned} \psi &= \Lambda u + f = \Lambda u - Lu \text{ у регулярних вузлах,} \\ \psi^* &= \Lambda^* u - Lu \text{ у нерегулярних вузлах.} \end{aligned} \tag{27}$$

Нехай $u \in C^4(\bar{\Omega})$, тоді аналогічно формулі (17) для регулярних вузлів маємо

$$|\psi| \leqslant M_4 h^2/6, \quad h^2 = h_1^2 + h_2^2. \tag{28}$$

У нерегулярних вузлах подамо похибку апроксимації у вигляді суми

$$\psi^* = \sum_{\alpha=1}^{2} \psi_\alpha^*, \quad \psi_\alpha^* = \Lambda_\alpha^* u - L_\alpha u.$$

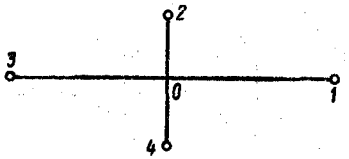


Рис. 14

Аналогічно формулі (10) дістанемо

$$\psi_\alpha^* = \frac{h_{\alpha+} - h_{\alpha-}}{2h_\alpha} \frac{\partial^2 u}{\partial x_\alpha^2} + O(h_\alpha) = O(1), \tag{28'}$$

тобто $\psi^* = O(1)$, а це означає, що у нерегулярних вузлах схема не апроксимує рівняння (1) з п. 5.1.

Зауваження. Якщо у пригранічних вузлах для апроксимації оператора Лапласа на нерегулярному шаблоні використати замість (8) апроксимацію (6), то дістанемо різницевий оператор, який може не мати таких важливих властивостей диференціального оператора як самоспряженість і знаковизначеність (від'ємна визначеність). Останнє, в свою чергу, затруднює застосування ефективних ітераційних методів для розв'язування сіткових рівнянь.

Різницеві рівняння (23), (24) можна записати в певній канонічній формі, яка пізніше дасть змогу знайти апріорну оцінку. Розглянемо регулярний вузол, який занумеруємо нулем, і регулярний п'ятиточковий шаблон, вузли якого занумеруємо числами 1, 2, 3, 4 (рис. 14). Відповідно нумеруватимемо значення функції в цих вузлах. Тоді рівняння (23) у цьому вузлі можна записати у вигляді

$$2(h_1^{-2} + h_2^{-2}) y_0 = h_1^{-2}(y_1 + y_3) + h_2^{-2}(y_2 + y_4) + f_0. \tag{29}$$

Для випадку, зображеного на рис. 13, *а*, маємо

$$\Lambda^* = \Lambda_1^* + \Lambda_2,$$

$$\Lambda_1^* y_0 = \frac{1}{h_1}\left(\frac{y_1 - y_0}{h_1} - \frac{y_0 - y_3}{h_1^*}\right) = \frac{1}{h_1}\left(\frac{y_1}{h_1} + \frac{y_3}{h_1^*} - \frac{2\hbar_1}{h_1 h_1^*} y_0\right), \tag{30}$$

$$\Lambda_2 y_0 = \frac{1}{h_2^2}(y_2 - 2y_0 + y_4),$$

де $\hbar_1 = 0{,}5(h_1 + h_1^*)$. Із рівняння $\Lambda^* y_0 = \Lambda_1^* y_0 + \Lambda_2 y_0 = -f$ знаходимо

$$\left(\frac{2\hbar_1}{h_1^2 h_1^*} + \frac{2}{h_2^2}\right) y_0 = \frac{1}{h_1^2} y_1 + \frac{1}{h_1 h_1^*} y_3 + \frac{1}{h_2^*}(y_2 + y_4) + f_0. \tag{31}$$

Для випадку, що відповідає рис. 13, *в*, матимемо і

$$\left(\frac{2\hbar_1}{h_1 h_{1-}^* h_{1+}^*} + \frac{2\hbar_2}{h_2^2 h_2^*}\right) y_0 = \frac{1}{h_1 h_{1+}^*} y_1 + \frac{1}{h_1 h_{1-}^*} y_3 + \frac{1}{h_2 h_2^*} y_2 + \frac{1}{h_2^2} y_4 + f_0, \tag{32}$$

де $\hbar_1 = 0{,}5(h_{1-}^* + h_{1+}^*)$, $\hbar_2 = 0{,}5(h_2 + h_2^*)$. Розглянувши також інші можливі випадки, дійдемо висновку, що рівняння (23), (24) можна записати у такій канонічній формі:

$$A(x) y(x) = \sum_{\xi \in Ш'(x)} B(x, \xi) y(\xi) + F(x), \quad x \in \omega_h, \tag{33}$$

де $Ш'(x)$ — множина чотирьох вузлів п'ятиточкового шаблону «хрест», з центром у точці x за виключенням самої точки x, тобто $\xi \neq x$, $A(x)$, $B(x, \xi)$ — задані коефіцієнти, $F(x)$ — задана функція. Множина $Ш'(x)$ називається *околом вузла* x. Із формул (29) — (32) бачимо, що

$$A(x) > 0,\ B(x, \xi) > 0,\ \sum_{\xi \in Ш'(x)} B(x, \xi) = A(x)\ \forall\, x \in \omega_h. \tag{34}$$

До рівняння (33) слід ще приєднати крайову умову

$$y(x) = \mu(x),\ x \in \gamma_h. \tag{35}$$

Якщо вузол x — пригранічний, то в одній з точок $\xi_0 \in Ш'(x)$ величина $y(\xi_0) = \mu(\xi_0)$ відома. Тоді (33) набере вигляду

$$A(x) y(x) = \sum_{\xi \in Ш'(x) \setminus \{\xi_0\}} B(x, \xi) y(\xi) + \tilde{F}(x),$$

де

$$\tilde{F}(x) = F(x) + B(x, \xi_0)\mu(\xi_0),$$

і для цього рівняння замість останньої рівності в (34) виконуватиметься нерівність $D(x) = A(x) - \sum_{\xi \in Ш'(x) \setminus \{\xi_0\}} B(x, \xi) > 0$. Різницева задача Діріхле є окремим випадком загальнішої задачі: знайти функцію $y(x)$, визначену на $\bar{\omega}_h = \omega_h \cup \gamma_h$, яка задовольняє рівність

$$\begin{aligned} A(x) y(x) &= \sum_{\xi \in Ш'(x)} B(x, \xi) y(\xi) + F(x), \quad x \in \omega_h, \\ y(x) &= \mu(x), \quad x \in \gamma_h, \end{aligned} \tag{36}$$

де

$$A(x) > 0,\ B(x, \xi) > 0,\ D(x) = A(x) - \sum_{\xi \in Ш'(x)} B(x, \xi) \geqslant 0 \quad \forall\, x \in \omega_h.$$

До обґрунтування схеми (23), (24), або що те саме, (33), (35), повернемося далі, а зараз зупинимося на відмінних від розглянутого методах апроксимації граничних умов. У багатовимірних задачах це є одним з багатьох утруднень.

Для апроксимації крайової умови Діріхле (2) з п. 5.1 не обов'язково вводити нерегулярні вузли, як це робилося вище. Можна розглядати лише регулярні вузли, а граничне значення з границі «перенести»

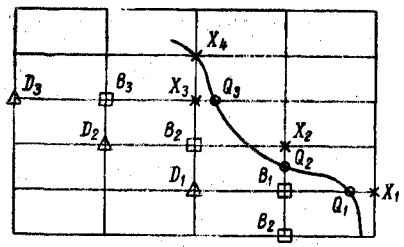


Рис. 15

на найближчий вузол регулярної сітки, який ми включимо до сіткової границі γ_h. Розглянемо, наприклад, ситуацію, зображену на рис. 15.

Вузол X називатимемо внутрішнім вузлом сітки $\bar{\omega}_h$, якщо він використовується лише при заміні диференціального рівняння (1) з 5.1. Всі інші вузли назвемо граничними. Ці вузли складають множину γ_h. На рис. 15 вузли X_i, $i = \overline{1,4}$ є граничними, а всі інші вузли — внутрішніми (причому всі регулярні). Нехай X_i — деякий граничний вузол, а Q_i — найближча до нього в певному розумінні (за відстанню від точки X_i або в напрямі якоїсь координатної осі) точка границі $\partial\Omega$. Тоді покладаємо

$$y(X_i) = \mu(Q_i). \tag{37}$$

Похибка апроксимації при цьому $O(h)$, $h = \max(h_1, h_2)$, що легко знайти з розвинення $u(X_i)$ у ряд Тейлора в околі точки Q_i. Дійсно,

$$u(X_i) = u(Q_i) \pm \delta_i \frac{\partial u(Q_i)}{\partial x} + \frac{\delta_i^2}{2!} \frac{\partial^2 u(Q_i)}{\partial x^2} \pm \cdots, \quad 0 \leqslant \delta_i < h,$$

тобто $u(X_i) = \mu(Q_j) + O(h)$.

Щоб підвищити порядок апроксимації, знайдемо точку B_i, яка є найближчим до Q (після X_i, якщо $X_i \in \Omega$) внутрішнім вузлом сітки у якомусь з координатних напрямів. За точками B_i, Q_i побудуємо лінійний інтерполяційний многочлен $P(x)$ і покладемо

$$y(X) = P(X), \quad X \in \gamma_h. \tag{38}$$

Наприклад, для точок Q_3, B_3 (рис. 15) такий многочлен має вигляд

$$P(x;\ y) \equiv P(x) = y(Q_3) + (x_1 - x_{1Q_3})\, y(Q_3;\ B_3),$$

де

$$y(Q_3;\ B_3) = \frac{y(Q_3) - y(B_3)}{x_{1Q_3} - x_{1B_3}},$$

x_{1Q_3}, x_{1B_3} — перші координати точок Q_3, B_3 відповідно. Позначимо через δ_i відстань між точками Q_i та X_i, тоді матимемо

$$y(X_3) = P(X_3) \equiv \mu(Q_3) - \frac{\delta_3}{h_1 + \delta_3}(\mu(Q_3) - y(B_3))$$

або

$$y(X_3) = \frac{h_1 \mu(Q_3) + \delta_3 y(B_3)}{h_1 + \delta_3}. \tag{39}$$

Похибка апроксимації виразу (39) має вигляд

$$\psi(X_3) \equiv u(X_3) - P(X_3;\ u) = u(X_3) - u(Q_3) + \frac{\delta_3}{h_1 + \delta_3}(u(Q_3) - u(B_3)).$$

Розвиваючи $u(Q_3)$ та $u(B_3)$ у ряд Тейлора в точці X_3, маємо

$$\psi(X_3) = u(X_3) - \left[u(X_3) + \delta_3 \frac{\partial u(X_3)}{\partial x_1} + \frac{\delta_3^2}{2} \frac{\partial^2 u(\xi_1)}{\partial x_1^2}\right] +$$

$$+ \frac{\delta_3}{h_1 + \delta_3}\left[u(X_3) + \delta_3 \frac{\partial u(X_3)}{\partial x_1} + \frac{\delta_3^2}{2} \frac{\partial^2 u(\xi_1)}{\partial x_1^2} - u(X_3) + \right.$$

$$\left. + h_1 \frac{\partial u(X_3)}{\partial x_1} - \frac{h_1^2}{2} \frac{\partial^2 u(\xi_2)}{\partial x_1^2}\right] = -\frac{\delta_3^2}{2} \frac{\partial^2 u(\xi_1)}{\partial x_1^2} +$$

$$+ \frac{\delta_3^3}{2(h_1 + \delta_3)} \frac{\partial^2 u(\xi_1)}{\partial x_1^2} - \frac{\delta_3 h_1^2}{2(h_1 + \delta_3)} \frac{\partial^2 u(\xi_2)}{\partial x_1^2},$$

де ξ_1 — точка на відрізку, що сполучає X_3 та Q_3, ξ_2 — точка на відрізку, що сполучає B_3 та X_3. Оскільки $\delta_3 < h_1$, то звідси маємо при $u \in C^2(\bar{\Omega})$

$$|\psi(X_3)| \leqslant \frac{3}{2} M_2 h^2, \tag{40}$$

де

$$M_2 = \max\left\{\max_{x \in \bar{\Omega}} \left|\frac{\partial^2 u(x)}{\partial x_1^2}\right|,\ \max_{x \in \Omega} \left|\frac{\partial^2 u(x)}{\partial x_2^2}\right|\right\}.$$

За допомогою більшої кількості внутрішніх вузлів, що лежать на одній прямій, паралельній якій-небудь координатній лінії, можна побудувати інтерполяційний многочлен вищого порядку, а це дасть змогу підвищити порядок похибки апроксимації.

Рівняння (1) з п. 5.1 у внутрішніх вузлах, наприклад у B_2, апроксимуємо звичайним чином:

$$\frac{y(X_2) - 2y(B_2) + y(D_2)}{h_1^2} + \frac{y(X_3) - 2y(B_2) + y(D_1)}{h_2^2} = -f(B_2).$$

Розглянемо тепер граничну умову, яка містить похідні, наприклад умову третього роду

$$lu \equiv a(x)\, u(x) + b(x) \frac{\partial u(x)}{\partial \vec{s}} = \mu(x), \quad x \in \partial\Omega, \tag{41}$$

де $\frac{\partial u(x)}{\partial \vec{s}}$ — навскісна похідна в напрямі $\vec{s}$ у точці $x \in \partial\Omega$ (вектор $\vec{s}$ напрямлено всередину області Ω). Нехай точка Q є точкою перетину

границі $\partial\Omega$ з однією з координатних ліній. Проведемо з цієї точки пряму в напрямі $\vec{s}$ до перетину з яким-небудь відрізком, що сполучає два найближчі вузли сіткової області ω_0. Нехай такою точкою перетину буде точка P. Покладемо

$$\frac{\partial u(Q)}{\partial \vec{s}} \approx \frac{u(P) - u(Q)}{r_{PQ}},$$

де r_{PQ} — відстань між точками P та Q. Тоді апроксимація граничного оператора (41) у точці Q матиме вигляд

$$l_h u \equiv a(Q)\, u(Q) + b(Q) \frac{u(P) - u(Q)}{r_{PQ}}. \qquad (42)$$

Точка P не належить сітці ω_h (якщо $P \in \omega_h$, то (42) дає апроксимацію (41) на сітці $\bar{\omega}_h$), тому виразимо значення $y(P)$ через значення інтерполяційного многочлена в точках B і D, тобто покладемо

$$u(P) \approx u(B) \frac{r_{DP}}{r_{BD}} + u(D) \frac{r_{PB}}{r_{BD}}.$$

Тоді (42) на сітці $\bar{\omega}_h$ набирає вигляду

$$l_h u \equiv a(Q)\, u(Q) + \frac{b(Q)}{r_{QP}} \left[u(B) \frac{r_{DP}}{r_{BD}} + u(D) \frac{r_{PB}}{r_{BD}} - u(Q) \right], \qquad (43)$$

причому за допомогою розвинення у ряд Тейлора можна довести, що похибка апроксимації

$$\psi_\nu = l_h u - lu(Q) = O(h), \quad h = \max(h_1, h_2).$$

Умова (41) апроксимується умовою

$$l_h y \equiv a(Q)\, y(Q) + \frac{b(Q)}{r_{QP}} \left[y(B) \frac{x_{DP}}{r_{BD}} + y(D) \frac{r_{PB}}{r_{BD}} - y(Q) \right] = \mu(Q). \qquad (44)$$

Якщо вважати, що $Q \in \gamma_h$, а вузли B, D нерегулярні, то диференціальний оператор (1) з п. 5.1 у них апроксимується так, як було вказано вище. Якщо ж вважати, що $B \in \gamma_h$, $D \in \gamma_h$, то в апроксимаціях (43), (44) $u(B)$ та $y(B)$ слід замінити на $u(Q)$ та $y(Q)$ відповідно (просте перенесення граничного значення) або ж на значення інтерполяційного многочлена, побудованого за внутрішніми вузлами сітки.

Коли $\partial\Omega$ складається з відрізків прямих, паралельних координатним осям, ситуація спрощується. Наприклад, для випадку $\frac{\partial u}{\partial s} = \frac{\partial u}{\partial x_1}$, користуючись односторонніми апроксимаціями першої похідної (див. п. 2.9 з ч. 1), можна покласти

$$l_h u|_{x_i} = a(X_i)\, u(X_i) + b(X_i) \frac{u(X_{i+1}) - u(X_i)}{h_1}, \qquad (45)$$

причому

$$(l_h u - lu)(X_i) = O(h)_1.$$

Якщо покласти

$$l_h u(X_i) \equiv a(X_i)\, u(X_i) + b(X_i) \frac{-3u(X_i) + 3u(X_{i+1}) - u(X_{i+2})}{2h_1}, \qquad (46)$$

то матимемо

$$l_h u(X_i) - lu(X_i) = O(h_1^2).$$

Використовуючи метод невизначених коефіцієнтів, можна побудувати апроксимації і вищих порядків, але при цьому використовується ще більше внутрішніх вузлів сітки і ускладнюється матриця всієї сіткової схеми. Згідно з (45), (46), гранична умова (41) у точці X_i в сітковій схемі відповідає рівнянню

$$a(X_i)\, y(X_i) + b(X_i) \frac{y(X_{i+1}) - y(X_i)}{h_1} = \mu(X_i)$$

або відповідно

$$a(X_i)\, y(X_i) + b(X_i) \frac{-3y(X_i) + 4y(X_{i+1}) - y(X_{i+2})}{2h_1} = \mu(X_i).$$

Підвищити порядок апроксимації крайових умов можна використанням для їхньої апроксимації диференціального рівняння, для якого поставлена крайова задача. Розглянемо, нарешті, використання фіктивних вузлів для підвищення порядку апроксимації крайових умов. Цей метод застосовується, коли границя $\partial\Omega$ складається з відрізків прямих, паралельних координатним осям. Сітка будується так, щоб границя $\partial\Omega$ проходила посередині між лініями сітки.

У сітковій ділянці $\bar{\omega}_h = \omega_h \cup \gamma_h$, в якій записується сіткова схема, вузли, позначені хрестиками (наприклад, X), належать границі γ_h, а вузли, позначені кружечками (наприклад, B), є внутрішніми і належать ω_h. Вузол X називається *фіктивним*, оскільки він не належить області $\bar{\Omega}$, але використовується для запису сіткової схеми. Якщо припустити, що розв'язок крайової задачі можна продовжити за область $\bar{\Omega}$ зі збереженням гладкості, то

$$\frac{\partial u(Q)}{\partial \vec{s}} \equiv \frac{\partial u(Q)}{\partial x_1} = \frac{u(B) - u(X)}{h_1} + O(h_1^2),$$

$$u(Q) = \frac{u(X) + u(B)}{2} + Q(h_1^2).$$

Тому, якщо покласти

$$l_h u = a(Q) \frac{u(X) + u(B)}{2} + b(Q) \frac{u(B) - u(X)}{h_1},$$

то похибка апроксимації при достатній гладкості u буде

$$\psi = l_h u(Q) - lu(Q) = O(h_1^2).$$

У сітковій схемі в точці X задаємо умову

$$a(Q)\frac{y(X)+y(B)}{2} + b(Q)\frac{y(B)-y(X)}{h_1} = \mu(Q).$$

Приклад 1. Пружна деформація жорстко закріпленої по периметру Γ квадратної пластини під дією сталої сили описується такою математичною моделлю:

$$\Delta u = -1, \quad (x, y) \in \Omega \equiv \{(x, y) : x \in (0, 1), \quad y \in (0, 1)\},$$

$$u(x, y) = 0, \quad (x, y) \in \Gamma, \quad \Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

Знайти розв'язок цієї задачі методом сіток, якщо крок сітки $h = 1/4$.

Розв'язання. Насамперед зауважимо, що в даному випадку має місце повна симетрія значення шуканої функції, оскільки права частина рівняння Пуассона $\Delta u = -1$ стала і граничні умови нульові. Тому скінченно-різницеві рівняння досить записати для чверті квадрата, тобто для вузлів з номерами (1, 1), (2, 1), (1, 2), (2, 2) (вузол (i, j) лежить на перетині i-ї вертикальної і j-ї горизонтальної ліній). Маємо

$$\frac{u_{21}-2u_{11}+u_{01}}{h^2} + \frac{u_{12}-2u_{11}+u_{10}}{h^2} = -1 \text{ (для вузла (1, 1))},$$

$$\frac{u_{31}-2u_{21}+u_{11}}{h^2} + \frac{u_{22}-2u_{21}+u_{20}}{h^2} = -1 \text{ (для вузла (2, 1))},$$

$$\frac{u_{22}-2u_{12}+u_{02}}{h^2} + \frac{u_{13}-2u_{12}+u_{11}}{h^2} = -1 \text{ (для вузла (1, 2))},$$

$$\frac{u_{32}-2u_{22}+u_{12}}{h^2} + \frac{u_{23}-2u_{22}+u_{21}}{h^2} = -1 \text{ (для вузла (2, 2))}.$$

Враховуючи геометрію розв'язку ($u_{12} = u_{21}$) і граничні умови ($u_{01} = 0$, $u_{02} = 0$, $u_{10} = 0$, $u_{20} = 0$), дістаємо систему трьох рівнянь:

$$\begin{aligned} -4u_{11} + 2u_{12} &= -0{,}0625, \\ 2u_{11} - 4u_{12} + u_{22} &= -0{,}0625, \\ 4u_{12} - 4u_{22} &= -0{,}0625. \end{aligned}$$

Розв'язавши цю систему, наприклад методом Гаусса, знайдемо

$$u_{11} = 0{,}0429, \quad u_{12} = u_{21} = 0{,}0547, \quad u_{22} = 0{,}0703.$$

Розглянемо схему для задачі (1), (2) з п. 5.1, знайдену методом усереднюючих операторів. Вважатимемо, що $f(x) \in L_2(\Omega)$, $u(x) \in W_{2,0}^2(\Omega)$, $\Omega = \{x = (x_1, x_2) : 0 < x_\alpha < l_\alpha, \alpha = 1, 2\}$ — прямокутник з границею $\partial\Omega$. Покриємо прямокутник Ω сіткою

$$\omega_h = \omega_1 \times \omega_2, \ \omega_\alpha = \{x_\alpha = i_\alpha h_\alpha : i_\alpha = \overline{1, N_\alpha - 1}, \ h_\alpha = l_\alpha / N_\alpha\}, \ \alpha = 1, 2,$$

а через γ_h позначимо границю сіткової області:

$$\gamma_h = \bigcup_{\alpha=1}^{2} (\gamma_{+\alpha} \cup \gamma_{-\alpha}), \quad \gamma_{-1} = \{x = (0, x_2) : x_2 = i_2 h_2,$$

$$i_2 = \overline{1, N_2 - 1}\}, \ \gamma_{+1} = \{x = (l_1, x_2) : x_2 = i_2 h_2, \ i_2 = \overline{1, N_2 - 1}\},$$

$$\gamma_{-2} = \{x = (x_1, 0) : x_1 = i_1 h_1, \ i_1 = \overline{1, N_1 - 1}\}, \ \gamma_{+2} = \{x = (x_1, l_2) : x_1 = i_1 h_1, \ i_1 = \overline{1, N_1 - 1}\}.$$

Поклавши у виразі (9)

$$v = \varphi_h(|\xi - x|), \quad x \in \omega_h, \quad \xi \in \Omega,$$

де $\varphi_h(|\xi|)$ — ядро усереднення (див. п. 4.3.6)

$$\varphi_h(|\xi|) = \prod_{i=2}^{2} \frac{1}{h_i} \varphi\left(\frac{|\xi|}{h_i}\right), \quad \varphi(t) = \begin{cases} 1 - |t|, & |t| \leqslant 1, \\ 0, & |t| \geqslant 1, \end{cases} \tag{47}$$

дістанемо рівність

$$-(T_2 u)_{\bar{x}_1 x_1} - (T_1 u)_{\bar{x}_2 x_2} = Tf(x), \quad x \in \omega_h. \tag{48}$$

Замінюючи в лівій частині функцію $u(\xi)$ (по ξ_i діє оператор T_i) її інтерполяційним многочленом нульового степеня по одній точці $x \in \omega_h$, знаходимо різницеву схему

$$\Lambda y \equiv -y_{\bar{x}_1 x_1} - y_{\bar{x}_2 x_2} = \varphi(x), \quad x \in \omega_h, \quad y(x) = 0, \quad x \in \gamma_h, \tag{49}$$

де $\varphi(x) = Tf(x)$. Ця схема складніша, ніж (14), (15), де $\varphi(x) = f(x)$, але її можна використовувати для негладких (узагальнених) розв'язків задачі (1), (2) з п. 5.1 (у наступному пункті ми доведемо збіжність схеми (49) в енергетичній нормі $W_2^1(\omega_h)$ при $u \in W_{2,0}^2(\Omega)$).

5.2.2. Збіжність методу сіток. Як і в одновимірному випадку, дослідження збіжності і оцінки точності опираються на загальну теорему Лакса (див. п. 2.2). Розглянемо два приклади оцінки швидкості збіжності сіткових схем для задачі Діріхле для рівняння Пуассона.

Приклад 2. Дослідити збіжність і знайти оцінку точності схеми (23) — (25) для задачі (1), (2) з п. 5.1 в нормі $C(\omega)$.

Розв'язання. а) Задача для похибки $z = y(x) - u(x)$, $x \in \omega$ має вигляд (26), причому для похибки апроксимації за умови $u \in C^4(\bar{\Omega})$ мають місце оцінки (28), (28′), точніше

$$\begin{aligned} &|\psi(x)| \leqslant M_4 h^2/6 \text{ для регулярних вузлів}, \\ &|\psi(x)| \leqslant 2M_2 \text{ для нерегулярних вузлів}; \end{aligned} \tag{50}$$

б) знайдемо апріорну оцінку для задачі (25). Введемо різницевий оператор $\bar{\Lambda}$, який збігається з Λ^* у нерегулярних вузлах і з Λ в інших внутрішніх вузлах. Тоді задача (25) для похибки матиме вигляд

$$\bar{\Lambda} z = -\psi(x), \quad x \in \omega_h, \tag{51}$$

$$z(x) = 0, \quad x \in \gamma_h. \tag{52}$$

Запишемо (51) у вигляді (див. (33))

$$A(z)\,z(x)=\sum_{\xi\in\widetilde{\mathcal{Ш}}'(x)}B(x,\xi)\,z(\xi)-\psi(x),\quad x\in\omega_h;\quad z(x)=0,\quad x\in\gamma_h,$$

причому для $A(x)$, $B(x,\xi)$ виконуються умови (34).

Подамо праву частину ψ у вигляді

$$\psi=\mathring{\psi}+\psi^*,$$

де $\mathring{\psi}=\psi$, $\psi^*=0$ у строго внутрішніх вузлах x, тобто $x\in\mathring{\omega}_h$ і $\mathring{\psi}=0$, $\psi^*=\psi$ у пригранічних вузлах $x\in\omega_h^*$. Зазначимо, що $z=v+w$, де v, w — розв'язки задач

$$\overline{\Lambda}v=-\mathring{\psi}(x),\quad x\in\omega_h;\quad v(x)=0,\quad x\in\gamma_h, \tag{53}$$

$$\overline{\Lambda}w=-\psi^*(x),\quad x\in\omega_h;\quad w(x)=0,\quad x\in\gamma_h, \tag{54}$$

і оцінимо кожну з функцій v і w окремо.

Для оцінки $v(x)$ побудуємо мажорантну функцію $Y(x)$. Вважатимемо, що початок координат лежить всередині області Ω, і мажоранту шукатимемо у вигляді

$$Y(x)=K(R^2-r^2),\quad r^2=r^2(x)=x_1^2+x_2^2,$$

де K — стала, яку виберемо пізніше, R — радіус круга з центром у початку координат, що містить область Ω.

Враховуючи, що $\Lambda_\alpha x_\beta^2=0$, $\Lambda_\alpha^* x_\beta^2=0$, при $\alpha\neq\beta$ і

$$\Lambda_\alpha x_\alpha^2=[(x_\alpha+h_\alpha)^2-2x_\alpha^2+(x_\alpha-h_\alpha)^2]/h_\alpha^2=2,$$

$$\Lambda_\alpha^* x_\alpha^2=h_\alpha^{-1}\left[\frac{(x_\alpha+h_{\alpha+})^2-x_\alpha^2}{h_{\alpha+}}-\frac{x_\alpha^2-(x_\alpha-h_{\alpha-})^2}{h_{\alpha-}}\right]=$$

$$=2\theta_\alpha,\quad \theta_\alpha=(h_{\alpha+}+h_{\alpha-})/(2h_\alpha),$$

знаходимо

$$\Lambda Y(x)=\sum_{\alpha=1}^{2}\Lambda_\alpha Y(x)=-4K,\quad x\in\mathring{\omega}_h,$$

$$\Lambda^*Y(x)=-4\theta K,\quad x\in\omega_h^*,$$

де $\theta=0{,}5\,(\theta_1+\theta_2)$, причому $\theta_\alpha=1$, якщо вузол $x\in\omega_h^*$ є регулярним за напрямом x_α. Таким чином, $Y(x)$ є розв'язком задачі

$$\begin{aligned}&\overline{\Lambda}Y(x)=-\overline{F}(x),\quad x\in\omega_h,\\ &Y(x)=K(R^2-r^2(x))\geqslant 0,\quad x\in\gamma_h,\end{aligned} \tag{55}$$

де

$$\overline{F}(x)=\begin{cases}4K, & x\in\mathring{\omega}_h,\\ 4\theta K, & x\in\omega_h^*.\end{cases}$$

Порівнюючи задачі (55) і (53), бачимо, що нерівність $|\mathring{\psi}(x)|\leqslant\overline{F}(x)$ буде виконуватись, якщо покласти

$$K=\frac{1}{4}\|\mathring{\psi}\|_{C(\mathring{\omega}_h)}=\frac{1}{4}\|\mathring{\psi}\|_{C(\mathring{\omega}_h)}=\frac{1}{4}\max_{x\in\mathring{\omega}_h}|\mathring{\psi}(x)|,$$

причому виконуються всі умови теореми 10, п. 2.2 (теореми порівняння). Тому

$$\|v\|_{C(\omega_h)}\leqslant\|Y\|_{C(\omega_h)},$$

а оскільки, як легко помітити, $\|Y\|_{C(\omega_h)}\leqslant KR^2$, то

$$\|v\|_{C(\omega_h)}\leqslant\frac{R^2}{4}\|\mathring{\psi}\|_{C(\mathring{\omega}_h)}=\frac{R^2}{4}\|\mathring{\psi}\|_{C(\mathring{\omega}_h)}. \tag{55'}$$

Щоб оцінити $w(x)$, виключимо з лівої частини рівнянь (54) в пригранічних вузлах значення функції за допомогою крайової умови $w(x)=0$, $x\in\gamma_h$, запишемо потім (54) у канонічній формі (33) і покажемо, що

$$D(x)\geqslant\frac{1}{h^2},\quad x\in\omega_h^*,\quad h=\max_\alpha h_\alpha, \tag{56}$$

$$D(x)=0,\quad x\in\mathring{\omega}_h. \tag{57}$$

Після цього можна буде скористатися теоремою 12 з п. 2.2.

Рівність (57) очевидна. Розглянемо тепер пригранічний вузол $x\in\omega_h^*$ і рівняння (54) у цьому вузлі (у канонічній формі)

$$A(x)\,w(x)=\sum_{\xi\in\widetilde{\mathcal{Ш}}'(x)}B(x,\xi)\,w(\xi)+F(x),\quad F(x)=\psi^*(x).$$

Якщо один із вузлів $\xi=\xi_0\in\mathcal{Ш}'(x)$ є граничним, то $w(\xi_0)=0$ і це рівняння можна переписати у вигляді

$$A(x)\,w(x)=\sum_{\xi\in\widetilde{\mathcal{Ш}}'(x)}B(x,\xi)\,w(\xi)+F(x), \tag{58}$$

де $\widetilde{\mathcal{Ш}}'(x)=\mathcal{Ш}'(x)\setminus\{\xi_0\}$. Тоді, оскільки для сіткової апроксимації рівняння Лапласа мають місце умови (34), то

$$D(x)=A(x)-\sum_{\xi\in\widetilde{\mathcal{Ш}}'(x)}B(x,\xi)=A(x)-$$

$$-\left[\sum_{\xi\in\mathcal{Ш}'(x)}B(x,\xi)-B(x,\xi_0)\right]=B(x,\xi_0)>0.$$

Якщо вузол x є пригранічним за двома напрямами, тобто існують дві точки ξ_0, $\xi_1\in\mathcal{Ш}'(x)$, які є граничними, то

$$D(x)=B(x,\xi_0)+B(x,\xi_1)>0.$$

Покажемо, що $D(x)\geqslant h^{-2}$.

Справді, нехай вузол $x\in\omega_h^*$ — пригранічний і нерегулярний лише у напрямі x_α, причому $\xi_0=x^{(+1\alpha)}\in\gamma_\alpha$, $x^{(-1\alpha)}\in\omega_h$. Тоді рівняння (54) має вигляд

$$\Lambda_\alpha^* w+\Lambda_\beta w=-\psi^*(x),$$

де

$$\Lambda_\beta y=y_{\bar{x}_\beta x_\beta},\quad \beta=3-\alpha,$$

$$\Lambda_\alpha^* w=\frac{1}{h_\alpha}\left(\frac{w^{(+1\alpha)}-w}{h_{\alpha+}}-\frac{w-w^{(-1\alpha)}}{h_\alpha}\right)=-\frac{1}{h_\alpha}\left(\frac{w}{h_{\alpha+}}+\frac{w-w^{(-1\alpha)}}{h_\alpha}\right).$$

Порівнюючи це з (56), бачимо, що

$$A(x) = \frac{1}{h_\alpha h_{\alpha+}} + \frac{1}{h_\alpha^2} + \frac{2}{h_\beta^2},$$

$$\sum_{\xi \in \tilde{\Pi}'(x)} B(x, \xi) = \frac{1}{h_\alpha^2} + \frac{2}{h_\beta^2},$$

$$D(x) = \frac{1}{h_\alpha h_{\alpha+}} \geqslant \frac{1}{h^2}, \quad h = \max_\alpha h_\alpha.$$

Якщо x — регулярний пригранічний лише по x_α вузол, то $D(x) = \frac{1}{h_\alpha^2} \geqslant \frac{1}{h^2}$. Якщо ж цей вузол пригранічний і по x_β, $\beta = 3 - \alpha$, то $D(x)$ може лише збільшитися. Тому оцінка (56) справедлива завжди.

Користуючись тепер для задачі (54) теоремою 12 з п. 2.2, дістаємо

$$\| w \|_{C(\omega_h)} \leqslant \| \psi^* / D \|_{C(\omega_h^*)} \leqslant h^2 \| \psi^* \|_{C(\omega_h^*)}. \tag{59}$$

Враховуючи, що $z = v + w$, а також (55′), (59) для розв'язування задачі (51), знаходимо оцінку

$$\| z \|_{C(\omega_h)} \leqslant \frac{R^2}{4} \| \mathring{\psi} \|_{C(\mathring{\omega}_h)} + h^2 \| \psi^* \|_{C(\omega_h^*)}. \tag{60}$$

Тепер з нерівностей (50), (60) випливає таке твердження.

Теорема 1. *Якщо розв'язок задачі* (1), (2) *з п.* 4.1 *належить класу* $C^4(\bar{\Omega})$, *то різницева схема* (23), (24) *рівномірно збігається, причому має місце оцінка*

$$\| y - u \|_{C(\omega_h)} \leqslant \left(\frac{R^2}{24} M_4 + 2M_2 \right) h^2,$$

де

$$M_k = \max_{\substack{x \in \Omega \\ \alpha_1 + \alpha_2 = k}} \left| \frac{\partial^k u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}} \right|, \quad h = \max_{\alpha = 1,2} h_\alpha.$$

Зауваження. Аналогічно доведенню оцінок (55), (59), користуючись додатково наслідком теореми 10 з п. 2.2 ч. I, для задачі (23), (24) можна знайти оцінку

$$\| y \|_{C(\bar{\omega}_h)} \leqslant \| \mu \|_{C(\gamma_h)} + \frac{R^2}{4} \| f \|_{C(\mathring{\omega})_h} + h^2 \| f \|_{C(\omega_h^*)},$$

яка виражає стійкість різницевої задачі Діріхле (23), (24) за крайовими умовами правою частиною.

Приклад 3. Дослідити на збіжність схему (49) і знайти її оцінку точності для задачі (1), (2) з п. 4.1 у нормі $W_2^1(\omega_h)$.

Р о з в' я з а н н я. а) Запишемо задачу для похибки $z(x) = y(x) - u(x)$, $x \in \omega_h$:

$$\Lambda z(x) = \psi(x), \quad x \in \omega_h, \tag{61}$$

$$z(x) = 0, \quad x \in \gamma_h,$$

де $\psi(x) = \Lambda y - \Lambda u = \varphi(x) - \Lambda u = Tf - \Lambda u$. Враховуючи рівність (48), перетворимо $\psi(x)$ таким чином:

$$-\psi(x) = -\eta_{1\bar{x}_1 x_1}(x) - \eta_{2\bar{x}_2 x_2}(x), \tag{62}$$

де функціонали від u (при фіксованому x) η_α мають вигляд

$$\eta_\alpha(x) \equiv \eta_\alpha(x; u) \equiv \eta_\alpha(u) = u(x) - T_{3-\alpha} u(x), \quad \alpha = 1, 2. \tag{63}$$

б) Знайдемо апріорну оцінку для задачі (61), враховуючи вигляд похибки апроксимації (62), (63). Для цього в просторі сіткових функцій двох змінних, які задано на $\bar{\omega}_h = \omega_h \cup \gamma$ і перетворюються на нуль на γ_y, введемо скалярний добуток

$$(y, v) = \sum_{x \in \omega_h} h_1 h_2 y(x) v(x) \equiv \sum_{i_1=1}^{N_1-1} h_1 \sum_{i_2=1}^{N_2-1} h_2 y_{i_1 i_2} v_{i_1 i_2}.$$

Помножимо рівняння (61) скалярно на z і в обох частинах знайденої рівності підсумуємо частинами. Матимемо

$$\begin{gathered} -(z_{\bar{x}_1 x_1}, z) - (z_{\bar{x}_2 x_2}, z) = -(\eta_{1\bar{x}_1 x_1}, z) - (\eta_{2\bar{x}_2 x_2}, z), \\ (z_{\bar{x}_1}, z_{\bar{x}_1}]_1 + (z_{\bar{x}_2}, z_{\bar{x}_2}]_2 = (\eta_{1\bar{x}_1}, z_{\bar{x}_1}]_1 + (\eta_{\bar{x}_2}, z_{\bar{x}_2}]_2, \\ |z|_{1,\omega_h}^2 \equiv \| \nabla z \|_{0,\omega_h}^2 = (\eta_{1\bar{x}_1}, z_{\bar{x}_1}]_1 + (\eta_{\bar{x}_2}, z_{\bar{x}_2}]_2, \end{gathered} \tag{64}$$

де

$$(y, v]_\alpha = \sum_{i_\alpha=1}^{N_\alpha} h_\alpha \sum_{i_{3-\alpha}=1}^{N_{3-\alpha}} h_{3-\alpha} y_{i_1 i_2} v_{i_1 i_2}, \quad \alpha = 1, 2; \quad \| v]|_\alpha^2 = (v, v]_\alpha,$$

$$|z|_{1,\omega_h} = ((z_{\bar{x}_1}, z_{\bar{x}_1}]_1 + (z_{\bar{x}_2}, z_{\bar{x}_2}]_2)^{1/2}$$

—напівнорма в сітковому просторі $\mathring{W}_2^1$ з нормою

$$\| v \|_{1,\omega_h} \equiv \| v \|_{\mathring{W}_2^1(\omega_h)} = (\| v \|_{0,\omega_h}^2 + |v|_{1,\omega_h}^2)^{1/2},$$

$$\| v \|_{0,\omega_h}^2 = (v, v).$$

Враховуючи зображення $z(x) = \sum_{x_1=h_1}^{x_1} h_1 z_{\bar{x}_1}(x) = \sum_{x_2=h_2}^{x_2} h_2 z_{\bar{x}_2}(x)$, неважко показати, що в просторі функцій, які дорівнюють нулю на γ_h, напівнорма $|z|_{1,\omega_h}$ і $\| z \|_{1,\omega_h}$ еквівалентні. Застосовуючи у правій частині (64) нерівність Коші — Буняковського, дістаємо

$$|z|_{1,\omega_h}^2 \leqslant \| \eta_{1\bar{x}_1}]|_1 \| z_{\bar{x}_1}]|_1 + \| \eta_{2\bar{x}_2}]| \| z_{\bar{x}_2}]|_2,$$

$$|z|_{1,\omega_h}^2 \leqslant (\| \eta_{1\bar{x}_1}]|_1 + \| \eta_{2\bar{x}_2}]|_2) |z|_{1,\omega_h}.$$

Таким чином, маємо шукані оцінки

$$|z|_{1,\omega_h} \leqslant \|\eta_{1\bar{x}_1}\|_1 + \|\eta_{2\bar{x}_2}\|_2,$$
$$\|z\|_{1,\omega_h} \leqslant M(\|\eta_{1\bar{x}_1}\|_1 + \|\eta_{2\bar{x}_2}\|_2), \tag{65}$$

де стала M не залежить від h та від u. Зауважимо, що за рахунок зображення похибки апроксимації у вигляді (62) і застосування методу енергетичних оцінок вдалося у правій частині (64) позбутися однієї різницевої похідної, яка входила до похибки апроксимації.

в) Оцінимо величини $|\eta_{x_\alpha}|$, $\alpha = 1, 2$. Зробивши заміну $\xi_\alpha = x_\alpha + s_\alpha h_\alpha$, $\alpha = 1, 2$, $s_\alpha \in (-1, 1)$, відобразимо елементарний прямокутник $e(x) = (x_1 - h_1, x_1 + h_1) \times (x_2 - h_2, x_2 + h_2)$, $x \in \omega_h$ на квадрат $E = (-1, 1) \times (-1, 1)$, $\xi \in e(x)$, $s \in E$, $u(\xi) = \tilde{u}(s)$. Тоді функціонали $\eta_\alpha(x, u)$, $x \in \omega_h$, $\alpha = 1, 2$, наберуть вигляду

$$\eta_1(u) = \eta_1(\tilde{u}) = \tilde{u}(0) - T_2 u(x) =$$
$$= \tilde{u}(0) - \int_{-1}^{1} (1 - |s_2|)\, u(x_1, x_2 + s_2 h_2)\, ds_2 = \tilde{u}(0) - \int_{-1}^{1} (1 - |s_2|)\, \tilde{u}(0, s_2)\, ds_2,$$

$$\eta_2(u) = \eta_2(\tilde{u}) = \tilde{u}(0) - \int_{-1}^{1} (1 - |s_1|)\, u(x_1 + s_1 h_1, x_2)\, ds_1 =$$
$$= \tilde{u}(0) - \int_{-1}^{1} (1 - |s_1|)\, \tilde{u}(s_1, 0)\, ds_1.$$

Використовуючи теореми вкладення, маємо

$$|\eta_1(\tilde{u})| \leqslant 2\|\tilde{u}\|_{C(\bar{E})} \leqslant 2M\|\tilde{u}\|_{W_2^2(E)},$$
$$|\eta_2(\tilde{u})| \leqslant 2M\|\tilde{u}\|_{W_2^2(E)},$$

тобто функціонали η_α, $\alpha = 1, 2$, обмежені в просторі $W_2^2(E)$. Неважко переконатися, що вони перетворюються на нуль на многочленах нульового та першого степенів:

$$\eta_1(1) = 1 - \int_{-1}^{1} (1 - |s_2|)\, ds_2 = 1 - 1 = 0,$$
$$\eta_1(s_1) = 0 - \int_{-1}^{1} (1 - |s_2|)\, s_1|_{s_1=0}\, ds_2 = 0,$$
$$\eta_1(s_2) = 0 - \int_{-1}^{1} (1 - |s_2|)\, s_2 ds_2 = 0$$

(останній інтеграл рівний нулю як інтеграл від непарної функції по симетричному відносно 0 проміжку). Але, очевидно,

$$\eta_1(s_2^2) = 0 - \int_{-1}^{1} (1 - |s_2|)\, s_2^2 ds_2 \neq 0,$$

бо інтеграл по симетричному відносно нуля проміжку береться від непарної функції.

Отже, за лемою Бремблa — Гільберта, якщо $u \in W_2^2(\Omega)$, то

$$|\eta_1(\tilde{u})| \leqslant 2M\overline{M}\,|\tilde{u}|_{W_2^2(E)}. \tag{66}$$

Переходячи до змінної ξ, матимемо ($ds = ds_1 ds_2 = (h_1 h_2)^{-1}\, d\xi$)

$$|\tilde{u}|_{W_2^2(E)} = \left(\int_E \left[\left(\frac{\partial^2 \tilde{u}}{\partial s_1^2}\right)^2 + \left(\frac{\partial^2 \tilde{u}}{\partial s_1 \partial s_2}\right)^2 + \left(\frac{\partial^2 \tilde{u}}{\partial s_2^2}\right)^2\right] ds\right)^{1/2} =$$
$$= \left(\int_{e(x)} \left[\left(h_1^2 \frac{\partial^2 u}{\partial \xi_1^2}\right)^2 + \left(h_1 h_2 \frac{\partial^2 u}{\partial \xi_1 \partial \xi_2}\right)^2 + \left(h_2^2 \frac{\partial^2 u}{\partial \xi_2^2}\right)^2\right] (h_1 h_2)^{-1}\, d\xi\right)^{1/2} \leqslant$$
$$\leqslant h^2 (h_1 h_2)^{-1/2} \left(\int_{e(x)} \left[\left(\frac{\partial^2 u}{\partial \xi_1^2}\right)^2 + \left(\frac{\partial^2 u}{\partial \xi_1 \partial \xi_2}\right)^2 + \left(\frac{\partial^2 u}{\partial \xi_2^2}\right)^2\right] d\xi\right)^{1/2} = h^2 (h_1 h_2)^{-1/2} |u|_{W_2^2(e(x))},$$
$$x \in \omega_h, \quad h = \max(h_1, h_2).$$

Тепер у формулі (66) маємо

$$|\eta_1(x)| \equiv |\eta_1(x; u)| \leqslant 2M\overline{M} h^2 (h_1 h_2)^{-1/2} |u|_{W_2^2(e(x))}, \quad x \in \omega_h, \tag{67}$$

де сталі $\overline{M}$, M не залежать від h та u. Аналогічно знаходимо оцінку

$$|\eta_2(x)| \leqslant 2M\overline{M} h^2 (h_1 h_2)^{-1/2} |u|_{W_2^2(e(x))}, \quad x \in \omega_h. \tag{68}$$

З останніх двох нерівностей

$$|\eta_{1,\bar{x}_1}(x)| \leqslant Mh (h_1 h_2)^{-1/2} (|u|_{W_2^2(e(x))} + |u|_{W_2^2(e(x_1 - h_1, x_2))}),$$
$$|\eta_{2\bar{x}_2}(x)| \leqslant Mh (h_1 h_2)^{-1/2} (|u|_{W_2^2(e(x))} + |u|_{W_2^2(e(x_1, x_2 - h_2))}),$$

а тому

$$\|\eta_{1\bar{x}_1}\|_1 = \left(\sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2 - 1} h_1 h_2 \eta_{1\bar{x}_1}^2(x_{1i_1}, x_{2i_2})\right)^{1/2} \leqslant Mh |u|_{W_2^2(\Omega)},$$
$$\|\eta_{2\bar{x}_2}\|_2 \leqslant Mh |u|_{W_2^2(\Omega)},$$

де M — деякі нові сталі, що не залежать від h та від u.

З оцінки (67), (68) дістаємо таку теорему збіжності та оцінку точності.

Теорема 2. *Якщо розв'язок задачі* (9) *належить простору* $W_2^2(\Omega)$, *то розв'язок різницевої схеми* (49) *при* $h = \max(h_1, h_2) \to 0$ *збігається*

до точного розв'язку в сітковій нормі простору $W_2^1(\omega_h)$, *причому має місце узгоджена оцінка швидкості збіжності*

$$|y-u|_{1,\omega_h} \leqslant \|y-u\|_{1,\omega_h} \leqslant Mh\,|u|_{M_2^2(\Omega)},$$

де стала M *не залежить від* h *та від* u.

Вправа 1. Довести, що теорема справедлива і для схеми

$$\Lambda y = Sf(x), \quad x \in \omega_h, \quad y(x) = 0, \quad x \in \gamma_h,$$

де $S = S_1S_2$ — оператор усереднення за Стекловим.

Вказівка. Замість рівності (48) використати таку:

$$-\sum_{\alpha=1}^{2} \frac{1}{h_\alpha} S_{3-\alpha}\left(\frac{\partial u^{(+0,5\alpha)}(x)}{\partial x_\alpha} - \frac{\partial u^{(-0,5\alpha)}(x)}{\partial x_\alpha}\right) = S_1S_2 f(x), \quad x \in \omega_h,$$

де

$$u^{(\pm 0,5_1)} = u\left(x_1 \pm \frac{h_1}{2},\ x_2\right), \quad u^{(\pm 0,5_2)} = u\left(x_1,\ x_2 \pm \frac{h_2}{2}\right).$$

Вправа 2. Застосовуючи такий самий метод, як у прикладі п. 4.3.7, для схеми (49) знайти узгоджену оцінку

$$\|y-u\|_{0,\omega_h} \leqslant Mh^2\,|u|_{W_2^2(\Omega)}.$$

5.3. Метод скінченних елементів для рівняння Пуассона

Ідея методу скінченних елементів для звичайних диференціальних рівнянь і для рівнянь з частинними похідними одна й та сама, хоча технічні труднощі для багатовимірного випадку значно зростають. Тому деякі деталі опускатимемо, адресуючи читача до спеціальної літератури.

Як ми бачили в одновимірному випадку, одна з основних ідей методу скінченних елементів (МСЕ) — це спеціальний вибір координатних функцій і застосування методу Гальоркіна.

5.3.1. Задача Діріхле в квадраті. Кусково-лінійні координатні функції. Розглянемо задачу (9) з п. 5.1 в квадраті $\Omega = \{x = (x_1, x_2) : 0 < x_\alpha < 1,\ \alpha = 1, 2\}$. Побудову координатних функцій слід, очевидно, почати з побудови сітки.

Вузли сітки ω_h — це точки перетину $(x_{1,i}, x_{2,j})$ прямих $x_1 = x_{1,i} = i_1h$, $x_2 = x_{2,j} = jh$, $h = 1/N$, $i, j = \overline{0, N}$. Кожну комірку сітки $\Pi_{i,j} = (x_{1,i} \leqslant x_1 \leqslant x_{1,i+1},\ x_{2,j} \leqslant x_2 \leqslant x_{2,j+1})$ розділимо на два трикутники діагоналлю, яка сполучає вершини $(x_{1,i}, x_{2,j})$ та $(x_{1,i+1}, x_{2,j+1})$. У результаті дістанемо розбиття квадрата, яке називається *тріангуляцією області* (рис. 16). Кожному вузлу $(x_{1,i}, x_{2,j})$ поставимо у відповідність функцію

$$\varphi_{ij}(x_1, x_2) = \varphi\left(\frac{x_1}{h} - i,\ \frac{x_2}{h} - j\right), \quad i, j = \overline{0, N}, \tag{1}$$

де

$$\varphi(s, t) = \begin{cases} 0, & (s, t) \notin e, \\ 1 - (|s| + |t| + |s-t|)/2, & (s, t) \in e, \end{cases}$$

яка дорівнює одиниці у вузлі $(x_{1,i}, x_{2,j})$, нулю в інших вузлах, є лінійною в кожному трикутнику тріангуляції і дорівнює нулю за межами області e (x), $x \in \omega_h$ (рис. 17). Неважко помітити, що $\varphi_{i,j}(x) \in \mathring{W}_2^1(\Omega)$.

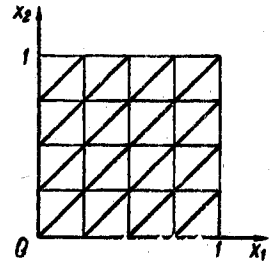


Рис. 16

Нехай $u(x)$ — неперервна функція, визначена в Ω. Обчисливши значення $u(x)$ у точках сітки ω_h, дістанемо сіткову функцію $u_h(x) = u_h(x_{1,i}, x_{2,j}) = u(x_{1,i}, x_{2,j}) = u_{i,j}$. Розглянемо функцію

$$\tilde{u}(x) = \sum_{i,j=1}^{N-1} u_{ij}\varphi_{ij}(x). \tag{2}$$

Неважко помітити, що $\tilde{u}(x_{1,i}, x_{2,j}) = u_{ij}$, $\tilde{u}(x)$ неперервна, кусково-диференційовна, $\tilde{u}(x) = 0$, $x \in \partial\Omega$, тобто $\tilde{u} \in \mathring{W}_2^1(\Omega)$. Функція $\tilde{u}(x)$ є кусково-лінійним інтерполянтом функції $u(x)$ (хвиля і надалі означатиме, що відповідна функція є кусково-лінійною).

Множина неперервних, лінійних у кожному трикутнику тріангуляції області Ω функцій, як випливає з (2), утворює $(N-1)^2$-вимірний підпростір простору $\mathring{W}_2^1(\Omega)$ з базисом $\varphi_{ij}(x)$. Цей підпростір позначимо $\mathring{H}_h$.

Наближений розв'язок задачі (1), (2) або ж (9) чи (10), п. 5.1 означимо як функцію з $\mathring{H}_h$ виду

$$\tilde{y}(x) = \sum_{i,j=1}^{N-1} y_{ij}\varphi_{ij}(x), \tag{3}$$

яка задовольняє рівняння

$$a(\tilde{y}, \tilde{v}) = (f, \tilde{v}) \quad \forall \tilde{v} \in \mathring{H}_h \tag{4}$$

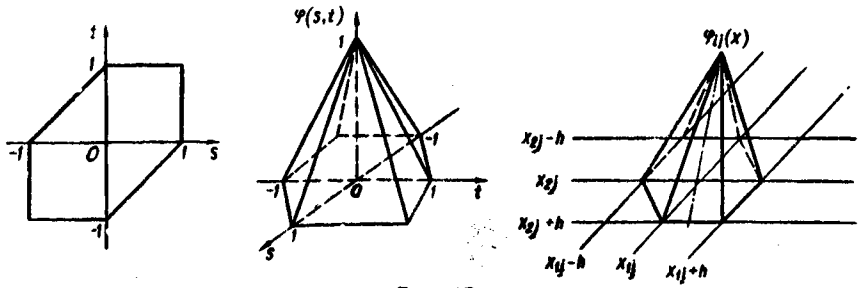


Рис. 17

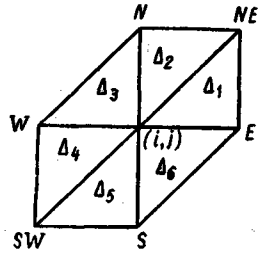


Рис. 18

(це рівняння є не що інше, як рівняння (п. 10.1), яке розглядається на підпросторі $\mathring{H}_h$). Рівняння (4) еквівалентне системі рівностей

$$a(\tilde{y}, \varphi_{ij}) = (f, \varphi_{ij}), \quad i, j = \overline{1, N-1}, \qquad (5)$$

(число рівнянь в (5) дорівнює кількості внутрішніх вузлів, значить і кількості невідомих y_{ij}). Кожна функція $\varphi_{ij}(x)$ відмінна від нуля лише в шести трикутниках з $\overline{\Omega}$, які прилягають до $(x_{1,i}, x_{2,j}) \in \omega_h$, тому інтегрування в (4) ведеться фактично лише по об'єднанню цих трикутників, яке позначимо через ω_{ij}. Вигляд області ω_{ij} показано на рис. 18, де введено й інші очевидні позначення.

Неважко помітити, що функції $\varphi_{ij}(x)$ та $\tilde{y}(x)$ в ω_{ij} мають вигляд

$$\varphi_{ij}(x) = \begin{cases} 1 - (x_1 - x_{1i})/h, & x \in \Delta_1, \\ 1 - (x_2 - x_{2j})/h, & x \in \Delta_2, \\ 1 + (x_1 - x_{1i})/h - (x_2 - x_{2j})/h, & x \in \Delta_3, \\ 1 + (x_1 - x_{1i})/h, & x \in \Delta_4, \\ 1 + (x_2 - x_{2j})/h, & x \in \Delta_5, \\ 1 - (x_1 - x_{1j})/h + (x_2 - x_{2j})/h; & x \in \Lambda_6, \end{cases} \qquad (6)$$

$$\tilde{y}(x) = \begin{cases} y_{ij} + y_{x_1}(x_1 - x_{1i}) + (y_{x_2})_E (x_2 - x_{2j}), & x \in \Delta_1, \\ y_{ij} + (y_{x_1})_N (x_1 - x_{1i}) + y_{x_2}(x_2 - x_{2j}), & x \in \Delta_2, \\ y_{ij} + y_{\bar{x}_1}(x_1 - x_{1i}) + y_{x_2}(x_2 - x_{2j}), & x \in \Delta_3, \\ y_{ij} + y_{\bar{x}_1}(x_1 - x_{1,i}) + (y_{\bar{x}_2})_W (x_2 - x_{2j}), & x \in \Delta_4, \\ y_{ij} + (y_{\bar{x}_1})_S (x_1 - x_{1i}) + y_{\bar{x}_2}(x_2 - x_{2j}), & x \in \Delta_5, \\ y_{ij} + y_{x_1}(x_1 - x_{1,i}) + y_{\bar{x}_2}(x_2 - x_{2j}), & x \in \Delta_6. \end{cases} \qquad (7)$$

Тут через $(y_{x_\alpha})_Q$, $(y_{\bar{x}_\alpha})_Q$, y_{x_α}, $y_{\bar{x}_\alpha}$ позначено відповідні різницеві похідні від функції u в точках Q та $x = (x_{1i}, x_{2j})$. Використовуючи (6), (7), перетворимо ліві частини системи рівнянь (5) таким чином:

$$a(\tilde{y}, \varphi_{ij}) = \int_{\omega_{ij}} \left[\frac{\partial \tilde{y}}{\partial x_1} \frac{\partial \varphi_{ij}}{\partial x_1} + \frac{\partial \tilde{y}}{\partial x_2} \frac{\partial \varphi_{ij}}{\partial x_2} \right] dx = \int_{\Delta_1 \cup \Delta_6} y_{x_1} \left(-\frac{1}{h} \right) dx +$$

$$+ \int_{\Delta_3 \cup \Delta_4} y_{\bar{x}_1} \left(\frac{1}{h} \right) dx + \int_{\Delta_2 \cup \Delta_3} y_{x_2} \left(-\frac{1}{h} \right) dx +$$

$$+ \int_{\Delta_5 \cup \Delta_6} y_{\bar{x}_2} \left(\frac{1}{h} \right) dx = -h^2 (y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2}),$$

тепер система рівнянь (5) набирає вигляду

$$-h^2 (y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2}) = f^h_{ij}, \quad (x_{1i}, x_{2j}) \in \omega_h, \qquad (8)$$

де

$$f^h_{ij} = \int_{\omega_{ij}} f(x) \varphi_{ij}(x)\, dx, \quad y(x) = 0, \quad x \in \gamma_h,$$

γ_h — вузли сітки, які лежать на $\partial\Omega$. Якщо поділити кожне з рівнянь на h^2, врахувати, що $\varphi_{ij}(x) \geqslant 0$ і

$$h^{-2} \int_{\omega_{ij}} \varphi_{ij}(x)\, dx = 1, \qquad (9)$$

то всі відмінності схеми (8) від схем п. 5.2 полягають у тому, що права частина обчислюється за іншою формулою (зауважимо, що рівність (9) є нормуючою умовою на ядро усереднення).

Схема (8) була побудована з використанням явних зображень (6), (7). Проте в більш загальних ситуаціях, наприклад, при використанні тріангуляцій, які складаються з нерегулярних трикутників, побудова явного вигляду сіткової схеми стає більш складною і неефективною. Тому, як про це вже говорилося в п. 4.4, конструктивніший шлях побудови матриці системи рівнянь методу скінченних елементів (або, як кажуть, глобальної матриці жорсткості) полягає у формуванні її з матриць жорсткості окремих елементів, які обчислюються досить просто. Проілюструємо це на прикладі побудови системи рівнянь (8).

На трикутнику Δ_1 (рис. 18) введемо локальну нумерацію вузлів: нехай вузол (x_{1i}, x_{2j}) має номер 1, вузол E — номер 2 і вузол NE — номер 3. Відповідну нумерацію введемо і для значень шуканої функції у вузлах та для значень координатних функцій. З усіх координатних функцій лише три мають ненульові значення на Δ_1. Введемо на Δ_1 локальну систему координат, покладаючи

$$s = (x_1 - x_{1i})/h, \quad t = (x_2 - x_{2i})/h. \qquad (10)$$

Беручи до уваги (6), неважко помітити, що на Δ_1 відмінні від нуля координатні функції мають вигляд

$$\varphi_1 = 1 - s, \quad \varphi_2 = s - t, \quad \varphi_3 = t. \qquad (11)$$

Складемо з них матрицю-рядок $\Phi^{(1)} \equiv \Phi^{(1)}(x) = [\varphi_1(x), \varphi_2(x), \varphi_3(x)]$, а зі значень наближеного розв'язку — вектор-стовпчик $\vec{y}^{(1)} = [y_1, y_2, y_3]^T$. Тоді наближений розв'язок $\tilde{y}(x)$ на Δ_1 матиме вигляд

$$\tilde{y}(x) = \Phi^{(1)} \vec{y}^{(1)}. \qquad (12)$$

Зобразимо білінійну форму $a(\tilde{y}, v)$ у вигляді суми інтегралів по скін-

ченних елементах. З урахуванням (12) матимемо

$$a_{\Delta 1}(\tilde{y}, v) = \int_{\Delta_1} \left(\frac{\partial \tilde{y}}{\partial x_1} \frac{\partial v}{\partial x_1} + \frac{\partial \tilde{y}}{\partial x_2} \frac{\partial v}{\partial x_2} \right) dx = \vec{v}^{(1)T} K^{(1)} \vec{y}^{(1)}, \qquad (13)$$

де знак T означає транспонування; $K^{(1)}$ — матриця жорсткості елемента Δ_1,

$$K^{(1)} = \int_{\Delta_1} \left(\frac{\partial \Phi^{(1)T}}{\partial x_1} \frac{\partial \Phi^{(1)}}{\partial x_1} + \frac{\partial \Phi^{(1)T}}{\partial x_2} \frac{\partial \Phi^{(1)}}{\partial x_2} \right) dx. \qquad (14)$$

Беручи до уваги (10), (11), неважко помітити, що

$$\frac{\partial \Phi^{(1)}}{\partial x_1} = h^{-1}[-1, 1, 0], \quad \frac{\partial \Phi^{(1)}}{\partial x_2} = h^{-1}[0, -1, 1],$$

а тому

$$K^{(1)} = \frac{1}{2} \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

Очевидно, що такими самими будуть матриці жорсткості елементів Δ_3, Δ_2, якщо на них ввести локальну нумерацію вузлів, починаючи з лівого нижнього вузла проти ходу годинникової стрілки.

Вводячи на елементі Δ_2 локальну нумерацію вузлів, починаючи з лівого нижнього проти ходу годинникової стрілки, неважко знайти, що в локальній системі координат (10) матимемо

$$\varphi_1 = 1 - t, \quad \varphi_2 = s, \quad \varphi_3 = t - s,$$

і далі матриця жорсткості елемента Δ_2 має вигляд

$$K^{(2)} = \frac{1}{2} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{bmatrix}.$$

Покажемо тепер, як побудувати глобальну матрицю жорсткості. Оскільки всі рівняння (4) мають однаковий вигляд, то досить знайти одне рівняння, побудувавши матрицю жорсткості для ω_{ij}.

Перенумеруємо вузли в ω_{ij} у порядку зліва направо і знизу вгору: $SW - 1$, $S - 2$, $W - 3$, $(i, j) - 4$, $E - 5$, $N - 6$, $NE - 7$. Встановимо зв'язки між локальними нумераціями вузлових значень на елементах і нумерацію на ω_{ij}. Коли позначити $\vec{Y} = [y_1, y_2, \ldots, y_7]^T$, то матриці $S^{(i)}$, $i = \overline{1, 6}$, які встановлюють відповідні зв'язки за формулою $\vec{y}^{(i)} = S^{(i)}\vec{Y}$, мають вигляд

$$S^{(1)} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad S^{(2)} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

і т. д. Ці матриці називаються *матрицями кінематичних зв'язків*. Кожна з них містить три рядки (за числом вузлів у елементі) і по сім стовпчиків (за числом вузлів у ω_{ij}). Неважко помітити, що номери ненульових стовпчиків матриць $S^{(i)}$ збігаються з глобальними номерами вузлів, а номери ненульових елементів цих сповпчиків збігаються з локальними номерами вузлових значень.

Вправа 1. Записати вигляд матриць $S^{(i)}$, $\overline{i = 3,6}$.

Підставляючи зображення $\vec{y}^{(1)} = S^{(1)}\vec{Y}$ в (12), знаходимо

$$\tilde{y} \equiv \tilde{y}(x) = \Phi^{(1)} S^{(1)} \vec{Y}, \quad x \in \Delta_1. \qquad (15)$$

Запишемо рівняння (5) для вузла (i, j) у вигляді

$$a(\tilde{y}, \varphi_{ij}) \equiv a_{\Delta_1}(\tilde{y}, \varphi_{ij}) + a_{\Delta_2}(\tilde{y}, \varphi_{ij}) + \cdots + a_{\Delta_6}(\tilde{y}, \varphi_{ij}) = (f, \varphi_{ij}). \qquad (16)$$

Як випливає з (13),

$$a_{\Delta_1}(\tilde{y}, \varphi_{ij}) = \vec{\varphi}_{ij}^{(1)T} K^{(1)} \vec{y}^{(1)},$$

а оскільки $\vec{\varphi}_{ij}^{(1)} = S^{(1)}\vec{\Phi}$, де $\vec{\Phi} = [(\varphi_{ij})_1, \ldots, (\varphi_{ij})_7]^T = [0, 0, 0, 1, 0, 0, 0]^T$, $\vec{\varphi}_{ij}^{(1)T} = \Phi^T S^{(1)T}$, то

$$a_{\Delta_1}(\tilde{y}, \varphi_{ij}) = \vec{\Phi}^T S^{(1)T} K^{(1)} S^{(1)} \vec{Y}. \qquad (17)$$

Аналогічно

$$a_{\Delta_2}(\tilde{y}, \varphi_{ij}) = \vec{\Phi}^T S^{(2)T} K^{(2)} S^{(2)} \vec{Y} \qquad (18)$$

і т. д. Як бачимо, вклад матриці жорсткості $K^{(1)}$ елемента Δ_1 в глобальну матрицю жорсткості сукупності елементів ω_{ij} визначається матрицею

$$S^{(1)T} K^{(1)} S^{(1)} = \frac{1}{2} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

елемента Δ_2—матрицею

$$S^{(2)T}K^{(2)}S^{(2)} = \frac{1}{2}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

і т. д.

Неважко помітити, що матриця $S^{(1)T}K^{(1)}S^{(1)}$ визначається з матриці $K^{(1)}$ розміщенням між другим і третім рядками, а також між другим і третім стовпчиками нульових рядка і стовпчика (це відповідає тому, що між другим і третім вузловими значеннями на елементі Δ_1 при глобальній нумерації відбувається «стрибок» на один номер) і наступним доповненням утвореної матриці зверху і зліва трьома нульовими рядками і стовпчиками (це відповідає тому, що при глобальній нумерації нульових значень перед першим вузловим значенням елемента Δ_1 пропускається три номери). Такий простий зв'язок між вказаними матрицями обумовлений тим, що локальна нумерація вузлових значень відносно глобальної не має інверсій, тобто якщо одне з двох вузлових значень на елементі має більший номер при глобальній нумерації, то це вузлове значення має більший номер і при локальній нумерації.

Розглядаючи елемент Δ_2, бачимо, що на ньому локальна нумерація містить одну інверсію відносно глобальної: вузлове значення y_7 (при глобальній нумерації) має номер 2 при локальній нумерації, в той час, коли y_6 має номер 3. І хоча розміщення нульових рядків і стовпчиків у $S^{(2)T}K^{(2)}S^{(2)}$ підкоряється тому самому правилу, що і в $S^{(1)T}K^{(1)}S^{(1)}$, після викреслювання їх з $S^{(2)T}K^{(2)}S^{(2)}$ ми не дістанемо $K^{(2)}$, а знайдемо матрицю

$$M = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix},$$

яка збігається з $K^{(1)}$. Проте якщо ми відповідно до наявності інверсій переставимо місцями в $K^{(2)}$ другий і третій рядки, а також другий і третій стовпчики, то дістаємо матрицю M, і далі за загальним правилом, про яке говорилося вище, можемо дістати матрицю $S^{(2)T}K^{(2)}S^{(2)}$.

Сказане вище дає змогу формувати матриці $S^{(i)T}K^{(i)}S^{(i)}$ без застосування операцій множення.

Вправа 2. Записати матриці $S^{(i)T}$, $K^{(i)}$, $S^{(i)}$, $i = \overline{3, 6}$. Із (16), (17), (18), а також із структури вектора $\vec{\Phi}$ випливає, що рядок глобальної матриці жорсткості для ω_{ij}, яка відповідає рівнянню (16) у вузлі (i, j), дорівнює сумі четвертих рядків матриць $S^{(i)T}$, $K^{(i)}$, $S^{(i)}$. Це відображено в наступній таблиці.

Таблиця

Δ_i	Четвертий рядок матриці $S^{(i)T}K^{(i)}S^{(i)}$, що відповідає Δ_i						
Δ_1	0	0	0	1/2	−1/2	0	0
Δ_2	0	0	0	1/2	0	−1/2	0
Δ_3	0	0	−1/2	1	0	−1/2	0
Δ_4	0	0	−1/2	1/2	0	0	0
Δ_5	0	−1/2	0	1/2	0	0	0
Δ_6	0	−1/2	0	1	−1/2	0	0
Рядок глобальної матриці жорсткості для ω_{ij}	0	−1	−1	4	−1	−1	0

На цьому закінчимо виклад методу скінченних елементів для задачі Діріхле у разі рівняння Пуассона. Зауважимо лише, що при складанні програми для ЕОМ цим методом виникає ще цілий ряд питань: як автоматизувати (алгоритмізувати) процес тріангуляції області, як автоматизувати нумерацію невідомих, щоб ширина смуги ненульових елементів матриці була мінімальною, як врахувати граничні умови, як розв'язувати системи алгебраїчних рівнянь МСЕ і т. п. З приводу цих питань знову адресуємо читача до спеціальної літератури, наприклад [Schwarz H. R. Methode der finiten Elemente. 2. Aufl. Studtgart; Teubner, 1984].

Ще одне зауваження стосується координатних функцій. Зрозуміло, що розглянутими кусково-лінійними функціями вибір не обмежується. Зазначимо, наприклад, що клас кусково-білінійних координатних функцій, які відповідають тріангуляції області Ω, складається з комірок сітки ω_h, тобто з квадратів. Кожному вузлу сітки (x_{1i}, x_{2j}) в $\overline{\Omega}$ ставиться у відповідність координатна функція $\psi_{ij}(x)$, яка задається за допомогою одновимірних кусково-лінійних координатних функцій

$$\psi_{ij}(x) = \varphi_i(x_1)\,\varphi_j(x_2).$$

Функція $\psi_{ij}(x)$ відмінна від нуля лише в комірках сіток, які прилягають до вузла (x_{1i}, x_{2j}), дорівнює 1 у цьому вузлі і 0 в усіх інших. Оскільки кожна кусково-білінійна функція «неортогональна» не більше, ніж восьми іншим координатним функціям, то кожне рівняння системи сіткових рівнянь містить не більше дев'яти невідомих. Використовують також інші координатні функції.

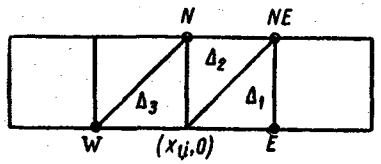


Рис. 19

Вправа 3. Для задачі

$$-\Delta u + qu = f, \quad q = \text{const} > 0, \quad x \in \Omega; \tag{19}$$

$$\frac{\partial u}{\partial n}(x) = 0, \quad x \in \partial\Omega$$

в одиничному квадраті Ω в узагальненій постановці

$$b(u, v) = (f, v), \tag{20}$$

де

$$b(u, v) = \int_{\Omega} \left[\frac{\partial u}{\partial x_1} \frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2} \frac{\partial v}{\partial x_2} + quv \right] dx,$$

$$(f, v) = \int_{\Omega} fv dx, \quad f(x) \in L_2(\Omega), \quad u \in W_2^1(\Omega),$$

n — зовнішня нормаль до $\partial\Omega$ (зауважимо, що тут рівняння (20) розглядається на множині функцій, які «більш вільні» в своїй поведінці на $\partial\Omega$; це пов'язано з тим, що крайова умова в (19) природна, тобто функція, що задовольняє (20), вже в силу цього задовольняє крайову умову з (19)) побудувати схему методу скінченних елементів (МСЕ), вибравши кусково-лінійні координатні функції. Як буде відрізнятися матриця жорсткості, що відповідає білінійній формі $b(u, v)$, від матриці жорсткості, що відповідає формі $a(u, v)$?

Вказівка. Наближений розв'язок шукати у просторі H_h функцій виду

$$\tilde{y} = \sum_{i,j=0}^{N} y_{ij} \varphi_{ij}(x),$$

які задовольняють рівняння

$$b(\tilde{y}, \tilde{v}) = (f, \tilde{v}) \ \forall\, \tilde{v} \in H_h.$$

Записати рівняння для внутрішнього вузла сітки у вигляді

$$-h^2\left(1 - \frac{h^2}{6} q\right)(y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2}) + h^2 qy + \frac{h^4}{12} q\,(y_{x_1 x_2} + y_{\bar{x}_1 \bar{x}_2}) = \int_{\omega_{ij}} f \varphi_{ij} dx$$

(на відміну від рівняння Пуассона шаблон цієї схеми не п'ятиточковий, а семиточковий, що пов'язане з наявністю члена qu). Для граничного вузла, наприклад $(x_{1,i}, 0)$, $0 < i < N$ (рис. 19), рівняння матиме вигляд

$$-h\left(1 - \frac{h^2 q}{6}\right) y_{x_2} - \frac{h^2}{2}\left(1 - \frac{h^2 q}{6}\right) y_{\bar{x}_1 x_1} + \frac{h^2 q}{2} y +$$

$$+ \frac{h^3 q}{12}(y_{x_1} + h y_{x_1 x_2}) = \int_{\Omega} f \varphi_{i0} dx.$$

Записавши $b(u, v) = a(u, v) + q \int_{\Omega} uv dx$, впевнитися, що матриця жорсткості, яка відповідає $b(u, v)$, дорівнює сумі матриці жорсткості $a(u, v)$ і так званої *матриці маси*, що відповідає білінійній формі $q \int_{\Omega} uv dx$ і має вигляд (для елемента Δ_l)

$$M^{(l)} = \frac{h^2}{24} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

5.3.2. Задача Діріхле для рівняння Пуассона у многокутнику. Якщо область Ω має криволінійну границю, то одна з можливостей розв'язування крайових задач у такій області зв'язана з апроксимацією області многокутником. Крім того, така форма області, як многокутник, сама досить поширена на практиці. Будь-який многокутник можна тріангулювати, тобто розбити на трикутники так, щоб будь-які трикутники або не перетиналися, або ж мали лише спільну вершину чи лише спільну сторону. Позначимо многокутник через Ω. Тоді вершини трикутників тріангуляції утворюють в області $\overline{\Omega}$ деяку нерегулярну сітку вузлів. Щоб застосувати МСЕ, вважатимемо, що ці вузли перенумеровано в деякому порядку. Кроками сітки будемо називати відстань між суміжними вузлами, тобто довжини сторін трикутників тріангуляції.

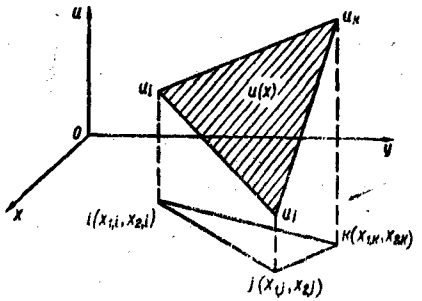


Рис. 20

Неперервні кусково-лінійні координатні функції на нерегулярній сітці будують так, як і на регулярній: кожному вузлу сітки $(x_{1,k}, x_{2,k})$ ставиться у відповідність неперервна кусково-лінійна функція $\varphi_k(x)$, яка дорівнює 1 у вузлі $(x_{1,k}, x_{2,k})$ і нулю в усіх інших. Функція $\varphi_k(x)$ відмінна від нуля лише на трикутниках, що прилягають до вузла (x_{1k}, x_{2k}). Якщо $u(x)$ — неперервна в $\overline{\Omega}$ функція, то її кусково-лінійну апроксимацію $\tilde{u}(x)$ можна зобразити у вигляді

$$\tilde{u}(x) = \sum_{k=1}^{N} u_k \varphi_k(x),$$

де $u_k = u(x_{1,k}, x_{2,k})$; N — число вузлів сітки в $\overline{\Omega}$. Очевидно, $\tilde{u}(x_k) = u_k$. У кожному трикутнику тріангуляції функція $\tilde{u}(x)$, з одного боку, може бути подана у вигляді

$$\tilde{u}(x) = ax_1 + bx_2 + c, \tag{21}$$

а з другого (рис. 20), у вигляді

$$\tilde{u}(x) = u_i \varphi_i(x) + u_j \varphi_j(x) + u_k \varphi_k(x). \tag{22}$$

Щоб визначити коефіцієнти a, b, c, треба розв'язати систему

$$ax_{1,i} + bx_{2,i} + c = u_i,$$
$$ax_{1,j} + bx_{2,j} + c = u_j,$$
$$ax_{1,k} + bx_{2,k} + c = u_k.$$

Розв'язавши її, матимемо

$$a = \frac{1}{2S}[(x_{2,j} - x_{2,k})u_i + (x_{2,k} - x_{2,i})u_j + (x_{2,i} - x_{2,j})u_k],$$

$$b = \frac{1}{2S}[(x_{1,k} - x_{1,j})u_i + (x_{1,i} - x_{1,k})u_j + (x_{1,j} - x_{1,i})u_k],$$

$$c = \frac{1}{2S}[(x_{1,j}x_{2,k} - x_{1,k}x_{2,j})u_i + (x_{1,k}x_{2,i} - x_{1,i}x_{2,k})u_j + (x_{1,i}x_{2,j} - x_{1,j}x_{2,i})u_k],$$

де S — площа трикутника, яку можна обчислити за формулою

$$S = \frac{1}{2}\left|\det\begin{bmatrix} 1 & x_{1,i} & x_{2,i} \\ 1 & x_{1,j} & x_{2,j} \\ 1 & x_{1,k} & x_{2,k} \end{bmatrix}\right|.$$

Якщо підставити тепер вирази для a, b, c в (21) і перетворити до вигляду (22), то дістанемо зображення для функцій $\varphi_i(x)$, $\varphi_j(x)$, $\varphi_k(x)$ у даному трикутнику, а саме:

$$\varphi_i(x) = \frac{1}{2S}(\alpha_i x_1 + \beta_i x_2 + \gamma_i), \qquad (23)$$

$$\alpha_i = x_{2j} - x_{2,k}, \quad \beta_i = x_{1,k} - x_{1,j}, \quad \gamma_i = x_{1,j}x_{2,k} - x_{1,k}x_{2,j}.$$

Функції $\varphi_j(x)$ і $\varphi_k(x)$ можна знайти з формули для $\varphi_i(x)$ циклічною перестановкою індексів $i \to j \to k \to i$ при α, β, γ, x_1, x_2 (наприклад, $\alpha_j = x_{2,k} - x_{1,i}$ і т. д.). Явні зображення функцій $\varphi_i(x)$, $\varphi_j(x)$, $\varphi_k(x)$ використовуються для побудови систем сіткових рівнянь.

Перейдемо тепер до задачі Діріхле для рівняння Пуассона у многокутнику $\overline{\Omega}$. Позначимо через $\mathring{H}_h$ підпростір простору $W_2^1(\Omega)$ неперервних кусково-лінійних функцій, які дорівнюють нулю в граничних вузлах, тобто підпростір функцій виду

$$\tilde{v}(x) = \sum_{k=1}^{N_0} y_k \varphi_k(x),$$

де N_0 — кількість внутрішніх вузлів в Ω. Наближений розв'язок задачі (1), (2) з п. 5.1 (див. також (9), (10), п. 5.1) шукатимемо у вигляді

$$\tilde{y}(x) = \sum_{k=1}^{N_0} y_k \varphi_k(x),$$

а наближені значення y_k шукатимемо з рівняння

$$a(\tilde{y}, \tilde{v}) = (f, \tilde{v}) \; \forall \tilde{v} \in \mathring{H}_h,$$

яке еквівалентне такій системі сіткових рівнянь:

$$a(\tilde{y}, \varphi_k) = (f, \varphi_k), \quad k = \overline{1, N_0}. \qquad (24)$$

Це і є система методу скінченних елементів. Для формування матриці цієї системи, як і раніше, зручно користуватися матрицями жорсткості.

Вправа 4. Впевнитися, що матриця жорсткості має вигляд

$$K^{(i)} = \frac{1}{4S}\begin{vmatrix} \alpha_1^2 + \beta_1^2 & \alpha_1\alpha_2 + \beta_1\beta_2 & \alpha_1\alpha_3 + \beta_1\beta_3 \\ \alpha_1\alpha_2 + \beta_1\beta_2 & \alpha_2^2 + \beta_2^2 & \alpha_2\alpha_3 + \beta_2\beta_3 \\ \alpha_1\alpha_3 + \beta_1\beta_3 & \alpha_2\alpha_3 + \beta_2\beta_3 & \alpha_3^2 + \beta_3^2 \end{vmatrix}$$

(вузли трикутника пронумеровано числами 1, 2, 3).

На цьому ми закінчимо виклад алгоритму МСЕ для многокутника, хоча для його практичного застосування не з'ясовано такі питання, як автоматизація тріангуляції області $\overline{\Omega}$, нумерування вузлів сітки, формування в деталях матриці системи (24), збільшення точності і т. д. Висловимо деякі зауваження.

Якщо побудовано деяку тріангуляцію області Ω, то тріангуляцію з трикутниками менших розмірів неважко побудувати (а це, очевидно, має привести до збільшення точності наближеного розв'язку). Наприклад, з'єднавши в кожному трикутнику середини сторін між собою, дістанемо нову тріангуляцію. Зауважимо, що тріангуляція визначає максимальне число невідомих, які входять в одне рівняння системи (27) (це залежить від кількості трикутників, що мають вузол спільною вершиною). Якщо поставити умову, щоб кути в трикутниках тріангуляції були меншими від деякого $\theta_0 > 0$ незалежно від кількості трикутників тріангуляції області Ω, то максимальне число невідомих в одному рівнянні (тобто максимальне число відмінних від нуля елементів у даному рядку матриці системи (24)) не залежатиме від загального числа невідомих. На структуру матриці системи впливає порядок глобальної нумерації невідомих, і тому треба прагнути до такої нумерації, при якій кількість ненульових діагоналей була б мінімальною. Це дає певні переваги при розв'язуванні системи рівнянь (24) на ЕОМ. Існують алгоритми оптимальної нумерації. При побудові глобальної матриці жорсткості вершини всіх трикутників нумеруються підряд від 1 до N, включаючи і ті, що лежать на границі області. Розглянемо, як при цьому можна врахувати, скажімо, граничні умови Діріхле. При такій суцільній нумерації значення функціонала $J(\tilde{u}) = a(\tilde{u}, \tilde{u}) - 2(f, \tilde{u})$ на наближеному розв'язку $\tilde{u}(x) = \sum_{k=1}^{N} u_k \varphi_k(x)$ зображається у вигляді

$$J(\tilde{u}) = U^T K U - 2U^T F = U^T K U - 2F^T U,$$

де глобальна матриця жорсткості K та глобальний вектор навантаження F виражено через суми добутків матриці кінематичних зв'язків

та матриць жорсткості і векторів навантажень окремих елементів відповідно до прийнятої нумерації, причому матриця K є симетричною, але сингулярною. Нехай у вузлі з номером j задано граничну умову Діріхле, тобто $u_j = \varphi_j$, де φ_j — відоме значення. Щоб врахувати її, зауважимо, що в цьому випадку квадратична частина U^TKU функціонала енергії містить лінійні члени $(k_{ij}u_i\varphi_j + k_{ji}\varphi_j u_i) = 2k_{ij}u_i\varphi_j$ $(i \neq j)$, які можна врахувати в лінійній частині $2U^TF$. Це означає, що j-й стовпчик матриці A, помножений на p_j, потрібно додати до вектора F. Після цього потрібно було б викреслити j-й стовпчик у матриці A та k-ту компоненту вектора F. Замість цього всі елементи j-го стовпчика, крім діагонального, матриці A заміняємо нулями, діагональний елемент k_{ij} покладаємо рівним 1 і j-й елемент вектора F рівним φ_j. Описана операція проводиться для всіх граничних вузлів. Після цього, як звичайно, потрібно розв'язати систему лінійних алгебраїчних рівнянь $KU = F$ з модифікованими матрицею K та вектором F. Ця система містить ряд тривіальних рівнянь, що відповідають граничним умовам. Розв'язок U містить також ті компоненти, які задано граничними умовами Діріхле, що буває зручно при подальшій обробці цього вектора, наприклад при побудові ліній рівня наближеного розв'язку. Для розв'язування системи $KU = F$ часто застосовуються модифікації методу Холецького та ітераційні методи.

При обчисленні елементів матриці системи (23) виникає необхідність знаходження інтегралів від поліномів по трикутнику. Це зручно зробити, ввівши так звану барицентричну систему координат.

Розглянемо трикутник з вершинами i $(x_{1,i}, x_{2,i})$, j $(x_{1,j}, x_{2,j})$, k $(x_{1,k}, x_{2,k})$ (рис. 21). Положення будь-якої точки (x_1, x_2) всередині трикутника можна однозначно задати за допомогою трьох чисел:

$$\zeta_i = \frac{S_{0jk}}{S}, \quad \zeta_j = \frac{S_{0ki}}{S}, \quad \zeta_k = \frac{S_{0ij}}{S},$$

де S — площа всього трикутника з вершинами i, j, k; S_{0jk} — площа трикутника Okj і т. д. Очевидно,

$$\zeta_i + \zeta_j + \zeta_k = 1.$$

Старі і нові координати зв'язані формулами (25), якщо

$$S_{0jk} = \frac{1}{2}\det\begin{bmatrix} 1 & x_1 & x_2 \\ 1 & x_{1,j} & x_{2,j} \\ 1 & x_{1,k} & x_{2,k} \end{bmatrix},$$

тобто

$$\zeta_i = \frac{1}{2S}(\alpha_i x_1 + \beta_i x_2 + \gamma_i),$$

де α_i, β_i, γ_i — ті самі, що і в (23). Координати ζ_j, ζ_k знаходяться з ζ_i

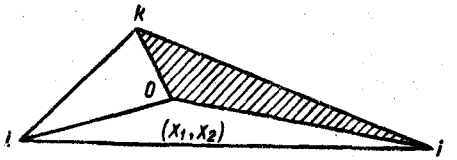


Рис. 21

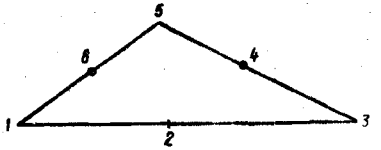


Рис. 22

шляхом циклічної перестановки індексів. Із формули (23) випливає, що координатні змінні ζ_i, ζ_j, ζ_k зв'язані з координатними функціями:

$$\zeta_i = \varphi_i(x), \quad \zeta_j = \varphi_j(x), \quad \zeta_k = \varphi_k(x).$$

Для обчислення інтегралів від многочленів по трикутнику Δ_{ijk} з вершинами i, j, k можна використати формулу

$$\int_{\Delta_{ijk}} \zeta_i^p \zeta_j^q \zeta_k^r dx = \frac{p!q!r!}{(p+q+r)!} 2S.$$

Точність методу скінченних елементів можна підвищувати не лише подрібненням тріангуляції, але й вибором базисних функцій.

Наведемо приклад ще одного класу неперервних кусково-поліноміальних базисних функцій. Для цього до вузлів, які лежать у вершинах трикутників, приєднаємо нові вузли — середини сторін трикутників (рис. 22).

Нехай $u(x)$ — неперервна в Ω. Визначимо функцію

$$\tilde{u}(x) = \sum_{k=1}^{6} u_k \varphi_k(x), \tag{25}$$

яка збігається з $u(x)$ у вузлах сітки, тобто $\tilde{u}(x_k) = u_k$, k — номер вузла і є поліномом другого степеня на кожному трикутнику тріангуляції. Тут $\varphi_k(x)$ — базисні функції на трикутнику, які є многочленами другого степеня, такі, що $\varphi_k(x_k) = 1$ і $\varphi_k(x_l) = 0$ при $k \neq l$ (шість умов однозначно визначають шість коефіцієнтів многочлена другого степеня). Неважко помітити, що в барицентричних координатах

$$\begin{gathered}\varphi_1 = \zeta_1(2\zeta_1 - 1), \quad \varphi_2 = 4\zeta_1\zeta_2, \quad \varphi_3 = \zeta_2(2\zeta_2 - 1), \\ \varphi_4 = 4\zeta_2\zeta_3, \quad \varphi_5 = \zeta_3(2\zeta_3 - 1), \quad \varphi_6 = 4\zeta_3\zeta_1.\end{gathered} \tag{26}$$

Функція $\tilde{u}(x)$ неперервна в Ω, бо при визначенні її на кожному трикутнику співвідношеннями (25), (26) вона є поліномом від однієї змінної другого степеня на кожній стороні і тому однозначно визначається в трьох вузлах на цій стороні.

Множина кусково-квадратичних функцій є підмножиною $W_2^1(\Omega)$, а множина кусково-квадратичних функцій, рівних нулю на $\partial\Omega$, є підмножиною $\mathring{W}_2^1(\Omega)$. Базис у кожному з цих підпросторів очевидно

будується за допомогою виразів (25), (26) і має ті самі властивості, що і базис у просторі кусково-лінійних неперервних функцій: кожна функція дорівнює одиниці лише в одному вузлі сітки і нулю в усіх інших, причому вона відмінна від нуля лише в трикутниках, які мають спільною вершиною даний вузол. А далі застосовується стандартна схема МСЕ, яка приводить до системи виду (24).

5.3.3. Збіжність методу скінченних елементів у многокутнику. Згадаємо, що точний і наближений розв'язки задачі Діріхле для рівняння Пуассона визначаються відповідно рівностями

$$a(v, u) = (f, v)\ \forall v \in \mathring{W}_2^1(\Omega), \tag{27}$$

$$a(\tilde{y}, \tilde{v}) = (f, \tilde{v})\ \ \forall \tilde{v} \in \mathring{H}_h. \tag{27'}$$

Вибравши $v = \tilde{v}$ і віднявши рівність (27′) від (27), дістанемо

$$a(u - \tilde{y}, \tilde{v}) = 0 \quad \forall \tilde{v} \in \mathring{H}_h.$$

Покладемо $\tilde{v} = \tilde{y} - \tilde{w}$, де $\tilde{w}$ — довільна функція з $\mathring{H}_h$:

$$a(u - \tilde{y}, \tilde{y} - \tilde{w}) = 0\ \forall\ \tilde{w} \in \mathring{H}_h.$$

Цю рівність перетворимо так:

$$a(u - \tilde{y}, u - \tilde{y}) = a(u - \tilde{y}, u - \tilde{w})\ \forall\ \tilde{w} \in \mathring{H}_h.$$

Неважко помітити, що $a(\varphi, \psi)$ можна взяти за скалярний добуток в $\mathring{W}_2^1(\Omega)$, тому з нерівності Коші — Буняковського

$$|a(\varphi, \psi)| \leqslant \sqrt{a(\varphi, \varphi)}\sqrt{a(\psi, \psi)} = |\varphi|\,|\psi|,$$

де $|\varphi| = \sqrt{a(\varphi, \varphi)}$ — норма, що породжується скалярним добутком $a(\varphi, \psi)$. Звідси маємо

$$a(u - \tilde{y}, u - \tilde{y}) \equiv |u - \tilde{y}|^2 \leqslant |u - \tilde{y}|\,|u - \tilde{w}|\ \forall\ \tilde{w} \in \mathring{H}_h,$$

або

$$|u - \tilde{y}| \leqslant |u - \tilde{w}|\ \forall\ \tilde{w} \in \mathring{H}_h, \tag{28}$$

що еквівалентно

$$|u - \tilde{y}| = \min_{\tilde{w} \in \mathring{H}_h} |u - \tilde{w}|. \tag{29}$$

Це і є основне співвідношення, яке, як і в одновимірному випадку (див. п. 3.2, ч. 1), показує, що розв'язок, знайдений у МСЕ, є найкращою апроксимацією точного розв'язку в енергетичній нормі функціями з $\mathring{H}_h$, і, таким чином, питання про збіжність зведене до питання про

апроксимацію. На відміну від п. 4.4 рівність (29) ми дістали без зв'язку з варіаційною постановкою задачі, що може бути застосовано і для задач, в яких диференціальне рівняння (1) з п. 5.1 містить перші похідні.

Далі вважатимемо, що точний розв'язок $u(x) \in W_2^2(\Omega) \cap \mathring{W}_2^1(\Omega)$, і тому, за теоремами вкладення, $u \in C(\bar{\Omega})$. Розглянемо МСЕ з кусково-лінійними координатними функціями. Нехай $\Omega = \bigcup_{i=1}^{M} e_i$, e_i — трикутник тріангуляції з номером i (вважаємо, що вузли сітки і трикутники мають відповідну нумерацію). Функції $u(x)$ поставимо у відповідність сіткову функцію $u_h = \{u(x_{1i}, x_{2i})\}_{i=1}^{N}$. Через $\tilde{u}(x)$ позначимо кусково-лінійну інтерполяцію сіткової функції u_h, яка є інтерполянтом $u(x)$ у кожному трикутнику. В силу (28)

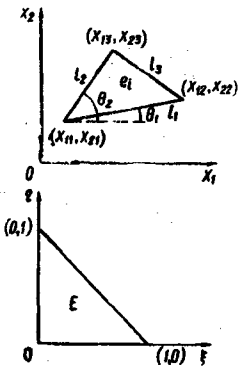


Рис. 23

$$|u - \tilde{y}| \leqslant |u - \tilde{u}|. \tag{30}$$

Оцінимо величину $|u - \tilde{u}|$. Маємо

$$|u - \tilde{u}|^2 = a(u - \tilde{u}, u - \tilde{u}) = \sum_{i=1}^{M} \int_{e_i} \left[\left(\frac{\partial (u - \tilde{u})}{\partial x_1} \right)^2 + \left(\frac{\partial (u - \tilde{u})}{\partial x_2} \right)^2 \right] dx = \sum_{i=1}^{M} a_i (u - \tilde{u}, u - \tilde{u}). \tag{31}$$

Щоб оцінити $a_i(u - \tilde{u}, u - \tilde{u})$, спочатку за допомогою заміни $x_1 = x_{11} + (x_{12} - x_{11})\,\xi + (x_{13} - x_{11})\,\eta$, $x_2 = x_{21} + (x_{22} - x_{21})\,\xi + (x_{23} - x_{21})\,\eta$ відобразимо трикутник l_i у площині координат (x_1, x_2) на трикутник E у площині координат (ξ, η) (рис. 23). Якобіан цього відображення

$$I(\xi, \eta) = \det \begin{vmatrix} x_{12} - x_{11} & x_{13} - x_{11} \\ x_{22} - x_{21} & x_{23} - x_{21} \end{vmatrix} = (x_{12} - x_{11})(x_{23} - x_{21}) - (x_{22} - x_{21})(x_{13} - x_{11}) = \tag{32}$$

$$= l_1 l_2 \left(\frac{x_{12} - x_{11}}{l_1} \cdot \frac{x_{23} - x_{21}}{l_2} - \frac{x_{22} - x_{21}}{l_1} \cdot \frac{x_{13} - x_{11}}{l_2} \right) =$$

$$= l_1 l_2 (\cos\theta_1 \sin\theta_2 - \sin\theta_1 \cos\theta_2) = l_1 l_2 \sin(\theta_2 - \theta_1),$$

де l_i — довжини сторін трикутника. Якщо θ_0 — мінімальний з усіх кутів усіх трикутників тріангуляції, то, очевидно,

$$h^2 \geqslant |I(\xi, \eta)| \geqslant h_{\min}^2 \sin\theta_0,$$

де $h_{\min}$ — мінімальна з довжин сторін трикутників; h — максимальна з цих довжин.

Обернене до (31) відображення має вигляд

$$\xi = \frac{(x_{23}-x_{21})(x_1-x_{11})-(x_{13}-x_{11})(x_2-x_{21})}{(x_{23}-x_{21})(x_{12}-x_{11})-(x_{13}-x_{11})(x_{22}-x_{21})}, \qquad (33)$$

$$\eta = \frac{(x_{22}-x_{21})(x_1-x_{11})-(x_{12}-x_{11})(x_2-x_{21})}{(x_{22}-x_{21})(x_{13}-x_{11})-(x_{22}-x_{11})(x_{23}-x_{21})}.$$

Як відомо, якобіан $\tilde{I}(x_1, x_2)$ цього відображення зв'язаний з якобіаном $I(\xi, \eta)$ співвідношенням

$$\tilde{I}(x_1, x_2)\, I(\xi, \eta) = 1. \qquad (34)$$

Позначимо $U(\xi, \eta) = u(x_1(\xi, \eta), x_2(\xi, \eta))$, $\tilde{U}(\xi, \eta) = \tilde{u}(x_1(\xi, \eta), x_2(\xi, \eta))$. Тоді маємо

$$\frac{\partial u}{\partial x_1} = \frac{\partial U}{\partial \xi}\xi_{x_1} + \frac{\partial U}{\partial \eta}\eta_{x_1}, \quad \frac{\partial u}{\partial x_2} = \frac{\partial U}{\partial \xi}\xi_{x_2} + \frac{\partial U}{\partial \eta}\eta_{x_2},$$

тому, користуючись нерівністю $(a+b)^2 \leqslant 2(a^2+b^2)$, знаходимо

$$\xi_{x_1} = \frac{(x_{23}-x_{21})}{(x_{23}-x_{21})(x_{12}-x_{11})-(x_{13}-x_{11})(x_{22}-x_{21})} =$$

$$= \frac{x_{23}-x_{21}}{l_1 l_2\left(\frac{x_{23}-x_{21}}{l_2}\frac{x_{12}-x_{11}}{l_1} - \frac{x_{13}-x_{11}}{l_2}\frac{x_{22}-x_{21}}{l_1}\right)} =$$

$$= \frac{x_{23}-x_{21}}{l_1 l_2(\sin\theta_2\cos\theta_1 - \cos\theta_2\sin\theta_1)} = \frac{x_{23}-x_{21}}{l_1 l_2 \sin(\theta_2-\theta_1)}, \qquad (35)$$

тобто

$$|\xi_{x_1}| \leqslant \frac{h}{h_{\min}^2 \sin\theta_0}.$$

Аналогічно визначаються оцінки

$$|\xi_{x_2}|, \quad |\eta_{x_1}|, \quad |\eta_{x_2}| \leqslant \frac{h}{h_{\min}^2 \sin\theta_0}.$$

Вважаючи, що тріангуляція квазірегулярна, тобто

$$h \leqslant c_0 h_{\min},$$

де $c_0 > 0$ — стала, незалежна від h, маємо

$$|\xi_{x_\alpha}|, \quad |\eta_{x_\alpha}| \leqslant \frac{c_0}{h_{\min}\sin\theta_0}, \quad \alpha = 1, 2. \qquad (36)$$

Використовуючи (32), (36), з (35) дістаємо

$$a_i(u-\tilde{u}, u-\tilde{u}) \leqslant \frac{4c_0^2 l_1 l_2 \sin(\theta_2-\theta_1)}{h_{\min}^2 \sin^2\theta_0} \int_E \left[\left(\frac{\partial(U-\tilde{U})}{\partial\xi}\right)^2 + \right.$$

$$\left. + \left(\frac{\partial(U-\tilde{U})}{\partial\eta}\right)^2\right] d\xi d\eta \leqslant \frac{4c_0^2 l_1 l_2 \sin(\theta_2-\theta_1)}{h_{\min}^2 \sin^2\theta_0} \|U-\tilde{U}\|^2_{W_2^2(E)}. \qquad (37)$$

Розглянемо такі лінійні функціонали від функції $v \in W_2^2(E)$

$$l_1(v) = v(0,0), \quad l_2(v) = v(1,0), \quad l_3(v) = v(0,1),$$

які в силу теореми вкладення є обмеженими в $W_2^2(E)$. Неважко також помітити, що ці функціонали не обертаються одночасно на нуль на жодному не рівному тотожно нулю многочлені, не вищому першого степеня. Тому за теоремою Соболєва про еквівалентне нормування простору $W_2^2(E)$ маємо

$$\|v\|^2_{W_2^2(E)} \leqslant c\left(|v|^2_{W_2^2(E)} + \sum_{i=1}^{3}|l_i(v)|^2\right),$$

де стала c не залежить від v,

$$\|v\|^2_{W_2^2(E)} = \int_E \sum_{|\alpha|=2}(D^\alpha v)^2 d\xi d\eta.$$

Звідси для функцій v, які дорівнюють нулю у вершинах трикутника E, маємо

$$\|v\|^2_{W_2^2(E)} \leqslant c\,|v|^2_{W_2^2(E)}.$$

Оскільки функція $U-\tilde{U}$ якраз і дорівнює нулю у вершинах трикутника E, то звідси для (37) дістаємо наступну оцінку

$$a_i(u-\tilde{u}, u-\tilde{u}) \leqslant \frac{4cc_0^2 l_1 l_2 \sin(\theta_2-\theta_1)}{h_{\min}^2 \sin^2\theta_0} \cdot |U-\tilde{U}|^2_{W_2^2(E)}. \qquad (38)$$

Перейдемо у виразі (38) до старих координат x_1, x_2 за допомогою оберненого відображення (33). При цьому врахуємо, що

$$\frac{\partial U}{\partial\xi} = \frac{\partial u}{\partial x_1}\frac{\partial x_1}{\partial\xi} + \frac{\partial u}{\partial x_2}\frac{\partial x_2}{\partial\xi}, \quad \frac{\partial U}{\partial\eta} = \frac{\partial u}{\partial x_1}\frac{\partial x_1}{\partial\eta} + \frac{\partial u}{\partial x_2}\frac{\partial x_2}{\partial\eta};$$

$$\frac{\partial^2 U}{\partial\xi^2} = \left(\frac{\partial x_1}{\partial\xi}\right)^2\frac{\partial^2 u}{\partial x_1^2} + 2\frac{\partial x_1}{\partial\xi}\frac{\partial x_2}{\partial\xi}\frac{\partial^2 u}{\partial x_1\partial x_2} + \left(\frac{\partial x_2}{\partial\xi}\right)^2\frac{\partial^2 u}{\partial x_2^2};$$

$$\frac{\partial^2 U}{\partial\eta^2} = \left(\frac{\partial x_1}{\partial\eta}\right)^2\frac{\partial^2 u}{\partial x_1^2} + 2\frac{\partial x_1}{\partial\eta}\frac{\partial x_2}{\partial\eta}\frac{\partial^2 u}{\partial x_1\partial x_2} + \left(\frac{\partial x_2}{\partial\eta}\right)^2\frac{\partial^2 u}{\partial x_2^2},$$

$$\frac{\partial^2 U}{\partial\xi\partial\eta} = \frac{\partial x_1}{\partial\xi}\frac{\partial x_1}{\partial\eta}\frac{\partial^2 u}{\partial x_1^2} + \left(\frac{\partial x_1}{\partial\xi}\frac{\partial x_2}{\partial\eta} + \frac{\partial x_2}{\partial\xi}\frac{\partial x_1}{\partial\eta}\right)\frac{\partial^2 u}{\partial x_1 \partial x_2} +$$

$$+ \frac{\partial x_2}{\partial\xi}\frac{\partial x_2}{\partial\eta}\frac{\partial^2 u}{\partial x_2^2},$$

$$\max\left(\frac{\partial x_\alpha}{\partial\xi}, \frac{\partial x_\alpha}{\partial\eta}\right) \leqslant h, \quad \alpha = 1, 2.$$

Дістанемо (див. (34))

$$|U - \tilde{U}|^2_{W_2^2(E)} \leqslant \left[\max_\alpha \max\left(\left|\frac{\partial x_\alpha}{\partial\xi}\right|, \left|\frac{\partial x_\alpha}{\partial\eta}\right|\right)\right]^4 \times$$

$$\times \int_{e_i} |\tilde{l}(x_1, x_2)| \sum_{|\alpha|=2} (D^\alpha u)^2 \, dx_1 dx_2 \leqslant \frac{h^4}{l_1 l_2 \sin(\theta_2 - \theta_1)} |u|^2_{W_2^2(e_i)}, \quad (39)$$

тому з (38) маємо

$$a_i(u - \tilde{u}, u - \tilde{u}) \leqslant \frac{4cc_0^2 h^4}{h_{\min}^2 \sin^2\theta_0} |u|^2_{W_2^2(e_i)} \leqslant \frac{4cc_0^4 h^2}{\sin^2\theta_0} |u|^2_{W_2^2(e_i)}.$$

Підсумовуючи по всіх $i = \overline{1, M}$, дістаємо

$$\|u - \tilde{u}\|^2 = \sum_{i=1}^{M} a_i(u - \tilde{u}, u - \tilde{u}) \leqslant c \frac{h^2}{\sin^2\theta_0} |u|^2_{W_2^2(\Omega)},$$

і з (30) випливає остаточна оцінка

$$\|u - \tilde{y}\| \leqslant c \frac{h^2}{\sin^2\theta_0} |u|^2_{W_2^2(\Omega)}, \quad (40)$$

де c — незалежна від h та u стала.

Таким чином, доведено таке твердження.

Теорема. *Нехай точний розв'язок u задачі Діріхле для рівняння Пуассона* (27) *у многокутнику Ω належить простору $W_2^2(\Omega) \cap \overset{\circ}{W}{}_2^1(\Omega)$, а $\tilde{y}$ — наближений розв'язок, знайдений з* (27), *причому тріангуляція області Ω квазірегулярна, тобто $h \leqslant c_0 h_{\min}$, де h, $h_{\min}$ — відповідно максимальна і мінімальна довжини сторін трикутників тріангуляції, c_0 — незалежна від h додатна стала, причому всі внутрішні кути цих трикутників обмежені знизу незалежно від h сталою $\theta_0 > 0$. Тоді при $h \to 0$ наближений розв'язок збігається до точного u і має місце оцінка точності* (40).

Прийом відображення e_i на E, який ми застосували при доведенні теореми, називається *масштабуванням* і має багато спільного з аналогічним прийомом, коли застосовується лема Брембла — Гільберта. Вибір теореми Соболєва про еквівалентне нормування чи леми Брембла — Гільберта залежить від зручності і обсягу обчислень.

Вправа 5. Аналогічно одновимірному випадку (див. п. 4.4.3) в припущенні, що в задачі (1), (2), п. 5.1 $f \in L_2(\Omega)$, $u(x) \in W_2^2(\Omega)$ і виконується оцінка $\|u\|_{W_2^2(\Omega)} \leqslant c\|f\|_{L_2(\Omega)}$, де c — стала, що не залежить від u, f, довести, що для схеми методу скінченних елементів у многокутнику з кусково-лінійними базисними функціями виконується оцінка

$$\|u - \tilde{y}\|_{0,\Omega} \equiv \|u - \tilde{y}\|_{L_2(\Omega)} \leqslant ch^2 \|f\|_{0,\Omega},$$

де стала c не залежить від h, u.

5.4. Розв'язування сіткових рівнянь

Розглянуті вище сіткові методи зводяться до розв'язування системи лінійних алгебраїчних рівнянь (СЛАР). СЛАР методу сіток, як правило, мають такі особливості: а) велика кількість невідомих (у двовимірному сітковому прямокутнику внутрішніх вузлів $(N_1 - 1)(N_2 - 1)$; для схеми точності $O(h^2)$ звичайно беруть N_1, $N_2 \sim (50...100)$, тобто число рівнянь у системі методу сіток має порядок $(10^3...10^4)$; б) специфічне розміщення ненульових елементів матриці системи і мала кількість їх порівняно з кількістю нульових елементів. Звичайно, такі системи можна розв'язувати і методом Гаусса, в якому кількість арифметичних операцій пропорціональна кубу кількості невідомих. При кількості невідомих $10^3...10^4$ це число дорівнює $10^9...10^{12}$ арифметичних операцій, що забагато навіть для сучасних ЕОМ. Зауважимо, що для розв'язку «великих» СЛАР застосовують ітераційні методи, але вони є найефективнішими тоді, коли відомі границі власних значень матриці СЛАР. Якщо ж спектр матриці досить широкий, що зустрічається не так рідко при застосуванні методу сіток, то ітераційні методи повільно збігаються, що робить їх практично неефективними. Крім того, багато сіткових схем погано обумовлені (див. п. 4.3).

Однак специфіка великого класу сіткових схем дає змогу побудувати для них ефективні алгоритми. Приймемо класифікацію методів розв'язування сіткових СЛАР (див. п. 1.2, 1.4 з ч. I), яку наведено на рис. 24.

5.4.1. Метод сумарних зображень та метод розділення змінних. Використання швидкого перетворення Фур'є. Ці два методи схожі між собою і мають аналоги в класичних методах математичної фізики: метод скінченних інтегральних перетворень, метод функції Гріна, метод розділення змінних. Головна ідея полягає в зведенні двовимірної задачі до одновимірної.

Розглянемо спочатку *метод сумарних зображень* (*МСЗ*). Нехай потрібно знайти розв'язок сіткової задачі

$$\begin{aligned} \Lambda y_{ij} &= f_{ij}, \quad (x_{1i}, x_{2j}) \in \omega_h, \\ l y_{ij} &= \gamma_{ij}, \quad (x_{1i}, x_{2j}) \in \gamma_h, \end{aligned} \quad (1)$$

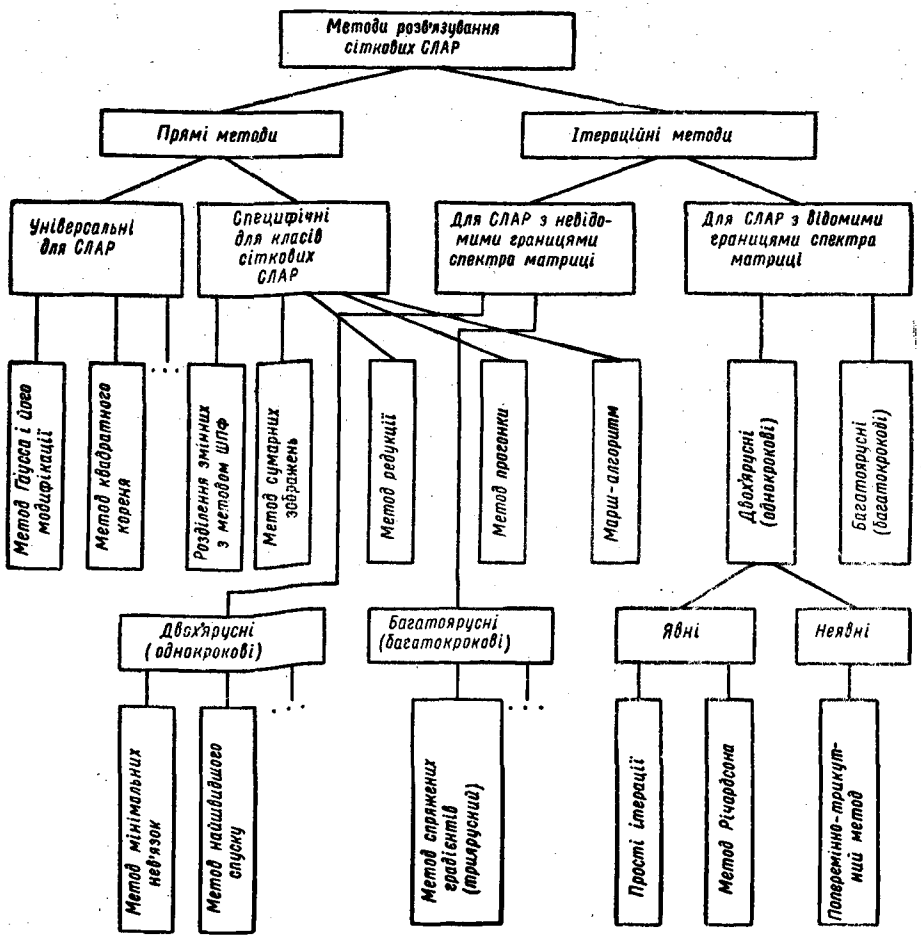


Рис. 24

де $\Lambda = \Lambda_1 + \Lambda_2$, $\Lambda_1 y_{ij} = a_i^{(1)} y_{i+1,j} - d_i^{(1)} y_{ij} + b_i^{(1)} y_{i-1,j}$, $\Lambda_2 y_{ij} = a_j^{(2)} y_{i,j+1} - d_j^{(2)} y_{ij} + b_j^{(2)} y_{i,j-1}$, $\omega_h = \{x = (x_{1j}, x_{2j})$: $x_{1i} = ih_1$, $x_{2j} = jh_2$, $i = \overline{1, N_1 - 1}$, $j = \overline{1, N_2 - 1}\}$ (сітковий прямокутник), $\gamma_h = \bigcup_{\alpha=1}^{2} \gamma_{\pm\alpha}$, $\gamma_{-1} = \{(x_{10}, x_{2j})$: $j = \overline{1, N_2 - 1}\}$, $\gamma_{+1} = \{(x_{1N_1}, x_{2j})$: $j = \overline{1, N_2 - 1}\}$, $\gamma_{-2} = \{(x_{1i}, x_{20})$: $i = \overline{1, N_1 - 1}\}$, $\gamma_{+2} = \{(x_{1i}, x_{2N2})$, $i = \overline{1, N_1 - 1}\}$, l—деякий сітковий граничний оператор. Візьмемо

для визначеності $ly_{ij} \equiv y_{ij}$. Запишемо задачу (1) у векторному вигляді:

$$\Pi^{(1)} \vec{y}_j + \Lambda_2 \vec{y}_j = \vec{f}_j - \vec{\omega}_j, \tag{2}$$

$$\vec{y}_0 = \vec{\gamma}_0, \quad \vec{y}_{N_2} = \vec{\gamma}_{N_2},$$

де

$$\vec{y}_j = (y_{ij})_{i=\overline{1,N_1-1}}, \quad \vec{f}_j = (f_{ij})_{i=\overline{1,N_1-1}},$$

$$\vec{\omega}_j = (b_1^{(1)} \gamma_{0j}, 0, \ldots, 0, \gamma_{N_1 j}),$$

$$\Pi^{(1)} = \begin{bmatrix} -d_1^{(1)} & a_1^{(1)} & 0 & 0 & \ldots & 0 & 0 & 0 \\ b_2^{(1)} & -d_2^{(1)} & a_2^{(1)} & 0 & \ldots & 0 & 0 & 0 \\ 0 & b_3^{(1)} & -d_3^{(1)} & a_3^{(1)} & \ldots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \ldots & b_{N_1-2}^{(1)} & -d_{N_1-2}^{(1)} & a_{N_1-2}^{(1)} \\ 0 & 0 & 0 & 0 & \ldots & 0 & b_{N_1-1}^{(1)} & -d_{N_1-1}^{(1)} \end{bmatrix}.$$

Нехай матриця $\Pi^{(1)}$ є *матрицею простої структури*, тобто подібна діагональній матриці $M = \operatorname{diag}(\mu_i)_{i=\overline{1,N_1-1}}$:

$$\Pi^{(1)} = PMP^{-1}.$$

Матриця P називається *фундаментальною*, і її стовпчиками є, очевидно, власні вектори матриці $\Pi^{(1)}$, які відповідають власним числам μ_i.

Перетворення

$$\hat{\vec{v}} = P^{-1}\vec{v} = (\hat{v}_i)_{i=\overline{1,N_1-1}} \tag{3}$$

називатимемо *P-трансформацією вектора* $\vec{v}$, а перетворення

$$\vec{v} = P\hat{\vec{v}}$$

оберненою P-трансформацією. Проведемо P-трансформацію задачі (2) множенням її зліва на матрицю P^{-1}:

$$M\hat{\vec{y}}_j + \Lambda_2 \hat{\vec{y}}_j = \hat{\vec{f}}_j - \hat{\vec{\omega}}_j, \tag{4}$$

$$\hat{\vec{y}}_0 = \hat{\vec{\gamma}}_0, \quad \hat{\vec{y}}_{N_2} = \hat{\vec{\gamma}}_{N_2}.$$

У скалярному вигляді (4) запишеться так:

$$a_j^{(2)} \hat{y}_{i,j+1} - (-\mu_i + d_j^{(2)}) \hat{y}_{ij} + b_j^{(2)} \hat{y}_{i,j-1} = \hat{f}_{ij} - \hat{\omega}_{ij}, \tag{5}$$

$$\hat{y}_{i0} = \hat{\gamma}_{i0}, \quad \hat{y}_{i,N_2} = \hat{\gamma}_{i,N_2}, \quad i = \overline{1, N_1 - 1}, \quad j = \overline{1, N_2 - 1}.$$

Система (5) являє собою СЛАР з тридіагональною матрицею і тому може бути, в принципі, ефективно розв'язана методом прогонки, після чого шукані вектори $\vec{y}_j$ можна знайти за формулою

$$\vec{y}_j = \hat{P}\hat{y}_j. \tag{6}$$

Мета методу сумарних зображень — побудова формул сумарних зображень (ФСЗ), тобто деякого аналітичного варіанта розв'язку задачі (1). Тому припустимо, що для кожного фіксованого i крайова задача (5) має сіткову функцію Гріна $G_{lj}(\mu_i)$, тобто функцію, за допомогою якої розв'язок задачі (5) можна записати у вигляді (при фіксованому i)

$$\hat{y}_{ij} = b_1^{(2)} G_{1j}(\mu_i)\hat{\gamma}_{i0} + a_{N_2}^{(2)} G_{N_2 j}(\mu_i)\hat{\gamma}_{i,N_2} - \\ - \sum_{l=1}^{N_2-1} G_{lj}(\mu_i)[\hat{f}_{il} - \hat{\omega}_{il}], \quad j = \overline{1,\ N_2-1}. \tag{7}$$

Перейшовши у виразі (7) до векторної формули запису і виконавши обернену P-трансформацію множенням зліва на P, дістанемо явну формулу для розв'язку задачі (1), тобто ФСЗ першого типу:

$$\vec{y}_j = b_1^{(2)} P G_{1j} P^{-1}\vec{\gamma}_0 + a_{N_2}^{(2)} P G_{N_2 j} P^{-1}\vec{\gamma}_{N_2} - P\sum_{l=1}^{N_2-1} G_{lj} P^{-1}[\vec{f}_l - \vec{\omega}_l], \\ j = \overline{1,\ N_2-1}, \tag{8}$$

де $G_{lj} = [G_{lj}(\mu_i)]_{i=\overline{1,N_1-1}}$ — діагональна матриця. Аналогічно можна отримати ФСЗ у вигляді, в якому матриця $\Pi^{(2)}$, що відповідає оператору Λ_2, є матрицею простої структури.

Якщо обидві матриці $\Pi^{(k)}$ є матрицями простої структури, тобто

$$\Pi^{(k)} = P^{(k)} M^{(k)} [P^{(k)}]^{-1}, \quad k = 1,\ 2,$$

то ФСЗ можна побудувати таким чином. Позначимо

$$Y = (y_{ij})_{i=\overline{1,N_1-1}}^{j=\overline{1,N_2-1}}, \quad F = (f_{ij})_{i=\overline{1,N_1-1}}^{j=\overline{1,N_2-1}},$$

$$W = \begin{bmatrix} b_1^{(1)}\gamma_{0,N_2-1} + a_{N_2-1}^{(2)}\gamma_{1N_2} & a_{N_2-1}^{(2)}\gamma_{2N_2} & \dots & a_{N_2-1}^{(2)}\gamma_{N_1-2,N_2} & a_{N_1-1}^{(1)}\gamma_{N_1N_2-1} + a_{N_2-1}^{(2)}\gamma_{N_1-1,N_2} \\ b_1^{(1)}\gamma_{0,N_2-2} & 0 & \dots & 0 & a_{N_1-1}^{(1)}\gamma_{N_1,N_2-1} \\ \dots & & & & \\ b_1^{(1)}\gamma_{0,2} & 0 & \dots & 0 & a_{N_1-1}^{(1)}\gamma_{N_1,2} \\ b_1^{(1)}\gamma_{0,1} + b_1^{(2)}\gamma_{1,0} & b_1^{(2)}\gamma_{2,0} & \dots & b_1^{(2)}\gamma_{N_1-2,0} & b_1^{(2)}\gamma_{N_1-1,0} + a_{N_1-1}^{(1)}\gamma_{N_1,1} \end{bmatrix}.$$

Тоді задача (1) запишеться в матричному вигляді

$$\Pi^{(1)}Y + Y\Pi^{(2)} = F - W. \tag{9}$$

Проведемо подвійну P-трансформацію рівняння (9) множенням його зліва на матрицю $[P^{(1)}]^{-1}$ і справа на матрицю $P^{(2)}$ і позначимо для будь-якої матриці V

$$\hat{V} = [P^{(1)}]^{-1} V P^{(2)} \equiv [\hat{v}_{ij}]_{i=\overline{1,N_1-1}}^{j=\overline{1,N_2-1}}.$$

Дістанемо

$$M^{(1)}\hat{Y} + \hat{Y}M^{(2)} = \hat{F} - \hat{W}$$

або в координатній формі

$$(\mu_i^{(1)} + \mu_j^{(2)})\hat{y}_{ij} = \hat{f}_{ij} - \hat{\omega}_{ij}, \quad i = \overline{1,\ N_1-1}, \quad j = \overline{1,\ N_2-1}.$$

Припустимо, що $\mu_i^{(1)} + \mu_j^{(2)} \neq 0\ \forall\ i,\ j$. Тоді

$$\hat{y}_{ij} = (\mu_i^{(1)} + \mu_j^{(2)})^{-1}\ (\hat{f}_{ij} - \hat{\omega}_{ij}).$$

Записавши цю рівність у матричному вигляді і виконавши обернену подвійну P-трансформацію (множенням зліва на $P^{(1)}$, справа на $[P^{(2)}]^{-1}$), дістанемо ФСЗ другого типу:

$$Y = P^{(1)}\{M^{(-1)} \otimes [(P^{(1)})^{-1}(F + W)P^{(2)}]\}(P^{(2)})^{-1},$$

де

$$M^{(-1)} = [(\mu_i^{(1)} + \mu_j^{(2)})^{-1}]_{i=\overline{1,N_1-1}}^{j=\overline{1,N_2-1}},$$

а знак $\otimes$ означає поелементне множення матриць.

Приклад. Побудувати ФСЗ для сіткової схеми, яка апроксимує задачу Діріхле для рівняння Пуассона в прямокутнику

$$\Lambda y \equiv \Lambda_1 y + \Lambda_2 y \equiv y_{\bar{x}_1 x_1}(x) + y_{\bar{x}_2 x_2}(x) = f(x), \\ y(x) = \gamma(x), \quad x \in \gamma_h. \tag{10}$$

Розв'язання. Запишемо цю задачу у вигляді

$$L_{1h}y_{ij} + \alpha^2 L_{2h}y_{ij} = h_1^2 f_{ij}, \quad (x_{1i}, x_{2j}) \in \omega_h, \quad y_{ij} = \gamma_{ij}, \quad (x_{1i}, x_{2j}) \in \gamma_h, \tag{11}$$

де $L_{1h}y_{ij} = y_{i+1,j} + y_{i-1,j}$, $L_{2h}y_{ij} = y_{i,j+1} - 2\beta y_{ij} + y_{i,j-1}$; $\beta = 1 + \alpha^{-2}$, $\alpha^2 = h_1^2/h_2^2$. Користуючись введеними вище позначеннями, запишемо задачу (11) у вигляді

$$\Pi^{(1)}\vec{y}_j + \alpha^2\vec{y}_{j+1} - 2\beta\vec{y}_j + \alpha^2\vec{y}_{j-1} = h_1^2\vec{f}_j - \vec{\omega}_j, \\ \vec{y}_0 = \vec{\gamma}_0, \quad \vec{y}_{N_2} = \vec{\gamma}_{N_2}, \tag{12}$$

де

$$\Pi^{(1)} = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & 0 \end{bmatrix}.$$

Знайдемо власні числа та власні вектори матриці $\Pi^{(1)}$. Задача

$$\Pi^{(1)}\vec{y} - 2\mu\vec{y} = 0$$

при переході до скалярної форми запису еквівалентна скінченнорізницевій задачі Штурма — Ліувілля

$$y_{k+1} - 2\mu y_k + y_{k-1} = 0, \quad k = \overline{1, N_1 - 1}, \quad y_0 = y_{N_1} = 0.$$

Неважко знайти (див. п. 1.4 з ч. 1), що

$$2\mu_k = 2\cos\frac{k\pi}{N_1}, \quad k = \overline{1, N_1 - 1},$$

а система ортонормованих власних векторів має вигляд

$$p_{kj} = \sqrt{\frac{2}{N_1}}\sin\frac{kj\pi}{N_1}, \quad k, j = \overline{1, N_1 - 1}.$$

Фундаментальна матриця

$$P = \sqrt{\frac{2}{N_1}}\left(\sin\frac{kj\pi}{N_1}\right)_{k,j=\overline{1,N_1-1}},$$

причому $P = P^{-1} = P^*$, тобто $\Pi^{(1)} = PMP$, де

$$M = \operatorname{diag}(2\mu_i)_{i=\overline{1,N_1-1}} = \operatorname{diag}\left(2\cos\frac{i\pi}{N_1}\right)_{i=\overline{1,N_1-1}}.$$

Виконавши P-трансформацію векторного рівняння (12), дістанемо

$$M\hat{\vec{y}}_i + \alpha^2[\hat{\vec{y}}_{j+1} - 2\beta\hat{\vec{y}}_j + \hat{\vec{y}}_{j-1}] = h_1^2\hat{\vec{f}}_j - \hat{\vec{\omega}}_j,$$

$$j = \overline{1, N_2 - 1}, \quad \hat{\vec{y}}_0 = \hat{\vec{\gamma}}_0, \quad \hat{\vec{y}}_{N_2} = \hat{\vec{\gamma}}_{N_2},$$

або в скалярній формі

$$\hat{y}_{i,j+1} - 2(\beta - \alpha^{-2}\mu_i)\hat{y}_{ij} + \hat{y}_{i,j-1} = h_2^2\hat{f}_{ij} - \alpha^{-2}\hat{\omega}_{ij},$$

$$i = \overline{1, N_1 - 1}, \quad j = \overline{1, N_2 - 1}; \quad \hat{y}_{i0} = \hat{\gamma}_{i0}, \quad \hat{y}_{i,N_2} = \hat{\gamma}_{i,N_2}, \quad i = \overline{1, N_1 - 1}. \quad (13)$$

Вправа 1. Враховуючи, що власні вектори матриць $\Pi = a\Pi^{(1)} - bI$ (де I — одинична матриця) однакові, а власні числа визначаються рівністю

$$2\tilde{\eta}_j = 2a\mu_j - b = 2\left(a\cos\frac{j\pi}{N_1} - \frac{b}{2}\right), \quad j = \overline{1, N_1 - 1},$$

довести, що функція Гріна оператора, який стоїть у лівій частині (13) і діє по j, має вигляд

$$G_{pj}(\eta_i) = \frac{U_{|p,j|-1}(\eta_i)\,U_{N_2-(p,j)-1}(\eta_i)}{U_{N_2-1}(\eta_i)}, \quad j = \overline{1, N_2 - 1},$$

де $U_l(\eta_i)$ — многочлени Чебишева другого роду степеня l;

$$|p, j| = \min(p, j), \quad (p, j) = \max(p, j), \quad \eta_i = 1 + \alpha^{-2}\left(1 - \cos\frac{i\pi}{N_1}\right).$$

Розв'язок задачі (13) має вигляд

$$\hat{y}_{ij} = G_{1j}(\eta_i)\hat{\gamma}_{i0} + G_{N_2-1,j}(\eta_i)\hat{\gamma}_{i,N_2} - \sum_{l=1}^{N_2-1} G_{lj}(\eta_i)[h_2^2\hat{f}_{il} - \alpha^{-2}\hat{\omega}_{il}].$$

Записавши це у векторному вигляді і виконавши обернену P-трансформацію, дістанемо ФСЗ першого типу

$$\vec{y}_j = PG_{1j}P\vec{\gamma}_0 + PG_{N_2-1,j}P\vec{\gamma}_{N_2} - P\sum_{l=1}^{N_2-1} G_{lj}P[h_2^2 f_l - \alpha^{-2}\vec{\omega}_l], \quad (14)$$

де $G_{lj} = \operatorname{diag}(G_{lj}(\eta_i))_{i=\overline{1,N_1-1}}$ — діагональні матриці. Щоб знайти ФСЗ другого типу для задачі (10), запишемо її у вигляді

$$\Pi^{(1)}Y + \alpha^2 Y\Pi^{(2)} - 2\beta Y = h_1^2 F - H - \alpha^2 Q, \quad (15)$$

де

$$H = (h_{ij})_{i=\overline{1,N_1-1}}^{j=\overline{1,N_2-1}} = \begin{bmatrix} \gamma_{01} & \gamma_{02} & \dots & \gamma_{0,N_2-1} \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \\ \gamma_{N_1,1} & \gamma_{N_1,2} & \dots & \gamma_{N_1,N_2-1} \end{bmatrix},$$

$$Q = (q_{ij})_{i=\overline{1,N_1-1}}^{j=\overline{1,N_2-1}} = \begin{bmatrix} \gamma_{10} & 0 & \dots & 0 & \gamma_{1,N_2} \\ \gamma_{20} & 0 & \dots & 0 & \gamma_{2,N_2} \\ \dots & \dots & \dots & \dots & \dots \\ \gamma_{N_1-1,0} & 0 & \dots & 0 & \gamma_{N_1-1,N_2} \end{bmatrix},$$

$\Pi^{(1)}$ — квадратна матриця розмірності $(N_1 - 1) \times (N_1 - 1)$ вигляду

$$\Pi^{(1)} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 & 1 \\ 0 & 0 & 0 & \dots & 0 & 1 & 0 \end{bmatrix},$$

$\Pi^{(2)}$ має такий самий вигляд, але розмірність її $(N_2 - 1) \times (N_2 - 1)$. Матриці $\Pi^{(1)}$, $\Pi^{(2)}$ визначаються з виразу

$$\Pi^{(\alpha)} = P^{(\alpha)}M^{(\alpha)}P^{(\alpha)}, \quad \alpha = 1, 2,$$

де

$$P^{(\alpha)} = (p_{ik}^{(\alpha)})_{i,k=\overline{1,N_\alpha-1}} = \sqrt{\frac{2}{N_\alpha}}\left(\sin\frac{ik\pi}{N_\alpha}\right)_{i,k=\overline{1,N_\alpha-1}},$$

$$M^{(\alpha)} = \operatorname{diag}(2\mu_i^{(\alpha)})_{i=\overline{1,N_\alpha-1}} = \operatorname{diag}\left(2\cos\frac{i\pi}{N_\alpha}\right)_{i=\overline{1,N_\alpha-1}},\quad P^{(\alpha)}P^{(\alpha)} = I.$$

Виконавши подвійну P-трансформацію рівняння (15) (тобто помноживши його на $P^{(1)}$ зліва і на $P^{(2)}$ справа) і позначивши $P^{(1)}VP^{(2)} = \hat{V}$, дістанемо

$$M^{(1)}\hat{Y} + \alpha^2\hat{Y}M^{(2)} - 2\beta\hat{Y} = h_1^2\hat{F} - \hat{H} - \alpha^2\hat{Q},$$

або в скалярній формі

$$2\mu_i^{(1)}\hat{y}_{ij} + 2\alpha^2\mu_j^{(2)}y_{ij} - 2\beta\hat{y}_{ij} = h_1^2\hat{f}_{ij} - \alpha^2\hat{q}_{ij} - \hat{h}_{ij}.$$

Якщо $\beta - \mu_i^{(1)} - \alpha^2\mu_j^{(2)} \neq 0$, то звідси

$$\hat{y}_{ij} = \frac{-h_1^2\hat{f}_{ij} + \alpha^2\hat{q}_{ij} + \hat{h}_{ij}}{2(\beta - \mu_i^{(1)} - \alpha^2\mu_j^{(2)})}.$$

Виконавши обернене перетворення, дістанемо ФСЗ другого типу

$$y_{ij} = \sum_{k=1}^{N_1-1}\sum_{l=1}^{N_2-1} p_{ik}^{(1)}p_{jl}^{(2)}\hat{y}_{kl} =$$

$$= \sum_{k=1}^{N_1-1}\sum_{l=1}^{N_2-1}\frac{p_{ik}^{(1)}p_{jl}^{(2)}}{2(\beta - \mu_k^{(1)} - \alpha^2\mu_l^{(2)})}\sum_{\mu=1}^{N_1-1}\sum_{\nu=1}^{N_2-1} p_{\mu k}^{(1)}p_{\nu l}^{(2)}(-h_1^2 f_{\mu\nu} + h_{\mu\nu} + \alpha^2 q_{\mu\nu}). \tag{16}$$

Метод сумарних зображень зручно використовувати для розв'язування сіткових крайових задач в областях складної форми, що складаються з канонічних областей, для яких відомі ФСЗ. Розглянемо, наприклад, різницеву схему вигляду (10) для задачі Діріхле для рівняння Пуассона в L-подібній області Ω, що складається з двох прямокутників Ω_1 і Ω_2. Відповідні сіткові області

$$\bar{\omega}_{1h} = \{(x_{1i}, x_{2j}) : x_{1i} = ih_1,\quad x_{2j} = jh_2,\quad i = \overline{0,\ N_1},\quad j = \overline{0,\ N_2}\},$$

$$\bar{\omega}_{2h} = \{(x_{1i}, x_{2j}) : x_{1i} = ih_1,\quad x_{2j} = jh_2,\quad i = \overline{N_1,\ N_1 + N_1'},\quad j = \overline{0,\ N_2'}\}$$

і нехай вони перетинаються по множині γ^* із $N_2' - 1$ вузлів. Для кожної з областей ω_{1h}, ω_{2h} можна записати ФСЗ (14) (або (16)), в якій невідомі значення шуканої функції у вузлах γ^* ввійдуть до векторів $\vec{\gamma}_0$ чи $\vec{\gamma}_N$. Щоб знайти ці невідомі значення, запишемо різницеві рівняння на γ^*:

$$h_1^{-2}[y_{N_1+1,j} - 2y_{N_1,j} + y_{N_1-1,j}] + h_2^{-2}[y_{N_1,j+1} - 2y_{N_1,j} + y_{N_1,j-1}] =$$

$$= f_{N_1,j},\quad j = \overline{1,\ N_2' - 1}.$$

Підставляючи замість $y_{N_1-1,j}$ ($j = \overline{1,\ N_2' - 1}$) значення, знайдені за ФСЗ для прямокутника ω_{1h}, а замість $y_{N_1+1,j}$ ($j = \overline{1,\ N_2' - 1}$) значення, знайдені за ФСЗ для прямокутника ω_{2h}, дістанемо відносно $y_{N_1,j}$ ($j = \overline{1,\ N_2' - 1}$) систему лінійних алгебраїчних рівнянь порядку $N_2' - 1$, яку можна привести до вигляду

$$\sum_{l=1}^{N_2'-1} c_{l,j}y_{N_1,l} = F_{N_1,j},\quad j = \overline{1,\ N_2' - 1}.$$

Цю систему можна розв'язати, наприклад, методом Гаусса за $O((N_2' - 1)^3)$ арифметичних операцій. ФСЗ зручно використовувати тоді, коли розв'язок задачі треба знайти не на всій області, а лише в деяких вузлах або обчислити деякі функціонали від розв'язку (інтеграли, похідні), що можна робити, перетворюючи саму формулу.

Неважко підрахувати, що обчислення одного значення за ФСЗ (8) потребує $O((N_i - 1)^2(N_1 - 1))$ операцій множення, якщо відомі матриці P^{-1}, G_{2j}, тобто при $N_1 \sim N_2 \sim N$ порядку $O(N^3)$ арифметичних операцій. Але для деяких задач це число можна значно скоротити. Наприклад, для задачі, розглянутої в прикладі 1 (див. (14), (16) — задача Діріхле для рівняння Пуассона), множення матриць на вектори зводиться до обчислення сум виду

$$w_i = \sum_{k=1}^{N-1}\gamma_k\sin\frac{k\pi i}{N},\quad i = \overline{1,\ N-1} \tag{17}$$

(за рахунок структури матриць P та G). Для обчислення таких сум розроблено спеціальний алгоритм швидкого перетворення Фур'є (ШПФ), який потребує лише $5N\log_2 N$ арифметичних дій (при $N = 2^n$, n — ціле). Якщо цей алгоритм застосувати при обчисленні за формулами (14) або (16), то замість $O(N^3)$ операцій розв'язок системи рівнянь (10) з $(N-1)^2$ невідомими (при $N_1 = N_2$) буде знайдено за $O(N^2\log_2 N)$ операцій, що лише в $\log_2 N$ гірше за оптимальну оцінку (по порядку N).

Для сіткових рівнянь з операторами розглянутого вище виду можна застосувати *метод розділення змінних* МРЗ, який проілюструємо на задачі (10). Приведемо спочатку задачу (10) до задачі з однорідними граничними умовами. Нехай $\mathring{y}$ — функція, задана на $\bar{\omega}_h$, яка обертається на нуль на γ_h і збігається з розв'язком задачі (10) на ω_h. Неважко

помітити, що $\overset{\circ}{y}$ є розв'язком задачі

$$\Lambda \overset{\circ}{y} \equiv \Lambda_1 y + \Lambda_2 \overset{\circ}{y} \equiv \overset{\circ}{y}_{\bar{x}_1 x_1} + \overset{\circ}{y}_{\bar{x}_2 x_2} = \varphi(x), \quad x \in \omega_h, \quad \overset{\circ}{y}(x) = 0, \quad x \in \gamma_h, \tag{18}$$

де

$$\varphi(x) = \begin{cases} f(x) - \dfrac{\gamma(l_1, x_2)}{h_1^2}, & x_1 = l_1 - h_1, \quad 0 < x_2 < l_2, \\ f(x) - \dfrac{\gamma(0, x_2)}{h_1^2}, & x_1 = h_1, \quad 0 < x_2 < l_2, \\ f(x) - \dfrac{\gamma(x_1, l_2)}{h_2^2}, & 0 < x_1 < l_1, \quad x_2 = l_2 - h_2, \\ f(x) - \dfrac{\gamma(x_1, 0)}{h_2^2}, & 0 < x_1 < l_1, \quad x_2 = h_2, \\ & l_1 = N_1 h_1, \quad l_2 = N_2 h_2. \end{cases}$$

Нехай $v_{k_2}(x_2)$, $\lambda_{k_2}^{(2)}$, $k_2 = \overline{1, N_2 - 1}$ — власні функції і власні значення оператора Λ_2, тобто розв'язок задачі

$$\Lambda_2 v + \lambda v = 0, \quad x \in \omega_h, \quad v(0) = v(l_2) = 0, \tag{19}$$

який ми знайшли в п. 1.4.4 з ч. 1. Розкладемо розв'язок задачі $\overset{\circ}{y}(x)$ і праву частину $\varphi(x)$ за власними функціями $\{v_{k_2}\}$:

$$\varphi(x) = \sum_{k_2=1}^{N_2-1} \psi_{k_2}(x_1) v_{k_2}(x_2), \qquad \overset{\circ}{y}(x) = \sum_{k_2=1}^{N_2-1} c_{k_2}(x_1) v_{k_2}(x_2), \tag{20}$$

де $\psi_{k_2}(x_1)$, $c_{k_2}(x_1)$ — коефіцієнти Фур'є, причому

$$\psi_{k_2}(x_1) = \sum_{i_2=1}^{N_2-1} h_2 \varphi(x_1, i_2 h_2) v_{k_2}(i_2 h_2), \tag{21}$$

а $c_{k_2}(x_1)$ поки що невизначені. Враховуючи, що λ_{kl}, $v_{k_2}(x_2)$ є розв'язком задачі (19), маємо

$$\Lambda c_{k_2}(x_1) v_{k_2}(x_2) = v_{k_2}(x_2) \Lambda_1 c_{k_2}(x_1) + c_{k_2}(x_1) \Lambda_2 v_{k_2}(x_2) = v_{k_2}(x_2) \Lambda_1 c_{k_2}(x_1) - \lambda_{k_2}^{(2)} c_{k_2}(x_1) v_{k_2}(x_2) = [\Lambda_1 c_{k_2}(x_1) - \lambda_{k_2}^{(2)} c_{k_2}(x_1)] v_{k_2}(x_2).$$

Враховуючи останню рівність і зображення (20), із виразу (18) маємо

$$\sum_{k_2=1}^{N_2-1} \{\Lambda_1 c_{k_2}(x_1) - \lambda_{k_2}^{(2)} c_{k_2}(x_1) - \psi_{k_2}(x_1)\} v_{k_2}(x_2) = 0.$$

В силу лінійної незалежності $\{v_{k_2}(x)\}$ це можливо лише тоді, коли

$$\Lambda_1 c_{k_2}(x_1) - \lambda_{k_2}^{(2)} c_{k_2}(x_1) = \psi_{k_2}(x_2), \quad k_2 = \overline{1, N_2 - 1}, \tag{22}$$

$$x_1 = i_1 h_1, \quad 0 < i_1 < N_1; \quad c_{k_2}(i_1 h_1) = 0 \text{ при } i_1 = 0, N_1.$$

Обчисливши $\psi_{k_2}(x_1)$ за першою формулою (20), задачі (22) можна розв'язати методом прогонки, застосувавши його $N_2 - 1$ раз для $k_2 = \overline{1, N_2 - 1}$. Знаючи $c_{k_2}(x)$, за другою з формул (20) знайдемо розв'язок задачі (18). З виразу (20) бачимо, що $\psi_{k_2}(x_1)$ і $y(x)$ обчислюються за формулами виду (17), тому для обчислення їх можна застосувати швидке перетворення Фур'є (ШПФ) і знайти розв'язок задачі (18) за $O(N_1 N_2 \log_2 N_2)$ арифметичних дій

Прямі методи є найекономнішими для крайових задач для рівняння Пуассона у прямокутнику (тепер такі методи і програми для них розроблені для крайових умов всіх трьох типів, а також для змішаних крайових умов). Але фактично прямі методи є економними лише для задач, в яких змінні розділяються. У випадку, коли область не є прямокутником або розглядаються рівняння зі змінними коефіцієнтами, в яких змінні не розділяються, застосовуються ітераційні методи.

5.4.2. Поперемінно-трикутний метод для різницевої задачі Діріхле для рівняння Пуассона. Запишемо розрахункові формули поперемінно-трикутного методу (див. п. 1.4.7 з ч. 1):

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + A y_k = f, \quad k = 0, 1, \ldots \quad \forall y_0 \in H,$$

$$B = (D + \omega R_1) D^{-1} (D + \omega R_2), \quad R_1 = R_2^*, \quad D = D^* > 0, \quad R = R_1 + R_2 = R^* > 0 \tag{23}$$

для задачі (18). Цю задачу розглядатимемо в просторі H_h функцій, які задано на сітці $\overline{\omega}_h = \{x = (x_1, x_2) : x_\alpha = i_\alpha h_\alpha, \ i_\alpha = \overline{0, N_\alpha}, h_\alpha = l_\alpha / N_\alpha, \ \alpha = 1, 2\}$ і перетворюються на нуль на границі $\gamma_h = \bigcup_{\alpha=1}^{2} \gamma_{\pm\alpha}$, $\gamma_{-1} = \{(0, i_2 h_2) : i_2 = \overline{0, N_2}\}$, $\gamma_{+1} = \{(l_1, i_2 h_2) : i_2 = \overline{0, N_2}\}$, $\gamma_{-2} = \{(i_1 h_1, 0) : i_1 = \overline{0, N_1}\}$, $\gamma_{+2} = \{(i_1 h_1, l_2) : i_1 = \overline{0, N_1}\}$. У просторі H введемо скалярний добуток за формулою

$$(y, v) = \sum_{x \in \omega_h} h_1 h_2 y(x) v(x)$$

для $\omega_h = \bar{\omega}_h \setminus \gamma_h$. Оператор A з областю визначення H, $A : H \to H$ визначимо формулою

$$v(x) = Ay = \begin{cases} -y_{\bar{x}_1 x_1}(x) - y_{\bar{x}_2 x_2}(x), & x \in \omega_h, \\ 0, & x \in \gamma_h \end{cases}$$

і подамо його у вигляді $A = A_1 + A_2$, де

$$A_1 y = \begin{cases} \dfrac{y_{\bar{x}_1}(x)}{h_1} + \dfrac{y_{\bar{x}_2}(x)}{h_2}, & x \in \omega_h, \\ 0, & x \in \gamma_h, \end{cases}$$

$$A_2 y = \begin{cases} -\dfrac{y_{x_1}(x)}{h_1} - \dfrac{y_{x_2}(x)}{h_2}, & x \in \omega_h, \\ 0, & x \in \gamma_h. \end{cases}$$

Покладемо в (23) $R = A$, $R_1 = A_1$, $R_2 = A_2$, $D = I$, де I — одиничний оператор. За допомогою формул підсумовування частинами (або першої різницевої формули Гріна) легко встановити, що

$$(A_1 y, v) = (y, A_2 v),$$

тобто $A_2 = A_1^*$ (це можна бачити і при порівнянні матриць, що відповідають A_1, A_2). Тоді в операторному вигляді попереміннo-трикутний метод запишеться так:

$$B\mathring{y}_{k+1} \equiv (I + \omega A_1)(I + \omega A_2)\,\mathring{y}_{k+1} = F_k,$$

$$F_k = B\mathring{y}_k + \tau_{k+1}(A\mathring{y}_k - \varphi),$$

де k — номер ітерації, $\mathring{y}_0$ — будь-який елемент H. Значення $\mathring{y}_{k+1}$ знаходимо послідовно розв'язуванням двох операторних рівнянь

$$(I + \omega A_1)\,\mathring{y}_k^{(1)} = F_k, \quad (I + \omega A_2)\,\mathring{y}_{k+1} = \mathring{y}_k^{(1)},$$

які в скалярному вигляді записуються таким чином

$$\mathring{y}_k^{(1)}(i_1, i_2) = \frac{\varkappa_1 \mathring{y}_k^{(1)}(i_1 - 1, i_2) + \varkappa_2 \mathring{y}_k^{(1)}(i_1, i_2 - 1) + F_k(i_1, i_2)}{1 + \varkappa_1 + \varkappa_2},$$

$$\mathring{y}_{k+1}(i_1, i_2) = \frac{\varkappa_1 \mathring{y}_{k+1}(i_1 + 1, i_2) + \varkappa_2 \mathring{y}_{k+1}(i_1, i_2 + 1) + \mathring{y}_k^{(1)}(i_1, i_2)}{1 + \varkappa_1 + \varkappa_2}, \tag{24}$$

де $\varkappa_1 = \dfrac{\omega}{h_1^2}$, $\varkappa_2 = \dfrac{\omega}{h_2^2}$, $y(i_1, i_2) \equiv y(x_{i1}, x_{i2}) = y(x)$, $x \in \omega_h$.

У даному разі не так очевидно, як в одновимірному, що за формулами (24) послідовно можна знайти розв'язок в усіх вузлах сітки. Проте це так, і розрахунок можна вести таким чином.

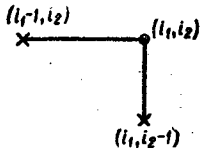


Рис. 25

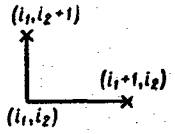


Рис. 26

Насамперед зауважимо, що шаблон s_1 $(i_1, i_2) = s_1 = \{(i_1, i_2), (i_1 - 1, i_2), (i_1, i_2 - 1)\}$, на якому записана перша з формул (24), показано на рис. 25, а шаблон s_2 $(i_1, i_2) = s_2 = \{(i_1, i_2), (i_1 + 1, i_2), (i_1, i_2 + 1)\}$ для другої формули (24) — на рис. 26. Вибравши $i_1 = 1$, $i_2 = 1$ (це вузол у нижньому лівому куті прямокутника), бачимо, що інші два вузли шаблона s_1 лежать на границі, тобто $\mathring{y}^{(1)}(0, i_2)$, $\mathring{y}^{(1)}(i_1, 0)$ відомі і за першою формулою (24) знаходимо $\mathring{y}^{(1)}(1, 1)$. Знаючи $\mathring{y}^{(1)}(1, 1)$, знаходимо $\mathring{y}^{(1)}(2, 1)$ і далі послідовно $\mathring{y}^{(1)}(3, 1)$, ..., $\mathring{y}^{(1)}(N_1 - 1, 1)$, тобто всі значення $\mathring{y}^{(1)}$ на першому рядку вузлів. Далі покладемо $i_2 = 2$ і аналогічно знаходимо $\mathring{y}^{(1)}$ на другому рядку і т. д. до визначення $\mathring{y}^{(1)}$ в усіх вузлах сітки. Неважко помітити, що розрахунок $\mathring{y}_k^{(1)}$ можна вести не по рядках, а по стовпчиках знизу вгору. Далі для визначення $\mathring{y}_{k+1}$ покладемо $i_1 = N_1 - 1$, $i_2 = N_2 - 1$, тобто беремо вузол шаблона s_2 у правому верхньому куті прямокутника. Два інших вузли тоді лежать на границі, і тому нам відомі нульові значення функції в них. За другою формулою (24) визначаємо $\mathring{y}_{k+1}(N_1 - 1, N_2 - 1)$ і далі послідовно $\mathring{y}_{k+1}(N_1 - 1, N_2 - 2)$, ..., $\mathring{y}_{k+1}(N_2 - 1, 1)$, $\mathring{y}_{k+1}(N_1 - 2, N_2 - 1)$, ..., $\mathring{y}_{k+1}(1, 1)$. Зауважимо, що розрахунок $\mathring{y}_{k+1}$ можна вести і по рядках справа наліво, тобто

$$\mathring{y}_{k+1}(N_1 - 1, N_2 - 1),\ \mathring{y}_{k+1}(N_1 - 2, N_2 - 1),\ \ldots,\ \mathring{y}_{k+1}(1, N_2 - 1),$$
$$\mathring{y}_{k+1}(N_1 - 1, N_2 - 2),\ \ldots,\ \mathring{y}_{k+1}(1, 1).$$

Легко підрахувати, що обчислення F_k потребує 10 дій додавання і 10 дій множення, а обчислення $\mathring{y}_{k+1}$ при вже обчислених F_k — 4 дії додавання і 6 дій множення. Всього для визначення y_{k+1} в одному вузлі треба виконати 14 дій додавання і 16 дій множення. Щоб зменшити число арифметичних дій, можна обчислення вести за таким алгоритмом:

$$(E + \omega A_1)\,\mathring{w}_{k+1/2} = \Lambda y_k - f, \quad (E + \omega A_2)\,\mathring{w}_{k+1} = \mathring{w}_{k+1/2},$$

$$y_{k+1} = y_k + \tau_{k+1}\mathring{w}_{k+1},$$

в якому на кожен вузол сітки виконується 10 дій додавання і 10 дій множення. Однак при цьому збільшується об'єм потрібної оперативної пам'яті ЕОМ: треба зберігати не один масив чисел, як у попередньому алгоритмі $\overset{\circ}{y}{}_k^{(1)}$, а два — y_k та $\overset{\circ}{w}_{k+1}$.

Щоб скористатися загальною теорією для методу (23) (див. п. 1.4.7 з ч. 1) і вибрати параметри у виразі (24), треба знайти сталі δ, Δ у нерівностях

$$A \geqslant \delta D, \quad A_1 D^{-1} A_2 \leqslant \frac{\Delta}{4} A, \quad \delta > 0, \quad \Delta > 0.$$

Тут $D = I$, тому δ — найменше власне значення оператора A, яке дорівнює сумі найменших власних значень операторів $\overline{A}_\alpha y = y_{\bar{x}_\alpha x_\alpha}$, $\alpha = 1, 2$,

$$\delta = 4\left(\frac{1}{h_1^2}\sin^2\frac{\pi h}{2l_1} + \frac{1}{h_2^2}\sin^2\frac{\pi h}{2l_2}\right). \tag{25}$$

Враховуючи, що $A = A_1 + A_2$, $A_1^* = A_2$, а також нерівність

$$(a_1 b_1 + a_2 b_2)^2 \leqslant (a_1^2 + a_2^2)(b_1^2 + b_2^2),$$

знаходимо

$$(A_1 A_2 y, y) = (A_2 y, A_2 y) = \left(\left(\frac{1}{h_1} y_{\bar{x}_1} + \frac{1}{h_2} y_{\bar{x}_2}\right)^2, 1\right) \leqslant$$

$$\leqslant \left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right)((y_{\bar{x}_1})^2 + (y_{\bar{x}_2})^2, 1) = \left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right) \times$$

$$\times \sum_{i_1=1}^{N_1-1}\sum_{i_2=1}^{N_2-1} [(y_{\bar{x}_1})^2 + (y_{\bar{x}_2})^2]_{i_1 i_2} h_1 h_2 \leqslant \left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right)(Ay, y)$$

(ми використали рівність

$$(Ay, y) = \sum_{i_2=1}^{N_2-1} h_2 \sum_{i_1=0}^{N_1-1} (y_{\bar{x}_1})^2_{i_1 i_2} h_1 + \sum_{i_1=1}^{N_1-1} h_1 \sum_{i_2=0}^{N_2-1} (y_{\bar{x}_2})^2_{i_1 i_2} h_2,$$

яка утворюється підсумовуванням частинами). Із цієї нерівності бачимо, що

$$\Delta = 4\left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right). \tag{26}$$

Маючи δ та Δ, знаходимо $\eta = \delta/\Delta$ і далі (див. п. 1.4.7 з ч. 1)

$$\overset{\circ}{\xi} = \frac{\gamma_1}{\gamma_2} = \frac{2\sqrt{\eta}}{1+\sqrt{\eta}}, \quad \overset{\circ}{\gamma}_1 = \frac{\delta}{2(1+\sqrt{\eta})}, \quad \overset{\circ}{\gamma}_2 = \frac{\delta}{4\sqrt{\eta}}, \quad \overset{\circ}{\omega} = \frac{2}{\sqrt{\Delta\delta}}$$

(сталі $\overset{\circ}{\gamma}_1$, $\overset{\circ}{\gamma}_2$ із нерівності $\overset{\circ}{\gamma}_1 B \leqslant A \leqslant \overset{\circ}{\gamma}_2 B$). Потім оцінюємо число ітерацій за формулою (ε — задано наперед)

$$n(\varepsilon) \approx \frac{\ln\frac{2}{\varepsilon}}{\ln\frac{1}{\rho_1}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}$$

і вибираємо по n (ε), $\overset{\circ}{\gamma}_1$, $\overset{\circ}{\gamma}_2$ стійкий набір чебишевських параметрів μ_k, τ_{k+1}. Нарешті, за формулою (24) проводимо описані вище розрахунки.

Якщо $h_1 = h_2 = h$, то $\delta = \frac{8}{h^2}\sin^2\frac{\pi h}{2}$, $\Delta = \frac{8}{h^2}$ і $\eta = \frac{\delta}{\Delta} = \sin^2\frac{\pi h}{2} \approx \frac{\pi^2 h^2}{4}$, тобто η для двовимірної задачі і далі одновимірної (див. п. 1.4.7 з ч. 1) збігається і кількість ітерацій, яка не залежить від розмірності задачі Діріхле для рівняння Пуассона, а тому оцінки числа ітерацій різних ітераційних методів, які ми дістали в п. 1.4.7 з ч. 1 для одновимірної модельної задачі, справедливі і для багатовимірного випадку. Якщо ж сітка не квадратна ($h_1 \neq h_2$), то число ітерацій слабо залежить від розмірності задачі.

5.4.3. Поперемінно-трикутний метод для випадку змінних коефіцієнтів. У прямокутнику $\Omega = \{x = (x_1, x_2) : 0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$ з границею Γ розглянемо задачу Діріхле для більш складного рівняння, ніж рівняння Пуассона:

$$\begin{aligned} Lu = L_1 u + L_2 u = -f(x), \quad x \in \Omega, \\ u(x) = \mu(x), \quad x \in \Gamma, \end{aligned} \tag{27}$$

де $L_\alpha u = \frac{\partial}{\partial x_\alpha}\left(k_\alpha(x)\frac{\partial v}{\partial x_\alpha}\right)$, $0 < c_1 \leqslant k_\alpha(x) \leqslant c_2$, $\alpha = 1, 2$, c_1, c_2 — сталі.

Наближений розв'язок задачі (27) шукатимемо за такою різницевою схемою:

$$\begin{aligned} \Lambda y = -f(x), \quad x \in \omega_h, \\ y(x) = \mu(x), \quad x \in \gamma_h, \end{aligned} \tag{28}$$

де

$$\Lambda y = \Lambda_1 y + \Lambda_2 y, \quad \Lambda_\alpha y = (a_\alpha y_{\bar{x}_\alpha})_{x_\alpha} =$$

$$= \frac{1}{h_\alpha}\left[\frac{a_\alpha^{(+1_\alpha)}(y^{(+1_\alpha)} - y)}{h_\alpha} - \frac{a_\alpha (y - y^{(-1_\alpha)})}{h_\alpha}\right],$$

$$y^{(\pm 1_1)} = y((i_1 \pm 1)h_1, i_2 h_2), \quad y^{(\pm 1_2)} = y(i_1 h_1, (i_2 \pm 1)h_2),$$

$$a_1(x_1, x_2) = k_1\left(x_1 - \frac{1}{2}h_1, x_2\right) = k_1^{(-0,5_1)},$$

$$a_2(x_1, x_2) = k_2\left(x_1, x_2 - \frac{1}{2}h_2\right) = k_2^{(-0,5)},$$

$$0 < c_1 \leqslant a_\alpha \leqslant c_2, \quad \alpha = 1, 2,$$

$$\omega_h = \omega_1 \times \omega_2, \quad \omega_\alpha = \{x_\alpha = i_\alpha h_\alpha : i_\alpha = \overline{1, N_\alpha - 1}, \ h_\alpha = l_\alpha/N_\alpha\},$$
$$\alpha = 1, 2,$$

$$\bar{\omega}_h = \bar{\omega}_1 \times \bar{\omega}_2, \quad \bar{\omega}_\alpha = \{x_\alpha = i_\alpha h_\alpha : i_\alpha = \overline{0, N_\alpha}, \ h_\alpha = l_\alpha/N_\alpha\}, \quad \alpha = 1, 2,$$

$$\gamma_h = \bar{\omega}_h \setminus \omega_h.$$

Вправа 2. Встановити, за яких умов гладкості даних задачі (27) матиме місце другий порядок локальної апроксимації, тобто

$$\Lambda_\alpha u - L_\alpha u = O(h_\alpha^2)$$

у кожній точці сітки.

Введемо простір H розмірності $N = (N_1 - 1)(N_2 - 1)$ сіткових функцій, заданих на сітці ω_h, зі скалярним добутком

$$(y, v) = \sum_{x \in \omega_h} h_1 h_2 y(x) v(x)$$

і в ньому оператори

$$Ay = -\Lambda \mathring{y}, \quad A = A_1 + A_2,$$

$$A_1 y = -\Lambda_1 \mathring{y}, \quad A_2 = -\Lambda_2 \mathring{y},$$

$$y \in H, \quad y(x) = \mathring{y}(x), \quad x \in \omega_h, \quad \mathring{y}(x) = 0, \quad x \in \gamma_h.$$

Запишемо задачу (28) в операторному вигляді

$$Ay = \varphi, \quad y, \varphi \in H, \tag{29}$$

де $\varphi(x) = f(x) + \frac{1}{h_1^2}\varphi_1(x) + \frac{1}{h_2^2}\varphi_2(x)$,

$$\varphi_1(x) = \begin{cases} a_1(h_1, x_2)\,\mu(0, x_2), & x_1 = h_1, \\ 0, \quad 2h_1 \leqslant x_1 \leqslant l_1 - 2h_1, \\ a_1(l_1, x_2)\,\mu(l_1, x_2) & x_1 = l_1 - h_1, \end{cases}$$

$$\varphi_2(x) = \begin{cases} a_2(x_1, h_2)\,\mu(x_1, 0), & x_2 = h_2, \\ 0, \quad 2h_2 \leqslant x_2 \leqslant l_2 - 2h_2, \\ a_2(x_1, l_2)\,\mu(x_1, l_2), & x_2 = l_2 - h_2. \end{cases}$$

Використовуючи різницеві формули Гріна, неважко знайти, що оператор A самоспряжений в H і має місце рівність

$$(Ay, y) = -(\Lambda \mathring{y}, \mathring{y}) = \sum_{\alpha=1}^{2} (a_\alpha \mathring{y}_{\bar{x}_\alpha}^2, 1)_\alpha,$$

де

$$(u, v)_\alpha = \sum_{x_\alpha = h_\alpha}^{l_\alpha} \sum_{x_\beta = h_\beta}^{l_\beta - h_\beta} h_\alpha h_\beta u(x) v(x), \quad \alpha = 1, 2, \quad \beta = 3 - \alpha.$$

Для наближеного розв'язування рівняння (29) застосуємо попереміннотрикутний метод

$$B\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = \varphi, \quad k = 0, 1, \ldots, \ y_0 \in H,$$
$$B = (I + \omega R_1)(I + \omega R_2), \quad R_1 = R_2^*, \quad R = R_1 + R_2, \tag{30}$$

в якому регуляризатор R виберемо так, що

$$Ry = -\mathscr{R}\mathring{y}, \quad \mathscr{R}y = \mathring{y}_{\bar{x}_1 x_1} + \mathring{y}_{\bar{x}_2 x_2}, \ \mathring{y} \in \mathring{H},$$

де $\mathring{H}$ — простір сіткових функцій, заданих на ω_h, які перетворюються на нуль на $\omega_h \setminus \omega_h$. Оператори R_1, R_2 визначимо за формулами

$$R_\alpha y = -\mathscr{R}_\alpha \mathring{y}, \quad \mathscr{R}_1 \mathring{y} = -\sum_{\alpha=1}^{2} \frac{1}{h_\alpha} \mathring{y}_{x_\alpha}, \quad \mathscr{R}_2 \mathring{y} = \sum_{\alpha=1}^{2} \frac{1}{h_\alpha} \mathring{y}_{x_\alpha}.$$

Підсумовуючи частинами і користуючись нерівностями $0 < c_1 \leqslant a_\alpha \leqslant c_2$, послідовно дістаємо

$$-\sum_{i_\alpha=1}^{N_\alpha - 1} (a_\alpha \mathring{y}_{\bar{x}_\alpha})_{x_\alpha, i_1 i_2} \mathring{y}_{i_1 i_2} h_\alpha = \sum_{i_\alpha=1}^{N_\alpha} (a_\alpha (\mathring{y}_{\bar{x}_\alpha})^2)_{i_1 i_2} h_\alpha,$$

$$(Ry, y) = \sum_{\alpha=1}^{2} (\mathring{y}_{\bar{x}_\alpha}^2, 1)_\alpha, \quad (Ay, y) = \sum_{\alpha=1}^{2} (a_\alpha \mathring{y}_{\bar{x}_\alpha}^2, 1)_\alpha,$$

тобто $c_1 R \leqslant A \leqslant c_2 R$, $c_1 > 0$. Вище було показано, що

$$\delta I \leqslant R, \quad R_1 R_2 \leqslant \frac{\Delta}{4} R,$$

де δ та Δ визначаються формулами (25), (26). Тому має місце нерівність (див. лему 4, п. 1.4 з ч. 1)

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B,$$

де $\gamma_1 = c_1 \mathring{\gamma}_1$, $\gamma_2 = c_2 \mathring{\gamma}_2$ і за загальною теорією для числа ітерацій маємо оцінку

$$n \geqslant n_0(\varepsilon), \quad n_0(\varepsilon) = \sqrt{\frac{c_2}{c_1}}\,\bar{n}_0(\varepsilon),$$

$$\bar{n}_0(\varepsilon) = \frac{1}{2\sqrt{2}\sqrt[4]{\eta}} \ln\frac{2}{\varepsilon}, \quad \eta = \frac{\delta}{\Delta} = \frac{\pi^2}{4N^2},$$

тобто для рівняння зі змінними коефіцієнтами потрібно в $\sqrt{\frac{c_2}{c_1}}$ раз більше ітерацій, ніж для рівняння Пуассона.

Якщо коефіцієнти в задачі (27) дуже коливаються, то це число може бути значним і дуже збільшуватиме обчислювальну роботу. У такому разі можна розглянути модифікований варіант попереміннo-трикутного методу

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=\varphi,\quad k=0,\ 1,\ \ldots,\ y_0\in H,$$
$$B=(D+\omega R_1)D^{-1}(D+\omega R_2),\quad R_1=R_2^*,\quad R_1+R_2=A, \tag{31}$$

де покладемо $Dy=d(x)\,y$, $x\in\omega$, $d(x)$ — деяка додатна на ω сіткова функція, яку визначимо пізніше. У цьому разі D — самоспряжений і додатно-визначений в H оператор. Сіткова функція $d(x)$ у виразі (31) відіграє роль додаткового ітераційного параметра і дає змогу врахувати особливості оператора A у кожному вузлі сітки ω.

Визначимо оператори R_α таким чином:

$$R_\alpha y=-\mathscr{R}_\alpha\mathring{y},\ y\in H,\ \mathring{y}\in\mathring{H},\ y(x)=\mathring{y}(x),\ x\in\omega,$$
$$\mathscr{R}_1y=-\sum_{\alpha=1}^{2}\left(\frac{a_\alpha}{h_\alpha}y_{\bar{x}_\alpha}+\frac{a_{\alpha x_\alpha}}{2h_\alpha}y\right),$$
$$\mathscr{R}_2y=\sum_{\alpha=1}^{2}\left(\frac{a_\alpha^{+1}}{h_\alpha}y_{x_\alpha}+\frac{a_{\alpha x_\alpha}}{2h_\alpha}y\right),\quad x\in\omega,$$
$$a_1^{\pm1}(x)=a_1(x_1\pm h_1,\ x_2),\ a_2^{\pm1}(x)=a_2(x_1,\ x_2\pm h_2).$$

Покажемо, що оператори R_1 і R_2 спряжені в H. Для цього досить довести, що $(\mathscr{R}_1\mathring{y},\mathring{v})=(\mathring{y},\mathscr{R}_2\mathring{v})$, $\mathring{y}\in\mathring{H}$, $\mathring{v}\in\mathring{H}$. Користуючись різницевими формулами Гріна для функцій з $\mathring{H}$ і формулою різницевого диференціювання добутку сіткових функцій $(yv)_{x_\alpha}=y^{+1}v_{x_\alpha}+y_{x_\alpha}v$, маємо

$$(\mathscr{R}_1\mathring{y},\mathring{v})=-\sum_{\alpha=1}^{2}\frac{1}{h_\alpha}(a_\alpha\mathring{y}_{\bar{x}_\alpha},\mathring{v})-\sum_{\alpha=1}^{2}\frac{1}{2h_\alpha}(a_{\alpha x_\alpha}\mathring{y},\mathring{v})=$$
$$=\sum_{\alpha=1}^{2}\left[\frac{1}{h_\alpha}(\mathring{y},(a_\alpha\mathring{v})_{x_\alpha})-\frac{1}{2h_\alpha}(a_{\alpha x_\alpha}\mathring{y},\mathring{v})\right]=$$
$$=\sum_{\alpha=1}^{2}\left[\frac{1}{h_\alpha}(\mathring{y},a_\alpha^{+1}\mathring{v}_{x_\alpha})+\frac{1}{2h_\alpha}(\mathring{y},a_{\alpha x_\alpha}\mathring{v})\right]=(\mathring{y},\mathscr{R}_2\mathring{v}),$$

що й треба було довести. Згідно із загальною теорією далі нам треба

довести нерівності (див. п. 1.4.7 з ч. 1)

$$\delta D\leqslant A,\quad R_1D^{-1}R_2\leqslant\frac{\Delta}{4}A,\quad \delta>0, \tag{32}$$

і оскільки відношення $\eta=\frac{\delta}{\Delta}$ визначає число ітерацій, то сіткову функцію $d(x)$ слід вибрати з умови максимальності цього відношення. Але спочатку доведемо таке допоміжне твердження.

Лема 1. *Нехай $p_\alpha(x)$, $q_\alpha(x)$, $u_\alpha(x)$, $v_\alpha(x)$, $\alpha=1,\ 2$ — сіткові функції, задані на ω. Тоді для будь-якого $x\in\omega$ має місце нерівність*

$$\left[\sum_{\alpha=1}^{2}(p_\alpha u_\alpha+q_\alpha v_\alpha)\right]^2\leqslant(1+\varepsilon)(|p_1|+\varkappa_1|q_1|)\left(|p_1|u_1^2+\right.$$
$$\left.+\frac{|q_1|}{\varkappa_1}v_1^2\right)+\frac{1+\varepsilon}{\varepsilon}(|p_2|+\varkappa_2|q_2|)\left(|p_2|u_2^2+\frac{|q_2|}{\varkappa_2}v_2^2\right), \tag{33}$$

де $\varepsilon(x)$, $\varkappa_1(x)$, $\varkappa_2(x)$ — довільні додатні на ω сіткові функції.

Доведення. Використовуючи ε-нерівність $2ab\leqslant\varepsilon a^2+\varepsilon^{-1}b^2$, $\varepsilon>0$, дістаємо

$$\left[\sum_{\alpha=1}^{2}(p_\alpha u_\alpha+q_\alpha v_\alpha)\right]^2=(p_1u_1+q_1v_1)^2+2(p_1u_1+q_1v_1)\times(p_2u_2+q_2v_2)+$$
$$+(p_2u_2+q_2v_2)^2\leqslant(1+\varepsilon)(p_1u_1+q_1v_1)^2+\frac{1+\varepsilon}{\varepsilon}(p_2u_2+q_2v_2)^2.$$

Із тієї самої нерівності маємо

$$(p_\alpha u_\alpha+q_\alpha v_\alpha)^2=p_\alpha^2u_\alpha^2+2q_\alpha p_\alpha u_\alpha v_\alpha+q_\alpha^2v_\alpha^2\leqslant$$
$$\leqslant p_\alpha^2u_\alpha^2+|p_\alpha||q_\alpha|\left(\varkappa_\alpha u_\alpha^2+\frac{1}{\varkappa_\alpha}v_\alpha^2\right)+q_\alpha^2v_\alpha^2=$$
$$=(|p_\alpha|+\varkappa_\alpha|q_\alpha|)\left(|p_\alpha|u_\alpha^2+\frac{|q_\alpha|}{\varkappa_\alpha}v_\alpha^2\right),\quad \varkappa_\alpha>0,\quad \alpha=1,\ 2.$$

Підставляючи цю нерівність у попередню, дістаємо твердження леми.

Підставляючи у (33) $p_\alpha=\frac{a_\alpha^{+1}}{h_\alpha}$, $q_\alpha=0{,}5a_{\alpha x_\alpha}$, $u_\alpha=\mathring{y}_{x_\alpha}$, $v_\alpha=\frac{1}{h_\alpha}\mathring{y}$, $\alpha=1,\ 2$, і користуючись означенням операторів R_1, R_2, маємо

$$(R_1D^{-1}R_2y,\ y)=(D^{-1}R_2\mathring{y},\ R_2\mathring{y})=$$
$$=\left(\frac{1}{d}\sum_{\alpha=1}^{2}\left(\frac{a_\alpha^{+1}}{h_\alpha}\mathring{y}_{x_\alpha}+\frac{a_{\alpha x_\alpha}}{2h_\alpha}\mathring{y}\right)^2,\ 1\right)\leqslant$$

$$\leqslant\left(\frac{(1+\varepsilon)}{dh_1^2}(a_1^{+1}+0{,}5h_1\varkappa_1\,|\,a_{1x_1}|)\right)\left(a_1^{+1}\overset{\circ}{y}{}^2_{x_1}+\frac{0{,}5h_1\,|\,a_{1x_1}|}{\varkappa_1 h_1^2}\overset{\circ}{y}{}^2,\ 1\right)+ \quad (34)$$

$$+\left(\frac{1+\varepsilon}{d\varepsilon h_2^2}(a_2^{+1}+0{,}5h_2\varkappa_2\,|\,a_{2x_2}|)\left(a_2^{+1}\overset{\circ}{y}{}^2_{x_2}+\frac{0{,}5h_2\,|\,a_{2x_2}|}{\varkappa_2 h_2^2}\overset{\circ}{y}{}^2\right),\ 1\right).$$

Вважатимемо, що в цій рівності $\varkappa_1$ є функцією лише від x_2, а $\varkappa_2$ — лише від x_1, тобто візьмемо

$$\varkappa_\alpha=\varkappa_\alpha(x_\beta),\quad \alpha=1,\ 2,\quad \beta=3-\alpha.$$

Покладемо

$$\varepsilon=\varepsilon(x)=\frac{a_2^{+1}+0{,}5h_2\varkappa_2\,|\,a_{2x_2}|}{a_1^{+1}+0{,}5h_1\varkappa_1\,|\,a_{1x_1}|}\,\frac{h_1^2\theta_2(x_1)}{h_2^2\theta_1(x_2)} \quad (35)$$

і визначимо $d(x)$ за формулою

$$d(x)=\sum_{\alpha=1}^{2}(a_\alpha^{+1}+0{,}5h_\alpha\varkappa_\alpha\,|\,a_{\alpha x_\alpha}|)\frac{\theta_\alpha}{h_\alpha^2}, \quad (36)$$

де $\theta_\alpha=\theta_\alpha(x_\beta)$, $\alpha=1,\ 2$, $\beta=3-\alpha$ — додатні на ω сіткові функції, які ще потрібно визначити.

Підставляючи (35), (36) в (34), маємо

$$(R_1D^{-1}R_2y,\ y)\leqslant\sum_{\alpha=1}^{2}\left(\frac{a_\alpha^{+1}}{\theta_\alpha}\overset{\circ}{y}{}^2_{x_\alpha},\ 1\right)+\sum_{\alpha=1}^{2}\left(\frac{|\,a_{\alpha x_\alpha}|}{2h_\alpha\theta_\alpha\varkappa_\alpha}\overset{\circ}{y}{}^2,\ 1\right).$$

Оскільки θ_α не залежить від x_α, то

$$\left(\frac{a_\alpha^{+1}}{\theta_\alpha}\overset{\circ}{y}{}^2_{x_\alpha},\ 1\right)\leqslant\left(\frac{a_\alpha}{\theta_\alpha}\overset{\circ}{y}{}^2_{x_\alpha},\ 1\right)_\alpha,$$

тому

$$(R_1D^{-1}R_2y,\ y)\leqslant\sum_{\alpha=1}^{2}\left(\frac{a_\alpha}{\theta_\alpha}\overset{\circ}{y}{}^2_{x_\alpha},\ 1\right)_\alpha+\sum_{\alpha=1}^{2}\left(\frac{|\,a_{\alpha x_\alpha}|}{2h_\alpha\theta_\alpha\varkappa_\alpha}\overset{\circ}{y}{}^2,\ 1\right). \quad (37)$$

Виберемо тепер θ_α та $\varkappa_\alpha$. Позначимо

$$\omega_\alpha=\{x_\alpha=i_\alpha h_\alpha,\ i_\alpha=\overline{1,\ N_\alpha-1},\ h_\alpha=l_\alpha/N_\alpha\},$$

$$\omega_\alpha^+=\{x_\alpha=i_\alpha h_\alpha:i_\alpha=\overline{1,\ N_\alpha},\quad h_\alpha=l_\alpha/N_\alpha\},\quad \alpha=1,\ 2,$$

а також введемо скалярні добутки

$$(u,\ v)_{\omega_\alpha}=\sum_{x_\alpha\in\omega_\alpha}u(x)\,v(x)\,h_\alpha,$$

$$(u,\ v)_{\omega_\alpha^+}=\sum_{x_\alpha\in\omega_\alpha^+}u(x)\,v(x)\,h_\alpha.$$

Тоді легко помітити, що

$$(u,\ v)=((u,\ v)_{\omega_1},\ 1)_{\omega_2}=((u,\ v)_{\omega_2},\ 1)_{\omega_1},$$
$$(u,\ v)_\alpha=((u,\ v)_{\omega_\alpha^+},\ 1)_{\omega_\beta},\quad \alpha=1,\ 2,\quad \beta=3-\alpha. \quad (38)$$

Нехай тепер $b_\alpha(x_\beta)=\max\limits_{x_\alpha\in\omega_\alpha}v^\alpha(x)$, $\alpha=1,\ 2$, $x_\beta\in\omega_\beta$, де $v^\alpha(x)$ для фіксованого x_β є розв'язком такої триточкової різницевої задачі:

$$(a_\alpha v^\alpha_{x_\alpha})_{x_\alpha}=-\frac{a_\alpha^{+1}}{h_\alpha^2},\quad h_\alpha\leqslant x_\alpha\leqslant l_\alpha-h_\alpha,$$
$$v^\alpha(x)=0,\quad x=0,\ l_\alpha,\quad x_\beta\in\omega_\beta. \quad (39)$$

Лема 2. *Нехай $\rho_i\geqslant 0$ — сіткова функція, задана на ω і не рівна тотожно нулю. Для будь-якої сіткової функції y_i, заданої на $\bar\omega=\{x_i=ih:i=\overline{0,\ N},\ h=l/N\}=\omega\cup\{0,\ l\}$ і такої, що $y_0=y_N=0$ має місце оцінка*

$$\gamma_1(\rho y,\ y)\leqslant(y_{\bar x}^2,\ 1)_{\omega+}, \quad (40)$$

де

$$\gamma_1^{-1}=\max_{i=\overline{1,N-1}}v_i,\quad (u,\ v)=\sum_{x\in\omega}hu(x)\,v(x),\quad (u,\ v)_{\omega+}=$$
$$=(u,\ v)+hu(x_N)\,v(x_N),$$

v_i — розв'язок крайової задачі

$$\Lambda v_i\equiv v_{\bar x x,i}=-\rho_i,\quad i=\overline{1,\ N-1},\quad v_0=v_N=0. \quad (41)$$

Доведення. Скористаємось властивостями функції Гріна G_{ik}, яка є розв'язком задачі

$$\Lambda G_{ik}=G_{\bar x x,ik}=-\frac{1}{h}\delta_{i,k},\quad i=\overline{1,\ N-1},\quad G_{0,k}=G_{N,k}=0,$$

де $\delta_{i,k}$ — символ Кронекера. Неважко довести (за допомогою різницевих формул Гріна), що

1) функція Гріна симетрична, тобто $G_{i,k}=G_{k,i}$. Крім того, $G_{i,k}$ як функція k при фіксованому $i=\overline{1,\ N-1}$ задовольняє умови

$$\Lambda G_{i,k}=-h^{-1}\delta_{i,k},\quad k=\overline{1,\ N-1},\quad G_{i,0}=G_{i,N}=0;$$

2) функція Гріна додатна, тобто $G_{i,k}>0$ при $i,\ k\neq\overline{0,\ N}$;

3) розв'язок задачі (41) можна подати у вигляді

$$v_i=-\sum_{k=1}^{N-1}G_{i,k}\Lambda v_k h=\sum_{k=1}^{N-1}G_{i,k}\rho_k h. \quad (42)$$

Використовуючи (42), матимемо

$$(\rho y, y) = \sum_{i=1}^{N-1} \rho_i y_i^2 h = -\sum_{i=1}^{N-1} \rho_i y_i h \left(\sum_{k=1}^{N-1} G_{i,k} \Lambda y_k h \right) =$$

$$= -\sum_{k=1}^{N-1} h \Lambda y_k \left(\sum_{i=1}^{N-1} \rho_i y_i G_{i,k} h \right) = -(\Lambda y, w),$$

де

$$w_k = \sum_{i=1}^{N-1} \rho_i y_i G_{i,k} h, \quad k = \overline{0, N}.$$

Застосовуючи узагальнену нерівність Коші — Буняковського для невід'ємного самоспряженого оператора — Λ:

$$|(-\Lambda u, v)| \leqslant (-\Lambda u, u)^{1/2} (-\Lambda v, v)^{1/2},$$

дістаємо

$$(\rho y, y) \leqslant (-\Lambda y, y)^{1/2} (-\Lambda w, w)^{1/2}$$

або в силу того, що $(-\Lambda y, y) = (y_{\bar{x}}^2, 1)_{\omega+}$, маємо

$$(\rho y, y)^2 \leqslant (y_{\bar{x}}^2, 1)_{\omega+} (-\Lambda w, w).$$

Із першої властивості функції Гріна

$$-\Lambda w_k = -\sum_{i=1}^{N-1} h \rho_i y_i \Lambda G_{i,k} = \sum_{i=1}^{N-1} \rho_i y_i \delta_{i,k} = \rho_k y_k,$$

тому

$$(-\Lambda w, w) = \sum_{k=1}^{N-1} h \rho_k y_k \left(\sum_{i=1}^{N-1} h \rho_i y_i G_{i,k} \right) = \sum_{i=1}^{N-1} \sum_{k=1}^{N-1} a_{i,k} y_i y_k,$$

де $a_{i,k} = h^2 \rho_i \rho_k G_{i,k}$, $1 \leqslant i,\ k \leqslant N-1$. Використовуючи нерівність $2y_i y_k \leqslant y_i^2 + y_k^2$, (42), а також симетрію і додатність функцій Гріна, знаходимо далі

$$(-\Lambda w, w) \leqslant \sum_{i=1}^{N-1} 0{,}5 y_i^2 \sum_{k=1}^{N-1} a_{ik} + \sum_{k=1}^{N-1} 0{,}5 y_k^2 \sum_{i=1}^{N-1} a_{ki} =$$

$$= \sum_{i=1}^{N-1} y_i^2 \sum_{k=1}^{N-1} a_{ik} = \sum_{i=1}^{N-1} \rho_i y_i^2 h \left(\sum_{k=1}^{N-1} \rho_k G_{i,k} h \right) =$$

$$= \sum_{i=1}^{N-1} \rho_i y_i^2 h v_i \leqslant \max_{i=\overline{1,N-1}} v_i (\rho y, y) = \frac{1}{\gamma_1} (\rho y, y),$$

що і доводить лему 2.

Цілком аналогічно доводиться наступне твердження.

Лема 3. *Нехай* $\rho_i \geqslant 0$, $d_i \geqslant 0$ *задані на* ω, *а функція* $a_i \geqslant c_1 > 0$ *задана на* ω^+. *Для довільної функції* y_i, $i = \overline{0, N}$, $y_0 = y_N = 0$, *має місце оцінка*

$$\gamma_1 (\rho y, y) \leqslant (a y_{\bar{x}}^2, 1)_{\omega+} + (dy, y), \quad \gamma_1^{-1} = \max_{i=\overline{1,N-1}} v_i,$$

де v_i — *розв'язок задачі*

$$\Lambda v_i \equiv (a v_{\bar{x}})_{x,i} - d_i v_i = -\rho_i, \quad i = \overline{1, N-1}, \quad v_0 = v_N = 0.$$

В силу леми 3 з розв'язку (39) маємо

$$\left(\frac{a_\alpha^{+1}}{h_\alpha^2} \mathring{y}^2, 1 \right)_{\omega_\alpha} \leqslant b_\alpha (x_\beta) (a_\alpha \mathring{y}_{\bar{x}_\alpha}^2, 1)_{\omega_\alpha^+}, \quad \alpha = 1, 2.$$

Помножимо цю нерівність на $\theta_\alpha (x_\beta)$ і просумуємо по ω_β. Тоді в силу (38) дістаємо

$$\left(\frac{\theta_\alpha a_\alpha^{+1}}{h_\alpha^2} \mathring{y}^2, 1 \right) \leqslant (b_\alpha a_\alpha \theta_\alpha \mathring{y}_{\bar{x}_\alpha}^2, 1)_\alpha, \quad \alpha = 1, 2. \tag{43}$$

Нехай $c_\alpha (x_\beta) = \max\limits_{x_\alpha \in \omega_\alpha} w^\alpha (x)$, $\alpha = 1, 2$, $x_\beta \in \omega_\beta$, де $w^\alpha (x)$ для фіксованого x_β є розв'язком задачі

$$(a_\alpha w_{\bar{x}_\alpha}^\alpha)_{x_\alpha} = -\frac{|a_{\alpha x_\alpha}|}{2h_\alpha}, \quad h_\alpha \leqslant x_\alpha \leqslant l_\alpha - h_\alpha,$$
$$w^\alpha (x) = 0, \quad x_\alpha = 0, l_\alpha, \quad x_\beta \in \omega_\beta. \tag{44}$$

Аналогічно попередньому дістаємо нерівності

$$\left(\frac{\varkappa_\alpha \theta_\alpha |a_{\alpha x_\alpha}|}{2h_\alpha} \mathring{y}^2, 1 \right) \leqslant (\varkappa_\alpha \theta_\alpha c_\alpha a_\alpha \mathring{y}_{\bar{x}_\alpha}^2, 1)_\alpha, \quad \alpha = 1, 2, \tag{45}$$

$$\left(\frac{|a_{\alpha x_\alpha}|}{2h_\alpha \theta_\alpha \varkappa_\alpha} \mathring{y}^2, 1 \right) \leqslant \left(\frac{c_\alpha}{\varkappa_\alpha \theta_\alpha} a_\alpha \mathring{y}_{\bar{x}_\alpha}^2, 1 \right)_\alpha, \quad \alpha = 1, 2. \tag{46}$$

Додамо нерівності (43) та (45) і просумуємо їх по α, тоді в силу (35) дістанемо

$$(d\mathring{y}^2, 1) = (Dy, y) \leqslant ((\varkappa_\alpha c_\alpha + b_\alpha) \theta_\alpha a_\alpha \mathring{y}_{\bar{x}_\alpha}^2, 1)_\alpha.$$

Вибираючи θ_α за формулою

$$\theta_\alpha (x_\beta) = \frac{1}{b_\alpha (x_\beta) + c_\alpha (x_\beta) \varkappa_\alpha (x_\beta)}, \quad \alpha = 1, 2, \quad \beta = 3 - \alpha, \tag{47}$$

і враховуючи рівність

$$(Ay,\ y) = -(\Lambda \mathring{y},\ \mathring{y}) = \sum_{\alpha=1}^{2} (a_\alpha \mathring{y}^2_{\bar{x}_\alpha},\ 1)_\alpha, \qquad (48)$$

дістаємо $(Dy,\ y) \leqslant (Ay,\ y)$. Отже, в (32) можна покласти $\delta = 1$.

Оцінимо тепер Δ. Для цього підставимо (46) в (37) і врахуємо формулу (47). У результаті визначимо оцінку

$$(R_1 D^{-1} R_2 y,\ y) \leqslant \sum_{\alpha=1}^{2} ((1 + c_\alpha/\varkappa_\alpha)(b_\alpha + c_\alpha \varkappa_\alpha)\, a_\alpha \mathring{y}^2_{\bar{x}_\alpha},\ 1)_\alpha.$$

Виберемо тепер $\varkappa_\alpha$ з умови мінімуму виразу $(1 + c_\alpha/\varkappa_\alpha)(b_\alpha + c_\alpha \varkappa_\alpha)$ по $\varkappa_\alpha$. Дістаємо $\varkappa_\alpha(x_\beta) = \sqrt{b_\alpha(x_\beta)}$, $\alpha = 1, 2$, $\beta = 3 - \alpha$, причому

$$(R_1 D^{-1} R_2 y,\ y) \leqslant \sum_{\alpha=1}^{2} ((c_\alpha + \sqrt{b_\alpha})^2 a_\alpha \mathring{y}^2_{\bar{x}_\alpha},\ 1).$$

Порівнюючи цю оцінку з (48), знаходимо, що в (32) можна покласти

$$\Delta = 4 \max_{\alpha=1,2} (\max_{x_\beta \in \omega_\beta} (c_\alpha(x_\beta) + \sqrt{b_\alpha(x_\beta)})^2), \quad \beta = 3 - \alpha. \qquad (49)$$

Підставляючи в (36) знайдені вирази для $\varkappa_\alpha$, θ_α, дістаємо

$$d(x) = \sum_{\alpha=1}^{2} \left(\frac{a_\alpha^{+1}}{h_\alpha^2 \sqrt{b_\alpha}} + \frac{|a_{\alpha x_\alpha}|}{2h_\alpha} \right) \frac{1}{c_\alpha + \sqrt{b_\alpha}}, \quad x \in \omega. \qquad (50)$$

Загальна теорія дає тепер для числа ітерацій оцінку

$$n \geqslant n_0(\varepsilon), \quad n_0(\varepsilon) = \frac{\sqrt[4]{\Delta}\, \ln(2/\varepsilon)}{2\sqrt{2}}.$$

В силу нерівностей $0 < c_1 \leqslant a_\alpha \leqslant c_2$ з (39), (44) дістаємо, що $b_\alpha = O(1/h_\alpha^2)$, $c_\alpha = O(1/h_\alpha)$, якщо число точок, в яких $a_{\alpha x_\alpha} = O(h_\alpha^{-1})$ скінченне і не залежить від h. Звідси

$$n_0(\varepsilon) = O(\sqrt{N} \ln(2/\varepsilon)),$$

тобто асимптотика числа ітерацій така сама, як і для методу (30). Проте на відміну від останнього тут оцінка числа ітерацій залежить не від екстремальних характеристик коефіцієнтів $a_\alpha(x)$, $\alpha = 1, 2$, а від їхніх інтегральних характеристик. Розрахунки на модельних прикладах показують, що число ітерацій незначно залежить від відношення c_2/c_1.

Метод (31) реалізується в такій послідовності. Спочатку для фіксованого $x_\beta \in [h_\beta,\ l_\beta - h_\beta]$ методом прогонки розв'язуються задачі (39), (44) і визначаються $b_\alpha(x_\beta)$, $c_\alpha(x_\beta)$, $\alpha = 1, 2$. Ці чотири одновимірні сіткові функції використовуються в процесі ітерацій для обчислення $d(x)$ за формулою (50). Простота цієї формули дає змогу

не зберігати двовимірну сіткову функцію $d(x)$, а обчислювати її при необхідності.

Далі за формулою (49) знаходять Δ і покладають $\delta = 1$. За цими величинами знаходять число ітерацій і згідно із загальною теорією визначають параметри ω і τ_k. Для визначення y_{k+1} за y_k можна використати таку схему:

$$(D + \omega_0 R_1) v = \varphi_k, \quad (D + \omega_0 R_2) \mathring{y}_{k+1} = Dv, \qquad (51)$$

$$\varphi_k = (D + \omega_0 R_1) D^{-1} (D + \omega_0 R_2) y_k - \tau_{k+1}(Ay_k - \varphi),$$

або в індексному вигляді

$$v(i,\ j) = \alpha_1(i,\ j) v(i-1,\ j) + \beta_1(i,\ j) v(i,\ j-1) +$$
$$+ \varkappa(i,\ j) \varphi_k(i,\ j), \quad i = \overline{1,\ N_1 - 1}, \quad j = \overline{1,\ N_2 - 1}; \qquad (52)$$
$$v(0,\ j) = 0, \quad j = \overline{1,\ N_2 - 1}, \quad v(i,\ 0) = 0, \quad i = \overline{1,\ N_1 - 1};$$

$$\mathring{y}_{k+1}(i,\ j) = \alpha_2(i,\ j) \mathring{y}_{k+1}(i+1,\ j) + \beta_2(i,\ j) \mathring{y}_{k+1}(i,\ j+1) +$$
$$+ \varkappa(i,\ j)\, d(i,\ j)\, v(i,\ j), \quad i = N_1 - 1, \ldots, 1; \quad j = N_2 - 1, \ldots, 1, \qquad (53)$$

де

$$\alpha_1 = \frac{\omega_0 a_1 \varkappa}{h_1^2}, \quad \beta_1 = \frac{\omega_0 a_2 \varkappa}{h_2^2}, \quad \alpha_2 = \frac{\omega_0 a_1^{+1} \varkappa}{h_1^2}, \quad \beta_2 = \frac{\omega_0 a_2^{+1} \varkappa}{h_2^2},$$

$$\frac{1}{\varkappa} = d + \omega_0 \left[\frac{a_1^{+1} + a_1}{2h_1^2} + \frac{a_2^{+1} + a_2}{2h_2^2} \right].$$

Права частина $\varphi_k(i,\ j)$ обчислюється за формулами

$$\varphi_k(i,\ j) = [P(i-1,\ j) + Q(i,\ j-1) + S(i,\ j)] \mathring{y}_k(i,\ j) +$$
$$+ R_1(i,\ j) \mathring{y}_k(i+1,\ j) + R_1(i-1,\ j) \mathring{y}_k(i-1,\ j) +$$
$$+ R_2(i,\ j) \mathring{y}_k(i,\ j+1) + R_2(i,\ j-1) \mathring{y}_k(i,\ j-1) +$$
$$+ G(i-1,\ j) \mathring{y}_k(i-1,\ j+1) + G(i,\ j-1) \mathring{y}_k(i+1,\ j-1) +$$
$$+ \tau_{k+1} \varphi(i,\ j), \quad i = \overline{1,\ N_1 - 1}, \quad j = \overline{1,\ N_2 - 1},$$

де

$$G = \frac{\omega_0^2 a_1^{+1} a_2^{+1}}{h_1^2 h_2^2 d}; \quad R_\alpha = \left(\tau_{k+1} - \frac{\omega_0}{d\varkappa} \right) \frac{a_\alpha^+}{h_\alpha^2}, \quad \alpha = 1, 2,$$

$$S = \frac{1}{\omega_0 \varkappa} \left[\frac{\omega_0}{d\varkappa} - 2\tau_{k+1}(1 - \varkappa d) \right], \quad P = \frac{\omega_0^2 (a_1^{+1})^2}{h_1^4 d}, \quad Q = \frac{\omega_0^2 (a_2^{+1})^2}{h_2^4 d},$$

причому $P(0, j) = 0$, $j = \overline{1, N_2 - 1}$, $Q(i, 0) = 0$, $i = \overline{1, N_1 - 1}$.

Зауважимо, що в силу (49), (50) справедливі оцінки

$$c_\alpha + \sqrt{b_\alpha} \leqslant \frac{\sqrt{\Delta}}{2} = \frac{1}{\omega_0}, \quad d \geqslant \omega_0 \sum_{\alpha=1}^{2} \frac{|a_{\alpha x_\alpha}|}{2h_\alpha},$$

звідки

$$\frac{1}{\varkappa} = d + \omega_0 \sum_{\alpha=1}^{2} \frac{a_\alpha^{+1} + a_\alpha}{2h_\alpha^2} \geqslant \omega_0 \sum_{\alpha=1}^{2} \left(\frac{|a_{\alpha x_\alpha}|}{2h_\alpha} + \frac{a_\alpha^{+1} + a_\alpha}{2h_\alpha^2} \right) =$$

$$= \omega_0 \left(\max\left(\frac{a_1^{+1}}{h_1^2}, \frac{a_1}{h_1^2} \right) + \max\left(\frac{a_2^{+1}}{h_2^2}, \frac{a_2}{h_2^2} \right) \right),$$

або

$$\max(\alpha_1, \alpha_2) + \max(\beta_1, \beta_2) \leqslant 1.$$

Звідси $\alpha_1 + \beta_1 \leqslant 1$, $\alpha_2 + \beta_2 \leqslant 1$, тому рахунок за формулами (52), (53) буде стійким.

Розглянемо, крім (27), ще задачу

$$\begin{gathered} \tilde{L}u \equiv Lu - q(x)\,u = -f(x), \quad x \in \Omega, \\ u(x) = \mu(x), \quad x \in \Gamma, \end{gathered} \tag{54}$$

де

$$0 \leqslant q(x) \leqslant c, \quad c = \text{const} \gg 1. \tag{55}$$

Для її наближеного розв'язування застосуємо різницеву схему

$$\begin{gathered} \tilde{\Lambda}y \equiv \Lambda y - q(x)\,y = -f(x), \quad x \in \omega_h, \\ y(x) = \mu(x), \quad x \in \gamma_h, \end{gathered} \tag{56}$$

де оператор Λ той самий, що і в (28). Аналогічно (29) схему (56) можна записати як операторне рівняння в просторі H:

$$\tilde{A}y = \varphi, \ y, \ \varphi \in H, \quad \tilde{A} = A + Q, \quad Q = q(x)\,I \tag{57}$$

і застосувати для його розв'язування попереміннo-трикутний метод

$$\tilde{B}\frac{y_{k+1} - y_k}{\tau_{k+1}} + \tilde{A}y_k = \varphi, \quad k = 0, 1, \ldots, \ \forall\, y_0 \in H, \tag{58}$$

$$\begin{gathered} \tilde{B} = (\tilde{D} + \omega\tilde{R}_1)\,\tilde{D}^{-1}(\tilde{D} + \omega\tilde{R}_2), \\ \tilde{R} = \tilde{R}_1 + \tilde{R}_2 = \tilde{R}^* > 0, \quad \tilde{R}_2 = \tilde{R}_1^*. \end{gathered} \tag{59}$$

Нехай відомо, що

$$\begin{gathered} A = A^* > 0, \quad c_1 R \leqslant A \leqslant c_2 R, \quad c_1 > 0, \\ R = R_1 + R_2 = R^* > 0, \quad R_1 = R_2^*, \end{gathered} \tag{60}$$

і, крім того, для діагонального оператора $D = d(x)\,I$ маємо

$$\delta D \leqslant R, \quad R_1 D^{-1} R_2 \leqslant \frac{\Delta}{4} R, \quad D = D^* > 0. \tag{61}$$

Тоді для оператора $\tilde{A} = \tilde{A}^* > 0$ неважко знайти оцінку

$$c_1 R \leqslant \tilde{A} \leqslant \left(c_2 + \frac{c}{d\delta} \right) R. \tag{62}$$

Якщо застосувати метод (58) для розв'язування рівняння $Ay = \varphi$ при $D \geqslant d(x)\,I$, $\tilde{R}_\alpha = R_\alpha$, $\alpha = 1, 2$, $\tilde{A} = A$, то загальна теорія дає таку оцінку кількості ітерацій:

$$n \geqslant n_0(\varepsilon) = \sqrt{\frac{c_2}{c_1}}\,\frac{\ln\frac{e}{2}}{2\sqrt{2}\sqrt[4]{\eta}}, \quad \eta = \frac{\delta}{\Delta},$$

а в силу (62) метод (58) для рівняння (57) при тому самому виборі $\tilde{D}$ і $\tilde{R}_\alpha$ потребує кількості ітерацій

$$n \geqslant \tilde{n}_0(\varepsilon) = \sqrt{1 + \frac{c}{c_2 d\delta}}\, n_0(\varepsilon),$$

що при великих c, тобто при сильно осцилюючому коефіцієнті $q(x)$ може бути значно більше n_0.

Тому розглянемо таку модифікацію методу (58), (59), яка використовує оператори

$$\tilde{R} = R + \tilde{Q}, \quad \tilde{Q} = \frac{1}{O_1} q(x), \quad \tilde{R}_i = R_i + \frac{1}{2}\tilde{Q}, \quad i = 1, 2, \tag{63}$$

$$\begin{gathered} \tilde{D} = D + s\tilde{Q}, \quad s \geqslant 0, \\ \tilde{B} = (\tilde{D} + \tilde{\omega}\tilde{R}_1)\,\tilde{D}^{-1}(\tilde{D} + \tilde{\omega}\tilde{R}_2). \end{gathered}$$

Неважко помітити, що

$$\begin{gathered} c_1\tilde{R} \leqslant \tilde{A} \leqslant c_2\tilde{R}, \\ D = D + s\tilde{Q} \leqslant \frac{1}{\delta} R + s\tilde{Q} \leqslant \max\left\{ \frac{1}{\delta}, s \right\} \tilde{R}. \end{gathered} \tag{64}$$

Далі

$$(\tilde{R}_1\tilde{D}^{-1}\tilde{R}_2 y, y) = (\tilde{D}^{-1}\tilde{R}_2, \tilde{R}_2 y) = (D^{-1}R_2 y, R_2 y) + (\tilde{D}^{-1}\tilde{Q}y, R_2 y) +$$

$$+ \frac{1}{4}(\tilde{D}^{-1}\tilde{Q}y, \tilde{Q}y) \leqslant (D^{-1}R_2 y, R_2 y) + (D^{-1}QD^{1/2}y, D^{-1/2}R_2 y) +$$

$$+\frac{1}{4}(\tilde{D}^{-1}\tilde{Q}\tilde{Q}^{1/2}y,\ \tilde{Q}^{1/2}y)\leqslant\frac{\Delta}{4}(Ry,\ y)+\|\tilde{D}^{-1}\tilde{Q}\|\times$$

$$\times\left\{(Dy,\ y)^{1/2}(D^{-1}R_2y,\ R_2y)^{1/2}+\frac{1}{4}(\tilde{Q}y,\ y)\right\}.$$

Оскільки $\|\tilde{D}^{-1}\tilde{Q}\|\leqslant\frac{c}{d+cs}$, то

$$(\tilde{R}_1\tilde{D}^{-1}\tilde{R}_2y,\ y)\leqslant\frac{\Delta}{4}(Ry,\ y)+\frac{c}{d+cs}\left\{\left(\frac{\Delta}{4\delta}\right)^{1/2}(Ry,\ y)+\frac{1}{4}(\tilde{Q}y,\ y)\right\}\leqslant$$

$$\leqslant\frac{1}{4}\left(\Delta+\frac{2c}{d+sc}\sqrt{\frac{\Delta}{\delta}}\right)(Ry,\ y)+\frac{1}{4}\frac{c}{d+sc}(\tilde{Q}y,\ y)\leqslant$$

$$\leqslant\frac{1}{4}\left(\Delta+\frac{2c\eta^{-1/2}}{d+sc}\right)(\tilde{R}y,\ y).$$

Отже (60), (61) для оператора $\tilde{A}$ виконується відповідно з константами

$$\tilde{c}_1=c_1,\quad \tilde{c}_2=c_2,\quad \tilde{\delta}=\min\left\{\delta,\ \frac{1}{s}\right\},\quad \tilde{\Delta}=\Delta+\frac{2c\eta^{-1/2}}{d+sc},$$

і тому загальна теорія ПТМ дає оцінку кількості ітерацій

$$n(\varepsilon,\ s)=\sqrt{\frac{c_2}{c_1}}\frac{\ln(2/\varepsilon)}{2\sqrt{2}\,(\tilde{\eta}(s))^{1/4}},\quad \tilde{\eta}=\frac{\tilde{\delta}}{\tilde{\Delta}}.$$

Мінімум $\tilde{n}(\varepsilon)$ функції $n(\varepsilon,\ s)$ досягається при максимальному $\tilde{\eta}(s)$.

Лема 4. *Функція* $\tilde{\eta}=\tilde{\eta}(s)$ *досягає максимуму*

$$\tilde{\eta}_{\max}=\frac{\eta}{1+2\sqrt{\eta}\left(1+\frac{\delta d}{c}\right)^{-1}}\geqslant\frac{\eta}{1+2\sqrt{\eta}},\quad \eta=\frac{\delta}{\Delta}$$

при $s_{\max}=\delta^{-1}$.

Доведення. Для $s\geqslant\delta^{-1}$

$$\tilde{\eta}(s)=\left(s\Delta+2\eta^{-1/2}\left(1+\frac{d}{2c}\right)\right)^{-1},$$

і це є монотонно спадна функція. Тому $s_{\max}\leqslant\delta^{-1}$. З іншого боку, для $s\leqslant\delta^{-1}$

$$\tilde{\eta}(s)=\delta\left(\Delta+2\eta^{-1/2}\left(s+\frac{d}{s}\right)^{-1}\right)^{-1}$$

і ця функція монотонно зростає, тобто має бути $s_{\max}\geqslant\delta^{-1}$. Це і доводить лему.

Таким чином, ПТМ (58), (59) при виборі операторів (63), де $s=\delta^{-}$, $\tilde{\omega}=1/\sqrt{\tilde{\delta}\tilde{\Delta}}$, потребує

$$\tilde{n}(\varepsilon)=a_0(\varepsilon)(1+2\eta^{1/2})^{1/4},\quad n_0(\varepsilon)=\sqrt{\frac{c_2}{c_1}}\frac{\ln\frac{2}{\varepsilon}}{2\sqrt{2}\sqrt{\eta}},\quad \eta=\frac{\delta}{\Delta}$$

ітерацій. Оскільки $\eta<1$, то число ітерацій не набагато більше від $n_0(\varepsilon)$ і не залежить від c. Якщо крім c також c_2/c_1 — велике число, тобто сильно осцилюють також коефіцієнти k_α, то за рахунок множника $\sqrt{c_2/c_1}$ число ітерацій може бути знову значним. У цьому випадку в методі (58), (59) можна вибрати замість $\tilde{D}_1\tilde{R}_\alpha$ оператори (див. (50))

$$\hat{D}\overset{\circ}{y}(x)=\hat{d}(x)\overset{\circ}{y}(x),\quad x\in\omega_h,$$

$$\hat{d}(x)=\sum_{\alpha=1}^{2}\left(\frac{a_\alpha^{+1\alpha}}{h_\alpha\sqrt{b_\alpha}}+\frac{\left|a_{\alpha x_\alpha}-h_\alpha\frac{q}{2}\right|}{2h_\alpha}\right)\frac{1}{c_\alpha+\sqrt{b_\alpha}},$$

$$\hat{R}_1\overset{\circ}{y}=\sum_{\alpha=1}^{2}\left(\frac{a_\alpha}{h_\alpha}\overset{\circ}{y}_{\bar{x}_\alpha}+\left(\frac{a_{\alpha x_\alpha}}{2h_\alpha}+\frac{q}{4}\right)\overset{\circ}{y}\right),\quad x\in\omega_h,$$

$$\hat{R}_2\overset{\circ}{y}=-\sum_{\alpha=1}^{2}\left(\frac{a_\alpha^{+1\alpha}}{h_\alpha}\overset{\circ}{y}_{x_\alpha}+\left(\frac{a_{\alpha x_\alpha}}{2h_\alpha}-\frac{q}{4}\right)\overset{\circ}{y}\right),\quad x\in\omega_h.$$

Оскільки тоді $\hat{A}=\hat{R}_1+\hat{R}_2$ і $\hat{R}_1=\hat{R}_2^*$, то нерівність $\hat{c}_1\hat{R}\leqslant A\leqslant\hat{c}_2\hat{R}$, $\hat{R}=\hat{R}_1+\hat{R}_2$ виконується при $\hat{c}_1=\hat{c}_2=1$. Коефіцієнти b_α, c_α, $\alpha=1,\ 2$, визначаються формулами

$$b_\alpha(x_\beta)=\max_{x_\alpha\in\omega_\alpha}v^\alpha(x),$$

$$c_\alpha(x_\beta)=\max_{x_\alpha\in\omega_\alpha}w^\alpha(x),\quad x_\beta\in\omega_\beta,\quad \alpha=1,\ 2,\quad \beta=3-\alpha,$$

де (див. (44), (39)) v^α, w^α — розв'язок задач

$$(a_\alpha v^\alpha_{\bar{x}_\alpha})_{x_\alpha}-\frac{q}{2}v^\alpha=-\frac{a_\alpha^{+1\alpha}}{h_\alpha^2},\quad x_\alpha\in[h_\alpha,\ l_\alpha-h_\alpha],$$

$$v^\alpha=0,\ x_\alpha=0,\ l_\alpha,\quad x_\beta\in\omega_\beta,$$

$$(a_\alpha w^\alpha_{\bar{x}_\alpha})_{x_\alpha}-\frac{q}{2}w^\alpha=-\frac{|a_{\alpha x_\alpha}-h_\alpha q/2|}{2h_\alpha},$$

$$x_\alpha\in[h_\alpha,\ l_\alpha-h_\alpha],\quad w^\alpha=0,\ x_\alpha=0,\ l_\alpha,\quad x_\beta\in\omega_\beta,\quad \beta=3-\alpha,\quad \alpha=1,\ 2.$$

Аналогічно попередньому для такого оператора $\hat{D}$ можна дістати такі константи в нерівностях (61):

$$\hat{\Delta} = 4 \max_{\alpha=1,2} \{\max (c_\alpha + \sqrt{b_\alpha})^2\}, \quad \beta = 3 - \alpha, \quad \hat{\delta} = 1.$$

Розрахунки показують, що при такому виборі $\hat{D}$, $\hat{R}_\alpha$, $\alpha = 1, 2$, кількість ітерацій ПТМ слабо залежить від чисел c_2/c_1 та c.

5.4.4. Багатосітковий метод. Розглянемо модельну задачу

$$-u''(x) = f(x), \quad x \in (0, 1), \quad u(0) = u(1) = 0, \tag{65}$$

яку на сітці $\bar{\omega}_h = \{x_i \equiv x_i^{(1)} = ih : i = \overline{0, 2N}, \ h = 1/(2N)\}$ замінимо скінченнорізницевою задачею в матричній формі

$$A_h u_h = f_h, \tag{66}$$

де $u_h = (u_h(h), u_h(2h), \ldots, u_h(1-h))^T$, $f_h = (f(h), f(2h), \ldots, f(1-h))^T$ — $(2N-1)$-вимірні вектори; A_h — матриця розмірності $(2N-1) \times (2N-1)$:

$$A_h = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & 0 & 2 & -1 \\ 0 & 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{bmatrix}.$$

Застосуємо для розв'язування (66) неявний метод простої ітерації (метод Якобі)

$$u_h^{(j+1)} = u_h^{(j)} - \bar{\omega} D_h^{-1} (A_h u_h^{(j)} - f_h), \quad 0 < \bar{\omega} \leqslant 1,$$

де $D_h = \operatorname{diag} A_h$, $A_h = D_h - P_h$. Враховуючи, що діагональні елементи матриці D_h^{-1} дорівнюють $h^2/2$, покладемо $\omega = \bar{\omega}/2$ і дістанемо

$$u_h^{(j+1)} = u_h^{(j)} - \omega h^2 (A_h u_h^{(j)} - f_h), \quad 0 < \omega \leqslant \frac{1}{2},$$

або

$$u_h^{(j+1)} = B_h u_h^{(j)} + g_h,$$
$$B_h = B_h(\omega) = I - \omega h^2 A_h, \quad g_h = \omega h^2 f_h. \tag{67}$$

Для похибки $v_h^{(j)} = u_h^{(j)} - u_h^*$, де u_h^* — точний розв'язок (66), маємо задачу

$$v_h^{(j+1)} = B_h v_h^{(j)}.$$

Неважко знайти (див. п. 1.4.4 з ч. 1), що власними значеннями матриці B_h є

$$\lambda_{i,h}(\omega) = 1 - 4\omega \sin^2\left(\frac{i\pi h}{2}\right), \quad i = 1, \overline{2N-1},$$

і їм відповідають власні вектори

$$w_h^i = \{\sqrt{2h} \sin(i\pi h), \sqrt{2h} \sin(2i\pi h), \ldots$$
$$\ldots, \sqrt{2h} \sin((2N-1) i\pi h)\}. \tag{67'}$$

Неважко помітити, що спектральний радіус матриці B_h (модуль максимального власного значення) є

$$\rho_\omega(B_h) = \left|1 - 4\omega \sin^2 \frac{\pi h}{2}\right| = 1 - \omega\pi^2 h^2 + O(h^4),$$

і він визначає швидкість збіжності процесу (67). Для малих h величина $\rho_\omega(B_h)$ близька до 1 і швидкість збіжності у цьому випадку невисока. Із збільшенням h зростає і швидкість збіжності ітераційного процесу (67). Розкладемо похибку $v_h^{(0)} = u_h^{(0)} - u_h^*$ за власними векторами w_h^i:

$$v_h^{(0)} = \sum_{i=1}^{2N-1} \alpha_{i,h} w_h^i.$$

Для похибки j-ї ітерації маємо

$$v_h^{(j)} = \sum_{i=1}^{2N-1} \alpha_{i,h} (B_h)^j w_h^i = \sum_{i=1}^{2N-1} \alpha_{i,h} (\lambda_{i,h}(\omega))^j w_k^i = \sum_{i=1}^{2N-1} \beta_{i,j,h}(\omega) w_h^i,$$

$$\beta_{i,j,h}(\omega) = (\lambda_{i,h}(\omega))^j \alpha_{i,h} = \left(1 - 4\omega \sin^2\left(\frac{i\pi h}{2}\right)\right)^j \alpha_{i,h}.$$

Для малих i (для низьких частот) маємо $\alpha_{i,h} \approx \beta_{i,j,h}$, а для великих i (для високих частот) $|\beta_{i,j,h}(\omega)| \ll |\alpha_{i,h}|$. Це означає, що високі частоти в похибці після невеликого числа ітерацій будуть сильно подавлені і не впливатимуть помітно на величину похибки. Це означає, в свою чергу, що похибка стане «згладженою», бо при малих i власні вектори w_h^i як сіткові функції змінюються повільно. Отже, незважаючи на незначну швидкість збіжності, метод Якобі (4) швидко згладжує похибку. Можна довести, що й інші ітераційні методи (наприклад, метод верхньої релаксації, метод Зейделя) мають таку саму властивість.

Перед тим як перейти до багатосіткового методу, розглянемо двосітковий метод.

Двосітковий метод.

Припустимо, що ітераційний метод

$$u_h^{(j+1)} = S_h u_h^{(j)} + g_h \tag{68}$$

має описану вище властивість згладжувати похибку. В методі (67) маємо $S_h = B_h = I - \omega h^2 A_h$, $g_h = \omega h^2 f_h$. Далі вважатимемо, що $\omega = 1/4$.

Нехай u_h^m, $m \geqslant 0$ — деяке наближення до розв'язку задачі

$$A_h u_h = f_h. \tag{69}$$

Вибравши u_h^m за початкове наближення, виконаємо декілька ітерацій за методом (68) і дістанемо деякий вектор $\bar{u}_h^m$, для якого

$$A_h \bar{u}_h^m = f_h + d_h. \tag{70}$$

Оскільки метод (67) згладжує похибку, то вектор похибки $v_h^m = \bar{u}_h^m - u_h^*$ матиме вищу «гладкість», ніж вектор $u_h^m - u_h^*$, де u_h^* — точний розв'язок вихідного рівняння (69). Віднявши (69) від (70), дістанемо

$$A_h v_h^m = d_h. \tag{71}$$

Оскільки функція v_h^m має високу «гладкість», то доцільно рівняння (71) розв'язувати (наближено) на більш грубій сітці, зменшуючи тим самим кількість невідомих. Для визначеності візьмемо сітку $\bar{\omega}_{2h}$ з кроком $2h$, вузлами $x_i^{(2)} = 2ih$, $i = \overline{0, N}$, і розглянемо на цій сітці систему рівнянь

$$A_{2h} v_{2h}^m = d_{2h}, \tag{72}$$

де матриця A_{2h} розмірності $(N-1) \times (N-1)$ має вигляд

$$A_{2h} = \frac{1}{4h^2} \begin{bmatrix} 2 & -1 & 0 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{bmatrix},$$

а v_{2h}^m — вектор розмірності $N-1$, який можна розглядати також і як сіткову функцію

$$v_{2h}^m = (v_{2h}^m(2h),\ v_{2h}^m(4h),\ \dots,\ v_{2h}^m(1-2h))^T.$$

Права частина d_{2h} є деяким «звуженням» сіткової функції d_h на сітку $\bar{\omega}_{2h}$, тобто $d_{2h} = Rd_h$, де R — деякий оператор (матриця) звуження. Цей оператор можна задати, наприклад, рівністю $d_{2h}(x) = d_h(x)$, $x \in \bar{\omega}_{2h}$, або матрицею розмірності $(N-1) \times (2N-1)$

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

Можливий і такий, наприклад, варіант:

$$d_{2h} = \frac{1}{4}(d_h(x-h) + 2d_h(x) + d_h(x+h)), \quad x \in \bar{\omega}_{2h},$$

$$R = \frac{1}{4} \begin{pmatrix} 1 & 2 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 & \dots & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 2 & 1 \end{pmatrix},$$

тобто

$$d_{2h}(2h) = \frac{1}{4}(d_h(h) + 2d_h(2h) + d_h(3h)),$$

$$d_{2h}(4h) = \frac{1}{4}(d_h(3h) + 2d_h(4h) + d_h(h)),$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

$$d_{2h}(1-2h) = \frac{1}{4}(d_h(1-3h) + 2d_h(1-2h) + d_h(1-h)).$$

Припустимо, що ми розв'язали систему (72) і дістали в результаті $v_{2h}^m = A_{2h}^{-1} d_{2h}$. Після цього потрібно сіткову функцію v_{2h}^m продовжити з сітки $\bar{\omega}_{2h}$ на сітку $\bar{\omega}_h$ за допомогою деякого оператора продовження

$$\tilde{v}_h^m = P v_{2h}^m.$$

Одне з можливих продовжень можна задати формулами

$$\tilde{v}_h^m(x) = \begin{cases} v_{2h}^m(x), & x \in \bar{\omega}_{2h}, \\ \frac{1}{2}[v_{2h}^m(x-h) + v_{2h}^m(x+h)], & x \in \bar{\omega}_h \setminus \bar{\omega}_{2h} \end{cases}$$

або матрицею розмірності $(2N-1) \times (N-1)$

$$P = \frac{1}{2} \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 2 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 \\ 0 & 2 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 0 & 2 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

(тут враховано, що $v_{2h}^m(0) = v_{2h}^m(1) = 0$). Отже, $\tilde{v}_h^m$ є деяким наближенням для $v_h^m = \bar{u}_h^m - u_h^*$, і тому $u_h^{m+1} = \bar{u}_h^m - \tilde{v}_h^m$ є наближенням

до точного розв'язку u_h^*. Враховуючи формули $v_{2h}^m = A_{2h}^{-1} d_{2h} = A_{2h}^{-1} R d_h = A_{2h}^{-1} R (A_h \bar{u}_h^m - f_h)$, маємо

$$u_h^{m+1} = \bar{u}_h^m - P A_{2h}^{-1} R (A_h \bar{u}_h^m - f_h). \tag{73}$$

Отже, двосітковий метод складається з таких кроків. Виходячи з початкового наближення u_h^0 для $m = 0, 1, 2, \ldots$ виконати:

1) ν згладжуючих ітерацій

$$u_h^{m(0)} = u_h^m, \quad u_h^{m(j+1)} = u_h^{m(j)} - \frac{h^2}{4} (A_h u_h^{m(j)} - f_h), \quad j = \overline{0, \nu - 1},$$

$$u^{m(\nu)} = \bar{u}_h^m;$$

2) обчислити нев'язку (дефект) $d_h = A_h \bar{u}_h^m - f_h$;

3) звузити d_h на ω_{2h} : $d_{2h} = R d_h$;

4) розв'язати рівняння $A_{2h} v_{2h}^m = d_{2h}$ на сітці $\bar{\omega}_{2h}$;

5) продовжити v_{2h}^m на сітку $\bar{\omega}_h$ і обчислити $(m+1)$-шу ітерацію

$$u_h^{m+1} = \bar{u}_h^m - P v_{2h}^m.$$

Звернемося тепер до швидкості збіжності двосіткового методу. Його можна записати у вигляді двох'ярусної ітераційної схеми

$$u_h^{m+1} = M_h(\nu)\, u_h^m + N_h(\nu)\, f_h. \tag{74}$$

Дійсно, для $j = \overline{0, \nu - 1}$ маємо

$$u_h^{m(j+1)} = \left(I - \frac{h^2}{4} A_h\right) u_h^{m(j)} + \frac{h^2}{4} f_h = S_h u_h^{m(j)} + \frac{h^2}{4} f_h.$$

Звідки методом математичної індукції виводимо

$$\bar{u}_h^m = u_h^{m(\nu)} = S_h^\nu u_h^m + \frac{h^2}{4} \left(\sum_{i=0}^{\nu-1} S_h^i\right) f_h. \tag{74'}$$

Спектральний радіус S_h менший за одиницю, тому $S_h - I$ — невироджена матриця. Звідси витікає зображення

$$\sum_{i=0}^{\nu-1} S_h^i = (S_h - I)^{-1} (S_h^\nu - I),$$

тобто з (74') маємо

$$\bar{u}_h^m = S_h^\nu u_h^m + \frac{h^2}{4} (S_h - I)^{-1} (S_h^\nu - I) f_h.$$

Підставляючи цей вираз у (73), матимемо

$$u_h^{m+1} = (I - P A_{2h}^{-1} R A_h) \left\{ S_h^\nu u_h^m + \frac{h^2}{4} (S_h - I)^{-1} (S_h^\nu - I) f_h \right\} + P A_{2h}^{-1} R f_h =$$

$$= (I - P A_{2h}^{-1} R A_h) S_h^\nu u_h^m +$$

$$+ \left\{ \frac{h^2}{4} (I - P A_{2h}^{-1} R A_h) (S_h - I)^{-1} (S_h^\nu - I) + P A_{2h}^{-1} R \right\} f_h.$$

Порівнюючи це з (74), дістаємо

$$M_h(\nu) = (I - P A_{2h}^{-1} R A_h) S_h^\nu,$$

$$N_h(\nu) = \frac{h^2}{4} (I - P A_{2h}^{-1} R A_h) (S_h - I)^{-1} (S_h^\nu - I) + P A_{2h}^{-1} R. \tag{75}$$

Щоб оцінити швидкість збіжності методу (74), (75), потрібно оцінити спектральний радіус або деяку норму матриці $M_h(\nu)$. Таку оцінку дає наступна теорема.

Теорема. *Для спектрального радіуса* $\rho(M_h(\nu))$ *та для спектральної норми* $\| M_h(\nu) \|_2$ *мають місце оцінки*

$$\rho(M_h(\nu)) = \rho_\nu - |O(h^2)|, \quad \| M_h(\nu) \|_2 = \sigma_\nu - |O(h^2)|,$$

де $\rho_\nu = \max\limits_{0 \leqslant r \leqslant 1/2} [r(1-r)^\nu + (1-r) r^\nu] < 1$, $\sigma_\nu = \max\limits_{0 \leqslant s \leqslant 1/2} \{2 [s^2 (1-s)^{2\nu} + (1-s)^2 s^{2\nu}]^2\} < 1$ — *сталі, що залежать лише від* ν *і не залежать від* h.

Доведення. Зауважимо, по-перше, що $S_h = B_h = I - \omega h^2 A_h = I - \frac{h^2}{4} A_h$. Матриці B_h та A_h мають спільну систему власних векторів (67'). Неважко знайти, що для i-го власного вектора

$$\| w_h^i \|_2^2 = 2h \sum_{\nu=1}^{2N-1} \sin^2(\nu i \pi h) = 2hN = 1,$$

причому

$$(w_h^i, w_h^k) = 2h \sum_{\nu=1}^{2N-1} \sin(\nu i \pi h) \sin(\nu k \pi h) = 0, \quad i \neq k,$$

тобто вектори w_h^i утворюють ортонормальну систему. Те саме можна сказати про вектори w_{2h}^i на сітці $\bar{\omega}_{2h}$. Введемо унітарну матрицю розмірності $(2N-1) \times (2N-1)$:

$$W_h = [w_h^1, w_h^{2N-1}, w_h^2, w_h^{2N-2}, \ldots, w_h^{N-1}, w_h^{N+1}, w_h^N],$$

стовпчиками якої є ортонормальні вектори w_h^i. Нехай

$$W_{2h} = [w_{2h}^1, w_{2h}^2, \ldots, w_{2h}^{N-1}]$$

— матриця, стовпчиками якої є власні вектори матриці A_{2h}. Покладемо

$$\bar{A}_h = W_h^* A_h W_h, \quad \bar{A}_{2h} = W_{2h}^* A_{2h} W_{2h},$$

$$\bar{B}_h = W_h^* B_h W_h, \quad \bar{P} = W_h^* P W_{2h}, \quad \bar{R} = W_{2h}^* R W_h,$$

де $*$ означає знак транспонування. Неважко перевірити, що $\overline{B}_h$ є блочно-діагональною матрицею з блоками

$$\overline{B}_h^{(\mu)} = \begin{pmatrix} c_\mu^2 & 0 \\ 0 & s_\mu^2 \end{pmatrix}, \quad \mu = \overline{1,\ N-1}, \quad \overline{B}_h^{(N)} = \frac{1}{2},$$

$\overline{A}_h$ — блочно-діагональною матрицею з блоками

$$\overline{A}_h^{(\mu)} = \frac{4}{h^2}\begin{pmatrix} s_\mu^2 & 0 \\ 0 & c_\mu^2 \end{pmatrix} = \frac{4}{h^2}(I - \overline{B}_h^{(\mu)}), \quad \mu = \overline{1,\ N-1}, \quad \overline{A}_h^{(N)} = \frac{2}{h^2},$$

$\overline{A}_{2h}$ — діагональною матрицею з елементами

$$\overline{A}_{2h}^{(\mu)} = \frac{4}{h^2} s_\mu^2 c_\mu^2, \quad \mu = \overline{1,\ N-1},$$

де $s_\mu = \sin\frac{\mu\pi h}{2}$, $c_\mu = \cos\frac{\mu\pi h}{2}$. Матриці «продовження» розмірності $(2N-1)\times(N-1)$ та звуження розмірності $(N-1)\times(2N-1)$ мають вигляд

$$\overline{P} = \begin{bmatrix} \overline{P}^{(1)} & & & \\ & \ddots & & \\ & & \overline{P}^{(N-1)} & \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \dots\ 0 \end{bmatrix}, \quad \overline{R} = \begin{bmatrix} \overline{R}^{(1)} & & & \dots\ 0 \\ & \ddots & & \dots\ \dots \\ & & \overline{R}^{(N-1)} & \dots\ 0 \\ & & & \dots\ \dots \\ & & & \dots\ 0 \end{bmatrix},$$

де $\overline{P}^{(\mu)} = \sqrt{2}\begin{pmatrix} c_\mu^2 \\ -s_\mu^2 \end{pmatrix}$, $\overline{R}^{(\mu)} = \frac{1}{\sqrt{2}}(c_\mu^2, -s_\mu^2)$, $\mu = \overline{1,\ N-1}$. Матриця $\overline{M}_h(\nu) = \overline{W}_h^* M_h(\nu)\overline{W}_h$ є також блочно-діагональною з блоками

$$\overline{M}_h^{(\mu)}(\nu) = \begin{pmatrix} s_\mu^2 & c_\mu^2 \\ s_\mu^2 & c_\mu^2 \end{pmatrix}\begin{pmatrix} c_\mu^2 & 0 \\ 0 & s_\mu^2 \end{pmatrix}^\nu, \quad \mu = \overline{1,\ N-1}, \quad \overline{M}_h^{(N)}(\nu) = 0.$$

Далі маємо

$$\rho(M_h(\nu)) = \rho(\overline{M}_h(\nu)) = \max_\mu\{\rho(\overline{M}_h^\mu(\nu))\}, \quad \mu = \overline{1,\ N},$$

$$\|M_h(\nu)\|_2 = \|\overline{M}_h(\nu)\|_2 = \max_\mu\{\|\overline{M}_h^{(\mu)}(\nu)\|_2\}, \quad \mu = \overline{1,\ N}.$$

Обчислюючи власні значення матриць $\overline{M}_h^{(\mu)}(\nu)$ та $(\overline{M}_h^{(\mu)}(\nu))^* \overline{M}_h^{(\mu)}(\nu)$, дістаємо твердження теореми.

Найважливішим результатом цієї теореми є незалежні від h оцінки

$$\rho(M_h(\nu)) \leqslant \rho_\nu, \quad \|M_h(\nu)\|_2 \leqslant \sigma_\nu,$$

які в силу нерівностей $\rho_\nu < 1$, $\sigma_\nu < 1$ означають швидку і незалежну від h (тобто від кількості рівнянь) збіжність двосіткового методу. Залежність ρ_ν та σ_ν від ν дає наступна таблиця:

ν	1	2	3	4	5
ρ_ν	0,5	0,25	0,125	0,0832	0,0671
σ_ν	0,5	0,25	0,150	0,1159	0,0947

Ця таблиця показує дуже швидку збіжність двосіткового методу. Так, для $\mu = 10$ ітерацій при $\nu = 2$ згладжуючих ітераціях маємо оцінку

$$\|u_h^{10} - u_h^*\|_2 \leqslant 10^{-6}\|u_h^0 - u_h^*\|_2,$$

а щоб такої самої точності досягти в методі верхньої релаксації для $h = 0{,}01$ (99 рівнянь), потрібно 200 ітерацій, для $h = 0{,}001$ (999 рівнянь) 3447 ітерацій.

Зауважимо, що записавши метод верхньої релаксації у вигляді

$$u_h^{m+1} = M_h u_h^m + N_h f_h,$$

для спектрального радіуса матриці M_h при оптимальному значенні параметра релаксації можна довести оцінку $\rho(M_h) = 1 - |O(h)|$, а для розглянутого вище методу простої ітерації $\rho(M_h) = 1 - |O(h^2)|$, що і обумовлює їхню повільну збіжність (тим повільнішу, чим менше h).

Ітераційний багатосітковий метод. У двосітковому методі система рівнянь (72) має бути розв'язана деяким прямим методом. Однак ця система може бути досить високої розмірності, тому доцільно розв'язати її також двосітковим методом, перейшовши до сітки з кроком, наприклад, $4h$. У результаті прийдемо до трисіткового методу. Продовживши таку процедуру, дістанемо багатосітковий метод. Щоб записати цей метод, розглянемо деяку послідовність сіток $\overline{\omega}_{h_l}$ з кроками сітки $h_0 > h_1 > h_2 > \dots > h_l > \dots$, наприклад, $h_l = 2^{-(l+1)}$, $l = 0, 1, \dots$ $\dots, N_l h_l = 1$. Для спрощення позначень сітку ω_{h_l} будемо записувати як ω_l. Крім граничних точок, в яких розв'язок задачі відомий, сітка ω_0 складається з однієї точки 1/2, а сітка ω_l включає внутрішні вузли вигляду sh_l, $s = 1, 2, \dots, 2^{l+1} - 1$. На сітці ω_l маємо аналогічно (69) рівняння

$$A_l u_l = f_l,$$

де $u_l = [u_l(h_l), u_l(2h_l), \dots, u_l(1-h_l)]^T$ — невідомий вектор (сіткова

функція), $f_l = [f(h_l),\ f(2h_l),\ \ldots,\ f(1-h_l)]^T$ — відомий вектор,

$$A_l = \frac{1}{h_l^2}\begin{bmatrix} 2 & -1 & 0 & 0 \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 \ldots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 \ldots & 0 & -1 & 2 \end{bmatrix}$$

матриця розмірності $(N_l - 1) \times (N_l - 1)$, $N_l = 2^{l+1}$. У двосітковому методі ми мали $u_l^{m+1} = \bar{u}_l^m - P_l v_{l-1}^m$. Далі можна було б покласти $\tilde{u}_l = \bar{u}_l^m - P_l v_{l-1}^m$, потім $\tilde{u}_l$ «покращити» за допомогою декількох згладжуючих ітерацій і результат позначити u_l^{m+1}. Цю процедуру в багатосітковому методі можна виконувати на кожній сітці. Далі вважатимемо, що при переході на більш грубу сітку кожен раз виконується ν_1 згладжуючих ітерацій, а при переході на дрібнішу сітку — ν_2 згладжуючих ітерацій.

Нехай u_l^m, $m \geqslant 0$ відоме. Тоді багатосітковий метод складається з таких кроків.

1. Виконати ν_1 згладжуючих ітерацій для $A_l u_l = f_l$, починаючи з $u_l^{(0)} = u_l^m$. У результаті дістанемо $\bar{u}_l$.

2. Для $i = 0, 1, \ldots, L-1$ обчислити $d_{l-i} = A_{l-i}\bar{u}_{l-i} - f_{l-i}$, $l-i \to l-i-1 = R_{l-i}d_{l-i}$. Для $i < L-1$ виконати ν_1 згладжуючих ітерацій для $A_{l-i-1}u_{l-i-1} = f_{l-i-1}$, виходячи з $u_{l-i-1}^{(0)}$. У результаті маємо $\bar{u}_{l-i-1}$.

3. Розв'язати систему $A_{l-L}u_{l-L} = f_{l-L}$ на сітці ω_{l-L}. Результат позначимо $\bar{\bar{u}}_{l-L}$.

4. Покласти для $k = 1, 2, \ldots, L$

$$\tilde{u}_{l-L+k} = \bar{u}_{l-L+k} - P_{l-L+k}\bar{\bar{u}}_{l-L+k-1}$$

і виконати ν_2 згладжуючих ітерацій для $A_{l-L+k}u_{l-L+k} = f_{l-L+k}$, виходячи з $u_{l-L+k}^{(0)} = \tilde{u}_{l-L+k}$. Позначимо результат через $\bar{\bar{u}}_{l-L+k}$.

5. Покласти $u_l^{m+1} = \bar{\bar{u}}_l$.

Алгоритм починає роботу з найдрібнішої сітки, поступово переходить до найгрубішої сітки і повертається знову до найдрібнішої. Кажуть при цьому, що він виконує V-цикл (рис. 27). Зрозуміло, що можливі інші варіанти переходу від сітки до сітки, наприклад так званий W-цикл (рис. 28).

Можна починати ітерації не з найдрібнішої сітки, а з деякої іншої, наприклад з найгрубішої.

На таких самих ідеях будуються багатосіткові методи для багатовимірних апроксимацій рівнянь з частинними похідними, хоча тех-

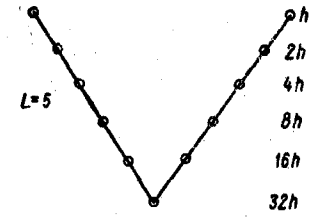


Рис. 27

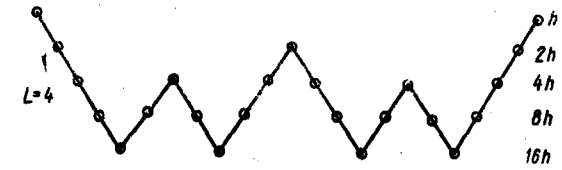


Рис. 28

нічна сторона в цьому випадку дещо складніша (так, має значення нумерація вузлів, складнішу структуру мають оператори звуження і продовження і т. п.).

Враховуючи, що матриця $M_h(\nu)$ в (74) має спеціальну структуру і є добутком розріджених матриць з кількістю ненульових елементів $O(2N) = O(1/h)$, дістаємо, що кожна ітерація виконується за $O(N) = O(1/h)$ арифметичних операцій, а оскільки кількість ітерацій не залежить від h, то загальна кількість арифметичних операцій в двосітковому методі є величиною того самого порядку, що і кількість рівнянь. Те саме можна довести і для багатосіткового методу. Для порівняння нагадаємо, що метод Гаусса для системи з n рівнянь вимагає $O(n^3)$ арифметичних операцій, а метод верхньої релаксації для розрідженої системи (69) — $O(n^2) = O(1/h^2)$. Багатосіткові методи можна застосувати і для більш складних сіткових апроксимацій і область його застосування весь час розширюється, хоча універсальним його назвати не можна. Одним з основних недоліків цього методу є те, що для його застосування важливою є вимога рівномірності сітки.

ГЛАВА 6

МЕТОДИ РОЗВ'ЯЗУВАННЯ ПОЧАТКОВО-КРАЙОВИХ ЗАДАЧ ДЛЯ НЕСТАЦІОНАРНИХ РІВНЯНЬ

6.1. Постановка та загальні ідеї апроксимації задач параболічного типу

Рівняння параболічного типу можна записати у вигляді

$$\frac{\partial u}{\partial t} = Lu + f(x, t), \quad (x, t) \in Q_T,$$

де $u = u(x, t)$, $x \in \Omega$, $t \in [0; T]$; Ω — деяка область евклідового простору $\mathbb{R}^n$ з границею $\partial\Omega$; $Q_T = \Omega \times (0, T]$ — циліндр, L — деякий еліптичний оператор. Щоб виділити єдиний розв'язок цього рівняння, крім

граничних (крайових) умов $lu = u(t)$, $x \in \partial\Omega$ для оператора L (l — деякий граничний оператор), очевидно, потрібно накласти ще одну умову, наприклад задати функцію і при деякому $t = t_0$. Задача вигляду

$$\frac{\partial u(x, t)}{\partial t} = Lu(x, t) + f(x, t), \quad (x, t) \in Q_T,$$

$$lu(x, t) = u_1(t), \quad x \in \partial\Omega, \tag{1}$$

$$u(x, t_0) = u_0(x), \quad x \in \Omega,$$

де $f(x, t)$, $u_1(t)$, $u_0(t)$ — задані функції, називається *крайовою (граничною) задачею для рівняння параболічного типу* або іноді *початково-крайовою задачею* (умова $u(t_0, x) = u_0(x)$ називається *початковою умовою*). Ми спеціально виділили змінну t, тому що на практиці в задачах виду (1) вона позначає час. Оскільки розв'язок задачі (1) залежить від часу, то такі задачі називають також *нестаціонарними*. Наприклад, процес поширення теплоти в одновимірному стержні $0 < x < l$ описується рівнянням теплопровідності

$$c\rho \frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(k \frac{\partial u}{\partial x}\right) + f_0(x, t), \tag{2}$$

де $u = u(x, t)$ — температура в точці x стержня в момент t; c — теплоємність одиниці маси; ρ — щільність; $c\rho$ — теплоємність одиниці довжини; k — коефіцієнт теплопровідності; f_0 — щільність теплових джерел (взагалі k, c, ρ, f_0 можуть залежати не лише від x, t, але й від температури u — тоді рівняння (2) називається *квазілінійним рівнянням теплопровідності;* якщо ж ці величини залежать ще від похідних $\partial u/\partial x$, $\partial^2 u/\partial x^2$, то рівняння (2) є повністю нелінійним). Якщо k, c, ρ — сталі, то (2) записується у вигляді

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2} + f, \tag{3}$$

де $f = f_0/(c\rho)$, $a^2 = k/(c\rho)$ — коефіцієнт температуропровідності. Якщо ввести нові змінні $x_1 = x/l$, $t_1 = a^2 t/l^2$, $f_1 = l^2 f/a^2$, то замість (3) дістанемо більш просте рівняння

$$\frac{\partial u}{\partial t_1} = \frac{\partial^2 u}{\partial x_1^2} + f_1, \quad x_1 \in (0, 1), \ t \in (0, T_1]. \tag{4}$$

Щоб визначити єдиний розв'язок рівняння (4) $u(x, t) \in C^1(I, C^2(\Omega)) \cap \cap C(\bar{\Omega})$, можна поставити таку крайову (початково-крайову) задачу (за виглядом крайових умов та рівняння вона також називається *задачею Діріхле для рівняння теплопровідності*)

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad x \in \Omega, \ t \in I^-,$$

$$u(x, 0) = u_0(x), \quad x \in \bar{\Omega}, \tag{5}$$

$$u(0, t) = u_1(t), \quad u(1, t) = u_2(t), \quad t \in I,$$

де $\bar{\Omega} = [0, 1]$; $\Omega = (0, 1)$; $I = [0, T]$; $I^- = (0, T]$, $C^1(I, C^2(\Omega)) \cap \cap C(\bar{\Omega})$ означає простір функцій, які мають неперервні похідні в I по t, неперервні другі похідні по x в Ω і є неперервними по x в $\bar{\Omega}$; u_0, u_1, u_2 — задані неперервні функції.

Розглянемо простір $\mathring{H} = C^2(\Omega) \cap C(\bar{\Omega})$ функцій від x, які дорівнюють нулю при $x = 0, 1$ зі скалярним добутком

$$(v, w) = \int_\Omega v(x)\, w(x)\, dx.$$

Покладемо також в (5) для зручності $u_1 = u_2 = 0$. Тоді при кожному t функція $u(x, t)$ буде елементом простору $\mathring{H}$, який позначатимемо $u(t) \in \mathring{H}$. Помноживши скалярно рівняння (5) на довільну функцію $v \in \mathring{H}$ і проінтегрувавши частинами, дістанемо (при кожному значенні параметра t)

$$\left(\frac{du(t)}{dt}, v\right)_{\mathring{H}} + \left(\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}\right)_{\mathring{H}} - (f(t), v)_{\mathring{H}} = 0$$

$$\forall t \in I, \quad \forall v \in \mathring{H}, u(0) = u_0. \tag{6}$$

У цій тотожності, як бачимо, по-перше, враховано крайові умови і, по-друге, містяться лише перші похідні від шуканої функції u, причому вони не обов'язково мають бути неперервними, бо досить, наприклад, інтегровності з квадратом, щоб тотожність (6) мала сенс. Тому (6) можна прийняти за означення узагальненого (тобто з меншою гладкістю) розв'язку задачі (5), наприклад з простору $\mathring{W}_2^1(0, 1)$. Ввівши певні додаткові математичні поняття, можна довести, що така постановка задачі коректна, тобто її розв'язок існує і єдиний. Проте і без цього тотожність (6) разом із (5) може бути використана для побудови наближених методів розв'язування задачі (5).

У нестаціонарному випадку можна і по-іншому узагальнити поняття розв'язку задачі (5). Для цього введемо простір $L(I, \mathring{H})$ функцій $u(t) : I \to \mathring{H}$, $u(0) = u_0$ зі скалярним добутком

$$(u, v) = \int_0^T (u, v)_{\mathring{H}}\, dt,$$

наприклад, для введеного вище простору $\mathring{H}$

$$(u, v) = \int_0^T \int_0^1 u(x, t)\, v(x, t)\, dx dt.$$

Тоді, інтегруючи рівність (6) по t, матимемо

$$\left(\frac{du}{dt}, v\right)+\left(\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}\right)-(f, v)=0 \quad \forall v \in L(I, \mathring{H}). \tag{7}$$

У певному розумінні (7) також є узагальненою постановкою задачі (5) (наприклад, для розв'язку з простору $L(I, \mathring{W}_2^1(0,1))$) і може використовуватись для побудови наближених методів.

Аналогічно можна розглядати різні постановки крайових задач для двовимірного рівняння теплопровідності. Наприклад, перша крайова задача має вигляд

$$\frac{\partial u}{\partial t}=Lu+f(x, t), \quad (x, t) \in Q_T=\Omega \times (0, T],$$

$$u(x, 0)=u_0(x), \quad x \in \bar{\Omega}, \tag{8}$$

$$u(x, t)=u_1(x, t), \quad x \in \partial\Omega,$$

де

$$Lu=L_1u+L_2u, \quad L_1u=\frac{\partial}{\partial x_1}\left(k_1(x, t)\frac{\partial u}{\partial x_1}\right),$$

$$L_2u=\frac{\partial}{\partial x_2}\left(k_2(x, t)\frac{\partial u}{\partial x_2}\right), \quad 0 \leqslant c_1 \leqslant k_\alpha(x, t) \leqslant c_2.$$

Постановка цієї задачі у вигляді (6) при $u_1=0$ записується так

$$\int_\Omega \frac{\partial u(x, t)}{\partial t} v(x)\,dx+\sum_{\alpha=1}^{2}\int_\Omega k_\alpha(x, t)\frac{\partial u(x, t)}{\partial x_1}\frac{\partial v(x)}{\partial x_2}\,dx-$$

$$-\int_\Omega f(x, t)\,v(x)\,dx=0, \quad u(0)=u_0(x). \tag{9}$$

Для апроксимації задач виду (1) можна застосувати метод сіток; причому можна дискретизувати як всі змінні, так і лише частину з них. Зауважимо, що в нестаціонарних задачах змінна t має певну особливість, яка пов'язана з тим фактом, що стан процесу, який описується нестаціонарним рівнянням, у даний момент залежить від його стану в минулому і не залежить від його стану в наступні моменти часу. Тому в нестаціонарних сіткових задачах значення функції на j-му ярусі (по t) визначається лише через значення функції на попередніх ярусах і не залежить від значення функції на наступних (у математичному плані це пов'язано, очевидно, з тим, що по t в задачі (1) задається лише початкова умова на відміну від крайової задачі, де умови задаються на обох кінцях одновимірного інтервалу).

Введемо сітки

$$\bar{\omega}_\tau=\{t_j=j\tau : j=\overline{0, L}, \ \tau=T/L\},$$

$$\omega_\tau=\{t_j=j\tau : j=\overline{1, L}; \ \tau=T/L\},$$

які покривають відповідно відрізок $[0, T]$ та інтервал $(0, T]$, а також деякі сітки $\bar{\omega}_h$, ω_h, які покривають області $\bar{\Omega}$ та Ω. Тоді задачу (1) можна апроксимувати такими способами.

Залишивши просторові змінні неперервними, а похідну $\frac{du}{dt}$ апроксимувавши якимось виразом на сітці $\bar{\omega}_\tau$, дістанемо схему

$$B_\tau v(x)=Lv(x)+F(x),$$

$$lv(x)=v_1, \quad x \in \partial\Omega, \tag{10}$$

$$v(x, t_0)=u_0(x), \quad x \in \bar{\Omega},$$

де B_τ — деякий сітковий оператор; $v(x)=\{v(x, t_0), v(x, t_1), \ldots, v(x, t_L)\}^T$ — вектор, який наближає $u(x, t)$ в точках сітки $\bar{\omega}_\tau$; $lv(x)=\{lv(x, t_0), \ldots, lv(x, t_L)\}^T$; $v_1=\{u_1(x, t_0), \ldots, u_1(x, t_L)\}^T$; $F(x)=\{f(x, t_0), \ldots, f(x, t_L)\}^T$.

Схема виду (10) називається *схемою методу прямих* або *схемою Роте*. Ця схема називається також *поперечною схемою методу прямих*. Вона є системою еліптичних рівнянь, і, таким чином, метод Роте зводить розв'язок нестаціонарної задачі до розв'язку системи еліптичних рівнянь (за умови, що задача (10) поставлена коректно).

Приклад 1. Побудувати схему Роте для задачі (5).

Розв'язання. Введемо вектори

$$v(x)=\{v(x, t_0), h(x, t_1), \ldots, v(x, t_L)\}^T=\{v(x, t)\}_{t \in \bar{\omega}_\tau}=$$

$$=\{v_i(x)\}_{i=\overline{0,L}};$$

$$F(x)=\{f(x, t)\}_{t \in \bar{\omega}_\tau}, \quad v_1=\{u_1(t)\}_{t \in \bar{\omega}_\tau}, \quad v_2=\{u_2(t)\}_{t=\bar{\omega}_\tau}.$$

Перша компонента вектора $v(x, t_0)=v(x, 0)=u_0(x)$, $x \in \bar{\Omega}$ відома з початкової умови. На часовому ярусі t_i апроксимуємо похідну по часу $\partial u/\partial t$ таким різницевим виразом

$$\frac{\partial u(x, t_i)}{\partial t} \sim \frac{v(x, t_i)-v(x, t_{i-1})}{\tau}$$

(тим самим ми задали оператор B_τ у схемі (10)). Для апроксимації $\frac{\partial^2 u}{\partial x^2}$ маємо різні можливості, наприклад

$$\frac{\partial^2 u(x, t_i)}{\partial x^2} \sim \frac{\partial^2 v(x, t_i)}{\partial x^2}$$

або

$$\frac{\partial^2 u(x, t_i)}{\partial x^2} \sim \frac{\partial^2 v(x, t_{i-1})}{\partial x^2}.$$

У першому випадку дістаємо таку систему Роте (двох'ярусну):

$$\frac{v_i-v_{i-1}}{\tau}=\frac{d^2 v_i(x)}{dx^2}+f_i(x), \quad x \in (0, 1), \quad i=\overline{1, L},$$

$$v_i(0) = u_{1i}, \quad v_i(1) = u_{2i}, \tag{11}$$
$$v_0(x) = u_0(x), \quad x \in [0, 1],$$

де $f_i(x) = f(x, t_i)$, $u_{1i} = u_1(t_i)$, $u_{2i} = u_2(t_i)$.

Маємо $v_0(x) = u_0(x)$, звідки після розв'язування еліптичної крайової задачі

$$-\frac{d^2 v_1(x)}{dx^2} + \frac{1}{\tau} v_1(x) = \frac{1}{\tau} v_0(x) + f_1(x), \tag{12}$$
$$v_1(0) = u_{11}, \quad v_1(1) = u_{21},$$

знаходимо $v_1(x)$, потім аналогічно $v_2(x)$ і т. д. Схема (11) називається *неявною схемою Роте*, оскільки еліптична частина рівняння береться на верхньому ярусі по часу і для отримання v треба розв'язувати еліптичні задачі виду (12).

Скориставшись іншою апроксимацією похідної по часу, дістаємо таку схему методу прямих (методу Роте):

$$\frac{v_i - v_{i-1}}{\tau} = \frac{d^2 v_{i-1}(x)}{dx^2} + f_i(x), \quad x \in (0, 1), \quad i = \overline{1, L},$$
$$v_i(0) = u_{1i}, \quad v_i(1) = u_{2i}, \tag{13}$$
$$v_0(x) = u_0(x), \quad x \in [0, 1].$$

Знаючи $v_0(x) = u_0(x)$, знаходимо

$$v_1(x) = v_0(x) + \tau \frac{d^2 v_0(x)}{dx^2} + \tau f_1(x), \tag{14}$$

потім аналогічно $v_2(x)$ і т. д. Оскільки еліптична частина в цій схемі апроксимується на попередньому часовому ярусі, то це дає змогу вести розрахунки за явними формулами виду (14), тому схема (13) називається *явною*. Зауважимо, що можливі не лише двох'ярусні апроксимації, але й багатоярусні (по часу).

Задачу (1) можна апроксимувати також за допомогою іншої схеми методу прямих, яка називається ще *поздовжньою схемою методу прямих*. При цьому змінна t залишається неперервною, а еліптичний оператор, який задається диференціальним виразом L та граничними умовами на сітці $\bar{\omega}_h$, апроксимується деяким сітковим оператором A_h (методи побудови таких апроксимацій ми розглянули в гл. 4, 5). У результаті задача (1) апроксимується схемою

$$\frac{\partial v(t)}{\partial t} = A_h v(t) + F(t),$$
$$v(t_0) = v_0, \tag{15}$$

де $v(t) = \{v(x, t), x \in \omega_h\}$, $v(x, t)$ наближення для $u(x, t)$ у точці сітки $x \in \omega_h$, $F(t) = \{f(x, t), x \in \omega_h\}$, $v_0 = \{u_0(x), x \in \bar{\omega}_h\}$. Таким чином, поздовжня схема методу прямих зводить параболічну задачу виду (1) до задачі Коші для системи звичайних диференціальних рівнянь виду (15).

Приклад 2. Побудувати поздовжню схему методу для задачі (5).

Розв'язання. На інтервалі (0; 1) введемо сітку $\omega_h = \{x_i = ih : i = \overline{1, N-1}, h = 1/N\}$ і позначимо $v(t) = \{v(x, t), x \in \omega_h\} = \{v_i(t)\}_{i=\overline{1,N-1}}$,

($v(x_0, t) = v(0, t) = v_0(t) = u_1(t)$ та $v(x_N, t) = v(1, t) = v_N(t) = u_2(t)$ відомі). Оператор A_h введемо таким чином:

$$(A_h v)_i = \begin{cases} \dfrac{v_2(t) - 2v_1(t) + u_1(t)}{h^2}, & i = 1, \\ \dfrac{v_{i+1}(t) - 2v_i(t) + v_{i-1}(t)}{h^2}, & i = \overline{1, N-2}, \\ \dfrac{u_2(t) - 2v_{N-1}(t) + v_{N-2}(t)}{h^2}, & i = N-1. \end{cases}$$

Тоді задача (5) відповідно до (15) апроксимується такою задачею Коші:

$$\frac{dv(t)}{dt} = A_h v + F(t), \quad v(0) = v_0 = \{u_0(x)\}_{x \in \omega_h},$$

де $F(t) = \{f(x, t)\}_{x \in \omega_h} = \{f_i(t)\}_{i=\overline{1,N-1}}$ або в скалярному вигляді

$$\frac{dv_1(t)}{dt} = \frac{v_2(t) - 2v_1(t)}{h^2} + f_1(t) + \frac{u_1(t)}{h^2};$$
$$\frac{dv_i(t)}{dt} = \frac{v_{i+1}(t) - 2v_i(t) + v_{i-1}(t)}{h^2} + f_i(t), \quad i = \overline{1, N-2};$$
$$\frac{dv_{N-1}(t)}{dt} = \frac{-2v_{N-1}(t) + v_{N-2}(t)}{h^2} + f_{N-1}(t) + \frac{u_2(t)}{h^2};$$
$$v_i(0) = u_0(ih), \quad i = \overline{1, N-1}.$$

Задачу (1) можна дискретизувати повністю. Для цього циліндр Q_T покривається сіткою $\bar{\omega}_{h\tau} = \bar{\omega}_h \times \bar{\omega}_\tau$ ($\times$ — знак декартового добутку множин) і задача (1) замінюється деякою схемою методу сіток

$$B_\tau y = A_h y, \quad (x, \tau) \in \omega_{h\tau} = \omega_h \times \omega_\tau;$$
$$y(x, t_0) = u_0(x), \quad x \in \bar{\omega}_h,$$

де $y(x, t)$ — апроксимація для $u(x, t)$ в точці сітки $(x, t) \in \bar{\omega}_{h\tau}$; B_τ — сітковий оператор, що апроксимує похідну $\partial u/\partial t$ на сітці $\bar{\omega}_\tau$; A_h — сітковий оператор, який апроксимує L та оператор крайових умов l на сітці $\bar{\omega}_h$.

Приклад 3. Апроксимувати задачу (5) за методом сіток.

Розв'язання. Виберемо сітки $\bar{\omega}_\tau = \{t_i = i\tau : i = \overline{0, L}, \tau = T/L\}$, $\omega_\tau = \{t_i = i\tau : i = \overline{1, L-1}, \tau = T/L\}$, $\omega_\tau^+ = \omega_\tau \cup \{T\}$, $\omega_\tau^- = \omega_\tau \cup \{0\}$, $\bar{\omega}_h = \{x_i = ih : i = \overline{0, N}, h = 1/N\}$.

Методом безпосередньої заміни похідних різницевими співвідношеннями дістаємо, наприклад, такі схеми:

$$y_t^j(x) = y_{\bar{x}x}^j + \varphi^j(x), \quad x \in \omega_h,$$

$$y^0(x) = u_0(x), \quad x \in \overline{\omega}_h, \tag{16}$$
$$y^j(0) \equiv y_0^j = u_1(t_j), \quad y^j(1) \equiv y_N^j = u_2(t_j),$$

де $\varphi^j(x) = f(x, t_j)$, $x \in \omega_h$, $y^j(x)$ — апроксимація для $u(x, t_j)$ (в індексному вигляді $\varphi_i^j = f(x_i, t_j)$, $y_i^j = y(x_i, t_j)$);

$$y_t^j(x) = y_{\overline{xx}}^{j+1}(x) + \varphi^j(x), \quad x \in \omega_h,$$

$$y^0(x) = u_0(x), \quad x \in \overline{\omega}_h; \tag{17}$$
$$y^j(0) = u_1(t_j), \quad y^j(1) = u_2(t_j).$$

Шаблони цих схем мають відповідно вигляд

$$\begin{pmatrix} \times \\ \times\times\times \end{pmatrix} \text{ і } \begin{pmatrix} \times\times\times \\ \times \end{pmatrix}.$$

В індексній формі запису схема (16) має вигляд

$$\frac{y_i^{j+1} - y_i^j}{\tau} = \frac{y_{i+1}^j - 2y_i^j + y_{i-1}^j}{h^2} + \varphi_i^j, \quad i = \overline{1, N-1}; \quad j = \overline{0, L-1};$$

$$y_i^0 = u_0(x_i), \quad i = \overline{0, N}, \tag{18}$$
$$y_0^j = u_1^j, \quad y_N^j = u_2^j, \quad j = \overline{1, L}$$

або

$$y_i^{j+1} = y_i^j + \frac{\tau}{h^2}(y_{i+1}^j - 2y_i^j + y_{i-1}^j) + \tau\varphi_i^j, \quad i = \overline{1, N-1};$$
$$j = \overline{0, L-1},$$

$$y_i^0 = u_0(x_i), \quad i = \overline{0, N}, \tag{19}$$
$$y_0^j = u_1^j, \quad y_N^j = u_2^j, \quad j = \overline{1/L}.$$

Звідси бачимо, що значення y_i^0, $i = \overline{0, N}$, y_0^j, y_N^j, $j = \overline{1, L}$, явно виражаються через вихідні дані задачі (5), а через них за першою з формул (19) обчислюємо y_i^1, $i = \overline{1, N-1}$; маючи y_i^1, за тією самою формулою обчислюємо y_i^2 і т. д.

Оскільки в цьому разі розв'язок схеми (16) знаходимо за явними формулами (19), то ця схема називається *явною*. В явній схемі похідна $\partial u/\partial t$ в точці t_j замінюється різницевою похідною y_t^j, а еліптична частина рівняння $\frac{\partial^2 u}{\partial x^2}$ апроксимується на ярусі t_j виразом $y_{\overline{xx}}^j$.

Різницеву схему (17) запишемо у вигляді

$$Ay^{j+1} = F^j, \quad y^0 = (u_0(x_1), \ldots, u_0(x_{N-2}))^T, \tag{20}$$

де $y^{j+1} = (y_i^{j+1})_{i=\overline{1,N-1}}$, $F^j = \left(\tau\varphi_1^j + y_1^j + \frac{\tau}{h^2}u_1^j, \tau\varphi_2^j + y_2^j, \ldots, \tau\varphi_{N-1}^j + y_{N-2}^j, \tau\varphi_{N-1}^j + y_{N-1}^j + \frac{\tau}{h^2}u_2^j\right)^T$,

$$A = \frac{\tau}{h^2}\begin{bmatrix} 2+\frac{h^2}{\tau} & -1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2+\frac{h^2}{\tau} & -1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2+\frac{h^2}{\tau} & -1 & \ldots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \ldots & -1 & 2+\frac{h^2}{\tau} & -1 \\ 0 & 0 & 0 & 0 & \ldots & 0 & -1 & 2+\frac{h^2}{\tau} \end{bmatrix}$$

(A — квадратна матриця розмірності $(N-1) \times (N-1)$). Вектор y^0 визначається вихідними даними задачі. Покладаючи в (20) $j = 0$, бачимо, що F^0 визначається через y^0 і вихідні дані задачі, отже, для знаходження y^1 треба розв'язати систему лінійних алгебраїчних рівнянь із тридіагональною матрицею. Маючи y^1, визначаємо F^1, знову розв'язуємо систему (20) і знаходимо y^2 і т. д. Оскільки для визначення вектора y^{j+1} на кожному часовому ярусі треба розв'язувати систему лінійних алгебраїчних рівнянь (20), то схема (17) називається *неявною*. Розв'язок систем виду (20) можна знаходити методом прогонки. Перевага явної схеми в алгоритмічній простоті, її недоліки і переваги неявної схеми стануть зрозумілими пізніше, коли ми розглядатимемо їхню стійкість.

Вправа 1. Довести, що похибка апроксимації схем (16), (17) має порядок $O(\tau + h^2)$. За яких умов гладкості розв'язку задачі (5) це має місце?

Сіткові схеми для нестаціонарних задач буває зручно записувати у такому канонічному вигляді:

$$By_t + Ay = F(t), \quad t \in \omega_\tau^-; \quad y^0 = u_0, \tag{21}$$

де $y = y(t) : \overline{\omega}_\tau \to H_h$ — сіткова функція, задана на деякій сітці $\overline{\omega}_\tau = \{t_j : j = \overline{1, L}\}$ із значеннями в деякому просторі H_h сіткових функцій $y(x, t)$, $x \in \omega_h$, заданих на сітці $\omega_h = \{x_i : i = \overline{1, N-1}\}$, A, B — оператори в H_h. Для досліджень зручно розглядати простір H_h зі скалярним добутком $(\cdot, \cdot)$.

Наприклад, введемо простір $\Omega_{N-1} = H_h$ сіткових функцій, заданих на сітці ω_h, зі скалярним добутком

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h,$$

а також простір $\mathring{\Omega}_{N+1}$ функцій, які задані на сітці $\overline{\omega}_h$ і дорівнюють нулю при $x = 0$ та при $x = 1$. Позначимо через $\mathring{y} \in \mathring{\Omega}_{N+1}$ функцію,

яка на сітці ω_h збігається з функцією $y \in \Omega_{N-1}$, і визначимо в Ω_{N-1} оператор A за формулою

$$Ay = -\mathring{y}_{xx}, \quad y \in \Omega_{N-1}, \quad \mathring{y} \in \mathring{\Omega}_{N+1}.$$

Тоді явна схема (16) може бути записана у вигляді

$$By_t(t) + Ay(t) = F(t), \quad t \in \omega_\tau^-, \quad y(0) = u_0, \tag{22}$$

де $B = I$, $u_0 = \{u_0(x)\, x \in \omega_h\}$,

$$F(t) = \begin{cases} \varphi(x_1, t) + \frac{1}{h^2} u_1(t), & x = x_1, \\ \varphi(x, t), & x = x_i, \quad i = 2, \overline{N-2}, \\ \varphi(x_{N-1}, t) + \frac{1}{h^2} u_2(t), & x = x_{N-1}. \end{cases}$$

Неявна схема (17) записується також у вигляді (21), але позначення матимуть такий зміст: $B = I + \tau A$,

$$F(t) = \begin{cases} \varphi(x_1, t) + \frac{1}{h^2} u_1(t+\tau), & x = x_1, \\ \varphi(x_i, t), & x = x_i, \ i = 2, N-2, \\ \varphi(x_{N-1}, t) + \frac{1}{h^2} u_2(t+\tau). \end{cases}$$

Ввівши індексні позначення $v(t_j) = v^j$, схему (22) запишемо у вигляді

$$B\frac{y^{j+1} - y^j}{\tau} + Ay^j = F^j, \quad y^0 = u_0. \tag{23}$$

Ми бачимо, що при такому запису різницевих схем (16), (17) для задачі (5) вони є окремим випадком абстрактної двох'ярусної однокрокової схеми в п. 1.4.5 з ч. 1. Тому для цих схем можна ввести поняття стійкості, яке було визначено в п. 1.4.5 з ч. 1, і для них будуть справедливі всі дістані там результати. Нагадаємо лише означення та наведемо окремі приклади.

Означення 1. Схема (21) називається *коректною*, якщо вона має єдиний розв'язок при будь-яких вхідних даних (тобто, u_0, F) і він неперервно залежить від вхідних даних (остання властивість називається *умовою стійкості*), тобто для будь-яких u_0, F і для всіх досить малих h, τ ($h \leqslant h_0$, $\tau \leqslant \tau_0$) справджується нерівність

$$\|y^{j+1}\|_{H_h} \leqslant M_1 \|y^0\|_{H_h} + M_2 \Phi(F), \tag{24}$$

де $\Phi(F)$ — функціонал від вектора $F = (F^0, F^1, \ldots, F^j)^T$ з властивостями норми; M_1, M_2 — додатні сталі, незалежні від h, τ, y^0, F.

Схеми для нестаціонарних рівнянь можуть бути багатоярусними. Схема виду

$$\sum_{k=0}^{m-1} B_k^j y^{j+1-k} = f^j, \quad j = m-2, m-1, m, \ldots, \tag{25}$$

де $m \geqslant 2$, $y^0, y^1, \ldots, y^{m-2}$, f^j — задані елементи простору H_h; B_k^j — задані лінійні оператори в H_h, причому існує B_0^{-1}, називається *лінійною m-ярусною схемою*. Двох'ярусна схема має вигляд $B_0 y^{j+1} + B_1 y^j = f^j$, $j = 0, 1, \ldots$; $y^0 = u_0$, де B_1, $B_0 : H_h \to H_h$, причому існує B_0^{-1}; f^j, u_0 — задані елементи H_h. Ця схема може бути записана в канонічному вигляді (21), де оператори B, A, B_0, B_1 зв'язані співвідношеннями $B = \tau B_0$, $A = B_1 + B_0$. Всяку m-ярусну схему можна записати у вигляді двох'ярусної і, таким чином, звести вивчення її до вивчення відповідної двох'ярусної схеми. Дійсно, введемо вектор

$$Y^j = (y^j, y^{j-1}, \ldots, y^{j-(m-2)})^T$$

і вважатимемо його розв'язком схеми (25) на ярусі t_j. Всі вектори такого типу утворюють лінійний простір $H_h^{(m-1)}$, в якому операції додавання і множення на число α визначаються покоординатно, тобто

$$V^j + Z^j = (v^j + z^j, v^{j-1} + z^{j-1}, \ldots, v^{j-(m-2)} + z^{j-(m-2)})^T,$$

$$\alpha V^j = (\alpha v^j, \ldots, \alpha v^{j-(m-2)})^T.$$

Нульовим елементом простору $H_h^{(m-1)}$ буде вектор, кожна компонента якого є нуль простору H_h. Простір $H_h^{(m-1)}$ називається *прямою сумою* $m-1$ *просторів* H_h, а операція утворення прямої суми позначається так:

$$H_h^{(m-1)} = H_h \oplus H_h \oplus \cdots \oplus H_h.$$

Якщо в H_h задано оператори $D_{\alpha\beta}$, $\alpha, \beta = \overline{1, m-1}$, $D_{\alpha\beta}$: $H_h \to H_h$, то матриця

$$D = (D_{\alpha\beta})_{\alpha,\beta=\overline{1,m-1}}$$

визначає оператор $D : H_h^{(m-1)} \to H_h^{(m-1)}$. Для таких операторних матриць справедливі звичайні правила додавання матриць та множення на число.

Тепер m-ярусну схему (25) можна записати у вигляді

$$Y^{j+1} + TY^j = F^j, \quad j = 0, 1, \cdots,$$

де $Y^0 = (y^0, \ldots, y^{m-2})$ — заданий елемент простору $H_h^{(m-1)}$,

$$T = \begin{pmatrix} B_0^{-1}B_1 & B_0^{-1}B_2 & \ldots & B_0^{-1}B_{m-2} & B_0^{-1}B_{m-1} \\ -I & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & & -I & 0 \end{pmatrix},$$

$$F^j = (B_0^{-1}f^j, \ldots, 0)^T$$

Слід однак зазначити, що дослідження стійкості багатоярусних схем у багатьох випадках зручніше проводити на основі означення цих схем, а не зведенням їх до двох'ярусних.

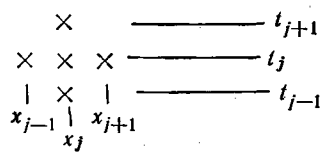
Приклад 4. Для рівняння теплопровідності (5) заміною похідних різницевими співвідношеннями можна дістати схему

$$y^j_{\overset{\circ}{t}} = y^j_{\overline{xx}}(x) + \varphi^j(x), \quad x \in \omega_h, \quad j = 1, 2, \ldots,$$

$$y^0(x) = u_0(x), \quad x \in \bar{\omega}_h, \tag{26}$$

$$y^j(0) = u_1(t_j), \quad y^j(1) = u_2(t_j),$$

де $y^j_{\overset{\circ}{t}} = \dfrac{y^{j+1} - y^{j-1}}{2\tau}$. Шаблон її має вигляд ромба:



Щоб вести розрахунки, недостатньо мати $y^0(x)$, а треба знати ще, наприклад, $y^1(x)$. Це окреме питання, яке розглянемо у наступному параграфі. Зазначимо лише, що y^1 можна знайти, виконавши один крок за схемою (16).

Введемо вектор $Y^j = (y^j, y^{j-1})^T$, який вважатимемо розв'язком схеми (26) на ярусі t_j. Ввівши оператор A у праву частину F аналогічно (22), запишемо схему (26) у вигляді

$$B_0 y^{j+1} + B_1 y^j + B_2 y^{j-1} = f^j, \quad j = 1, 2, \ldots,$$

де y^0, y^1, f^j — задані елементи з H_h; $B_0 = I$, $B_1 = 2\tau A$, $B_2 = -I$; I — одиничний оператор. Запишемо цю схему як двох'ярусну:

$$Y^{j+1} + TY^j = F^j, \quad j = 1, 2, \ldots, \tag{27}$$

де $Y^1 = (y^1, y^0)^T$ — заданий елемент $H_h^{(2)}$, $F^j = (2\tau f^j, 0)$,

$$T = \begin{pmatrix} 2\tau A & -I \\ -I & 0 \end{pmatrix}.$$

У канонічному вигляді схема (27) запишеться так:

$$\overline{B}\,\frac{Y^{j+1} - Y^j}{\tau} + \overline{A}Y^j = \overline{F}^j, \quad j = 1, 2, \ldots, \tag{28}$$

де $\overline{B} = I$, $\overline{A} = \tau^{-1}(I + T)$, $\overline{F}^j = \tau^{-1}F^j$.

Аналогічно означенню 1 багатоярусна схема (25) називається *стійкою,* якщо для будь-яких початкових даних $Y^0 = (y^0, \ldots, y^{m-2})^T$ та для будь-яких правих частин справджується нерівність

$$\|Y^{j+1}\|_{H_h^{(m-1)}} \leqslant M_1 \|Y^0\|_{H_h^{(m-1)}} + M_2 \Phi(F),$$

де $\|\cdot\|_{H_h^{(m-1)}}$ — норма у просторі $H_h^{(m-1)}$; $\Phi(F)$ — функціонал від $F = (F^0, \ldots, F^j)^T$ із властивостями норми; M_1, M_2 — додатні сталі, незалежні від h, τ, Y^0, F.

Означення 2. Схема, яка є стійкою при будь-яких h, τ (при яких вона має зміст), називається *абсолютно стійкою.* Схема, яка є стійкою при деяких додаткових умовах, які зв'язують h, τ, називається *умовно стійкою.*

Розв'язок схеми (21) можна зобразити у вигляді $y^j = y^j_{(1)} + y^j_{(2)}$, де $y^j_{(1)}$ — розв'язок схеми

$$B(y^j_{(1)})_t + Ay^j_{(1)} = 0, \quad j = 0, 1, \ldots, \quad y^0_{(1)} = u_0,$$

а $y^j_{(2)}$ — розв'язок схеми

$$B(y^j_{(2)})_t + Ay^j_{(2)} = F^j, \quad j = 0, 1, \ldots,$$

$$y^0_{(2)} = 0.$$

Тому природно розрізняють стійкість схеми (21) за початковими умовами та правою частиною.

Означення 3. Якщо для розв'язку схеми (21) має місце оцінка (24) при $y^0 = 0$, тобто

$$\|y^{j+1}\|_{H_h} \leqslant M_2 \Phi(F), \tag{29}$$

то схема (21) називається *стійкою за правою частиною.*

Означення 4. Якщо для розв'язку схеми (21) має місце оцінка (24) при $F = 0$, тобто

$$\|y^{j+1}\|_{H_h} \leqslant M_1 \|y^0\|_{H_h}, \tag{30}$$

то схема (21) називається *стійкою за початковими умовами.*

Якщо ж для схеми (21) при $F = 0$ має місце оцінка

$$\|y^{j+1}\|_{H_h} \leqslant M_1 \|y^k\|_{H_h}, \qquad k = \overline{0, j}, \quad j = 0, 1, \ldots, \tag{31}$$

для будь-яких y^k, $k = \overline{0, j}$, де M_1 не залежить від h, τ, y^k, $k = \overline{0, j}$, $j = 0, 1, \ldots$, то схема (21) називається *рівномірно стійкою за початковими умовами.*

Зауважимо, що рівномірна стійкість при $M_1 = 1$ гарантує незростання похибок, які вносяться на кожному проміжному ярусі. Якщо в H_h діє деякий додатний самоспряжений оператор $D = D^* > 0$, то можна ввести скалярний добуток і норму за формулами

$$(u, v)_D = (Du, v), \quad \|u\|_D = \sqrt{(Du, u)}.$$

Простір H_h з таким скалярним добутком і нормою називається *енергетичним простором оператора* D і позначається H_D.

Означення 5. Схема (21) у просторі H_D називається *ρ-стійкою*, якщо при $F \equiv 0$ для будь-яких $y^j \in H_h$ і для всіх j має місце оцінка

$$\| y^{j+1} \|_{H_D} \leqslant \rho \| y^j \|_{H_D}. \tag{32}$$

Зауважимо, що коли $\rho^j \leqslant M_1$ для всіх $j\tau \leqslant T_0$, де M_1 — додатна стала, незалежна від h, τ, j, то з (32) маємо

$$\| y^{j+1} \|_{H_D} \leqslant \rho^{j+1} \| y^0 \|_{H_D} \leqslant M_1 \| y^0 \|_{H_D},$$

тобто з ρ-стійкості в H_D випливає стійкість за початковими умовами в H_D. Стала ρ може взагалі не залежати від h, τ. Наприклад, при $\rho = 1$ схема (21) буде абсолютно стійкою в H_D.

Означення 6. Якщо в оцінці (32) $\rho = 1 + c\tau$, де c — додатна стала, незалежна від h, τ, то схема (21) називається *слабо стійкою в* H_0. Якщо $\rho = 1 - \beta(h, \tau, T_0) > 0$, де $\beta > 0$ і $\lim\limits_{h,\tau\to 0} \beta(h, \tau, T_0) = 0$, то схема (21) називається *сильно стійкою за початковими умовами в* H_D. При $\rho = e^{-\tau\delta(h,\tau)}$, де $\delta(h, \tau) > 0$, $\lim\limits_{h,\tau\to 0} \delta(h, \tau) = \delta_0$, δ_0 — незалежна від h, τ додатна стала, схема (21) називається *асимптотично стійкою в* H_D при $t_j \to \infty$.

Неважко помітити, що всі означення стійкості вказують на певний характер росту похибок під час практичних розрахунків y^j при переході з одного ярусу на інший. Наприклад, в абсолютно стійких схемах ці похибки не нагромаджуються з ростом j, тобто залишаються обмеженими. У слабо стійких схемах допускається зростання таких похибок із ростом j, якщо розглядається необмежений інтервал $t \in (0, \infty)$, але не швидше, ніж за законом $\rho^j = (1 + c\tau)^j$, тобто поліноміально за τ, і їх можна зменшити, вибравши меншим τ. Сильно стійкі схеми, очевидно, зменшують вплив похибок із ростом j (а також $t_j = t_0 + j\tau$). Асимптотично стійкі схеми дають змогу зменшенням τ зменшувати похибки розрахунків при необмеженому зростанні t_j. Можна давати й інші означення стійкості залежно від мети, яку ставить перед собою дослідник, вибираючи схему (зрозуміло, що ця мета зумовлена практичною задачею).

Приклад 5. Дослідити на стійкість за початковими умовами в H_A явну та неявну схеми (16), (17) для одновимірного рівняння теплопровідності.

Розв'язання. Як ми бачили, ці схеми записуються в канонічному вигляді (22), причому для явної схеми $B = I$, а для неявної $B = I + \tau A$. Використаємо теорему 1 з п. 1.4.5 з ч. 1, згідно з якою необхідною і достатньою умовою стійкості за початковими умовами є нерівність $B \geqslant \frac{\tau}{2} A$. За допомогою формул Гріна (див. п. 1.4 з ч. 1) неважко показати, що $A = A^* > 0$, і для обох схем $B = B^* > 0$. Для явної схеми в силу нерівності $0 < A \leqslant \| A \| I$ маємо

$$B - \frac{\tau}{2} A = I - \frac{\tau}{2} A \geqslant \left(\frac{1}{\| A \|} - \frac{\tau}{2} \right) A.$$

Нерівність $B - \frac{\tau}{2} A \geqslant 0$ виконуватиметься, коли $\frac{1}{\| A \|} - \frac{\tau}{2} \geqslant 0$. Оскільки A самоспряжений додатний оператор, то його норма дорівнює максимальному власному значенню (див. п. 1.4.4 з ч. 1)

$$\| A \| = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}.$$

Тобто явна схема буде стійкою за умови

$$\tau \leqslant \frac{h^2}{2 \cos^2 \frac{\pi h}{2}} < h^2, \quad h \leqslant 1/2,$$

отже є умовно стійкою.

Неявна схема є безумовно стійкою, оскільки для неї

$$B - \frac{\tau}{2} A = I + \tau A - \frac{\tau}{2} A = I + \frac{\tau}{2} A \geqslant 0$$

для всіх h, τ.

Приклад 6. Дослідити на стійкість схему з прикладу 4.

Розв'язання. Застосовуючи теореми п. 1.4.5 з ч. 1, бачимо, що достатньою умовою стійкості за початковими умовами в H_A і за правою частиною H_B є $\bar{B} \geqslant \frac{\tau}{2} \bar{A}$ (якщо $\bar{B} = \bar{B}^* > 0$, $\bar{A} = \bar{A}^* > 0$).

Вводячи в просторі $H_h^{(2)}$, який складається з елементів виду $v = (u_1, u_2)^T$, u_1, $u_2 \in H_h$, скалярний добуток за формулою

$$(u, v)_{H_h^{(2)}} = (u_1, v_1)_{H_h} + (u_2, v_2)_{H_h},$$

$$(u, v)_{H_h} = \sum_{i=1}^{N-1} h u_i v_i = \sum_{i=1}^{N-1} h u(x_i) v(x_i),$$

дістаємо $(Y = (y_1, y_2))^T : (BY, Y) > 0$, $B = B^*$, $A = A^*$,

$$\tau(\bar{A}Y, Y) = ((I + 2\tau A) y_1, y_1) - (y_2, y_1)_{H_h} - (y_1, y_2)_{H_h} + (y_2, y_2)_{H_h} =$$

$$= (y_1, y_1) - 2(y_1, y_2) + (y_2, y_2) + 2\tau (Ay_1, y_1) = (y_1 - y_2, y_1 - y_2) +$$

$$+ 2\tau (Ay_1, y_1) > 0, \quad (\bar{B}Y, Y) - \frac{\tau}{2} (\bar{A}Y, Y) = (y_1, y_1) -$$

$$- \frac{\tau}{2} (\tau^{-1} (I + 2\tau A) y_1 - \tau^{-1} y_2, y_1) + (y_2, y_2) - \frac{\tau}{2} (-\tau^{-1} y_1 + \tau^{-1} y_2, y_2) =$$

$$= (y_1, y_1) - \frac{1}{2} ((I + 2\tau A) y_1, y_1) + \frac{1}{2} (y_2, y_1) + (y_2, y_2) -$$

$$- \frac{1}{2} (y_1, y_2) - \frac{1}{2} (y_2, y_2).$$

Звідси бачимо, що умова $\bar{B} \geqslant \frac{\tau}{2} \bar{A}$ виконується, коли

$$I - \frac{1}{2} (I + 2\tau A) \geqslant 0 \text{ або } \frac{1}{2} I - \tau A \geqslant 0.$$

Оскільки $0 < A \leqslant \| A \| I$, то остання нерівність виконується при

$$\left(\frac{1}{2\| A \|} - \tau\right) A \geqslant 0 \tag{33}$$

або при $\frac{1}{2\| A \|} - \tau \geqslant 0$. Оскільки A — самоспряжений додатний в H_h оператор (що легко доводиться застосуванням формул Гріна), то його норма дорівнює максимальному власному значенню, тобто (див. п. 1.4.4 з ч. 1) $\| A \| = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}$. Отже, триярусна схема (26) для рівняння теплопровідності буде стійкою, коли

$$\frac{h^2}{8 \cos^2 \frac{\pi h}{2}} - \tau \geqslant 0 \quad \text{або} \quad \tau \leqslant \frac{h^2}{8 \cos^2 \frac{\pi h}{2}}, \tag{34}$$

тобто ця схема є умовно стійкою.

Як зазначалося, в рамках даних означень деякі результати щодо стійкості двох'ярусних схем знайдено в п. 1.4.5 з ч. 1.

Якщо маємо достатні умови стійкості (деякі доведено в п. 1.4.5 з ч. 1), то вони можуть використовуватися для побудови нових схем із заданими якостями. Такий метод побудови нових схем називається *методом регуляризації (принципом регуляризації)*.

Нехай схема вигляду (21) є стійкою (за початковими даними чи за правою частиною, що забезпечується достатньою умовою $B \geqslant \frac{\tau}{2} A$ (див. п. 1.4.5 з ч. 1)). Очевидно, якщо взяти схему такого самого виду з оператором $\tilde{B} \geqslant B$, то вона також буде стійкою. Якщо, наприклад, оператор B зображається у вигляді $B = I + \tau R$, $R = R^* > 0$, то, вибираючи $R \geqslant \left(\frac{1}{2} - \frac{1}{\tau \| A \|}\right) A$, дістаємо стійку схему, бо

$$B = I + \tau R \geqslant \frac{\tau}{2} A.$$

Оператор R називається *регуляризатором*. Очевидно, якщо вибрати $\hat{R} = \hat{R}^* \geqslant R$, то знову матимемо стійку схему. Вибір оператора регуляризації відбувається таким чином, щоб дістати у певному розумінні кращу схему, ніж початкова, наприклад, яка була б абсолютно стійкою при заданому порядку апроксимації. Зауважимо, що метод регуляризації не є формальним методом побудови схем, а вимагає від дослідника творчого підходу.

Приклад 7. На основі явної умовно стійкої схеми прикладу 3 знайти безумовно стійку схему.

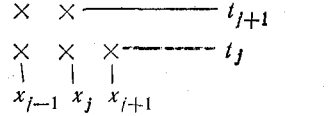
Розв'язання. Ідея методу регуляризації приводить до так званої *схеми Саульєва*

$$y_t^j + \frac{\alpha\tau}{h} y_{\bar{t}x}^j = y_{\bar{x}x}^j + \varphi^j(x), \quad x \in \omega_h, \quad j = 0, 1, \ldots,$$

$$y^0(x) = u_0(x), \quad x \in \bar{\omega}_h, \tag{35}$$

$$y^j(0) = u_1(t_j), \quad y^j(1) = u_2(t_j), \quad \alpha = \text{const},$$

яка записується на шаблоні



і відрізняється від розглянутої вище умовно стійкої схеми «малою» поправкою $\frac{\alpha\tau}{h} y_{\bar{t}x}^j$ $\left(\text{при } \frac{\tau}{h} \to 0\right)$. Ця схема також явна і в цьому її перевага. В індексній формі запису вона має вигляд

$$y_i^{j+1} = \frac{1}{\alpha + \beta^{-1}} \left[\alpha y_{i-1}^{j+1} + (1 - \alpha) y_{i-1}^j + (\beta^{-1} + \alpha - 2) y_i^j + y_{i+1}^j\right],$$

$$\alpha > 0, \quad \beta = \frac{\tau}{h^2}.$$

Якщо на j-му ярусі всі значення вже знайдено, то розрахунок на $(j+1)$-му ярусі починаємо з $i = 1$ (при цьому $y_{i-1}^{j+1} = y_0^{j+1} = u_1(t_{j+1})$ — задане значення), далі, маючи y_1^{j+1}, аналогічно знаходимо y_2^{j+1} і т. д.

Дослідимо цю схему на стійкість за початковими умовами, яка і була метою регуляризації. Введемо, як і вище, простори $\mathring{\Omega}_{N+1}$ і Ω_{N-1} і покладемо $Dy = \mathring{y}_{\bar{x}}$, $Ay = -\mathring{y}_{\bar{x}x}$; $y \in \Omega_{N-1}$, $\mathring{y} \in \Omega_{N+1}$. Тоді схему (35) при $\varphi = 0$, $u_1 = u_2 = 0$ можна подати у вигляді

$$By_t^j + Ay^j = 0,$$

де $B = I + \frac{\alpha\tau}{h} D$.

Зауважимо, що $\mathring{y}_{\bar{x}} - \mathring{y}_x = -h\mathring{y}_{\bar{x}x}$ і в силу формули підсумовування частинами $(\mathring{y}_{\bar{x}}, y) = -(\mathring{y}_x, y)$. Тому $(Dy, y) = -(\mathring{y}_x, y) = -\frac{h}{2}(\mathring{y}_{\bar{x}x}, y) > 0$, оскільки оператор $Ay = -\mathring{y}_{\bar{x}x}$ — додатно визначений. Отже, при $\alpha > 0$ оператори $B = I + \frac{\alpha\tau}{h} D > 0$, A задовольняють умови теореми 1 з п. 1.4.5, ч. 1, і схема стійка при $B - \frac{\tau}{2} A = I + \frac{\alpha\tau}{h} D - \frac{\tau}{2} A \geqslant 0$. Користуючись тим, що $(Dy, y) = \frac{h}{2}(Ay, y)$, та нерівністю $0 < A \leqslant \| A \| I$, маємо $B - \frac{\tau}{2} A = I + \frac{\alpha\tau}{h} D - \frac{\tau}{2} A \geqslant I + \frac{\tau}{2}(\alpha - 1) A \geqslant \left(\frac{1}{\| A \|} + \frac{1}{2}\tau(\alpha - 1)\right) A$, звідки видно, що схема стійка в H_A при

$$\frac{1}{\| A \|} + \frac{1}{2}\tau(\alpha - 1) \geqslant 0,$$

або $\alpha \geqslant 1 - \frac{2}{\tau \| A \|} \geqslant 1 - \frac{h^2}{2\tau}$. Бачимо також, що схема Саульєва при $\alpha \geqslant 1$ є безумовно стійкою в H_A за початковими умовами.

Вправа 2. 1. Довести, що схема Саульєва апроксимує рівняння теплопровідності з похибкою апроксимації

$$\psi = 0\left(\tau + h^2 + \frac{\alpha\tau}{h}\right)$$

$\left(\text{точніше кажучи, умовно апроксимує при } \tau, h, \frac{\tau}{h} \to 0\right)$. За яких умов гладкості на шуканий розв'язок це має місце?

2. Метод регуляризації від триярусної схеми з прикладу 6 призводить до явної схеми «ромб» виду $\left(\text{при } \frac{\tau}{h} \to 0\right)$

$$y_{\overset{\circ}{t}}^j + \frac{\tau^2}{h^2} y_{tt}^j = y_{\bar{x}x}^j(x) + \varphi^j(x), \quad x \in \omega_h$$

або в індексній формі запису

$$y_i^{j+1} = \frac{1}{1+2\beta}\left[2\beta\left(y_{i+1}^j + y_{i-1}^j\right) + (1-2\beta)\, y_i^{j-1}\right], \quad \beta = \tau/h^2.$$

Довести, що ця схема є безумовно стійкою і має порядок апроксимації

$$\psi = 0\left(\tau^2 + h^2 + \frac{\tau^2}{h^2}\right).$$

За яких умов гладкості це має місце?

Вказівка. Використати приклад 9 з п. 1.4.5, ч. 1.

Ми бачили в п. 1.4.5 з ч. 1, що для триярусних схем

$$B_0 y^{j+1} + B_1 y^j + B_2 y^{j-1} = F^j, \quad j = 1, 2, \ldots$$

(y^0, y^1 — задані елементи простору H_h), записаних в канонічному вигляді

$$B y_{\overset{\circ}{t}}^j + \tau^2 R y_{\bar{t}t}^j + A y = \varphi,$$

достатньою умовою стійкості в енергетичній нормі вигляду

$$\| y^j \| = \left(\frac{1}{4}\| y^{j+1} + y^j \|_A^2 + \| y^{j+1} - y^j \|_{R - \frac{1}{4}A}^2\right)^{1/2}$$

є виконання нерівностей

$$A > 0, \quad R \geqslant \frac{1}{4} A.$$

У п. 1.4.5 з ч. 1 показано, що поняття стійкості за початковими умовами та за правою частиною зв'язані між собою (див., наприклад, теореми 5, 6, п. 1.4 з ч. 1). Повернемося ще раз до цього питання і

розглянемо двох'ярусну схему в канонічному вигляді

$$B y_{\overset{\circ}{t}}^j + A y^j = \varphi^j(x), \quad x \in \omega_h, \quad j = 0, 1, \ldots, \; t \leqslant t_j \leqslant T,$$
$$y^0(x) = 0, \quad y^j(x) \in H_h. \tag{36}$$

Якщо оператор B^{-1} існує, то двох'ярусну неявну схему (36) можна зобразити у вигляді явної схеми

$$y^{j+1} = T y^j + \tau B^{-1}\varphi^j, \quad T = I - \tau B^{-1} A,$$

звідки в силу $y^0(x) = 0$ маємо

$$y^{j+1} = \sum_{k=0}^{j} \tau T^{j-k} B^{-1} \varphi^k. \tag{37}$$

Розглянемо схему (36) на скінченному за t проміжку, тобто $j\tau \leqslant T$, і припустимо, що

$$\| T \| \leqslant \rho, \tag{38}$$

де $\rho = \rho(\tau) > 0$, $\rho^j \leqslant M_1$ для всіх j таких, що $j\tau \leqslant T$; M_1 — додатна стала, незалежна від h, τ, j. Наприклад, якщо $\rho = 1 + c\tau$, де $c > 0$ — абсолютна стала і розглядається скінченний інтервал $t \in [t, T]$, $\tau = \frac{T-t}{L}$, то

$$\rho^j = (1 + c\tau)^j = \left(1 + \frac{c(T-t)}{L}\right)^j \leqslant \left(1 + \frac{c(T-t)}{L}\right)^L < e^{c(T-t)} \equiv M_1.$$

За умови (38) з формули (37) матимемо оцінку

$$\| y^{j+1} \|_{H_h} \leqslant \sum_{k=0}^{j} \tau \rho^{j-k} \| B^{-1}\varphi^k \|_{H_h}, \tag{39}$$

або

$$\| y^{j+1} \|_{H_h} \leqslant M_1 \sum_{k=0}^{j} \tau \| B^{-1}\varphi^k \|_{H_h}.$$

Ця оцінка означає стійкість за правою частиною згідно з означенням 3. Отже, ми довели таке твердження.

Теорема 1. *Якщо виконується умова* (38), *то схема* (36) *стійка за правою частиною і виконується оцінка* (39).

Згідно з означенням 4 (див. (31)) умова (38) означає рівномірну стійкість схеми за початковими умовами, тобто існує зв'язок між рівномірною стійкістю за початковими умовами та за правою частиною. Причому раніше ми не вимагали, щоб оператор переходу T не залежав від j. Якщо ж T не залежить від j, то має місце більш сильне твердження.

Теорема 2. *У випадку сталого оператора переходу T умова рівномірної стійкості за початковими умовами є необхідною і достатньою для стійкості виду* (39) *за правою частиною.*

Доведення. *Необхідність.* Нехай для схеми (36) має місце (39) для будь-яких правих частин φ^k. Виберемо φ^k так, щоб

$$\tau B^{-1}\varphi^k = \beta\delta_{kl},$$

де δ_{kl} — символ Кронекера; β — довільний фіксований елемент з H_h. Тоді в (37) матимемо $y^{j+1} = T^{j-l}\beta$ і в силу оцінки (39)

$$\|y^{j+1}\|_{H_h} = \|T^{j-l}\beta\|_{H_h} \leqslant M_1\|\beta\|_{H_h}.$$

У силу довільності j і l з останньої нерівності маємо

$$\|T^i\| \leqslant M_1, \quad i = \overline{1, j},$$

тобто схема (36) рівномірно стійка за початковими умовами.

Достатність. Нехай схема стійка за початковими умовами, тобто

$$\|y^{j+1}\|_{H_h} \leqslant M_1\|y^k\|_{H_h}, \quad k = \overline{0, j}, \quad j = 1, 2, \ldots.$$

Це означає, що для оператора переходу виконується умова

$$\|T^j\| \leqslant M_1,$$

тому з виразу (37) дістаємо нерівність (39), що і треба було довести.

Якщо виконується умова (38), то для розв'язку задачі (21) з попередніх теорем випливає оцінка

$$\|y^{j+1}\|_{H_h} \leqslant M_1\left(\|y^0\|_{H_h} + \sum_{k=0}^{j}\tau\|B^{-1}\varphi^k\|_{H_h}\right),$$

яка одночасно означає стійкість і за початковими умовами, і за правою частиною. Якщо оператори A, B сталі і $A = A^* > 0$, $B = B^* > 0$, то з теорем 3 і 5 з п. 1.4, ч. 1 випливає таке твердження.

Теорема 3. *Нехай $A = A^* > 0$, $B = B^* > 0$ і виконується умова $B \geqslant \frac{\tau}{2}A$. Тоді для розв'язку схеми* (21) *має місце оцінка*

$$\|y^{j+1}\|_{H_B} \leqslant \|y^0\|_{H_B} + \sum_{k=0}^{j}\tau\|\varphi^k\|_{H_B^{-1}},$$

яка означає стійкість і за початковими умовами, і за правою частиною.

Роль асимптотичної стійкості при розв'язуванні практичних задач розглянемо в наступному параграфі.

6.2. Різницева схема з ваговими коефіцієнтами для одновимірного рівняння теплопровідності

Виберемо сітки так, як у прикладі 3 попереднього параграфа, і в задачі (5) з п. 6.1 поставимо у відповідність похідній по часу на

ярусі t_j різницеве співвідношення

$$\frac{\partial u(x, t_j)}{\partial t} \sim \frac{y(x, t_{j+1}) - y(x, t_j)}{\tau} = \frac{y^{j+1}(x) - y^{(j)}(x)}{\tau} = y_t^j(x)$$

або в точці x_i в індексній формі запису

$$\frac{\partial u(x_i, t_j)}{\partial t} \sim \frac{y_i^{j+1} - y_i^j}{\tau} = (y_t)_i^j,$$

а похідній $\frac{\partial^2 u}{\partial x^2}$ поставимо у відповідність лінійну комбінацію других різницевих похідних по x при $t = t_j$ (на j-му ярусі) і при $t = t_{j+1}$ (на $(j+1)$-му ярусі):

$$\frac{y_i^{j+1} - y_i^j}{\tau} = \sigma\Lambda y_i^{j+1} + (1 - \sigma)\Lambda y_i^j + \varphi_i^j, \tag{1}$$

де $\Lambda y_i^{j+1} = y_{i+1}^{j+1} - 2y_i^{j+1} + y_{i-1}^{j+1}$, σ — параметр; φ_i^j — деяка права частина, наприклад $\varphi_i^j = f_i^j$ або $\varphi_i^j = f_i^{j+1/2}$. Врахувавши також крайові та початкову умови

$$y_0^j = u_1(t_j), \quad y_N^j = u_2(t), \quad j = \overline{1, L}, \tag{2}$$
$$y_i^0 = u_0(x_i), \quad i = \overline{0, N},$$

дістанемо так звану різницеву схему з ваговими коефіцієнтами для задачі (5) з п. 6.1. Схема (1), (2) визначена на 6-точковому шаблоні:

$$\begin{array}{ccc} (x_{i-1}, t_{j+1}) & (x_i, t_{j+1}) & (x_{i+1}, t_{j+1}) \\ \times & \times & \times \\ \times & \times & \times \\ (x_{i-1}, t_j) & (x_i, t_j) & (x_{i+1}, t_j). \end{array}$$

У випадку $\sigma = 0$ дістанемо явну схему (16) (або (18)), п. 6.1, на 4-точковому шаблоні, а при $\sigma = 1$ дістаємо повністю неявну схему (17) (або (19)), п. 6.1, яка також називається схемою з випередженням на шаблоні:

$$\begin{array}{ccc} (x_{i-1}, t_{j+1}) & (x_i, t_{j+1}) & (x_{i+1}, t_{j+1}) \\ \times & \times & \times \\ & \times & \\ & (x_i, t_j) & \end{array}$$

Часто використовується також симетрична неявна схема, яку називають *схемою Кранка — Нікольсона,* утворена зі схеми (1) при $\sigma = 1/2$. Вона має вигляд

$$\frac{y_i^{j+1} - y_i^j}{\tau} = \frac{1}{2}\left(\frac{y_{i-1}^{j+1} - 2y_i^{j+1} + y_{i-1}^{j+1}}{h^2} + \frac{y_{i-1}^j - 2y_i^j + y_{i-1}^j}{h^2}\right) + \varphi_i^j \tag{3}$$

і, як бачимо, записується на симетричному 6-точковому шаблоні. Якщо y_i^j, $i = \overline{0, N}$, тобто значення шуканої функції на j-му ярусі вже відомо, то значення y_i^{j+1} на $(j+1)$-му ярусі у випадку схеми (3) знаходять методом прогонки з крайової задачі

$$\frac{\tau}{2h^2} y_{i-1}^{j+1} - \left(1 + \frac{\tau}{h^2}\right) y_i^{j+1} + \frac{\tau}{2h^2} y_{i+1}^{j+1} = -F_i^j, \quad i = \overline{1, N-1},$$

$$y_0^{j+1} = u_1(t_{j+1}), \quad y_N^{j+1} = u_2(t_{j+1}), \tag{4}$$

$$F_i^j = \left(1 - \frac{\tau}{h^2}\right) y_i^j + \frac{\tau}{2h^2}(y_{i-1}^j + y_{i+1}^j) + \tau\varphi_i^j.$$

Неважко помітити, що при $\sigma \neq 0$ схема (1) з ваговими коефіцієнтами є неявною і y_i^{j+1} визначається методом прогонки як розв'язок задачі

$$\sigma\tau\Lambda y_i^{j+1} - y_i^{j+1} = -F_i^j, \quad i = \overline{1, N-1},$$

$$y_0^{j+1} = u_1(t_{j+1}), \quad y_N^{j+1} = u_2(t_{j+1}), \quad j = 0, 1, \ldots \tag{5}$$

Щоб обгрунтувати застосування схеми (1), (2) для розв'язування задачі (5) з п. 6.1, можна скористатися загальною теорією методу сіток (див. п. 2.1), тобто записати сіткову задачу для похибки, встановити нерівність коректності (нерівність стійкості), величину похибки апроксимації і на основі цього зробити висновок про збіжність схеми при $h, \tau \to 0$.

6.2.1. Стійкість схеми (1), (2). Насамперед зауважимо, що розглянуті вище схеми для рівняння теплопровідності можна записати в канонічному операторному вигляді. Дійсно, введемо простір сіткових функцій, заданих на сітці ω_h, зі скалярним добутком

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h,$$

а також простір $\mathring{\Omega}_{N+1}$ функцій, заданих на сітці $\bar{\omega}_h$ і рівних нулю при $x = 0$ та $x = 1$. Функцію з $\mathring{\Omega}_{N+1}$, яка на сітці ω_h збігається з функцією $y \in \Omega_{N-1}$, позначимо $\mathring{y}$. У просторі Ω_{N-1} введемо оператор A за формулою

$$Ay = -\Lambda\mathring{y}, \quad y \in \Omega_{N-1}, \quad \mathring{y} \in \mathring{\Omega}_{N+1}.$$

Введемо для зручності такі безіндексні позначення:

$$y = y_i^j, \ \hat{y} = y_i^{j+1}, \quad \Lambda y = y_{\bar{x}x} = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2},$$

$$y_t = \frac{y_i^{j+1} - y_i^j}{\tau}, \quad y^{(\sigma)} = \sigma y_i^{j+1} + (1-\sigma) y_i^j = \sigma\hat{y} + (1-\sigma) y.$$

У цих позначеннях схема з ваговими коефіцієнтами (1), (2) запишеться у вигляді

$$y_t = \Lambda y^{(\sigma)} + \varphi, \ (x_i, t_j) \in \omega_{h\tau}, \ y(x, 0) = u_0(x), \quad x \in \bar{\omega}_h,$$
$$y_0 = u_1(t), \ y_N = u_2(t), \quad t \in \omega_\tau. \tag{6}$$

Неважко помітити, що схему (6) можна також записати у такому канонічному вигляді:

$$By_t + Ay = F(t), \quad y(0) = u_0, \tag{7}$$

де $B = I + \sigma\tau A$; I — одиничний оператор;

$$F(t) = \begin{cases} \varphi(x_1, t) + \frac{\sigma}{h^2} u_1(t+\tau) + \frac{(1-\sigma)}{h^2} u(t_1), & x = x_1, \\ \varphi(x, t), \quad x = x_i, \quad i = \overline{2, N-2}, \\ \varphi(x_{N-1}, t) + \frac{\sigma}{h^2} u_2(t+\tau) + \frac{(1-\sigma)}{h^2} u_2(t), & x = x_{N-1}, \end{cases}$$

або

$$B\frac{y^{j+1} - y^j}{\tau} + Ay^j = F^j, \quad y^0 = u_0, \quad j = \overline{0, L},$$
$$y^j = y(t_j). \tag{8}$$

Порівнюючи схему (8) і однокрокову схему в канонічному вигляді з п. 1.4.5, бачимо, що (8) є конкретним прикладом абстрактної схеми (4), п. 1.4 з ч. 1. За допомогою формул підсумовування частинами неважко переконатись, що $A = A^* > 0$ в Ω_{N-1}, отже, $B = I + \sigma\tau A = B^* > 0$ при $\sigma \geqslant 0$, тобто існує B^{-1}. За теоремою 1, п. 1.4 з ч. 1 для стійкості схеми в H_A (див. п. 1.4.4 з ч. 1) необхідно і достатньо, щоб $B \geqslant \frac{\tau}{2} A$, тобто $B - \frac{1}{2}\tau A \geqslant 0$. Враховуючи вигляд B, маємо

$$B - \frac{1}{2}\tau A = I + \left(\sigma - \frac{1}{2}\right)\tau A \geqslant \left(\frac{1}{\|A\|} + \left(\sigma - \frac{1}{2}\right)\tau\right) A \geqslant 0,$$

якщо

$$\frac{1}{\|A\|} + \left(\sigma - \frac{1}{2}\right)\tau \geqslant 0 \text{ або } \sigma \geqslant \tilde{\sigma}_0 \equiv \frac{1}{2} - \frac{1}{\tau\|A\|}. \tag{9}$$

У силу самоспряженості оператора A його норма дорівнює максимальному за модулем власному значенню, тобто (див. п. 1.4.4 з ч. 1)

$$\|A\| = \frac{4}{h^2}\cos^2\frac{\pi h}{2} < \frac{4}{h^2}. \tag{10}$$

Із нерівності (9) бачимо, що схема з ваговими коефіцієнтами **стійка**

при будь-яких τ, h, якщо $\sigma \geqslant \frac{1}{2}$, а при $\sigma < \frac{1}{2}$ стійка, коли

$$\sigma \geqslant \sigma_0 = \frac{1}{2} - \frac{h^2}{4\tau} > \tilde{\sigma}_0. \tag{11}$$

Сіткові схеми називатимемо безумовно стійкими в певній нормі, якщо стійкість у цій нормі має місце без додаткових умов щодо величин h і r і умовно стійкими в протилежному разі.

Отже, схема з ваговими коефіцієнтами безумовно стійка в нормі простору H_A при $\sigma \geqslant \frac{1}{2}$ і умовно стійка за умови (11) у випадку $0 \leqslant \sigma < \frac{1}{2}$.

Запишемо задачу для похибки $z = y - u$, де y — розв'язок схеми (1), (2), u — розв'язок задачі (5), п. 6.1. Підставляючи у схему (6) $y = z + u$, дістаємо

$$\begin{gathered} z_t = \Lambda z^{(\sigma)} + \psi, \quad (x, t) \in \omega_{h\tau}; \\ z(x, 0) = 0, \quad x \in \bar{\omega}_h; \\ z(0, t) = 0, \quad z(1, t) = 0, \quad t \in \omega_\tau, \end{gathered} \tag{12}$$

де

$$\psi = \Lambda u^{(\sigma)} + \varphi - u_t. \tag{13}$$

В операторному вигляді вираз (12) запишеться так:

$$Bz_t + Az = \psi(t), \quad t \in \bar{\omega}_\tau, \quad z(0) = 0. \tag{14}$$

Скориставшись теоремою 5 (п. 1.4 з ч. 1) при $B \geqslant \frac{\tau}{2} A$, маємо оцінку

$$\| z^j \|_B \leqslant \sum_{k=0}^{j-1} \tau \| \psi^k \|_{B^{-1}}, \tag{15}$$

яка виражає стійкість схем (8) і (14) за правою частиною, де $B = I + \sigma\tau A$, $z^j = z(t_j)$, $\psi^k = \psi(t_k)$. Якщо ж $\sigma \geqslant \frac{1}{2}$, то, очевидно, виконується нерівність (24) (п. 1.4 з ч. 1) з теореми 6 для випадку зі сталою $\varepsilon = 1$. Тому з цієї теореми для безумовно стійкої схеми $\left(\sigma \geqslant \frac{1}{2}\right)$ маємо нерівність стійкості

$$\| z^j \|_A^2 \leqslant \frac{1}{2} \sum_{k=0}^{j-1} \tau \| \psi^k \|^2. \tag{16}$$

Користуючись тим, що $Az = -\overset{\circ}{z}_{\bar{x}x}$, і різницевими формулами Гріна, знаходимо

$$\| z \|_A^2 = (Az, z) = -(\overset{\circ}{z}_{\bar{x}x}, \overset{\circ}{z}) = \sum_{i=1}^{N} h (z_{\bar{x},i})^2 \equiv | z |_{\overset{\circ}{W}{}_2^1(\omega_h)}^2.$$

Далі скористуємося сітковою теоремою вкладання (див. п. 1.5 з ч. 1), згідно з якою маємо

$$\| z \|_C = \max_{x \in \omega_h} | z | \leqslant \frac{1}{2} \left(\sum_{i=1}^{N} h (z_{\bar{x},i})^2 \right)^{\frac{1}{2}} = \frac{1}{2} \| z \|_A,$$

що разом із (16) для безумовно стійкої схеми (6) при $\sigma \geqslant \frac{1}{2}$ дає оцінку

$$\| z^j \|_{C(\omega_h)}^2 \leqslant \frac{1}{8} \sum_{k=0}^{j-1} \tau \| \psi^k \|^2. \tag{17}$$

З теореми 1 (п. 1.4 з ч. 1) випливає, що схема (6) стійка за початковими умовами в H_A при виконанні нерівності $B \geqslant \frac{\tau}{2} A$. Аналогічно оцінці (17) дістаємо, що за цієї умови

$$\| y^j \| \leqslant M_0 \| y^0 \| \quad (\varphi \equiv 0). \tag{17'}$$

Якщо ж $0 \leqslant \sigma < \frac{1}{2}$, то в силу нерівності $B = I + \sigma\tau A > I$ з нерівності (15) дістаємо оцінку в слабшій нормі L_2 (ω_h)

$$\| z^j \| \leqslant \sum_{k=0}^{j-1} \tau \| \psi^k \|_{B^{-1}}. \tag{18}$$

За нерівністю Коші — Буняковського

$$\| \psi^k \|_{B^{-1}} = \sqrt{(B^{-1}\psi^k, \psi^k)} \leqslant \sqrt{\| B^{-1}\psi^k \| \| \psi^k \|} \leqslant \sqrt{\| B^{-1} \|} \| \psi^k \|.$$

У силу самоспряженості оператора B буде також самоспряженим і B^{-1}. Його норма дорівнює максимальному власному значенню оператора B^{-1}, яке, як легко помітити, дорівнює $(1 + \sigma\tau\lambda_1)^{-1}$, де $\lambda_1 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2}$ — найменше власне значення оператора A. Якщо $\sigma \geqslant 0$, $\tau > 0$, то $\| B^{-1} \| \leqslant 1$, тобто $\| \psi^k \|_{B_{j-1}^{-1}} \leqslant \| \psi^k \|$, і з виразу (18) дістаємо оцінку

$$\| z^j \| \leqslant \sum_{k=0}^{j-1} \tau \| \psi^k \|, \quad 0 < \tilde{\sigma}_0 \leqslant \sigma < \frac{1}{2}. \tag{19}$$

Скориставшись принципом максимуму, можна дістати нерівність стійкості за правою частиною в нормі C (ω_h), в якій і права частина оцінюватиметься в нормі C (ω_h) (зрозуміло, що слід прагнути знайти оцінки з якомога сильнішою нормою зліва і якомога слабшою справа).

існо, запишемо схему (6) у вигляді

$$\frac{\sigma\tau}{h^2}y_{i-1}^{j+1}-\left(1+\frac{2\sigma\tau}{h^2}\right)y_i^{j+1}+\frac{\sigma\tau}{h^2}y_{i+1}^{j+1}=-F_i^j,\quad i=\overline{1,\ N-1},$$

$$y_0^{j+1}=0,\quad y_N^{j+1}=0,\quad y_i^0=0,\quad i=\overline{0,\ N}, \tag{20}$$

$$F_i^j=\left(1-\frac{2(1-\sigma)\tau}{h^2}\right)y_i^j+\frac{(1-\sigma)\tau}{h^2}(y_{i-1}^j+y_{i+1}^j)+\tau\varphi_i^j.$$

що коефіцієнт при y_i^j невід'ємний, тобто

$$\tau\leqslant\frac{h^2}{2(1-\sigma)},\quad \sigma\geqslant 1-\frac{h^2}{2\tau}, \tag{21}$$

$$\|F^j\|_{C(\omega_h)}\leqslant\|y^j\|_{C(\omega_h)}+\tau\|\varphi^j\|_{C(\omega_h)},$$

скориставшись теоремою з п. 2.2, зі схеми (20) дістанемо

$$\|y^{j+1}\|_{C(\omega_h)}\leqslant\|y^j\|_{C(\omega_h)}+\tau\|\varphi^j\|_{C(\omega_h)}$$

алі

$$\|y^j\|_{C(\omega_h)}\leqslant\sum_{k=0}^{j-1}\tau\|\varphi^k\|_{C(\omega_h)}. \tag{22}$$

ри $\sigma\geqslant\frac{1}{2}$ слід надати перевагу оцінці (17), оскільки в порівнянні (22) права частина в ній знайдена в слабшій нормі.

Із виразу (21) випливає, що для явної схеми ($\sigma=0$) оцінка (22) обто стійкість за правою частиною в $C(\omega_h)$) має місце за умови $\tau\leqslant\frac{h^2}{2}$, а для симетричної схеми $\sigma=\frac{1}{2}$ за умови $\tau\leqslant h^2$. З іншого боку, зумовна стійкість виду (17) (тобто при використанні іншої норми правій частині) має місце при $\sigma\geqslant\frac{1}{2}$, тобто і для симетричної схеми. ей приклад показує певну умовність поняття умовної та безумовної ійкості, яке може залежати від вибору норм (а також і від методу ведення нерівності стійкості).

6.2.2. Оцінка похибки апроксимації і швидкості збіжності. Приустимо, що $u\in W_2^3(Q_T)$, тобто функція має треті похідні, інтегровні квадратом. Виберемо $\varphi=f$ і подамо похибку апроксимації у вигляді

$$\psi=\Lambda u^{(\sigma)}+\varphi-u_t=u_{\bar{x}x}^{(\sigma)}(x,\ t)-\frac{\partial^2u(x,\ t)}{\partial x^2}+\frac{\partial u(x,\ t)}{\partial t}-$$

$$-u_t(x,\ t)=\eta_1(x,\ t)+\eta_2(x,\ t),$$

$$(x,\ t)\in\omega_{h\tau},$$

де

$$\eta_1(x,\ t)\equiv\eta_1(x,\ t,\ u)\equiv\eta_1(u)=u_{\bar{x}x}^{(\sigma)}-\frac{\partial^2u(x,\ t)}{\partial x^2},$$

$$\eta_2(x,\ t)=\eta_2(x,\ t;\ u)\equiv\eta_2(u)=\frac{\partial u(x,\ t)}{\partial t}-u_t(x,\ t).$$

Теорема вкладення дає $W_2^3(Q_T)\subset C(\bar{Q}_T)$, тому неважко після масштабування знайти обмеженість функціоналів η_1, η_2 в $W_2^3(Q_T)$. Неважко також переконатися, що функціонал η_1 перетворюється на нуль на многочленах (від x і t) до другого степеня включно, а функціонал η_2 — на многочленах до першого степеня. Користуючись лемою Бремблa — Гільберта і припускаючи для простоти, що $h=c\tau$, де c — стала, стандартним методом дістаємо оцінки

$$|\eta_1(x,\ t)|\leqslant Mh(h\tau)^{-1/2}|u|_{W_2^3(e(x,t))},$$

$$|\eta_2(x,\ t)|\leqslant Mh(h\tau)^{-1/2}|u|_{W_2^2(e(x,t))},\quad (x,\ t)\in\omega_{h\tau}, \tag{23}$$

де $e(x,\ t)=(x-h,\ x+h)\times(t,\ t+\tau)$, M — стала, що не залежить від h, τ, u. Звідси маємо

$$\sum_{k=0}^{j-1}\tau\|\psi^k\|^2=\sum_{k=0}^{j-1}\tau\left(\sum_{j=1}^{N-1}h(\eta_1(x_i,\ t_k)+\eta_2(x_i,\ t_k))^2\right)\leqslant$$

$$\leqslant 2\sum_{k=0}^{L-1}\tau\sum_{i=1}^{N-1}h(\eta_1^2(x_i,\ t_k)+\eta_2^2(x_i,\ t_k))\leqslant 4h^2M^2(|u|_{W_2^3(Q_T)}^2+|u|_{W_2^2(Q_T)}^2).$$

За допомогою цієї оцінки і нерівності (17) дістаємо таке твердження.

Теорема 1. *Нехай розв'язок задачі* (5) *з п.* 6.1 *належить простору* $W_2^3(Q_T)$, $Q_T=(0,\ 1)\times(0,\ T)$. *Тоді розв'язок схеми з ваговими коефіцієнтами* (6) *при* $\varphi=f$, $\sigma\geqslant\frac{1}{2}$, $h=c\tau$, *де* c *— стала, збігається при* $h,\ \tau\to 0$ *до точного розв'язку, причому має місце оцінка швидкості збіжності*

$$\|y-u\|_{C(\omega_{h\tau})}\leqslant Mh(|u|_{W_2^3(Q_T)}+|u|_{W_2^2(Q_T)}),$$

де M *— стала, що не залежить від* h, τ, u.

Зауважимо, що теорема 1 досить груба, оскільки з рівняння (5), п. 6.1, видно, що вимоги гладкості по t і по x мають бути різні, і, крім того, для безумовно стійкої схеми $\left(\sigma\geqslant\frac{1}{2}\right)$ ми наклали умову $h=c\tau$. Від цих недоліків можна позбутись, розглядаючи задачу (5) з п. 6.1 в так званих анізотропних просторах $W_p^{k,m}(Q_T)$, де k вказує на гладкість розв'язку u по x, а m — по t. Теорема 1 є лише прикладом, як за допомогою стандартної методики можна знизити вимоги до гладкості розв'язку.

Розглянемо задачу (5) з п. 6.1 в просторах $C_k^m(\bar{Q}_T)$, де k показує, скільки неперервних похідних має $u(x, t)$ по x, а m — скільки по t. Вважаючи, що $u \in C_6^3(\bar{Q}_T)$, розкладемо $u(x, t)$ в околі точки $(x_i, \bar{t})$, $\bar{t} = t_j + \frac{\tau}{2}$. Позначаючи $\bar{v} = v(x, \bar{t})$, маємо

$$\hat{v} \equiv v(t+\tau) = \bar{v} + \frac{1}{2}\tau\frac{\partial\bar{v}}{\partial t} + \frac{\tau^2}{8}\frac{\partial^2\bar{v}}{\partial t^2} + \frac{\tau^3}{48}\frac{\partial^3 v}{\partial t^3} + O(\tau^4),$$

$$v \equiv v(t) = \bar{v} - \frac{1}{2}\tau\frac{\partial\bar{v}}{\partial t} + \frac{\tau^2}{8}\frac{\partial^2\bar{v}}{\partial t^2} - \frac{\tau^3}{48}\frac{\partial^3\bar{v}}{\partial t^3} + O(\tau^4).$$

Враховуючи, крім того, рівності

$$u^{(\sigma)} = \sigma\hat{u} + (1-\sigma)u = \frac{u+\hat{u}}{2} + \left(\sigma - \frac{1}{2}\right)\tau u_t,$$

$$\Lambda u = u_{\bar{x}x} = Lu + \frac{h^2}{12}u^{IV} + O(h^4), \quad Lu = \frac{\partial^2 u}{\partial x^2},$$

маємо

$$\psi = \left(L\bar{u} + \bar{f} - \frac{\partial\bar{u}}{\partial t}\right) + \varphi - \bar{f} + \left(\sigma - \frac{1}{2}\right)\tau L\frac{\partial\bar{u}}{\partial t} + \frac{h^2}{12}L^2\bar{u} + O(\tau^2 + h^4).$$

Оскільки в силу рівняння (5), п. 6.1, $L\bar{u} + \bar{f} - \frac{\partial\bar{u}}{\partial t} = 0$, то

$$L\frac{\partial u}{\partial t} = L^2 u + Lf,$$

$$\psi = \left(\varphi - \bar{f} + \left(\sigma - \frac{1}{2}\right)\tau L\bar{f}\right) + \left(\frac{h^2}{12} + \left(\sigma - \frac{1}{2}\right)\tau\right)L^2 u + O(\tau^2 + h^4).$$

Звідси

$\psi = O(\tau + h^2)$ при $\varphi = \bar{f}$, $\sigma \neq \frac{1}{2}$, $u \in C_k^m(\bar{Q}_T)$, $k \geqslant 4$, $m \geqslant 2$,

$\psi = O(\tau^2 + h^2)$ при $\varphi = \bar{f}$, $\sigma = \frac{1}{2}$, $u \in C_k^m(\bar{Q}_T)$, $k \geqslant 4$, $m \geqslant 3$.

Якщо ж вибрати при $u \in C_6^3$

$$\sigma = \sigma_* = \frac{1}{2} - \frac{h^2}{12\tau}, \quad \varphi = \bar{f} - \left(\sigma_* - \frac{1}{2}\right)\tau L\bar{f} = \bar{f} + \frac{h^2}{12}L\bar{f},$$

то дістанемо схему підвищеного порядку апроксимації $O(h^4 + \tau^2)$. Ця схема неявна і тому y_i^{j+1} визначається з рівняння $\sigma_*\tau\Lambda\hat{y} - \hat{y} = -F$ і відповідних крайових умов методом прогонки. Зауважимо,

що коли взяти $\sigma = \sigma_*$ і $\varphi = \bar{f} + \frac{h^2}{12}\Lambda\bar{f}$, то також дістанемо схему порядку апроксимації $O(h^4 + \tau^2)$, бо $\Lambda f - Lf = O(h^2)$ при $f \in C_4^0(Q_T)$. Із викладеного вище і оцінки (17) випливає така теорема.

Теорема 2. *Якщо в схемі* (1), (2) $\sigma > \frac{1}{2}$, *то при* $u \in C_k^m(\bar{Q}_T)$, $k \geqslant 4$, $m \geqslant 2$ *розв'язок цієї схеми збігається в нормі* $C(\omega_{h\tau})$ *до точного розв'язку и задачі* (5) *з п.* 6.1 *і має місце оцінка швидкості збіжності*

$$\|y - u\|_{C(\omega_{h\tau})} \leqslant M(\tau + h^2), \tag{24}$$

якщо $\sigma = \frac{1}{2}$, $u \in C_k^m(\bar{Q}_T)$, $k \geqslant 4$, $m \geqslant 3$, *то має місце оцінка*

$$\|y - u\|_{C(\omega_{h\tau})} \leqslant M(\tau^2 + h^2),$$

де стала M *не залежить від* τ, h.

Враховуючи оцінку (22), дістаємо таке твердження.

Теорема 3. *Якщо в схемі* (1), (2) $\sigma \in \left[2\sigma_0, \frac{1}{2}\right] \equiv \left[1 - \frac{h^2}{2\tau}, \frac{1}{2}\right]$, $u \in C_k^m(\bar{Q}_\tau)$, $k \geqslant 4$, $m \geqslant 2$, *то має місце оцінка* (24); *якщо ж* $\sigma = \sigma_* \in \left[2\sigma_0, \frac{1}{2}\right]$, $u \in C_k^m(\bar{Q}_T)$, $k \geqslant 6$, $m \geqslant 3$, *то*

$$\|y - u\|_{C(\omega_{h\tau})} \leqslant M(\tau^2 + h^4),$$

де стала M *не залежить від* h, τ.

Оцінка (19) приводить до наступної теореми, в якій використовується слабша норма, але розширюється інтервал для σ_*.

Теорема 4. *Якщо в схемі* (1), (2) $\sigma = \sigma_0 \in [\tilde{\sigma}_0, 1/2)$, $u \in C_k^m(\bar{Q}_T)$, $k \geqslant 6$, $m \geqslant 3$, *то розв'язок цієї схеми збігається до точного розв'язку и задачі* (5), *п.* 6.1, *причому має місце оцінка швидкості збіжності*

$$\|y - u\| \leqslant M(\tau^2 + h^4),$$

де стала M *не залежить від* τ, h.

6.2.3. Асимптотична стійкість схеми з ваговими коефіцієнтами. Розглянемо задачу

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad x \in (0, 1), \quad t > 0,$$

$$u(0, t) = u(1, t) = 0, \quad t \geqslant 0, \tag{25}$$

$$u(x, 0) = u_0(x), \quad x \in [0, 1].$$

Методом розділення змінних розв'язок цієї задачі можна знайти у вигляді

$$u(x, t) = \sum_{k=1}^{\infty} c_k e^{-\lambda_k t} X_k(x), \tag{26}$$

де λ_k, $X_k(x)$ — власні значення та ортонормовані власні функції задачі

$$X'' + \lambda X = 0, \quad x \in (0, 1); \quad X(0) = X(1) = 0.$$

Неважко переконатись, що

$$\lambda_k = k^2\pi^2, \quad X_k(x) = \sqrt{2}\sin k\pi x,$$

причому

$$(X_k, X_m) = \int_0^1 X_k(x) X_m(x)\, dx = \delta_{km},$$

δ_{km} — символ Кронекера. Дійсно, вираз (26) є розв'язком задачі (25), бо всі частинні розв'язки (або гармоніки) $u_k(x, t) = c_k e^{-\lambda_k t} X_k(x)$ задовольняють рівняння (25) і крайові умови, а з початкової умови

$$u(x, 0) = u_0(x) = \sum_{k=1}^{\infty} c_k X_k(x)$$

визначаються коефіцієнти $c_k = (u_0, x_k)$. Оскільки

$$\|u_0\|^2 = \sum_{k=1}^{\infty} c_k^2, \quad \ldots, \quad \lambda_k > \lambda_{k-1} > \cdots > \lambda_1 = \pi^2,$$

то з виразу (26) маємо

$$\|u(t)\|^2 = \sum_{k=1}^{\infty} c_k^2 e^{-2\lambda_k t} \|x_k\|^2 \leqslant e^{-2\lambda_1 t} \sum_{k=1}^{\infty} c_k^2 = e^{-2\lambda_1 t} \|u_0\|^2. \qquad (27)$$

Ця оцінка означає асимптотичну при $t \to \infty$ стійкість задачі (25) за початковими умовами. Оскільки $\lambda_k = k^2\pi^2 \to \infty$ при $k \to \infty$, то з ростом t у сумі (26) почне переважати перший доданок (перша гармоніка), тобто приблизно матимемо (при великих t)

$$u(x, t) \approx c_1 e^{-\lambda_1 t} X_1(x). \qquad (26')$$

Ця стадія процесу, який описується математичною моделлю (25), називається *регулярним режимом*.

При розв'язуванні практичних задач методом сіток природно вимагати, щоб сітковий розв'язок відображав аналітичні властивості точного, зокрема, у випадку задачі (25), щоб він мав властивість (26'). Як видно з означення 6 попереднього параграфа, для цього схема має бути асимптотично стійкою.

Розглянемо для задачі (25) схему з ваговими коефіцієнтами (1), (2):

$$y_t(x) = \sigma\Lambda\hat{y}(v) + (1 - \sigma)\Lambda y(x), \quad x \in \omega_h, \quad t \in \omega_\tau;$$

$$y(0, t) = 0, \quad y(1, t) = 0, \quad t \in \omega_\tau; \quad y(x, 0) = u_0(x), \quad x \in \omega_h,$$

або в канонічному операторному вигляді

$$By_t + Ay = 0, \quad t \in \omega_\tau,$$
$$y(0) = u_0, \quad B = I + \sigma\tau A,$$

$$A : \Omega_{N-1} \to \Omega_{N-1}, \quad Ay = -\overset{\circ}{y}_{\bar{x}x}, \quad \overset{\circ}{y} \in \overset{\circ}{\Omega}_{N+1}, \quad \Omega_{N-1} —$$

простір сіткових функцій $y(x)$, $x \in \omega_h$ зі скалярним добутком $(y, v) = \sum_{x \in \omega_h} hy(x) v(x)$. Оскільки оператор A самоспряжений і додатний, то

$$\delta I \leqslant A \leqslant \Delta I, \qquad (28)$$

де δ, Δ — його відповідно найменше і найбільше власні значення. Відомо (див. п. 1.4.4 з ч. 1), що

$$\delta = \frac{4}{h^2}\sin^2\frac{\pi h}{2}, \quad \Delta = \frac{4}{h^2}\cos^2\frac{\pi h}{2}.$$

Користуючись теоремою 4, п. 1.4 з ч. 1, при

$$\|y^j\|_A \leqslant \rho \|y^{j-1}\|_A \leqslant \rho^j \|y^0\|_A,$$

матимемо

$$\rho = |1 - \gamma\tau| < 1, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \tau \leqslant \tau_0, \qquad (29)$$

де γ_1, γ_2 — сталі з нерівності

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B.$$

Обчислимо ці сталі для схеми з ваговими коефіцієнтами. Маємо (з урахуванням (28))

$$B = I + \sigma\tau A \geqslant \left(\frac{1}{\Delta} + \sigma\tau\right) A = \frac{1}{\gamma_2} A,$$

$$B \leqslant \left(\frac{1}{\delta} + \sigma\tau\right) A = \frac{1}{\gamma_1} A,$$

звідки

$$\gamma_1 = \frac{\delta}{1 + \sigma\tau\delta}, \quad \gamma_2 = \frac{\Delta}{1 + \sigma\tau\Delta}. \qquad (30)$$

Із виразу (29) знаходимо оцінку, яка виражає властивість асимптотичної стійкості (при $t_j \to \infty$)

$$\|y^j\|_A \leqslant e^{-\gamma_1 t_j} \|y^0\|_A. \qquad (31)$$

Розглянемо найбільш вживані схеми.

Для явної схеми $\sigma = 0$, $\gamma_1 = \delta$, $\gamma_2 = \Delta$ і умова асимптотичної

стійкості (чи нерівності (29)) має вигляд

$$\tau \leqslant \tau_0 = \frac{2}{\delta + \Delta} = \frac{h^2}{2}, \tag{32}$$

тобто умова асимптотичної стійкості збігається з умовою звичайної стійкості.

При $\sigma \neq 0$ умова $\tau \leqslant \tau_0 = \frac{2}{\gamma_1 + \gamma_2}$ приводить до нерівності

$$2 + 2\left(\sigma - \frac{1}{2}\right)\tau(\delta + \Delta) - 2\sigma(1 - \sigma)\tau^2\delta\Delta \geqslant 0. \tag{33}$$

Для неявної схеми $\sigma = 1$ ця нерівність виконується для всіх τ, тобто неявна схема є і асимптотично безумовно стійкою.

Для симетричної схеми $\sigma = \frac{1}{2}$ із (33) маємо умову асимптотичної стійкості

$$\tau \leqslant \tau_0 = \frac{2}{\sqrt{\delta_\Delta}} = \frac{h^2}{\sin \pi h} \approx \frac{h}{\pi}, \tag{33'}$$

тобто симетрична схема умовно асимптотично стійка, хоча вона безумовно стійка в звичайному розумінні.

Розв'язок симетричної схеми

$$y_t = \frac{1}{2}\Lambda\hat{y} + \frac{1}{2}\Lambda y, \quad y(0, t) = y(1, t) = 0,$$

$$y(x, 0) = u_0(x) \tag{34}$$

можна знайти явно методом розділення змінних. Шукатимемо його у вигляді

$$y^j = \sum_{k=1}^{N-1} c_k X_k^h(x)\rho_k^j, \tag{35}$$

де $X_k^h(x) = \sqrt{2}\sin \pi k x$ — власна функція оператора $-\Lambda y = -y_{\bar{x}x}$, $y(0) = y(1) = 0$, яка відповідає власному значенню $\lambda_k = \frac{4}{h^2}\sin^2\frac{\pi k h}{2}$ (див. п. 1.4.4 з ч. 1). Підставляючи (35) в (34), маємо

$$\tau^{-1}\sum_{k=1}^{N-1} c_k X_k^h(x)(\rho_k^{j+1} - \rho_k^j) + \sum_{k=1}^{N-1} c_k \frac{2}{h^2}\sin^2\frac{\pi k h}{2} X_k^h(x)(\rho_k^{j+1} + \rho_k^j) = 0,$$

або

$$\sum_{k=1}^{N-1} c_k X_k^h(x)\left[\frac{\rho_k^{j+1} - \rho_k^j}{\tau} + \frac{2}{h^2}\sin^2\frac{\pi k h}{2}(\rho_k^{j+1} + \rho_k^j)\right] = 0.$$

Звідси для визначення ρ_k маємо рівняння

$$\left(1 + \frac{\tau}{2}\lambda_k\right)\rho_k^{j+1} - \left(1 - \frac{\tau}{2}\lambda_k\right)\rho_k^j = 0,$$

що дає

$$\rho_k = \frac{1 - \frac{\tau}{2}\lambda_k}{1 + \frac{\tau}{2}\lambda_k} = \frac{1 - \frac{2\tau}{h^2}\sin^2\frac{\pi k h}{2}}{1 + \frac{2\tau}{h^2}\sin^2\frac{\pi k h}{2}}. \tag{36}$$

Сталі c_k визначаються з початкової умови

$$y_0 = \sum_{k=1}^{N-1} c_k X_k^h(x) = u_0(x) = \sum_{k=0}^{N-1} \bar{c}_k X_k^h(x),$$

$$c_k = \bar{c}_k = (u_0, X_k^h).$$

Далі нам потрібне буде таке допоміжне твердження.

Лема. *Нехай* $\varphi(\tau, x) = \frac{1 - \tau x}{1 + \tau x}$ *— функція дійсних змінних з областю визначення* $0 < a \leqslant x \leqslant b < \infty$, $\tau > 0$ *і нехай* $\bar{q} = \min\limits_{\tau}\max\limits_{x\in[a,b]} |\varphi(\tau, x)|$. *Тоді*

$$\bar{q} = \varphi(\tau_0, a) = |\varphi(\tau_0, b)| = \frac{1 - \sqrt{a/b}}{1 + \sqrt{a/b}},$$

$$\tau_0 = \frac{1}{\sqrt{ab}}.$$

Крім того, якщо $0 < \tau < \tau_0$, то $|\varphi(\tau, a)| \geqslant |\varphi(\tau, b)|$, а при $\tau > \tau_0$ виконується нерівність $|\varphi(\tau, a)| < |\varphi(\tau, b)|$.

Доведення. Для будь-якого $\tau > 0$ функція $\varphi(\tau, x)$ строго монотонна при $x \in [a, b]$, причому $\varphi(\tau, 0) = 1$, $\varphi(\tau, +\infty) = -1$, значить $\varphi(\tau, x) < 1$. Якщо τ зростає, то $\varphi(\tau, x)$ спадає для всіх $x > 0$ і $\varphi(\tau, x) = 0$ при $\tau x = 1$. Звідси зрозуміло, що $\max\limits_{x\in[a,b]} |\varphi(\tau, x)|$ буде мінімальним для таких τ, для яких

$$\varphi(\tau, a) = -\varphi(\tau, b),$$

або

$$\frac{1 - \tau a}{1 + \tau a} = -\frac{1 - \tau b}{1 + \tau b},$$

тобто $\tau_0 = \frac{1}{\sqrt{ab}}$. Щоб довести друге твердження леми, розглянемо функцію

$$\psi(\tau) = |\varphi(\tau, a)| - |\varphi(\tau, b)| = \frac{|1 - \tau a|}{1 + \tau a} - \frac{|1 - \tau b|}{1 + \tau b}.$$

На проміжку $\tau \in (0, +\infty)$ вона, очевидно, неперервна і її можна подати у вигляді

$$\psi(\tau) = \begin{cases} \dfrac{1-\tau a}{1+\tau a} - \dfrac{1-\tau b}{1+\tau b}, & \tau \in \left(0; \dfrac{1}{b}\right), \\ \dfrac{1-\tau a}{1+\tau a} + \dfrac{1-\tau b}{1+\tau b}, & \tau \in \left[\dfrac{1}{b}, \dfrac{1}{a}\right), \\ \dfrac{1-\tau b}{1+\tau b} - \dfrac{1-\tau a}{1+\tau a}, & \tau \in \left[\dfrac{1}{a}, \infty\right). \end{cases}$$

Крім того, неважко впевнитись, що $\psi(\tau)$ перетворюється на нуль лише в одній точці $\tau_0 = \dfrac{1}{\sqrt{ab}} \in \left(\dfrac{1}{b}, \dfrac{1}{a}\right)$. Оскільки $\psi\left(\dfrac{1}{b}\right) = \dfrac{1-a/b}{1+a/b} > 0$, то $\psi(\tau) \geqslant 0$ для $0 < \tau \leqslant \tau_0$ і $\psi(\tau) < 0$ для $\tau > \tau_0$; а це і доводить твердження леми.

З леми випливає, що при виконанні умови (33′) для симетричної схеми матимемо $\max\limits_k \rho_k = |\rho_1|$ і в сумі (35) переважатиме перша гармоніка (при великих j або t_j), тобто

$$y_i^j \approx c_1 \rho_1^j \sin \pi x_i = c_1 \left(\frac{1-\tau\delta/2}{1+\tau\delta/2}\right)^j \sin \pi x_i.$$

Оскільки

$$\frac{1-\tau\delta/2}{1+\tau\delta/2} = \frac{1-\dfrac{\tau}{2}\dfrac{4}{h^2}\sin^2\dfrac{\pi h}{2}}{1+\dfrac{\tau}{2}\dfrac{4}{h^2}\sin^2\dfrac{\pi h}{2}} = \frac{1-\dfrac{\tau\pi^2}{2}+O(\tau h^3)}{1+\dfrac{\tau\pi^2}{2}+O(\tau h^3)} =$$

$$= \left(1 - \frac{\tau\pi^2 + O(\tau h^3)}{1+\tau\pi^2/2 + O(\tau h^3)}\right) = e^{-\pi^2\tau}(1 + O(\tau h^2)),$$

то

$$y_i^j \approx c_1 e^{-\pi^2 t_j} \sin \pi x_i,$$

і ми бачимо, що симетрична схема правильно передає поведінку точного розв'язку при $t \to \infty$. Якщо ж умова (33) не виконується, то, як випливає з леми 1, $|\rho_{N-1}| > |\rho_k|$, $k = \overline{1, N-2}$, і тому в сумі (35) переважатиме остання гармоніка, тобто

$$y_i^j \approx c_{N-1}\rho_{N-1}^j \sin \pi (N-1) x_i \approx c_{N-1}\rho_{N-1}^j (-1)^i \sin \pi x_i,$$

де

$$\rho_{N-1} = \frac{1-\tau\Delta/2}{1+\tau\Delta/2} < e^{-\pi^2\tau},$$

і ми бачимо, що цей розв'язок не має нічого спільного з точним розв'язком задачі (25).

Розглянутий приклад показує, що при чисельному розв'язуванні рівняння теплопровідності на великих проміжках часу вимога асимптотичної стійкості схеми є суттєвою і звичайної стійкості може бути не досить. Зауважимо, що умова $\tau \approx \dfrac{h}{\pi}$ для симетричної схеми не є жорсткою. Більше того, можна довести, що чисто неявна схема ($\sigma = 1$), яка є безумовно асимптотично стійкою, при великих t може забезпечити потрібну точність лише при кроках τ того самого порядку, що і в явній схемі. Ця обставина при розрахунках для великих t позбавляє чисто неявну схему її основної переваги — стійкості для будь-яких τ, h.

6.2.4. Триярусна схема з ваговими коефіцієнтами для рівняння теплопровідності.

Задачу (5) з п. 6.1 можна апроксимувати таким чином:

$$(1+\sigma)\, y_t^j - \sigma y_{\bar{t}}^j = y_{\bar{x}x}^{j+1}(x) + \varphi^{j+1}(x), \quad x \in \omega_h, \quad j = 1, 2, \ldots,$$

$$y^j(0) = u_1(t_j), \quad y^j(1) = u_2(t_j), \tag{37}$$

$$y^0(x) = u_0(x), \quad x \in \bar{\omega}_h.$$

Така різницева схема містить три яруси по часу: t_{j-1}, t_j, t_{j+1}, і для того щоб задача (37) розв'язувалася однозначно, треба ще задати $y_i^1 = y^1(x_i, \tau)$, $i = \overline{1, N-1}$. Функцію $y^1(x)$, $x \in \omega_h$ можна визначати по-різному, але природно шукати для неї таку апроксимацію, порядок якої узгоджується з порядком апроксимації співвідношень (37). Тому дослідимо спочатку похибку апроксимації рівняння теплопровідності схеми (37). Для похибки апроксимації маємо вираз

$$\psi_i^{j+1} = \varphi_i^{j+1} + \Lambda u^{j+1} - (1+\sigma)\, u_t^j + \sigma u_{\bar{t}}^j,$$

$$\Lambda u^{j+1} = u_{\bar{x}x}^{j+}$$

Користуючись формулою Тейлора, дістанемо в околі точки (x_i, t_{j+1})

$$\psi_i^{j+1} = \varphi_i^{j+1} + \left(\frac{\partial^2 u^{j+1}}{\partial x^2} + \frac{h^2}{12}\frac{\partial^4 u_i^{j+1}}{\partial x^4}\right) -$$

$$-(1+\sigma)\left(\frac{\partial u_i^{j+1}}{\partial t} - \frac{\tau}{2}\frac{\partial^2 u_i^{j+1}}{\partial t^2}\right) +$$

$$+\sigma\left(\frac{\partial u_i^{j+1}}{\partial t} - \frac{3}{2}\tau\frac{\partial^2 u_i^{j+1}}{\partial t}\right) + O(\tau^2 + h^4) =$$

$$= \varphi_i^{j+1} - f_i^{j+1} + \frac{h^2}{12}\frac{\partial^4 u_i^{j+1}}{\partial x^4} + \tau\left(\frac{1}{2} - \sigma\right)\frac{\partial^2 u_i^{j+1}}{\partial t^2} + O(\tau^2 + h^4). \tag{38}$$

Диференціюючи рівняння (5) з п. 6.1 по t, дістанемо рівність

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^4 u}{\partial x^4} + \frac{\partial^2 f}{\partial x^2} + \frac{\partial f}{\partial t},$$

з урахуванням якої вираз (38) набере вигляду

$$\psi_i^{j+1} = \varphi_i^{j+1} - f_i^{j+1} + \left[\frac{h^2}{12} + \tau\left(\frac{1}{2} - \sigma\right)\right]\frac{\partial^2 u_i^{j+1}}{\partial t^2} - \\ - \frac{h^2}{12}\left(\frac{\partial^2 f_i^{j+1}}{\partial x^2} + \frac{\partial f_i^{j+1}}{\partial t}\right) + O(\tau^2 + h^4).$$

Звідси маємо:

1) якщо $\sigma = \frac{1}{2} + \frac{h^2}{12\tau}$, $\varphi_i^{j+1} = f_i^{j+1} + \frac{h^2}{12}\left(\frac{\partial^2 f_i^{j+1}}{\partial x^2} + \frac{\partial f_i^{j+1}}{\partial t}\right)$, $u \in C_x^2(Q_T)$, то

$$\psi_i^{j+1} = O(\tau^2 + h^4); \tag{39}$$

2) якщо $\sigma = \frac{1}{2}$, $\varphi_i^{j+1} = f_i^{j+1}$, $u \in C_4^3$, то

$$\psi_i^{j+1} = O(\tau^2 + h^2); \tag{40}$$

3) якщо $\sigma \in (-\infty; \infty)$, $\sigma \neq \frac{1}{2}$, $\sigma \neq \frac{1}{2} + \frac{h^2}{12\tau}$, $\varphi_i^{j+1} = f_i^{j+1}$, $u \in C_4^2$, то

$$\psi_i^{j+1} = O(\tau + h^2). \tag{41}$$

Щоб знайти y^1, скористаємося таким розвиненням за формулою Тейлора:

$$\frac{\partial u_i^1}{\partial t} = \frac{\partial u_i^0}{\partial t} + \tau\frac{\partial^2 u_i^0}{\partial t^2} + O(\tau^2) = -\frac{\partial^2 u_i^0}{\partial x^2} + f_i^0 + \\ + \tau\left(\frac{\partial^4 u_i^0}{\partial x^4} + \frac{\partial^2 f_i^0}{\partial x^2} + \frac{\partial f_i^0}{\partial t}\right) + O(\tau^2)$$

$$(u \in C_4^3(\bar{Q}_T), \quad f \in C_2^1(\bar{Q}_T)).$$

Апроксимуючи

$$\frac{\partial u_i^1}{\partial t} = \frac{u_i^1 - u_i^0}{\tau} + O(\tau) \quad (u \in C_0^2(\bar{Q}_T))$$

і використовуючи попередню рівність, матимемо

$$u_i^1 = u_i^0 + \tau\frac{\partial^2 u_i^0}{\partial x^2} + \tau f_i^0 + \tau^2\left(\frac{\partial^4 u_i^0}{\partial x^4} + \frac{\partial^2 f_i^0}{\partial x^2} + \frac{\partial f_i^0}{\partial t}\right) + O(\tau^2).$$

Оскільки $u_i^0 = u_0(x)$ — відома функція, то, поклавши

$$y_i^1 = u_{0i} + \tau u_{0i}'' + \tau^2 u_{0i}^{(4)} + \tau f_i^0 + \tau^2\frac{\partial^2 f_i^0}{\partial x^2} + \tau^2\frac{\partial f_i^0}{\partial t} \tag{42}$$

і враховуючи (39) — (41), для задачі (5) дістанемо триярусну різницеву схему (37), (42), яка при $\sigma = \frac{1}{2}$, $\varphi_i^{j+1} = f_i^{j+1}$, $u \in C_4^3(Q_T)$, $f \in C_2^1(\bar{Q}_T)$ матиме порядок апроксимації $O(\tau^2 + h^2)$, а при $\sigma = \frac{1}{2} + \frac{h^2}{12\tau}$, $\varphi_i^{j+1} = f_i^{j+1} + \frac{h^2}{12}\left(\frac{\partial^2 f_i^{j+1}}{\partial x^2} + \frac{\partial f_i^{j+1}}{\partial t}\right)$, $u \in C_6^3(\bar{Q}_T)$, $f \in C_2^1(\bar{Q}_T)$ — порядок апроксимації $O(\tau^2 + h^4)$.

Значення y_i^{j+1}, $i = \overline{1, N-1}$ на $(j+1)$-му часовому ярусі визначається методом прогонки із системи триточкових рівнянь. Наприклад, при $\sigma = 1/2$, $\varphi_i^{j+1} = f_i^{j+1}$ схема (37), (42) матиме вигляд

$$y_{i-1}^{j+1} - \left(2 + \frac{3}{2}\frac{h^2}{\tau}\right)y_i^{j+1} + y_{i-1}^{j+1} = F_i^{j+1}, \quad i = \overline{1, N-1}; \quad j = 1, 2, \ldots,$$

$$F_i^{j+1} = -h^2 f_i^{j+1} - \frac{3h^2}{2\tau}y_i^j + \frac{h^2}{2\tau}y_i^{j-1},$$

$$y_0^{j+1} = u_1(t_{j+1}), \quad y_N^{j+1} = u_2(t_{j+1}); \quad y_2^0 = u_0(x_i), \quad i = \overline{0, N},$$

а y_i^1, $i = \overline{1, N-1}$ визначаються за формулою (42).

6.3. Метод сіток для багатовимірних задач параболічного типу

Як і у випадку еліптичних задач, підвищення розмірності задачі параболічного типу (тобто збільшення кількості незалежних змінних), як правило, не потребує нових чисельних методів. Але зі збільшенням розмірності ростуть технічні труднощі при побудові та дослідженні цих методів. Багатовимірність задачі приводить до збільшення обсягу обчислень (кількості арифметичних операцій), а також вносить і деякі притаманні лише таким задачам риси чисельних методів, без врахування яких ці методи не ведуть до шуканого результату. Зазначимо, що багатовимірність задачі в цілому є, так би мовити, «сильним ворогом» для дослідника і в боротьбі з ним поки що досягнуто лише окремих успіхів, а не «широкого прориву».

6.3.1. Різницева схема з ваговими коефіцієнтами для двовимірного рівняння теплопровідності. На площині $x = (x_1, x_2)$ розглянемо область Ω з границею $\partial\Omega$. Розглянемо початково-крайову задачу першого роду (задачу Діріхле для рівняння теплопровідності) для двовимірного рівняння теплопровідності в циліндрі $\bar{Q}_T = \bar{\Omega} \times [0, T]$:

$$\frac{\partial u}{\partial t} = Lu + f(x, t), \quad (x, t) \in Q_T \equiv \Omega \times (0, T], \tag{1}$$

$$u(x, t) = \mu(x, t), \quad x \in \partial\Omega, \quad t \in [0, T], \tag{2}$$

$$u(x, 0) = u_0(x), \quad x \in \bar{\Omega}, \quad Lu = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}. \tag{3}$$

Тут $u(x,\ t)$ — шукана функція; f, μ, u_0 — задані функції; умова (2) називається крайовою; умова (3) — початковою. Далі вважатимемо, що $\bar{\Omega}$ — прямокутник,

$$\bar{\Omega} = \{x = (x_1,\ x_2) : 0 \leqslant x_\alpha \leqslant l_\alpha,\quad \alpha = 1,\ 2\}.$$

Покриємо $\bar{\Omega}$ прямокутною сіткою $\bar{\omega}_h = \bar{\omega}_1 \times \bar{\omega}_2$, де

$$\bar{\omega}_\alpha = \left\{x_\alpha = i_\alpha h_\alpha : i_\alpha = \overline{0,\ N_\alpha},\ h_\alpha = \frac{l_\alpha}{N_\alpha}\right\},\quad \alpha = 1,\ 2.$$

Границею сітки $\bar{\omega}_h$ називатимемо множину вузлів

$$\gamma_h = \{x_i = (i_1 h_1,\ i_2 h_2) : i_1 = 0,\ N_1,\ 0 < i_2 < N_2;\ i_2 = 0,\ N_2,\ 0 < i_1 < N_1\}.$$

Введемо також сітку із внутрішніх вузлів $\omega_h = \omega_1 \times \omega_2$,

$$\omega_\alpha = \left\{x_\alpha = i_\alpha h_\alpha : i_\alpha = \overline{1,\ N_\alpha - 1};\ h_\alpha = \frac{l_\alpha}{N_\alpha}\right\},\quad \alpha = 1,\ 2.$$

На відрізку $t \in [0,\ T]$ введемо сітки з кроком τ:

$$\bar{\omega}_\tau = \left\{t_j = j\tau : j = \overline{0,\ L},\ \tau = \frac{T}{L}\right\},$$

$$\omega_\tau = \left\{t_j = j\tau : j = \overline{1,\ L},\ \tau = \frac{T}{L}\right\}.$$

Замінюючи похідні різницевими співвідношеннями, дістанемо для задачі (1) — (3) таку сіткову (різницеву) схему з ваговими коефіцієнтами:

$$\frac{y^{j+1} - y^j}{\tau} = \Lambda(\sigma y^{j+1} + (1 - \sigma)\, y^j) + \varphi^j,\quad j = 0,\ 1,\ \ldots,$$

$$y(x,\ t) = \mu(x,\ t),\quad x \in \gamma_h,\quad t \in \omega_\tau, \tag{4}$$

$$y^0 = y(x,\ 0) = u_0(x),\quad x \in \bar{\omega}_h,$$

де $y^j = y(x,\ t_j)$ — наближення для $u(x,\ t_j)$, $x \in \omega_h$, $\Lambda y = y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2}$. Щоб знайти розв'язок схеми (4), на новому часовому ярусі t_{j+1} (позначимо цей розв'язок $\hat{y} = y^{j+1}$), маючи $y = y^j$, потрібно розв'язати різницеве рівняння

$$\hat{y} - \sigma\tau\Lambda\hat{y} = F,\quad F = y + (1 - \sigma)\,\tau\Lambda y + \tau\varphi,\quad x \in \omega_h,$$

$$\hat{y} = \mu(x,\ t_{j+1}),\quad x \in \gamma_2. \tag{5}$$

Щоб дослідити це рівняння, введемо відповідні «інструменти» — простори сіткових функцій зі скалярним добутком. Позначимо через Ω_h простір розмірності $(N_1 + 1)(N_2 + 1)$ сіткових функцій, заданих на $\omega_h \cup \gamma_h$, які перетворюються на нуль на границі γ_h, а через

Ω_h — простір сіткових функцій, заданих на ω_h, зі скалярним добутком

$$(y,\ v) = \sum_{x \in \omega_h} h_1 h_2 y(x)\, v(x).$$

Функцію з простору Ω_h, яка на ω_h збігається з функцією $y(x) \in \Omega_h$, позначимо $\overset{\circ}{y}(x)$. Введемо в просторі Ω_h оператор A за формулою

$$Ay(x) = -\Lambda\overset{\circ}{y}(x),\quad x \in \omega_h.$$

Задачу (5) можна, очевидно, записати у вигляді аналогічної задачі з однорідними крайовими умовами, але з іншою правою частиною (див. також п. 5.3)

$$\hat{y} - \sigma\tau\Lambda\hat{y} = F_1(x),\quad x \in \omega_h, \tag{6}$$

$$\hat{y}(x) = 0,\quad x \in \gamma_h,$$

де

$$F_1(x) \equiv F_1^j(x) =$$

$$= \begin{cases}
F(h_1,\ h_2) + \dfrac{\sigma\tau}{h_1^2}\mu(0,\ h_2,\ t_{j+1}) + \dfrac{\sigma\tau}{h_2^2}\mu(h_1,\ 0,\ t_{j+1}),\ x = (h_1,\ h_2),\\
F(h_1,\ x_2) + \dfrac{\sigma\tau}{h_1^2}\mu(0,\ x_2,\ t_{j+1}),\ x = (h_1,\ x_2),\ x_2 = i_2 h_2,\ i_2 = \overline{2,\ N_2 - 2},\\
F(h_1,\ l_2 - h_2) + \dfrac{\sigma\tau}{h_1^2}\mu(0,\ l_2 - h_2,\ t_{j+1}) + \dfrac{\sigma\tau}{h_1^2}\mu(h_1,\ l_2,\ t_{j+1}),\\
x = (h_1,\ l_2 - h_2),\\
F(x_1,\ l_2 - h_2) + \dfrac{\sigma\tau}{h_2^2}\mu(x_1,\ l_2,\ t_{j+1}),\ x = (x_1,\ l_2 - h_2),\quad x_1 = i_1 h_1,\\
i_1 = \overline{2,\ N_1 - 2},\\
F(l_1 - h_1,\ l_2 - h_2) + \dfrac{\sigma\tau}{h_2^2}\mu(l_1 - h_1,\ l_2,\ t_{j+1}) + \dfrac{\sigma\tau}{h_1^2}\mu(l_1,\ l_2 - h_2,\\
t_{j+1}),\quad x = (l_1 - h_1,\ l_2 - h_2),\\
F(l_1 - h_1,\ x_1) + \dfrac{\sigma\tau}{h_1^2}\mu(l_1,\ x_2,\ t_{j+1}),\ x = (l_1 - h_1,\ x_2),\\
x_2 = i_2 h_2,\quad i_2 = \overline{2,\ N_2 - 2},\\
F(l_1 - h_1,\ h_2) + \dfrac{\sigma\tau}{h_1^2}\mu(l_1,\ h_2,\ t_{j+1}) + \dfrac{\sigma\tau}{h_2^2}\mu(l_1 - h_1,\ 0,\ t_{j+1}),\\
x = (l_1 - h_1,\ h_2),\\
F(x_1,\ h_2) + \dfrac{\sigma\tau}{h_2^2}\mu(x_1,\ 0,\ t_{j+1}),\ x = (x_1,\ h_2),\ x_1 = i\,h_1,\ i_1 = \overline{2,\ N_1 - 2},\\
F(x),\ x = (x_1,\ x_2),\ x_\alpha = i_\alpha h_\alpha,\ i_\alpha = \overline{2,\ N_\alpha - 2},\quad \alpha = 1,\ 2.
\end{cases}$$

В операторному вигляді рівність (6) набере вигляду

$$(I + \sigma\tau A)\,\hat{y} = F_1, \quad \hat{y} \in \Omega_h, \quad F_1 \in \Omega_h. \tag{7}$$

Неважко помітити, що власними значеннями оператора A є числа $\lambda_{1k} + \lambda_{2m}$, $k = \overline{1,\ N_1 - 1}$, $m = \overline{1,\ N_2 - 1}$, де $\lambda_{1k} = \frac{4}{h_1^2}\sin^2\frac{\pi k h_1}{2l_1}$, $\lambda_{2m} = \frac{4}{h_2^2}\sin^2\frac{\pi m h_2}{2l_2}$ — власні числа відповідно операторів $A_\alpha v = -v_{\bar{x}_\alpha x_\alpha}(x_\alpha)$, $v(0) = v(l_\alpha) = 0$, $\alpha = 1, 2$, які відповідають ортонормованим власним функціям $v_{1k}(x_1) = \sqrt{2}\sin\frac{k\pi x_1}{l_1}$, $v_{2m}(x_2) = \sqrt{2}\sin\frac{m\pi x_2}{l_2}$. Власними функціями оператора A є функції (ортонормовані)

$$v_{km} = v_{1k}(x_1)\,v_{2m}(x_2) = 2\sin\frac{k\pi x_1}{l_1}\sin\frac{m\pi x_2}{l_2}.$$

Враховуючи нерівності $(Ay, y) \leqslant \|Ay\|\,\|y\| \leqslant \|A\|\,\|y\|^2$,

$$\|A\| = \Delta_0 = \lambda_{1,N_1-1} + \lambda_{2,N_2-1} \text{ та } (Ay, y) \geqslant \delta_0\|y\|^2 > 0,$$

$$\delta_0 = \lambda_{11} + \lambda_{21} = \frac{4}{h_1^2}\sin^2\frac{\pi h_1}{2l_1} + \frac{4}{h_2^2}\sin^2\frac{\pi h_2}{2l_2},$$

маємо

$$((I + \sigma\tau A)\,y,\ y) = \|y\|^2 + \sigma\tau\,(Ay,\ y) \geqslant \left(\frac{1}{\|A\|} + \sigma\tau\right)(Ay,\ y) > 0,$$

тому при $\sigma > -\frac{1}{\tau\|A\|}$ оператор $I + \sigma\tau A$ є додатно визначеним. Це означає, що рівняння (7) (отже, і (5)) має розв'язок. В індексному вигляді (5) має вигляд

$$\sigma\nu_1(\hat{y}_{i_1-1,i_2} + \hat{y}_{i_1+1,i_2}) - (1 + 2\sigma(\nu_1 + \nu_2))\,\hat{y}_{i_1,i_2} +$$
$$+ \sigma\nu_2(\hat{y}_{i_1,i_2-1} + \hat{y}_{i_1,i_2+1}) = -F_{i_1,i_2}, \quad i_\alpha = \overline{1,\ N_{\alpha-1}}, \quad \alpha = 1, 2,$$
$$\hat{y}_{i_1,i_2} = \hat{\mu}_{i_1,i_2}, \quad x_i = (i_1h_1,\ i_2h_2) \in \gamma_h,$$

де

$$\hat{y}_{i_1,i_2} = y(i_1h_1,\ i_2h_2,\ t_{j+1}), \quad \nu_1 = \tau/h_1^2, \quad \nu_2 = \tau/h_2^2,$$
$$F_{i_1,i_2} = (1 - 2(1-\tau)(\nu_1 + \nu_2))\,y_{i_1,i_2} + (1-\sigma)\nu_1(y_{i_1+1,i_2} +$$
$$+ y_{i_1-1,i_2}) + (1-\sigma)\nu_2(y_{i_1,i_2-1} + y_{i_1i_2+1}) + \varphi_{i_1,i_2}. \tag{8}$$

Задача (5) (або в іншому вигляді (6), (7), (8)) схожа на різницеву задачу Діріхле для рівняння Пуассона в прямокутнику (див. п. 5.3), для її розв'язку можуть бути застосовані ті самі методи. Оскільки коефіцієнти тут сталі, а область Ω — прямокутник, то можна застосувати прямі методи, які є найбільш економними.

6.3.2. Стійкість та збіжність схеми з ваговими коефіцієнтами. Запишемо схему (4) в операторному канонічному вигляді:

$$B\frac{y^{j+1} - y^j}{\tau} + Ay^j = \tilde{\varphi}^j, \quad j = 0, 1, \ldots, \tag{9}$$
$$y^0 = u_0, \quad B = E + \sigma\tau A,$$

де $u_0 = u_0(x)$, $x \in \omega_h$, $\tilde{\varphi}^j$ збігається з φ^j при $x = \mathring{\omega}_h = \{x = (x_1, x_2) : x_1 = i_1h;\ x_2 = i_2h_2,\ i_1 = \overline{2,\ N_1 - 2},\ i_2 = \overline{2N_2 - 2}\}$, а в інших (пригранічних) вузлах сітки ω_h виражається через φ^j, μ^j та μ^{j+1}, оператор A, визначений вище. Як бачимо (9) по формі не відрізняється від абстрактних схем п. 1.4 з ч. 1 чи різницевої схеми з ваговими коефіцієнтами для розв'язування одновимірного рівняння теплопровідності. За допомогою формул Гріна (або формул підсумовування частинами) неважко впевнитися, що $(Ay, v) = (y, Av)$, тобто $A = A^*$, та в тому, що

$$\delta_0 I \leqslant A \leqslant \Delta_0 I,$$

$$\delta_0 = \frac{4}{h_1^2}\sin^2\frac{\pi h_1}{2l_1} + \frac{4}{h_2^2}\sin^2\frac{\pi h_2}{2l_1}, \quad \Delta_0 = \frac{4}{h_1^2}\sin^2\frac{\pi(N_1-1)h_1}{2l_1} +$$
$$+ \frac{4}{h_2^2}\sin^2\frac{\pi(N_2-1)h_2}{2l_2} = \frac{4}{h_1^2}\cos^2\frac{\pi h_1}{2l_1} + \frac{4}{h_2^2}\cos^2\frac{\pi h_2}{2l_2}, \quad \Delta_0 = \|A\|.$$

В силу загальної теорії стійкості (див. п. 1.4 з ч. 1, теорема 1) схема (9) стійка в H_A при $B \geqslant \frac{\tau}{2}A$ або $B - \frac{\tau}{2}A \geqslant 0$. Звідси маємо

$$B - \frac{\tau}{2}A = I + \left(\sigma - \frac{1}{2}\right)\tau A \geqslant \left(\frac{1}{\|A\|} + \left(\sigma - \frac{1}{2}\right)\tau\right)A \geqslant 0.$$

Ця умова виконується при

$$\frac{1}{\|A\|} + \left(\sigma - \frac{1}{2}\right)\tau \geqslant 0$$

або

$$\sigma \geqslant \sigma_0 = \frac{1}{2} - \frac{1}{\tau\|A\|}. \tag{10}$$

Зокрема, для явної схеми ($\sigma = 0$) маємо умову $\tau \leqslant \frac{2}{\Delta_0}$ або

$$\tau \leqslant \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right)^{-1}, \tag{11}$$

яка на квадратній сітці ($h_1 = h_2 = h$) набирає вигляду

$$\tau \leqslant \frac{h^2}{4}. \qquad (12)$$

Порівнюючи (12) з відповідною умовою одновимірної задачі ($\tau \leqslant h^2/2$), бачимо вплив двовимірності, а саме: умова (12) більш жорстка. Із умови (10) бачимо, що як і в одновимірному випадку, схеми з $\sigma \geqslant 1/2$, в тому числі чисто неявна ($\sigma = 1$) і симетрична ($\sigma = 1/2$), є безумовно стійкими.

Далі обмежимося конспективним викладом, пропонуючи читачеві порівняти сказане з одновимірним випадком.

Запишемо задачу для похибки $z = y - u$:

$$B\frac{z^{j+1} - z^j}{\tau} + Az^j = \psi^j, \quad j = 0, 1, \ldots; \; z^0 = 0, \qquad (13)$$

де $\psi = \Lambda(\sigma\hat{u} + (1-\sigma)u) + \varphi - u_t$ — похибка апроксимації. Аналогічно одновимірному випадку за допомогою розвинення в ряд Тейлора, дістанемо

$$\psi = O(|h|^2 + \tau^2) + \left(\sigma - \frac{1}{2}\right) O(\tau), \quad |h|^2 = h_1^2 + h_2^2, \qquad (14)$$

якщо $u \in C^3_{4,4}(\bar{Q}_T)$ (нижні індекси вказують на гладкість по x_1, x_2, верхній — по t), і $\psi = O(|h|^2 + \tau)$, якщо $u \in C^2_{4,4}(Q_T)$. Теорема 6, п. 1.4 з ч. 1, застосована до (13), дає апріорну оцінку (при $\varepsilon = 1$, $\sigma \geqslant 1/2$)

$$\|z^j\|_A \leqslant \sum_{k=0}^{j-1} \tau \|\psi^j\|, \qquad (15)$$

звідки з урахуванням (14) дістаємо таку теорему збіжності.

Теорема. *Нехай розв'язок $u(x, t)$ задачі (1) — (3) належить простору $C^3_{4,4}(\bar{Q}_T)$. Тоді при $h \to 0$ розв'язок схеми (4) при $\sigma = \frac{1}{2}$ прямує до точного розв'язку, причому має місце оцінка швидкості збіжності*

$$\|y^j - u^j\| \leqslant M(|h|^2 + \tau^2);$$

якщо ж $u \in C^2_{4,4}(\bar{Q}_T)$, то має місце оцінка

$$\|y^j - u^j\|_A \leqslant M(|h|^2 + \tau),$$

де M — стала, незалежна від $|h|$, τ.

Зауважимо, що $\|z\|_A^2 = \|z\|_{A_1}^2 + \|z\|_{A_2}^2$, $A_\alpha z = -z_{\bar{x}_\alpha x_\alpha}$, $\alpha = 1, 2$, і за допомогою підсумовування частинами дістанемо

$$\|z\|_A^2 = \sum_{i_1=1}^{N_1} h_1 \sum_{i_2=1}^{N_2-1} h_2 (z_{\bar{x}_1}(i_1h_1, i_2h_2))^2 + \sum_{i_1=1}^{N_1-1} h_1 \times$$

$$\times \sum_{i_2=1}^{N_2} h_2 (z_{\bar{x}_2}(i_1h_1, i_2h_2))^2 \equiv |z|^2_{W_2^1(\omega_h)} \leqslant \|z\|^2_{W_2^1(\omega_h)},$$

тобто теорема дає оцінки в сітковій нормі $W_2^1(\omega_h)$. На відміну від одновимірного випадку для функцій двох сіткових змінних немає вкладення $W_2^1(\omega_h) \subset C(\bar{\omega}_h)$ (тобто немає оцінки виду $\|\cdot\|_{C(\omega_h)} \leqslant M\|\cdot\|_{W_2^1}$) і з теореми не випливають відповідні оцінки в нормі C.

Оцінку в нормі C легко дістати для явної схеми ($\sigma = 0$), яка в індексному вигляді записується таким чином:

$$y^{j+1}_{i_1,i_2} = (1 - 2(\gamma_1 + \gamma_2)) y^j_{i_1,i_2} + \gamma_2 (y^j_{i_1-1,i_2} + y^j_{i_1+1,i_2}) +$$

$$+ \gamma_2 (y^j_{i_1,i_2-1} + y^j_{i_1,i_2+1}) + \tau\varphi^j_{i_1,i_2}, \; \gamma_\alpha = \frac{\tau}{h_\alpha^2}, \quad \alpha = 1, 2, \qquad (16)$$

$y^0_{i_1,i_2}$ задається з початкових $iy^j_{i_1,i_2}$ в граничних вузлах — з граничних умов. Неважко помітити, що сума коефіцієнтів при y в правій частині (16) дорівнює 1. Якщо вони невід'ємні (тобто виконана умова $\gamma_1 + \gamma_2 \leqslant 1/2$), то з (16) маємо співвідношення

$$\|y^{j+1}\|_{C(\omega_h)} \leqslant \|y^j\|_{C(\omega_h)} + \tau\|\varphi^j\|_{C(\omega_h)}.$$

Звідси при умові $\gamma_1 + \gamma_2 \leqslant 1/2$ випливає нерівність

$$\|y^j\|_{C(\omega_h)} \leqslant \|y^0\|_{C(\omega_h)} + \sum_{k=0}^{j-1} \tau\|\varphi^k\|_{C(\omega_h)}, \qquad (17)$$

яка еквівалентна умові (11). Можна довести, що насправді оцінка (17) має місце не лише при $\sigma = 0$. На основі (17) і (14) можна сформулювати теорему про швидкість збіжності явної схеми в нормі $C(\omega_h)$. Якщо схема (4) застосовується для розв'язування задачі (1) — (3) на великих проміжках часу, то важливо, як і в одновимірному випадку, щоб вона була асимптотично стійкою. Користуючись нерівністю $l\delta_0 \leqslant A \leqslant \Lambda_0 l$ і виразами для δ_0, Λ_0, аналогічно одновимірному випадку можна дістати умову асимптотичної стійкості схеми з ваговими коефіцієнтами, з якої випливатиме, що чисто неявна схема ($\sigma = 1$) є безумовно асимптотично стійкою, явна ($\sigma = 0$) — асимптотично стійкою при умові

$$\tau \leqslant \tau_0^{(0)} = 2\left(\frac{4}{h_1^2} + \frac{4}{h_2^2}\right)^{-1},$$

симетрична схема ($\sigma = 1/2$) — при умові

$$\tau \leqslant \tau_0^{(1/2)} = \frac{2}{\sqrt{\delta_0\Lambda_0}}.$$

Зокрема, при $h_1 = h_2 = h$, $l_1 = l_2 = l$ маємо

$$\delta_0 = \frac{8}{h^2}\sin^2\frac{\pi h}{2l}, \quad \Lambda_0 = \frac{8}{h^2}\cos^2\frac{\pi h}{2l},$$

$$\tau_0^{(0)} = \frac{h^2}{4}, \quad \tau_0^{(1/2)} = \frac{h^2}{2}\left(\sin\frac{\pi h}{l}\right)^{-1} \approx \frac{hl}{2\pi}.$$

Бачимо, що двовимірність задачі приводить до того, що граничне значення $\tau_0^{1/2}$ вдвічі менше, ніж в одновимірному випадку.

Вправа. Припускаючи відповідну гладкість розв'язку q-вимірної задачі

$$\frac{\partial u}{\partial t} = Lu + f(x, t), \quad (x, t) \in Q_T,$$

$$Q_T = \Omega \times (0, T], \quad \Omega = \{x = (x_1, \ldots, x) : 0 < 0 < x_\alpha < l_\alpha, \ \alpha = \overline{1, q}\},$$

$$Lu = \sum_{k=1}^{q} L_k y, \quad L_k u = \frac{\partial^2 u}{\partial x_k^2}; \tag{18}$$

$$u(x, t) = \mu(x, t), \quad y \in \partial\Omega,$$

$$u(x, 0) = u_0(x), \quad x \in \overline{\Omega} \ (f, u_0, u_l \text{ задані}),$$

користуючись рівностями

$$\Lambda u = Lu + \frac{h^2}{12} \sum_{k=1}^{q} L_k L_k u + O(h^4), \quad \Lambda = \sum_{k=1}^{q} \Lambda_k, \ \Lambda_k u = u_{\bar{x}_k x_k},$$

$$u_t = \frac{\partial u}{\partial t} + \frac{\tau}{2} \frac{\partial^2 u}{\partial t^2} + O(\tau^2),$$

$$\frac{u^{j+1} + u^j}{2} = u^j + \frac{\tau}{2} \frac{\partial u^j}{\partial t} + O(\tau^2),$$

довести, що двох'ярусна схема

$$y_t^j = \left(\Lambda + \frac{h^2}{6} \sum_{l=1}^{q-1} \sum_{m=l+1}^{q} \Lambda_l \Lambda_m\right) \frac{y^{j+1} + y^j}{2} -$$

$$- \left[\frac{h^2}{12} \Lambda + \frac{\tau^2}{4} (1 + \sigma) \sum_{l=1}^{q-1} \sum_{m=l+1}^{q} \Lambda_l \Lambda_m\right] y_t^j = \varphi^{j+1/2}$$

при $\varphi^{j+1/2} = f^{j+1/2} + \frac{h^2}{12} \Lambda f^{j+1/2}$, $\sigma = \frac{h^2}{6\tau}$ і точно апроксимованих початкових і граничних умовах має підвищений порядок апроксимації $O(\tau^2 + h^4)$.

Вказівка. Припустити, що розв'язок задачі (18) має обмежені похідні виду

$$\frac{\partial^{\beta_l + \lambda} u}{\partial x_l^{\beta_l} \partial t^\gamma}, \quad l = \overline{1, q}, \quad \beta_l \geqslant 6, \quad \gamma \geqslant 3,$$

і подати похибку апроксимації таким чином:

$$\psi_h = \left(I + \frac{\tau}{2} \frac{\partial}{\partial t} + \frac{h^2}{12} L + \frac{h^2 \tau}{24} L \frac{\partial}{\partial t}\right)\left(Lu + f - \frac{\partial u}{\partial t}\right) +$$

$$+ O(\tau^2 + h^4).$$

6.3.3. Різницеві схеми для рівняння теплопровідності зі змінними коефіцієнтами. Розглянемо задачу виду (1) — (3), в якій еліптичний оператор L має змінні коефіцієнти і не містить змішаних похідних, тобто

$$Lu = L_1 u + L_2 u, \quad L_\alpha u = \frac{\partial}{\partial x_\alpha}\left(k_\alpha(x, t) \frac{\partial u}{\partial x_\alpha}\right),$$

$$0 < c_1 \leqslant k_\alpha(x, t) \leqslant c_2, \quad \alpha = 1, 2. \tag{19}$$

Кожний з операторів L_α за допомогою вивчених раніше методів (див. гл. 4) можна апроксимувати різницевими триточковими операторами $L_\alpha \sim \Lambda_\alpha$, $\Lambda_\alpha y = (a_\alpha y_{\bar{x}_\alpha})_{x_\alpha}$, $\alpha = 1, 2$, де $a_\alpha = a_\alpha(i_1 h_1, i_2 h_2, t)$ — деякі функціонали від k_α, $\alpha = 1, 2$. Неважко довести, що коли взяти, наприклад, $a_1 = k_1((i_1 - 0{,}5) h_1, i_2 h_2, t)$, $a_2 = k_2(i_1 h_1, (i_2 - 0{,}5) h_2, t)$, тоді при достатній гладкості u матимемо $\Lambda_\alpha u - L_\alpha u = O(h_\alpha^2)$, $\alpha = 1, 2$,

$$\Lambda u - Lu = O(|h|^2), \quad \Lambda = \Lambda_1 + \Lambda_2, \tag{20}$$

$|h|^2 = h_1^2 + h_2^2$ (або $|h| = \max(h_1, h_2)$), тобто оператор Λ апроксимує L з другим порядком. Задачу (1) — (3) з оператором L виду (19) можна апроксимувати схемою з ваговими коефіцієнтами (4), де Λ визначається формулами (20). Аналогічно п. 6.1 вводиться простір сіткових функцій Ω_h, заданих на ω_h, з таким самим, як у п. 6.1, скалярним добутком.

Вводячи оператор

$$\Lambda y = -\Lambda \mathring{y} \equiv -(a_1 \mathring{y}_{\bar{x}_1})_{x_1} - (a_2 \mathring{y}_{\bar{x}_2})_{x_2},$$

у цьому просторі схему з ваговими коефіцієнтами можна записати у канонічному вигляді (9), причому

$$c_1 \mathring{A} \leqslant A \leqslant c_2 \mathring{A}, \quad \mathring{A} y = -\mathring{y}_{\bar{x}_1 x} - \mathring{y}_{\bar{x}_2 x_2}.$$

Звідси видно, що

$$\delta I \leqslant A \leqslant \Delta I,$$

$$\delta = c_1 \delta_0, \quad \Delta = c_2 \Delta_0, \tag{21}$$

де δ_0, Δ_0 визначені у п. 6.1. Для визначення $y^{j+1} = \hat{y}$ на новому часовому ярусі, якщо $y^j = y$ вже визначено, дістаємо задачу виду (6) або (7), де Λ визначається формулами (20).

Якщо ж схема явна ($\sigma = 0$), то $\hat{y}$ визначається за явною формулою

$$\hat{y}(x) = y(x) + (1 - \sigma) \tau \Lambda y(x) + \tau \varphi(x), \quad x \in \omega_h.$$

Для неявних схем ($\sigma \neq 0$) задача являтиме собою систему п'ятиточкових сіткових рівнянь на шаблоні (рис. 29) зі змінними коефіцієнта-

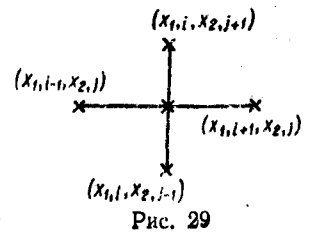


Рис. 29

ми. Враховуючи (21), тут можна застосувати ітераційні методи, найекономнішим з яких є попереміннo-трикутний (див. п. 5.4) з числом ітерацій при $\tau = O(h)$ і $h_1 = h_2 = h$

$$O\left(\frac{1}{\sqrt[4]{h}} \ln \frac{1}{\varepsilon}\right).$$

Для рівняння виду (6) з оператором Λ виду (20) цей метод стосовно методу, описаного в п. 5.4.3, треба дещо змінити.

6.4. Економні схеми для параболічних рівнянь

Якщо порівнювати розглянуті вище явні та неявні схеми для рівнянь параболічного типу по деякому критерію, то кожна з них має свої переваги і недоліки. Візьмемо за критерій порівняння такі дві характеристики: обсяг обчислень (кількість арифметичних операцій) для визначення y^{j+1} по заданому y^j та жорсткість обмежень на крок τ. Тоді на прикладі сіткової схеми з ваговими коефіцієнтами для двовимірного рівняння теплопровідності з п. 6.3 можна зазначити такі особливості. Для явної схеми ($\sigma = 0$): а) визначення $y^{j+1}(x)$, $x \in \omega_h$, потребує арифметичних дій, число яких пропорційне числу вузлів сітки ω_h, тобто число дій на один вузол сітки не залежить від сітки ω_h (це є перевагою схеми); б) недоліком явної схеми є жорстке обмеження на крок τ ($\tau \leqslant h^2/4$ при $h_1 = h_2 = h$), яке є умовою її стійкості. Для неявної схеми при $\sigma \geqslant 1/2$: а) визначення y^{j+1} потребує розв'язування системи з $(N_1 - 1)(N_2 - 1)$ п'ятиточкових різницевих рівнянь, що, наприклад, при змінних коефіцієнтах і застосуванні попереміннo-трикутного методу потребує на один вузол сітки ω_h числа дій, яке збільшується при $h \to 0$ (недолік неявної схеми); б) неявна схема безумовно стійка (при будь-яких τ, h), що є її перевагою.

Природно поставити за мету побудову таких схем, які б мали кращі властивості явних і неявних схем, тобто були б безумовно стійкими і число дій на один вузол сітки ω_h при обчислюванні розв'язку на наступному ярусі не залежало б від h. Такі схеми називаються *економними*. Деякі з них ми і розглянемо нижче.

6.4.1. Метод змінних напрямів. Схеми цього методу історично були першими економними схемами і з'явилися в 1955—1956 р. Суть його полягає в тому, що перехід з ярусу t_j на ярус t_{j+1} відбувається за допомогою розв'язування одновимірних задач.

Розглянемо, наприклад, так звану поздовжньо-поперечну схему Пісмена — Рекфорда для задачі

$$\frac{\partial u}{\partial t} = Lu + f(x, t), \quad (x, t) \in Q_T \equiv \Omega \times (0, T],$$
$$u(x, t) = \mu(x, t), \quad x \in \Omega, \tag{1}$$
$$u(x, 0) = u_0(x), \quad x \in \bar{\Omega},$$

де $L = L_1 + L_2$, $\Omega = \{x = (x_1, x_2) : 0 < x_\alpha < l_\alpha, \alpha = 1, 2\}$. Нехай $L_\alpha u = \frac{\partial^2 u}{\partial x_\alpha^2}$ або $L_\alpha u = \frac{\partial}{\partial x_\alpha}\left(k_\alpha(x, t)\frac{\partial u}{\partial x_\alpha}\right)$, $\alpha = 1, 2$, а Λ_α, $\alpha = 1, 2$ — відповідні триточкові різницеві апроксимації цих операторів на сітках $\omega_\alpha = \{x_\alpha = i_\alpha h_\alpha : i_\alpha = \overline{1, N_\alpha - 1}, h_\alpha = l_\alpha/N_\alpha\}$, $\Lambda = \Lambda_1 + \Lambda_2$. Введемо проміжне значення $\bar{y} = y^{j+1/2} = y\left(t_i + \frac{\tau}{2}\right)$ і запишемо таку різницеву схему методу змінних напрямків (схему Пісмена — Рекфорда)

$$\frac{y^{j+1/2} - y^j}{0{,}5\tau} = \Lambda_1 y^{j+1/2} + \Lambda_2 y^j + \varphi^j, \quad x \in \omega_h;$$
$$y^{j+1/2}|_{t_1 = 0, N_1} = \bar{\mu}|_{t_1 = 0, N_1}; \tag{2}$$
$$\frac{y^{j+1} - y^{j+1/2}}{0{,}5\tau} = \Lambda_1 y^{j+1/2} + \Lambda_2 y^{j+1} + \varphi^j, \quad x \in \omega_h; \tag{3}$$
$$y^{j+1}|_{t_2 = 0, N_2} = \mu^{j+1}|_{t_2 = 0, N_2}, \quad y(x) = u_0(x), \quad x \in \bar{\omega}_h,$$

де $\bar{\mu}$ — проміжне значення функції $\mu(x, t)$, яке обчислюється за формулою

$$\bar{\mu} = \frac{\mu^j + \mu^{j+1}}{2} - \frac{\tau}{4}\Lambda_2(\mu^{j+1} - \mu^j)$$

(ця формула забезпечує схемі (2) — (3) порядок апроксимації $O(\tau^2 + |h|^2)$). Щоб визначити y^{j+1} за відомим y^j, послідовно розв'язуємо такі триточкові одновимірні задачі

$$0{,}5\tau\Lambda_1 y^{j+1/2}(x_1, x_2) - y^{j+1/2}(x_1, x_2) = -F^j(x_1, x_2), \quad x_1 \in \omega_1;$$
$$F^j = y^j + 0{,}5\tau(\Lambda_2 y^j + \varphi^j); \quad y^{j+1/2}(0, x_2) = \bar{\mu}(0, x_2),$$
$$y^{j+1/2}(l_1, x_2) = \bar{\mu}(l_1, x_2) \tag{4}$$

для кожного $x_2 \in \omega_2$, а потім для кожного $x_1 \in \omega_1$ — задачі

$$0{,}5\tau\Lambda_2 y^{j+1}(x_1, x_2) - y^{j+1}(x_1, x_2) = -F^{j+1/2}(x_1, x_2), \quad x_2 \in \omega_2;$$
$$y^{j+1}(x, 0) = \mu^{j+1}(x_1, 0), \quad y^{j+1}(x_1, l_2) = \mu^{j+1}(x_1, l_2); \tag{5}$$
$$F^{j+1/2} = y^{j+1/2} + 0{,}5\tau(\Lambda_1 y^{j+1/2} + \varphi^j).$$

Розв'язування кожної з $N_2 - 1$ задач (4) методом прогонки потребує порядку $O(N_1)$ арифметичних дій, а кожної з $N_1 - 1$ задач (5) — $O(N_2)$ арифметичних дій. Отже, число дій на один вузол сітки $\omega_h = \omega_1 \times \omega_2$ з $(N_1 - 1)(N_2 - 1)$ вузлами не залежить від числа вузлів і, таким чином, схема (2) — (3) є економною. Щоб переконатися в стійкості схеми (2) — (3), запишемо її в операторному вигляді. Для цього введемо простір $H = \Omega_h$ сіткових функцій, заданих на сітці $\omega_h =$

$= \omega_1 \times \omega_2$, зі скалярним добутком

$$(y, v) = \sum_{x \in \omega_h} h_1 h_2 y(x) v(x)$$

та простір функцій $\overset{\circ}{H} = \overset{\circ}{\Omega}_h$, заданих на сітці $\bar{\omega}_h = \bar{\omega}_1 \times \bar{\omega}_2$, $\bar{\omega}_\alpha = \{x_\alpha = i_\alpha h_\alpha : i_\alpha = \overline{0, N_\alpha}, h_\alpha = l_\alpha / N_\alpha\}$, $\alpha = 1, 2$, які перетворюються на нуль на границі γ_h. Замінивши функції φ^j в пригранічних вузлах (див. п. 6.3) та ввівши оператори $A_\alpha : H \to H$, за формулами

$$A_\alpha y = -\Lambda_\alpha \overset{\circ}{y} = -\overset{\circ}{y}_{\bar{x}_\alpha x_\alpha}, \quad \alpha = 1, 2, \quad y \in H, \quad \overset{\circ}{y} \in \overset{\circ}{H},$$

$$y(x) = \overset{\circ}{y}(x), \quad x \in \omega_h$$

запишемо (2) — (3) у вигляді

$$\frac{y^{j+1/2} - y^j}{0{,}5\tau} = -A_1 y^{j+1/2} - A_2 y^j + \bar{\varphi}^j,$$

$$\frac{y^{j+1} - y^{j+1/2}}{0{,}5\tau} = -A_1 y^{j+1/2} - A_2 y^{j+1} + \bar{\bar{\varphi}}^j.$$

Виключаючи звідси $y^{j+1/2}$, прийдемо до еквівалентної для (2) — (3) схеми

$$B \frac{y^{j+1} - y^j}{\tau} + A y^j = \Phi^j, \quad j = 0, 1, \ldots; \quad y^0 = u_0 \in H, \qquad (6)$$

де $A = A_1 + A_2$, $B = \left(I + \frac{\tau}{2} A_1\right)\left(I + \frac{\tau}{2} A_2\right)$,

$$\Phi^j = 0{,}5 (I - 0{,}5\tau A_1) \bar{\varphi}^j + 0{,}5 (I + 0{,}5\tau A_1) \bar{\bar{\varphi}}^j.$$

Очевидно, що $A_\alpha = A_\alpha^* > 0$, $\alpha = 1, 2$, $A_1 A_2 = A_2 A_1$, а тому

$$B = I + 0{,}5\tau A + 0{,}25\tau^2 A_1 A_2 \geqslant I + 0{,}5\tau A > \frac{\tau}{2} A,$$

тобто схема (6), отже, і схема (2) — (3) стійка в H_A за початковими умовами і за правою частиною (див. теорему 6, п. 1.4 з ч. 1) і в нормі H_A має точність $O(\tau^2 + h^2)$.

6.4.2. Факторизовані схеми. Оператор B, зображений у вигляді добутку операторів $B = B_1 \ldots B_p$, називається факторизованим, а відповідна схема в канонічному вигляді

$$B \frac{y^{j+1} - y^j}{\tau} + A y^j = \varphi^j, \quad j = 0, 1, \ldots; \quad y^0 = u_0 \in H \qquad (7)$$

називається *факторизованою схемою*. Згідно з цим означенням схема (6) є факторизованою.

Очевидно, якщо для розв'язування кожної з p двовимірних задач (p не залежить від N_1, N_2)

$$B_\alpha v = F_\alpha, \quad \alpha = \overline{1, p}$$

при заданих F_α потрібно $O(N_1 N_2)$ арифметичних операцій, то і для визначення y^{j+1} по відомому y^j з (7) потрібно буде $O(N_1 N_2)$ операцій, тобто схема (7) буде економною. Такі оператори B_α також називаються економними. Отже, якщо схема (7) є факторизованою схемою з економними операторами B_α, $\alpha = \overline{1, p}$, то вона є економною.

Опираючись на теорію стійкості двох'ярусних схем і виходячи зі схеми з ваговими коефіцієнтами для двовимірного рівняння теплопровідності (див. п. 6.3), методом регуляризації сіткових схем можна дістати економну факторизовану схему. Отже, нехай в (9), п. 6.3,

$$A = A_1 + A_2, \quad B = I + \sigma\tau A = I + \sigma\tau (A_1 + A_2), \quad A_\alpha = A_\alpha^*, \quad \alpha = 1, 2.$$

Тоді схема (9) з п. 6.3 з такими операторами A, B буде стійкою при

$$\sigma \geqslant \sigma_0 = \frac{1}{2} - \frac{1}{\tau \| A \|}.$$

Замінимо B факторизованим оператором

$$\tilde{B} = (I + \sigma\tau A_1)(I + \sigma\tau A_2),$$

який відрізняється від B додатком $\sigma^2\tau^2 A_1 A_2$, тобто $\tilde{B} = B + \sigma^2\tau^2 A_1 A_2$. У результаті дістанемо факторизовану схему

$$\tilde{B} \frac{y^{j+1} - y^j}{\tau} + A y^j = \tilde{\varphi}^j, \quad j = 0, 1, \ldots; \quad y^0 = u_0 \in H, \qquad (8)$$

причому, якщо вихідна схема мала по τ, а порядок апроксимації не вище 2, то схема (8) матиме такий самий порядок апроксимації, як і (9) з п. 6.3. Оскільки при $\sigma \geqslant \sigma_0$ вихідна схема стійка, то і факторизована схема (8) стійка за умови

$$\tilde{B} > B \geqslant \frac{\tau}{2} A,$$

яка виконується через те, що оператори A_α, $\alpha = 1, 2$ додатні та переставні. Для визначення y^{j+1} з (8) маємо рівняння

$$\tilde{B} y^{j+1} = F^j, \quad F^j = \tilde{B} y^j + \tau (\tilde{\varphi}^j - A y^j),$$

яке розв'язується за два етапи:

$$(I + \sigma\tau A_1) \bar{y} = F^j, \quad (I + \sigma\tau A_2) y^{j+1} = \bar{y}.$$

Економнішим за кількістю операцій є алгоритм

$$\begin{aligned} (I + \sigma\tau A_1) W^{j+1/2} &= \tilde{F}^j \equiv \tilde{\varphi}^j - A y^j, \\ (I + \sigma\tau A_2) W^{j+1} &= W^{j+1/2}, \quad y^{j+1} = y^j + \tau W^{j+1}. \end{aligned} \qquad (9)$$

Тут економія досягається при обчисленні $\tilde{F}^j$, але треба зберігати не один, а два вектори y^j та W^{j+1} (або $W^{j+1/2}$). При $\sigma = 1$ схема (9) називається схемою змінних напрямів або схемою Дугласа — Рекфорда, яку можна записати також у вигляді

$$\frac{y^{j+1/2} - y^j}{\tau} + A_1 y^{j+1/2} + A_2 y^j = \tilde{\varphi}^j;$$

$$(I + \tau A_2)\frac{y^{j+1} - y^{j+1/2}}{\tau} = \frac{y^{j+1/2} - y^j}{\tau}.$$

Зауважимо, що в розглянутому випадку ми обминули проблему вибору крайових умов для проміжної функції $W^{j+1/2}$, розглядаючи схему (9) з п. 6.3 у просторі $H = \Omega_h$ (тобто ми зробили граничні умови нульовими, змінивши праву частину в приграничних вузлах). Інший спосіб пояснимо на прикладі неявної двох'ярусної схеми для двовимірного рівняння теплопровідності

$$\begin{gathered} y_t^j(x) = \Lambda y^{j+1}(x) + f^j(x), \quad x \in \omega_h, \\ y^j(x) \equiv y(x, t_j) \equiv y(x, j\tau) = \mu^j(x), \quad x \in \gamma_h, \\ y^0(x) = u_0(x), \quad x \in \bar{\omega}_h, \end{gathered} \tag{10}$$

де $\Lambda = \Lambda_1 + \Lambda_2$, $\Lambda_\alpha y = y_{\bar{x}_\alpha x_\alpha}$, $\alpha = 1, 2$,

$$\omega_h = \omega_1 \times \omega_2, \quad \bar{\omega}_h = \bar{\omega}_1 \times \bar{\omega}_2, \quad \gamma_h = \bar{\omega}_h \setminus \omega_h,$$

$$\omega_\alpha = \{x_\alpha = i_\alpha h_\alpha : i_\alpha = \overline{1, N_\alpha - 1},\ h_\alpha = l_\alpha/N_\alpha\};$$

$$\bar{\omega}_\alpha = \{x_\alpha = i_\alpha h_\alpha : i_\alpha = \overline{0, N_\alpha},\ h_\alpha = l_\alpha/N_\alpha\}, \quad \alpha = 1, 2.$$

Порядок апроксимації цієї схеми $O(\tau + h^2)$. Запишемо схему у вигляді

$$\begin{gathered} Dy^{j+1} = y^j + \tau f^j, \\ Dy^{j+1} = (I - \tau(\Lambda_1 + \Lambda_2))\, y^{j+1} \end{gathered} \tag{11}$$

і поряд з (11) розглянемо факторизовану схему

$$\begin{gathered} By^{j+1}(x) = y^j(x) + \tau f^j(x), \quad x \in \omega_h, \\ B = (I - \tau\Lambda_1)(I - \tau\Lambda_2), \\ y^j(x) = \mu^j(x), \quad x \in \gamma_h, \\ y^0(x) = u_0(x), \quad x \in \bar{\omega}_h. \end{gathered} \tag{12}$$

Введемо проміжну функцію $y^{j+1/2}$, яка задовольнятиме рівняння

$$(I - \tau\Lambda_1)\, y^{j+1/2}(x) = y^j(x) + \tau f^j(x), \quad x \in \omega_h. \tag{13}$$

При яких граничних умовах звідси можна однозначно знайти $y^{j+1/2}(x)$, $x \in \bar{\omega}_h$? Функцію y^{j+1} ми знайдемо із задачі

$$\begin{gathered} (I - \tau\Lambda_2)\, y^{j+1}(x) = y^{j+1/2}(x), \quad x \in \omega_h, \\ y^{j+1}(x) = \mu^{j+1}(x), \quad x \in \gamma_h. \end{gathered} \tag{14}$$

Але функція $y^{j+1/2}(x) = (I - \tau\Lambda_2)\, y^{j+1}$ визначена і для граничних точок $(x_{1,0}, x_{2,k}) = (0, x_{2,k})$, $(x_{1,N_1}, x_{2,k}) = (l_1, x_{2,k})$, $k = \overline{1, N_2 - 1}$, причому для цих точок

$$y^{j+1/2}(x) = (I - \tau\Lambda_2)\, \mu^{j+1}(x). \tag{15}$$

Таким чином, система для визначення допоміжної функції $y^{j+1/2}(x)$, $x \in \bar{\omega}_h$ має вигляд (13), (15) і складається з $N_2 - 1$ рівнянь (для кожного $x_2 \in \omega_2$ — одне рівняння), кожне з яких з урахуванням умов (15) можна розв'язати методом прогонки. Потім методом прогонки розв'язуються $N_1 - 1$ задач (14) (для кожного $x_1 \in \omega_1$ — одне рівняння).

6.4.3. Метод сумарної апроксимації. Нехай перехід від ярусу j до ярусу $j + 1$ при реалізації сіткової схеми здійснюється в декілька етапів, на кожному з яких використовується звичайна двох'ярусна схема виду

$$B\frac{y^{j+k/q} - y^{j+(k-1)/q}}{\tau} + \sum_{\alpha=1}^{q} D_{k\alpha} y^{j+\alpha/q} = \varphi_k^j, \quad k = \overline{1, q}; \quad j = 0, 1, 2, \ldots, \tag{16}$$

$y^0 \in H_h$ задано; $y^{j+\alpha/q} \in H_h$ — проміжні значення функції. Нехай ψ_k — похибка апроксимації окремого рівняння (16) на точному розв'язку диференціальної задачі, причому похибка апроксимації схеми (16) для цілих j (тобто після виключення проміжних значень) зображається у вигляді

$$\psi = \sum_{k=1}^{q} \psi_k.$$

Така схема називається *адитивною*. Будемо говорити, що адитивна схема має сумарну апроксимацію, якщо

$$\|\psi\| \to 0 \text{ при } h, \tau \to 0.$$

Зауважимо, що поняття сумарної апроксимації є більш слабким, ніж звичайне поняття апроксимації, бо не обов'язково $\|\psi_k\| \to 0$ при $h, \tau \to 0$. Це поняття було введене О. А. Самарським і дало змогу обґрунтувати відомі раніше методи (зокрема, метод змінних напрямків), а також побудувати ряд нових ефективних методів для різних класів задач.

Розглянемо задачу для q-вимірного рівняння параболічного типу:

$$\frac{\partial u}{\partial t} = Lu + f(x, t), \quad (x, t) \in Q_T = \Omega \times (0, T], \tag{17}$$

$$u(x, t) = \mu(x, t), \quad x \in \partial\Omega, \tag{18}$$

$$u(x, 0) = u_0(x), \quad x \in \bar{\Omega}, \tag{19}$$

де $Lu = \sum_{k=1}^{q} L_k u$ — еліптичний оператор; Ω — q-вимірна область змінних x_α, $\alpha = \overline{1, q}$. Запишемо рівняння (17) у вигляді

$$\sum_{k=1}^{q} W_k(u) = 0,$$

де

$$W_k(u) = \frac{1}{q}\frac{\partial u}{\partial t} - L_k u - f_k(x, t),$$

$$\sum_{k=1}^{q} f_k(x, t) = f(x, t).$$

Відрізок $[0, T]$ покриємо основною системою вузлів $t_j = j\tau$, $j = \overline{0, M}$, $\tau = T/M$, а кожний з основних проміжків $(t_j; t_{j+1}]$ покриємо допоміжною системою вузлів

$$t_{j+l/q} = t_j + \frac{l\tau}{q}, \quad l = \overline{1, q}.$$

Рівняння (17) на кожному відрізку $[t_j, t_{j+1}]$ заміняється системою диференціальних рівнянь (з відповідними граничними умовами)

$$W_l(v) = 0 \quad (l = \overline{1, q}) \tag{20}$$

таким чином, щоб на l-му допоміжному напівінтервалі $\left(t_j + \frac{l-1}{q}, t_j + \frac{l}{q}\right]$ виконувалося l-те рівняння системи (20), причому

$$v_l(x, t_{j+\frac{l-1}{q}}) = v_{l-1}(x, t_{j+\frac{l-1}{q}}), \quad l = \overline{2, q}; \quad j = 0, 1, \ldots,$$

$$v_q(x, t_j) = v_1(x, t_j), \quad j = 1, 2, \ldots, \tag{21}$$

$$v_1(x, 0) = u_0(x), \quad j = 0.$$

Якщо $t^* \in [t_j, t_{j+1}]$, то при достатній гладкості функції v

$$W_l(v(x, t^*)) = W_l(v(x, t_{j+1/2})) + O(\tau), \tag{22}$$

$$l = \overline{1, q}; \quad j = 0, 1, \ldots$$

Похибка апроксимації ψ рівняння (17) системою (20) дорівнює

$$\psi = \sum_{l=1}^{q} \psi_l(t), \quad t \in [t_j, t_{j+1}],$$

де $\psi_l(t)$ — похибка апроксимації для рівняння (20) номера l на розв'язку рівняння (17). Очевидно,

$$\psi_l(t) = W_l(u(x, t_{j+1/2})) + O(\tau),$$

тому в силу рівності $\sum_{k=1}^{q} W_k(u(x, t)) = 0 \; \forall x, t$ маємо

$$\psi = \sum_{l=1}^{q} \psi_l(t) = \sum_{l=1}^{q} [W_l(u(x, t_{j+1/2})) + O(\tau)] = O(\tau).$$

Таким чином, адитивна система диференціальних рівнянь (20) апроксимує у смислі сумарної апроксимації рівняння (17) з першим порядком по τ.

На кожному з допоміжних напівінтервалів $(t^{j+\frac{l-1}{q}}, t^{j+\frac{l}{q}}]$ l-те рівняння системи (20) апроксимується сітковою схемою

$$A_{hl}(y_l^j) = 0, \quad l = \overline{1, q}, \quad j = \overline{0, L}, \tag{23}$$

де $y_l^j = y_l^j(x)$, $x \in \omega_h$, ω_h — сітка, що покриває область Ω. Покажемо, що адитивна сіткова схема (23) має сумарну апроксимацію, якщо кожне l-те рівняння системи (20) апроксимується l-м сітковим рівнянням (23) у звичайному розумінні, тобто коли при $h, \tau \to 0$

$$\|\tilde{\psi}_l^j(v)\| \equiv \|\tilde{\psi}_l^j\| = \|-[W_l(\tilde{v}^j)]_{hl} + A_{hl}(\tilde{v}_{hl}^j)\| \to 0,$$

де $\tilde{v}^j = \tilde{v}(x, t^*)$, $t^* \in [t_j, t_{j+1}]$, а $\tilde{v}_{hl}^j$ позначає проекцію цієї функції на сітку ω_h (по x) та на точки $t^{j+\frac{l-1}{q}}$, $t^{j+\frac{l}{q}}$ допоміжної сітки по t. Справді, нехай ψ_{hl}^j — похибка, з якою апроксимує рівняння (23) номера l вихідне рівняння (17) на його розв'язку, тобто

$$\psi_{hl}^j = A_{hl}(u_{hl}^j).$$

Очевидно, що

$$\psi_{hl}^j = A_{hl}(u_{hl}^j) - (W_l(u^j))_{hl} + (W_l(u^j))_{hl} =$$

$$= \tilde{\psi}_{h,l}^j(u) + (W_l(u^j))_{hl},$$

тоді, враховуючи (22), маємо

$$\psi_{hl}^j = \tilde{\psi}_{hl}^j + (W_l(u(x, t_{j+1/2})))_h + O(\tau),$$

або ж

$$\psi_{hl}^j = \psi_{hl}^{*j} + (W_l(u(x, t_{j+1/2})))_h, \quad \|\psi_{hl}^{*j}\|_{h,\sigma\to 0} \to 0.$$

У цій рівності позначення $(\;)_h$ означає проекцію на сітку ω_h. Далі

маємо

$$\psi_h^j = \sum_{l=1}^{q} \psi_{hl}^j = \sum_{l=1}^{q} [\psi_{hl}^{*j} + (W_l(u(x, t_{j+1/2}))|_h] = \sum_{l=1}^{q} \psi_{hl}^{*j},$$

бо

$$\sum_{l=1}^{q} W_l(u(x, t)) = 0 \quad \forall\, x, t.$$

Тому при $h, \tau \to 0$ має місце $\|\psi_h^j\| \to 0$, тобто адитивна схема (23) має сумарну апроксимацію, хоча кожне окреме рівняння (23) може не апроксимувати рівняння (17). Адитивні схеми називають також *локально-одновимірними* або *схемами розщеплення*.

Приклад 1. Записати адитивну схему для задачі Коші для звичайного диференціального рівняння

$$\frac{du}{dt} + au = f(t), \quad t > 0, \quad u(0) = u_0, \tag{24}$$

де $a = \text{const} > 0$.

Розв'язання. Зобразимо $a = a_1 + a_2$, $a_1 > 0$, $a_2 > 0$, $f(t) = f_1(t) + f_2(t)$ (таке зображення, очевидно, не єдине і завжди можливе).

Введемо сітку $\omega_\tau = \{t_j = j\tau,\ j = 0, 1, \ldots\}$ і на кожному інтервалі (t_j, t_{j+1}) замість (24) розв'язуватимемо послідовно задачі

$$\frac{1}{2}\frac{dv_{(1)}}{dt} + a_1 v_{(1)} = f_1(t), \quad t \in (t_j, t_{j+1/2}]; \quad v_{(1)}(t_j) = v(t_j);$$

$$\frac{1}{2}\frac{dv_{(2)}}{dt} + a_2 v_{(2)} = f_2(t), \quad t \in (t_{j+1/2}, t_{j+1}]; \quad v_{(2)}(t_{j+1/2}) = \tag{25}$$

$$= v_{(1)}(t_{j+1/2}), \quad j = 0, 1, \ldots, \quad v_{(1)}(0) = u_0, \quad t_{j+1/2} = t_j + \frac{\tau}{2}.$$

Наближеним розв'язком задачі (24) вважатимемо функцію

$$v(t) = \begin{cases} v_{(1)}(t), & t \in [t_j, t_{j+1/2}), \\ v_{(2)}(t), & t \in [t_{j+1/2}, t_{j+1}]. \end{cases}$$

Кожне з рівнянь в (25) апроксимуємо двох'ярусною схемою з кроком $\tau/2$, наприклад, неявною схемою

$$\frac{y^{j+1/2} - y^j}{\tau} + a_1 y^{j+1/2} = f_1^j,$$

$$\frac{y^{j+1} - y^{j+1/2}}{\tau} + a_2 y^{j+1/2} = f_2^j. \tag{26}$$

Обчислимо похибки апроксимації. Підставивши в (26) $y^j = z^j + u^j$, $y^{j+1/2} = z^{j+1/2} + u^{j+1/2}$, $y^{j+1} = z^{j+1} + u^{j+1}$, матимемо задачі для похибки

$$\frac{z^{j+1/2} - z^j}{\tau} + a_1 z^{j+1/2} = -\psi_1^j,$$

$$\frac{z^{j+1} - z^{j+1/2}}{\tau} + a_2 z^{j+1} = -\psi_2^j, \quad j = 0, 1, \ldots, \quad z^0 = 0, \tag{26'}$$

де

$$\psi_1^j = \frac{u^{j+1/2} - u^j}{\tau} + a_1 u^{j+1/2} - f_1^j,$$

$$\psi_2^j = \frac{u^{j+1} - u^{j+1/2}}{\tau} + a_2 u^{j+1} - f_2^j.$$

За допомогою співвідношень

$$u^{j+1} = \left(u + \frac{\tau}{2}\frac{du}{dt}\right)^{j+1/2} + O(\tau^2), \quad u^j = \left(u - \frac{\tau}{2}\frac{du}{dt}\right)^{j+1/2} + O(\tau^2)$$

дістаємо

$$\psi_1^j = \left(\frac{1}{2}\frac{du}{dt} + a_1 u - f_1\right)^{j+1/2} + O(\tau),$$

$$\psi_2^j = \left(\frac{1}{2}\frac{du}{dt} + a_2 u - f_2\right)^{j+1/2} + O(\tau).$$

Звідси видно, що $\psi_1^j = O(1)$, $\psi_2^j = O(1)$, але при $\tau \to 0$

$$\psi_1^j + \psi_2^j = O(\tau) \to 0. \tag{27}$$

Зауважимо, що замість (25) можна було взяти таку систему рівнянь:

$$\frac{dv_{(1)}}{dt} + a_1 v_{(1)} = f_1(t), \quad t \in [t_j, t_{j+1}]; \quad v_{(1)}(t_j) = v(t_j);$$

$$\frac{dv_{(2)}}{dt} + a_2 v_{(2)} = f_2(t), \quad t \in [t_j, t_{j+1}], \quad v_{(2)}(t_j) = v_{(1)}(t_{j+1});$$

$$j = 0, 1, \ldots; \quad v_{(1)}(0) = u_0,$$

розв'язком якої матимемо $v(t) = v_{(2)}(t)$. Апроксимуючи ці рівняння з кроком τ, отримаємо схеми (26). Головним при зведенні задачі (24) до систем (25) чи (26) є властивості $a = a_1 + a_2$ та $f = f_1 + f_2$.

Приклад 2. Записати адитивну схему для двовимірної задачі (17) — (19), якщо $L_\alpha u = \dfrac{\partial^2 u}{\partial x_\alpha^2}$, $\alpha = 1, 2$, $\Omega = \{x = (x_1, x_2) : 0 < x_\alpha < l_\alpha, \alpha = 1, 2\}$.

Розв'язання. Система (20) набере вигляду

$$\frac{1}{2}\frac{\partial v_{(1)}}{\partial t} = L_1 v_{(1)} + f_1, \quad t_j \leqslant t \leqslant t_{j+1/2}; \quad v_{(1)}^j = v^j;$$

$$v_{(1)}(x, t) = \mu(x, t), \quad x \in \partial\Omega;$$

$$\frac{1}{2}\frac{\partial v_{(2)}}{\partial t} = L_2 v_{(2)} + f_2; \quad t_{j+1/2} \leqslant t \leqslant t_{j+1}; \quad v_{(2)}^{j+1/2} = v_{(1)}^{j+1/2}, \tag{28}$$

$$v_2(x, t) = \mu(x, t), \quad x \in \partial\Omega,$$

$$(v_{(i)} = v_{(i)}(x, t), \quad f_i = f_i(x, t)),$$

причому $v^{j+1} = v_{(2)}^{j+1}$. Апроксимуючи кожну з задач (28), наприклад, двох'ярусною схемою з ваговими коефіцієнтами (крок по t дорівнює $\tau/2$), дістаємо таку адитивну

схему (див. (23)):

$$\frac{y^{j+1/2}-y^j}{\tau}=\Lambda_1(\sigma_1 y^{j+1/2}+(1-\sigma_1)\,y^j)+\varphi_1^j,\quad x\in\omega_h,$$

$$y^{j+1/2}(x)=\mu^{j+1/2}(x),\quad x\in\gamma_h,\quad y^0(x)=u_0(x),\quad x\in\bar\omega_h, \tag{29}$$

$$\frac{y^{j+1}-y^{j+1/2}}{\tau}=\Lambda_2(\sigma_2 y^{j+1}+(1-\sigma_2)\,y^{j+1/2})+\varphi_2^j,\quad x\in\bar\omega_h,$$

$$y^{j+1}(x)=\mu^{j+1}(x),\quad x\in\gamma_h,\quad j=0,\ 1,\ \ldots,$$

де $\omega_h=\omega_1\times\omega_2$, $\bar\omega_h=\bar\omega_1\times\bar\omega_2$, $\omega_\alpha=\{x_\alpha=i_\alpha h_\alpha\colon i_\alpha=\overline{1,\ N_\alpha-1},$

$h_\alpha=l_\alpha/N_\alpha\}$, $\bar\omega_\alpha=\{x_\alpha=i_\alpha h_\alpha\colon i_\alpha=\overline{0,\ N_\alpha},\ h_\alpha=l_\alpha/N_\alpha\}$,

$\alpha=1,\ 2,\ \gamma_h=\bar\omega_h\setminus\omega_h$, $\Lambda_\alpha y=y_{\bar x_\alpha x_\alpha}$, $\alpha=1,\ 2,$

параметри σ_α, $\alpha=1,\ 2$ визначаються з умов стійкості та апроксимації. Наприклад, при $\sigma_1=\sigma_2=1$ дістаємо схему з випередженням:

$$\frac{y^{j+1/2}-y^j}{\tau}=\Lambda_1 y^{j+1/2}+\varphi_1^j,\quad y^0=u_0,$$

$$\frac{y^{j+1}-y^{j+1/2}}{\tau}=\Lambda_2 y^{j+1}+\varphi_2^j,\quad j=0,\ 1,\ \ldots,$$

при відповідних граничних умовах. Підставляючи сюди

$$y^j=z^j+u^j,\quad y^{j+1/2}=z^{j+1/2}+(u^j+u^{j+1})/2,\quad y^{j+1}=z^{j+1}+u^{j+1},$$

дістаємо для похибки z рівняння

$$\frac{z^{j+1/2}-z^j}{\tau}=\Lambda_1 z^{j+1/2}+\psi_1^j,$$

$$\frac{z^{j+1}-z^{j+1/2}}{\tau}=\Lambda_2 z^{j+1}+\psi_2^j,$$

де похибки апроксимації визначаються формулами

$$\psi_1^j=\Lambda_1\frac{u+\hat u}{2}-\frac{1}{2}\frac{\hat u-u}{\tau}+\varphi_1,$$

$$\psi_2^j=\Lambda_2\hat u-\frac{1}{2}\frac{\hat u-u}{\tau}+\varphi_2,\quad \hat u=u^{j+1},\quad u=u^j.$$

Бачимо, що $\psi_1=O(1)$, $\psi_2=O(1)$, тобто кожне з рівнянь (30) окремо не апроксимує рівняння (17). Сума нев'язок має вигляд

$$\psi=\psi_1+\psi_2=\Lambda_1\frac{u+\hat u}{2}+\Lambda_2\hat u-\frac{\hat u-u}{\tau}+\varphi_1+\varphi_2=$$

$$=(L_1+L_2)\,\bar u-\frac{\partial u}{\partial t}+\varphi_1+\varphi_2+O(\tau+h^2),\quad \bar u=u^{j+1/2}.$$

Враховуючи рівняння (17) при $t=t_j+\frac{\tau}{2}$, дістаємо

$$\psi=\varphi_1+\varphi_2-f^{j+1/2}+O(\tau+h^2)=O(\tau+h^2),$$

якщо $\varphi_1+\varphi_2=f^{j+1/2}+O(\tau)$. Останнє можливо, наприклад, при

$$\varphi_1=0,\quad \varphi_2=f^{j+1/2}\quad\text{або}\quad \varphi_1=\varphi_2=f^j/2.$$

Зауважимо, що замість (28) можна було взяти таку систему

$$\frac{\partial v_{(1)}}{\partial t}=L_1 v_{(1)}+f_1,\quad t_j\leqslant t\leqslant t_{j+1},\quad v_{(1)}^j=v^j,$$

$$\frac{\partial v_{(2)}}{\partial t}=L_2 v_{(2)}+f_2,\quad t_j\leqslant t\leqslant t_{j+1},\quad v_{(2)}^j=v_{(1)}^{j+1},$$

причому вважаємо $v^{j+1}=v_{(2)}^{j+1}$. Апроксимуючи її схемами з ваговими коефіцієнтами і з кроком τ по t, дістаємо (29).

Щоб довести збіжність адитивних схем, потрібно знайти оцінки для похибки, які враховують властивість сумарної апроксимації. Розглянемо, наприклад, схему (26). Задача для похибки $z^{j+1}=y^{j+1}-u^{j+1}$ має вигляд (26′). Зобразимо

$$\psi_\alpha=\psi_\alpha^0+\psi_\alpha^*,\quad \alpha=1,\ 2,\quad z^{j+1/2}=\eta_{j+1/2}+\xi_{j+1/2},\quad z^{j+1}=\eta_{j+1}+\xi_{j+1},$$

де

$$\mathring\psi_\alpha=\left(\frac{1}{2}\frac{du}{dt}+a_\alpha u-f_\alpha\right)^{j+1/2},\quad \psi_\alpha^*=O(\tau),$$

η_{j+1}, ξ_{j+1} — розв'язки задач

$$\eta_{j+1/2}=\eta_j+\tau\mathring\psi_1,\quad \eta_{j+1}=\eta_{j+1/2}+\tau\mathring\psi_2,\quad j=0,\ 1,\ \ldots;\quad \eta_0=0$$

$$(1+\tau a_1)\,\xi_{j+1/2}=\xi_j+\tau_1\tilde\psi_1,\quad (1+\tau a_2)\,\xi_{j+1}=\xi_{j+1/2}+\tau\tilde\psi_2,\quad j=0,\ 1,\ \ldots,$$

$$\xi_0=0,$$

$$\tilde\psi_1^j=\psi_1^{*j}-a_1\tau\eta_{j+1/2},\quad \tilde\psi_2^j=\psi_1^{*j}-a_2\tau\eta_{j+1}.$$

Звідси маємо

$$\eta_{j+1}=\eta_j+\tau(\mathring\psi_1^j+\mathring\psi_2^j)=\eta_j=\cdots=\eta_0=0,$$

тобто $z^j=\xi^j$, $\eta_{j+1/2}=\eta_j+\tau\mathring\psi_1=\tau\mathring\psi_1=O(\tau)$, і тому $\tilde\psi_\alpha=O(\tau)$, $\alpha=1,\ 2$. Далі з рівностей для ξ_j, $\xi_{j+1/2}$ знаходимо

$$|\xi_{j+1/2}|\leqslant|\xi_j|+\tau|\tilde\psi_1^j|,\quad |\xi_{j+1}|\leqslant|\xi_{j+1/2}|+\tau|\tilde\psi_2^j|\leqslant$$

$$\leqslant|\xi_j|+\tau(|\tilde\psi_1^j|+|\tilde\psi_2^j|),$$

тобто має місце оцінка

$$|z^{j+1}|\leqslant\sum_{k=0}^{j}\tau(|\tilde\psi_1^k|+|\tilde\psi_2^k|)=O(\tau),$$

яка означає збіжність адитивної схеми (26) для задачі (24) з швидкістю $O(\tau)$.

Вправа. Записати адитивні схеми з прикладів 1, 2 у операторному вигляді (16).

У загальному випадку для схеми (16) має місце теорема, за допомогою якої встановлюється збіжність у випадку сумарної апроксимації.

Теорема. *Якщо в схемі* (16) *оператор* B *та операторна матриця* $D = (D_{\alpha\beta})_{\alpha=\overline{1,q}}^{\beta=\overline{1,q}}$ *задовольняють умови*

$$B = B^* > 0, \quad \sum_{\alpha,\beta=1}^{q} (D_{\beta\alpha}\eta_\alpha, \ \eta_\beta) \geqslant 0, \quad \forall\, \eta_\alpha \in H_h,$$

то для розв'язку схеми (16) *має місце оцінка*

$$\|y^{j+1}\|_B \leqslant \exp(1)\left(\|y^0\|_B^2 + \sum_{m=1}^{j} \tau t_{j+1} \left\| \sum_{k=1}^{q} \varphi_k^m \right\|_{B^{-1}}^2 + \right.$$
$$\left. + \sum_{m=1}^{j} \sum_{k=1}^{q} \frac{\tau^2}{2} \sum_{k=1}^{q} \|\psi_k^m\|_{B^{-1}}^2 \right).$$

Доведення. Позначимо $y^{j+k/q} = w^k$, $\varphi_k^j = \Phi_k$, тоді (16) запишеться у вигляді

$$Bw_{\bar t}^k + \sum_{\alpha=1}^{q} D_{k\alpha} w^\alpha = \Phi_k, \quad k = \overline{1, q}. \tag{31}$$

Помножимо (31) скалярно на τw^k:

$$\tau\,(Bw_{\bar t}^k, w^k) + \tau \sum_{\alpha=1}^{q} (D_{k\alpha} w^\alpha, w^k) = \tau\,(\Phi_k, w^k). \tag{32}$$

Мають місце такі очевидні рівності:

$$\tau\,(Bw_{\bar t}^k, w^k) = \tau\left(Bw_{\bar t}^k, \frac{\tau}{2} w_{\bar t}^k + \frac{w^k + w^{k-1}}{2}\right) =$$
$$= \frac{\tau^2}{2}(Bw_{\bar t}^k, w_{\bar t}^k) + \frac{\tau}{2}(Bw_{\bar t}^k, w^k + w^{k-1}) =$$
$$= \frac{1}{2}[\tau^2 (Bw_{\bar t}^k, w_{\bar t}^k) + (Bw^k, w^k) - (Bw^{k-1}, w^{k-1})],$$
$$\tau\,(\Phi_k, w^k) = \tau\left(\Phi_k, w^0 + \sum_{i=1}^{k} (w^i - w^{i-1})\right) =$$
$$= \tau\,(\Phi_k, w^0) + \tau^2 \sum_{i=1}^{k} (\Phi_k, w_{\bar t}^i).$$

Підставивши їх у (32) і підсумувавши по $k = \overline{1, q}$, дістанемо основну

енергетичну тотожність

$$\sum_{k=1}^{q} \frac{\tau^2}{2} (Bw_{\bar t}^k, w_{\bar t}^k) + (Bw^q, w^q) - (Bw^0, w^0) +$$
$$+ \tau \sum_{k=1}^{q} \sum_{\alpha=1}^{q} (D_{k\alpha} w^\alpha, w^k) = \tau \sum_{k=1}^{q} (\Phi_k, w^0) + \tau^2 \sum_{k=1}^{q} \sum_{i=1}^{k} (\Phi_k, w_{\bar t}^i). \tag{33}$$

Перед оцінкою правої частини тотожності (33) зауважимо, що

$$\tau^2 \sum_{k=1}^{q} \sum_{i=1}^{k} (\Phi_k, w_{\bar t}^i) = \tau^2 \sum_{k=1}^{q} \left(w_{\bar t}^k, \sum_{i=k}^{q} \Phi_i\right) = \tau^2 \sum_{k=1}^{q} \left(w_{\bar t}^k, BB^{-1} \sum_{i=k}^{q} \Phi_i\right) =$$
$$= \tau^2 \sum_{k=1}^{q} \left(Bw_{\bar t}^k, B^{-1} \sum_{i=k}^{q} \Phi_i\right) \leqslant \frac{\tau^2}{2} \sum_{k=1}^{q} \left(\|w_{\bar t}^k\|_B^2 + \left\| \sum_{i=k}^{q} \Phi_i \right\|_{B^{-1}}^2\right). \tag{34}$$

За допомогою нерівності Коші — Буняковського та ε-нерівності

$$ab \leqslant \frac{\varepsilon}{2} a^2 + \frac{1}{2\varepsilon} b^2, \quad \varepsilon > 0$$

дістаємо оцінку для першого доданка в правій частині (33):

$$\tau \sum_{k=1}^{q} (\Phi_k, w^0) = \tau \left(Bw^0, B^{-1} \sum_{k=1}^{q} \Phi_k\right) \leqslant$$
$$\leqslant \frac{\tau\varepsilon}{2} \|w^0\|_B^2 + \frac{\tau}{2\varepsilon} \left\| \sum_{k=1}^{q} \Phi_k \right\|_{B^{-1}}^2. \tag{35}$$

Підставляючи (34), (35) в (33) і враховуючи (31), маємо

$$\frac{\tau^2}{2} \sum_{k=1}^{q} \|w_{\bar t}^k\|_B + \|w^q\|_B^2 \leqslant \|w^0\|_B^2 + \frac{\tau\varepsilon}{2} \|w^0\|_B^2 +$$
$$+ \frac{\tau}{2\varepsilon} \left\| \sum_{k=1}^{q} \Phi_k \right\|_{B^{-1}}^2 + \frac{\tau^2}{2} \sum_{k=1}^{q} \left(\|w_{\bar t}^k\|_B^2 + \left\| \sum_{i=k}^{q} \Phi_i \right\|_{B^{-1}}^2\right),$$

або після переходу до старих позначень

$$\|y^{j+1}\|_B^2 \leqslant \left(1 + \frac{\tau\varepsilon}{2}\right) \|y^j\|_B^2 + \frac{\tau}{2} \left[\sum_{k=1}^{q} \tau \left\| \sum_{i=k}^{q} \varphi_i^j \right\|_{B+1}^2 + \right.$$
$$\left. + \frac{1}{\varepsilon} \left\| \sum_{k=1}^{q} \varphi_k^j \right\|_{B^{-1}}^2 \right]. \tag{36}$$

Позначивши

$$q_{j+1} = \|y^{j+1}\|_B^2 > 0, \quad \psi_j = \frac{\tau}{2} \left[\sum_{k=1}^{q} \tau \left\| \sum_{i=k}^{q} \varphi_i^j \right\|_{B^{-1}}^2 + \frac{1}{\varepsilon} \left\| \sum_{k=1}^{q} \varphi_k^j \right\|_B^2 \right] > 0,$$
$$p = \left(1 + \frac{\tau\varepsilon}{2}\right),$$

нерівність (36) можна записати у вигляді

$$q_{j+1} = pq_j + \psi_j = p^{j+1}q_0 + \sum_{l=1}^{j} p^{j+1-l}\psi_l \leqslant$$

$$\leqslant p^{j+1}\left(q_0 + \sum_{l=1}^{j}\psi_l\right),$$

звідки

$$\|y^{j+1}\|_B^2 \leqslant \left(1 + \frac{\tau\varepsilon}{2}\right)^{j+1}\left(\|y^0\|_B^2 + \frac{\tau}{2}\sum_{l=1}^{j}\sum_{k=1}^{q}\tau\left\|\sum_{i=k}^{q}\varphi_i^l\right\|_{B^{-1}}^2 + \right.$$

$$\left. + \frac{\tau}{2\varepsilon}\sum_{l=1}^{j}\left\|\sum_{k=1}^{q}\varphi_k^l\right\|_{B^{-1}}^2\right).$$

Поклавши $\varepsilon = 2/t_{j+1}$, $t_{j+1} = (i+1)\tau$ і враховуючи нерівність $(1 + \tau\varepsilon/2) < \exp(\tau\varepsilon(j+1)/2)$, остаточно дістаємо твердження теореми.

Із теореми випливає, що коли $y^0 = 0$, $\|\varphi\| = \left\|\sum_{k=1}^{q}\varphi_k^m\right\|_{B^{-1}} \to 0$ при $h \to 0$, $\tau \to 0$ і величини $\sum_{k=1}^{q}\|\varphi_k^m\|_{B^{-1}} = O(1)$, тобто обмежені, то $\|y^{j+1}\|_B \to 0$, $j = 1, 2, \ldots$. Іншими словами, якщо записати задачу виду (16) для похибки, то при умовах теореми 1 із сумарної апроксимації і обмеженості норм окремих похибок апроксимації $\left(\sum_{k=1}^{q}\|\varphi_k^m\|_{B^{-1}} = O(1)\right)$ випливатиме збіжність адитивної схеми.

Отже, метод сумарної апроксимації дає змогу розщеплювати складні багатовимірні задачі на послідовність більш простих і тим самим значно спрощувати розв'язок багатовимірних задач.

6.5. Сіткові схеми для рівнянь гіперболічного типу

Найпростішим прикладом рівняння гіперболічного типу є одновимірне рівняння коливань струни

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad (x, t) \in Q_T = (0, a) \times (0, T], \tag{1}$$

єдиний розв'язок якого виділяється за допомогою граничних умов

$$u(0, t) = \gamma_2(t), \quad u(a, t) = \gamma_3(t) \tag{2}$$

та початкових умов

$$u(x, 0) = \gamma_0(x), \quad \frac{\partial u(x, 0)}{\partial t} = \gamma_1(x), \quad y \in [0, a]. \tag{3}$$

Введемо сітки $\bar{\omega}_\tau = \{t_j = j\tau : j = 0, 1, \ldots L, \tau = T/L\}$,

$$\omega_\tau = \{t_j = j\tau : j = \overline{1, L}, \ \tau = T/L\} \text{ та } \bar{\omega}_h = \{x_i = ih:$$

$$i = \overline{0, N}, \ h = a/N\}, \quad \omega_h = \{x_i = ih: \ i = \overline{1, N-1}, \ h = a/N\}.$$

Апроксимуючи на сітці ω_h еліптичну частину рівняння (1) (тобто покладаючи $\frac{\partial^2 u(x_i t)}{\partial x^2} \sim v_{\bar{x}x,i}(t)$) і залишаючи змінну t неперервною, тобто застосовуючи до задачі (1) — (2) поздовжню схему методу прямих, зведемо її наближений розв'язок до задачі Коші для системи звичайних диференціальних рівнянь другого порядку

$$\frac{d^2 v_i(t)}{dt} = v_{\bar{x}x,i}(t) + f(x_i, t), \quad i = \overline{1, N-1},$$

$$v_0(t) = \gamma_2(t), \quad v_N(t) = \gamma_3(t), \tag{4}$$

$$v_i(0) = \gamma_0(x_i), \quad v_i'(0) = \gamma_1(x_i), \quad i = \overline{0, N},$$

або в матричному вигляді

$$\frac{d^2\vec{V}(t)}{dt^2} = A\vec{V} + \vec{F},$$

$$\vec{V}(0) = \vec{\Gamma}_0, \quad \vec{V}(0) = \vec{\Gamma}_1, \tag{5}$$

де $\vec{V}(t) = (v_1(t), \ldots, v_{N-1}(t))^T$, $\vec{F}(t) = (f(x_1, t) + h^{-2}\gamma_2(t),$

$$f(x_2, t), \ldots, f(x_{N-2}, t), \quad f(x_{N-1}, t) + h^{-2}\gamma_3(t))^T,$$

$$\vec{\Gamma}_0 = (\gamma_0(x_1), \ldots, \gamma_0(x_{N-1}))^T, \quad \vec{\Gamma}_1 = (\gamma_1(x_1), \ldots, \gamma_1(x_{N-1}))^T,$$

$$A = \begin{bmatrix} -2 & 1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 & -2 \end{bmatrix}.$$

Застосовуючи поперечну схему методу прямих, тобто залишаючи неперервною змінну x і покладаючи $\frac{\partial^2 u(x, t_j)}{\partial t^2} \sim v_{\bar{t}t}^j(x) = v_{\bar{t}t}(x, t_j)$, дістанемо таку схему Роте, яка являє собою систему еліптичних крайових задач

$$v_{\bar{t}t}^j(x) = \frac{d^2 v^j(x)}{dx^2} + f(x, t_j),$$

$$v^j(0) = \gamma_2(t_j), \quad v^j(a) = \gamma_3(t_j), \qquad (6)$$
$$v^0(x) = \gamma_0(x).$$

Щоб повністю визначити систему (6), треба задати ще $y^1(x)$. Для цього можна, наприклад, використати рівності

$$\frac{u(x,\tau) - u(x,0)}{\tau} = \frac{\partial u(x,0)}{\partial t} + \frac{\tau}{2}\frac{\partial^2 u(x,0)}{\partial t^2} + O(\tau^2) =$$
$$= \frac{\partial u(x,0)}{\partial t} + \frac{\tau}{2}\left(\frac{\partial^2 u(x,0)}{\partial x^2} + f(x,0)\right) + O(\tau^2) =$$
$$= \gamma_1(x) + \frac{\tau}{2}(\gamma_0''(x) + f(x,0)) + O(\tau^2)$$

і покласти

$$v^1(x) = \gamma_0(x) + \tau\gamma_1(x) + \frac{\tau^2}{2}(\gamma_0''(x) + f(x,0)). \qquad (7)$$

Система (6), (7) повністю визначає функції $v^j(x)$, $j = \overline{0, L}$. Якщо в схемі (6), (7) дискретизувати і змінну x, то дістанемо таку схему методу сіток:

$$y_{\bar{t}t}^j = y_{\bar{x}x}^j + f_i^j, \quad (x_i, t_j) \in \omega_\tau \times \omega_h, \quad y_0^j = \gamma_2^j, \quad y_N^j = \gamma_3^j, \qquad (8)$$
$$y_i^0 = \gamma_0(x_i), \quad y_i^1 = \gamma_0(x_i) + \tau\gamma_1(x_i) + \frac{\tau^2}{2}(\gamma_0''(x_i) + f(x_i, 0)), \quad i = \overline{0, N}.$$

Неважко довести, що при гладкості розв'язку $u(x, t) \in C_4^4(Q_T)$ схема (8) має порядок апроксимації $O(h^2 + \tau^2)$. Ця схема є явною і значення шуканої функції $y^{j+1}(x)$ на $(j+1)$-му ярусі визначається через її значення на попередніх двох ярусах за формулою

$$y_i^{j+1} = 2y_i^j - y_i^{j-1} + \frac{\tau^2}{h^2}(y_{i-1}^j - 2y_i^j + y_{i-1}^j) + \tau^2 f_i^j, \quad i = \overline{1, N-1}, \qquad (9)$$
$$y_0^{j+1} = \gamma_2^{j+1}, \quad y_N^{j+1} = \gamma_3^{j+1}, \quad j = 1, 2, \ldots .$$

Зрозуміло, що ця схема є триярусною.

Введемо простір H функцій, заданих на сітці ω_h, а через $\mathring{H}$ позначимо простір функцій, заданих на $\bar{\omega}_h = \omega_h \cup \{0, l\}$, які перетворюються на нуль при $x = \overline{0, l}$. У просторі H введемо скалярний добуток

$$(y, v) = \sum_{x \in \omega_h} hy(x)v(x)$$

і оператор $A : H \to H$ такий, що $Ay = -\mathring{y}_{\bar{x}x}$, $y \in H$, $\mathring{y} \in \mathring{H}$, $y(x) = \mathring{y}(x)$, $x \in \omega_h$. Через y^j позначимо функцію $t_j \to H$. Тоді схема (8)

може бути записана в такому канонічному вигляді:

$$Dy_{\bar{t}t} + Ay = F, \; y^0, \; y^1 \text{ — задані}, \qquad (10)$$

де $D = I$, $Ay = -\mathring{y}_{\bar{x}x}$, $y \in H$, $\mathring{y} \in \mathring{H}$,

$$y(x) = \mathring{y}(x), \quad x \in \omega_h, \quad F = F^j = (f_1^j, \ldots, f_{N-1}^j)^T.$$

Похибка $z = y - u$ задовольняє рівняння

$$Dz_{\bar{t}t} + Az = \psi, \qquad (11)$$

де $|\psi_i^j| \leqslant M(h^2 + \tau^2)$; M — стала, незалежна від h, τ за умови, що точний розв'язок $u(x, t) \in C_4^4(\bar{Q}_T)$.

Схеми (10), (11) є триярусними, в силу теореми 11 (п. 1.4 з ч. 1) має місце оцінка

$$\|z^n\| \leqslant \sqrt{\frac{1+\varepsilon}{\varepsilon}}\left(\|z(0)\| + \|z_t(0)\|_{A^{-1}} + \sum_{s=1}^{k-1}\tau\|\psi^s\|_{A^{-1}}\right),$$
$$z^n = z(t_n) = z(n\tau) = (z_1(n\tau), \ldots, z_{N-1}(n\tau))^T, \qquad (12)$$

якщо виконуються умови $A = A^* > 0$, $D = D^* > 0$, $D \geqslant \frac{1+\varepsilon}{4}\tau^2 A$.

Перші дві нерівності очевидні, а остання в силу того, що $I \geqslant \frac{A}{\|A\|}$ виконується, коли

$$\left(\frac{1}{\|A\|} - \frac{1+\varepsilon}{4}\tau^2\right)A \geqslant 0,$$

тобто при

$$\frac{1}{\|A\|} - \frac{1+\varepsilon}{4}\tau^2 \geqslant 0.$$

Оскільки $\|A\| = \frac{4}{h^2}\cos^2\frac{\pi h}{2} < \frac{4}{h^2}$, то звідси дістаємо таку умову стійкості (див. також приклад 1, п 1.4.5 з ч.1):

$$\tau \leqslant \frac{h}{\sqrt{1+\varepsilon}}, \qquad (13)$$

де $\varepsilon > 0$ — довільне число. Крім того, $\|y\|_{A^{-1}}^2 = (A^{-1}y, y) \leqslant \lambda_{\max}\|y\|^2$, де $\lambda_{\max} = \frac{h^2}{4}\sin^{-2}\frac{\pi h}{2} = O(1)$ — максимальне власне значення оператора A^{-1}. Із виразів (7), (12), (13) випливає таке твердження.

Теорема. *Нехай розв'язок задачі* (1) — (3) *належить класу* $C_4^4(\bar{Q}_T)$, $\gamma_1(x) \in C[0, a]$, $\gamma_0(x) \in C^2[0, a]$, *і виконується умова* (13). *Тоді при*

$h, \tau \to 0$ розв'язок схеми (8) збігається до точного розв'язку і має місце така оцінка швидкості збіжності:

$$\| y^n - u^n \| \leqslant M(\varepsilon)(h^2 + \tau^2),$$

де $M(\varepsilon)$ — стала, що не залежить від h, τ.

ГЛАВА 7

ІНШІ МЕТОДИ

Однією з найскладніших і найактуальніших задач обчислювальної математики була і залишається «боротьба», як кажуть напівжартома, з «прокляттям розмірності». Ми вже могли впевнитися на простих прикладах, що двовимірні задачі значно складніші одновимірних, причому з дальшим ростом розмірності складність наростає ще стрімкіше. Однією з причин цієї складності є ускладнення геометрії задачі (наприклад, форми області, в якій визначається невідома функція, взаємозв'язків між невідомими). Іншою причиною є стрімке збільшення кількості співвідношень і невідомих при дискретизації задачі. Можна виділити ще багато різних особливостей багатовимірних задач, але ми коротко зупинимося на двох вищеназваних і розглянемо два методи, які полегшують (далеко не вирішують!) боротьбу з цими аспектами «прокляття розмірностей». Йдеться про *метод фіктивних областей* та *метод граничних елементів*.

Одним із шляхів «боротьби з розмірністю» є також створення продуктивніших ЕОМ і математичного забезпечення для них. Пошуки таких ЕОМ — суперкомп'ютерів — ведуться в основному розпаралелюванням їхньої роботи та створенням гнучкої архітектури ЕОМ, тобто ЕОМ, які пристосовані до особливостей конкретної задачі. Це вимагає створення паралельних алгоритмів і навіть перебудови способу мислення математиків-прикладників, який є традиційно послідовним. Тому торкнемося особливостей паралельних ЕОМ і створення «паралельних» чисельних методів. І хоч глава називається «Інші методи», не слід сподіватися, що в ній ми зможемо хоча б згадати всі інші методи. Різноманітність сучасних чисельних методів і темпи їхнього розвитку настільки великі, що це практично неможливо. Така назва підкреслює, що в книзі розглянуто лише частину найпоширеніших, але далеко не всіх методів чисельного аналізу. Поширеність однієї частини з них є наслідком їхніх достоїнств, іншої частини — наслідком зручностей їхнього програмування, деяких — наслідком мінливої математичної моди. Те, що широко застосовується сьогодні, може бути забуте завтра. Тому ця глава має переглядатися набагато частіше, ніж інші. Щоб не відстати від сучасного рівня розвитку чисельних методів, зацікавлений читач, безперечно, повинен сам неперервно брати в цьому участь.

7.1. Метод фіктивних областей

Як вже зазначалося, труднощі чисельного розв'язування багатовимірних задач математичної фізики значною мірою залежать від складності геометрії області, в якій задана шукана функція. Для областей канонічного вигляду, наприклад для прямокутника, можна побудувати ефективні методи (див. п. 5.4, 6.4). Метод фіктивних областей (МФО) полягає в тому, щоб побудувати за вихідною задачею в області складної форми іншу задачу (наприклад, у прямокутнику), розв'язок якої при прямуванні деяких її параметрів до нуля чи до нескінченності прямував би до розв'язку вихідної задачі. Переваги такого методу: а) універсальність по відношенню до геометрії області; б) простота і ефективність чисельних алгоритмів і програм для ЕОМ, оскільки вони будуються для прямокутної області. Покажемо на простих одновимірних прикладах, що таку ідею можна здійснити.

Розглянемо крайову задачу

$$-\frac{d^2u}{dx^2} = f(x), \quad x \in \Omega \equiv \left(0, \frac{1}{2}\right), \quad f(x) \equiv 2,$$
$$u(0) = u\left(\frac{1}{2}\right) = 0, \tag{1}$$

точним розв'язком якої є функція $u(x) = x(1/2 - x)$. Поряд з (1) розглянемо задачу в області $\Omega_0 \equiv (0, 1)$, $\Omega \subset \Omega_0$, $\Omega_1 = \Omega_0 \setminus \Omega = \equiv (1/2, 1)$

$$-\frac{d}{dx}\left(a^\varepsilon(x)\frac{du_\varepsilon}{dx}\right) = f_1(x), \quad x \in \Omega_0, \tag{2}$$

$$u_\varepsilon(0) = u_\varepsilon(1) = 0, \tag{3}$$

$$[u_\varepsilon]_{x=1/2} = 0, \quad \left[a\frac{du_\varepsilon}{dx}\right]_{x=1/2} = 0, \tag{4}$$

де

$$[v]_{x=\xi} = v(\xi + 0) - v(\xi - 0),$$

$$a^\varepsilon(x) = \begin{cases} 1, & x \in \Omega, \\ \varepsilon^{-2}, & x \in \Omega_1, \end{cases} \qquad f_1(x) = \begin{cases} f(x), & x \in \Omega, \\ 0, & x \in \Omega_1, \end{cases}$$

ε — параметр.

Під розв'язком задачі (2) — (4) розумітимемо неперервну функцію $u_\varepsilon(x)$, яка на кожному з інтервалів (0, 1/2) та (1/2, 1) має другі неперервні похідні і задовольняє диференціальне рівняння (2), на границі $\partial\Omega = \{1/2\}$ області Ω задовольняє умови (4) і на границі $\partial\Omega_0 = \{0, 1\}$ області Ω_0 — умови (3). Неважко знайти, що точний розв'язок задачі

(2) — (4) має вигляд

$$u_\varepsilon(x) = \begin{cases} x\left(\frac{1+2\varepsilon^2}{2(1+\varepsilon^2)} - x\right), & x \in \left[0, \frac{1}{2}\right], \\ \frac{\varepsilon^2}{2(1+\varepsilon^2)}(1-x), & x \in \left[\frac{1}{2}, 1\right], \end{cases}$$

причому

$$\lim_{\varepsilon\to 0} u_\varepsilon(x) = u(x), \quad x \in \bar{\Omega} \equiv \left[0, \frac{1}{2}\right],$$

$$\lim_{\varepsilon\to 0} u_\varepsilon(x) = 0, \quad x \in \Omega_1 \equiv \left(\frac{1}{2}, 1\right),$$

тобто при досить малих ε область Ω_1 можна розглядати як фіктивну, тобто таку, що «майже не впливає» на розв'язок задачі (2) — (4). Звідси і походить назва методу. МФО має під собою і фізичну основу. Йдеться про добре відомий у фізиці факт, що в середовищі з відносно великим коефіцієнтом теплопровідності температура змінюється відносно мало. Задача (1) описує процес стаціонарної теплопровідності з коефіцієнтом теплопровідності $a(x) \equiv 1$. Задача (2) описує той самий процес в області Ω_0, причому в області Ω коефіцієнти теплопровідності в задачах (1) та (2) — (4) однакові, а в доповненні Ω до Ω_0, тобто в Ω_1, коефіцієнт теплопровідності задачі (2) — (4) прямує до нескінченності при $\varepsilon \to 0$. Із фізичних міркувань слід чекати близькості розв'язків вказаних задач в області Ω, що і підтверджено математично.

Можливі також інші варіанти методу фіктивних областей, один з яких розглянемо знову на прикладі задачі (1). На цей раз замість (2) — (4) поставимо задачу

$$-\frac{d^2u_\varepsilon}{dx} + q^\varepsilon(x)\,u_\varepsilon(x) = f_1(x), \quad x \in \Omega_0 \equiv (0, 1), \tag{5}$$

$$u_\varepsilon(0) = u_\varepsilon(1) = 0, \tag{6}$$

$$[u_\varepsilon]_{x=1/2} = 0, \quad \left[\frac{du_\varepsilon}{dx}\right]_{x=1/2} = 0, \tag{7}$$

де

$$q^\varepsilon(x) = \begin{cases} 0, & x \in \Omega \equiv \left(0, \frac{1}{2}\right), \\ \varepsilon^{-2}, & x \in \Omega_1 \equiv \left(\frac{1}{2}, 1\right), \end{cases} \qquad f_1(x) = \begin{cases} f(x), & x \in \Omega, \\ 0, & x \in \Omega_1. \end{cases}$$

У цьому випадку умови (7) є наслідком рівняння (5), тому фактично розглядається задача (5), (6). Побудуємо точний розв'язок задачі (5) —

(7). Для цього зауважимо, що рівняння (5) в області Ω має вигляд

$$-\frac{d^2u_\varepsilon}{dx^2} = 2,$$

і його загальний розв'язок, який задовольняє першу з крайових умов (6), є

$$u_\varepsilon(x) = x(\alpha - x), \quad x \in \Omega, \tag{8}$$

де α — стала, яку визначимо пізніше. В області Ω_1 рівняння (5) має вигляд

$$-\frac{d^2u_\varepsilon}{dx^2} + \frac{1}{\varepsilon^2}u_\varepsilon = 0,$$

і неважко знайти його загальний розв'язок

$$u_\varepsilon(x) = a\,\mathrm{ch}\,\frac{x}{\varepsilon} + b\,\mathrm{sh}\,\frac{x}{\varepsilon}, \quad x \in \Omega_1, \tag{9}$$

де a, b — довільні сталі; $\mathrm{ch}\,z = (e^z + e^{-z})/2$ — гіперболічний косинус; $\mathrm{sh}\,z = (e^z - e^{-z})/2$ — гіперболічний синус. Щоб функція (9) задовольняла другу з крайових умов (6), знайдемо сталу b з рівняння

$$u_\varepsilon(1) = a\,\mathrm{ch}\,\frac{1}{\varepsilon} + b\,\mathrm{sh}\,\frac{1}{\varepsilon} = 0,$$

звідки

$$b = -a\,\mathrm{cth}\,\frac{1}{\varepsilon}, \tag{10}$$

і тому

$$u_\varepsilon(x) = a\left(\mathrm{ch}\,\frac{x}{\varepsilon} - \mathrm{cth}\,\frac{1}{\varepsilon}\,\mathrm{sh}\,\frac{x}{\varepsilon}\right), \tag{11}$$

$\mathrm{cth}\,z = \mathrm{ch}\,z/\mathrm{sh}\,z$ — гіперболічний котангенс. «Зшиємо» розв'язки (8) та (9) в точці $x = 1/2$ за допомогою умов (7). Неважко знайти

$$u_\varepsilon\left(\frac{1}{2} - 0\right) = \frac{2\alpha - 1}{4}, \quad u'_\varepsilon\left(\frac{1}{2} - 0\right) = \alpha - 1,$$

$$u_\varepsilon\left(\frac{1}{2} + 0\right) = a\,\mathrm{ch}\,\frac{1}{2\varepsilon} + b\,\mathrm{sh}\,\frac{1}{2\varepsilon} = a\left(\mathrm{ch}\,\frac{1}{2\varepsilon} - \mathrm{cth}\,\frac{1}{\varepsilon}\,\mathrm{sh}\,\frac{1}{2\varepsilon}\right),$$

$$u'_\varepsilon\left(\frac{1}{2} + 0\right) = \frac{a}{\varepsilon}\,\mathrm{sh}\,\frac{1}{2\varepsilon} + \frac{b}{\varepsilon}\,\mathrm{ch}\,\frac{1}{2\varepsilon} = \frac{a}{\varepsilon}\left(\mathrm{sh}\,\frac{1}{2\varepsilon} - \mathrm{cth}\,\frac{1}{\varepsilon}\,\mathrm{ch}\,\frac{1}{2\varepsilon}\right).$$

Перша з умов (7) дає рівняння

$$2\alpha - 1 = 4a\,\mathrm{ch}\,\frac{1}{2\varepsilon} - 4a\,\mathrm{cth}\,\frac{1}{\varepsilon}\,\mathrm{sh}\,\frac{1}{2\varepsilon},$$

звідки

$$\alpha = 2a\,\mathrm{ch}\,\frac{1}{2\varepsilon} - 2a\,\mathrm{cth}\,\frac{1}{\varepsilon}\,\mathrm{sh}\,\frac{1}{2\varepsilon} + \frac{1}{2}. \tag{12}$$

Друга з умов (7) приводить до рівняння

$$\alpha - 1 = \frac{a}{\varepsilon}\left(\operatorname{sh}\frac{1}{2\varepsilon} - \operatorname{cth}\frac{1}{\varepsilon}\operatorname{ch}\frac{1}{2\varepsilon}\right),$$

звідки, врахувавши (12), знаходимо

$$a = \frac{1}{4}\left[\operatorname{ch}\frac{1}{2\varepsilon} - \frac{1}{2\varepsilon}\operatorname{sh}\frac{1}{2\varepsilon} - \operatorname{cth}\frac{1}{\varepsilon}\left(\operatorname{sh}\frac{1}{2\varepsilon} - \frac{1}{2\varepsilon}\operatorname{ch}\frac{1}{2\varepsilon}\right)\right]^{-1}. \quad (13)$$

За допомогою правила Лопіталя неважко знайти, що

$$\lim_{\varepsilon\to 0}(4a) = \lim_{z\to\infty}\frac{1 - \operatorname{cth} 2z\,\dfrac{\operatorname{sh} z - 2z\operatorname{ch} z}{\operatorname{ch} z - 2z\operatorname{sh} z}}{(\operatorname{ch} z - 2z\operatorname{sh} z)^{-1}} =$$

$$= \lim_{z\to\infty}\frac{\dfrac{2}{\operatorname{sh}^2 2z}\,\dfrac{\operatorname{sh} z - 2z\operatorname{ch} z}{\operatorname{ch} z - 2z\operatorname{sh} z} - \operatorname{cth} 2z\,\dfrac{-(\operatorname{ch}^2 z + (2z\operatorname{sh} z)^2) + (\operatorname{sh}^2 z - (2z\operatorname{ch} z)^2)}{(\operatorname{ch} z - 2z\operatorname{sh} z)^2}}{\dfrac{\operatorname{sh} z + 2z\operatorname{ch} z}{(\operatorname{ch} z - 2z\operatorname{sh} z)^2}} =$$

$$= \lim_{z\to\infty}\frac{2(1 - 2z\operatorname{cth} z)(\operatorname{cth} z - 2z) + (1 + 4z^2)\operatorname{cth} 2z}{\operatorname{sh} z + 2z\operatorname{ch} z} = 0,$$

а тому з (10) маємо, що $\lim\limits_{\varepsilon\to 0} b = 0$, а з (12) — $\lim\limits_{\varepsilon\to 0}\alpha = \frac{1}{2}$. Це означає, що і для розв'язку задачі (5) — (7) мають місце рівності

$$\lim_{\varepsilon\to 0} u_\varepsilon(x) = u(x), \quad x\in\bar{\Omega},$$

$$\lim_{\varepsilon\to 0} u_\varepsilon(x) = 0, \quad x\in\bar{\Omega}_1,$$

тобто у вихідній області $\bar{\Omega}$ розв'язок задачі МФО (5) — (7) прямує до розв'язку задачі (1) при $\varepsilon\to 0$. Дещо подібне має місце і в багатовимірному випадку.

Сформулюємо задачі методу фіктивних областей для вихідної задачі

$$Lu \equiv -\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right) + a(x)u = f(x), \quad x\in\Omega,$$
$$u(x) = 0, \quad x\in\partial\Omega, \quad (14)$$

де Ω — обмежена n-вимірна область з кусково-гладкою границею $\partial\Omega$; $a_{ij}(x) = a_{ji}(x)$, $a_{ij}(x)\in C^1(\bar{\Omega})$, $a(x)$ — невід'ємна, вимірна, обмежена функція, $f(x)\in L_2(\Omega)$, і, крім того, виконується умова еліптичності

$$\sum_{i,j=1}^{n} a_{ij}(x)\,t_i t_j \geqslant \mu\sum_{i=1}^{n} t_i^2, \quad \mu > 0, \quad \mu = \text{const}.$$

Перший з розглянутих варіантів МФО для задачі (14) має вигляд

$$L_\varepsilon u_\varepsilon \equiv -\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}\left(a_{ij}^\varepsilon\frac{\partial u_\varepsilon}{\partial x_j}\right) + a^\varepsilon(x)u_\varepsilon = f_1(x), \quad x\in\Omega_0,$$
$$u_\varepsilon(x) = 0, \quad x\in\partial\Omega_0, \quad (15)$$
$$[u_\varepsilon(x)]_{x\in\partial\Omega} = 0, \quad \left[\sum_{i,j=1}^{n} a_{ij}^\varepsilon\cos(n, x_i)\frac{\partial u_\varepsilon}{\partial x_j}\right]_{x\in\partial\Omega} = 0,$$

де Ω_0 — паралелепіпед, який містить у собі область Ω, $[v]$ означає стрибок функції $v(x)$ на поверхні $\partial\Omega$, тобто стрибок при переході через $\partial\Omega$ у напрямі зовнішньої нормалі n до границі $\Gamma_1 = \Omega_0 \setminus \Omega$,

$$a_{ij}^\varepsilon(x) = \begin{cases} a_{ij}(x), & x\in\Omega,\\ 0, & x\in\Omega_1, \quad i\neq j,\\ \varepsilon^{-2}, & x\in\Omega_1, \quad i = j,\end{cases}$$

$$a^\varepsilon(x) = \begin{cases} a(x), & x\in\Omega,\\ 0, & x\in\Omega_1,\end{cases} \qquad f_1(x) = \begin{cases} f(x), & x\in\Omega,\\ 0, & x\in\Omega_1.\end{cases}$$

Можна довести, що при певних умовах для вихідних даних задачі (14) має місце така оптимальна по ε оцінка:

$$\|u_\varepsilon - u\|_{C(\bar{\Omega})} \leqslant c\varepsilon^2,$$

де c — незалежна від ε додатна стала.

Другий варіант МФО для задачі (14) має вигляд

$$L_\varepsilon u_\varepsilon = f_1(x), \quad x\in\Omega_0, \quad u_\varepsilon(x) = 0, \quad x\in\partial\Omega_0, \quad (16)$$

де коефіцієнти a_{ij} продовжені в область Ω_1 так, щоб вони залишались обмеженими і виконувалась умова еліптичності в області Ω_0, $f(x)$ продовжена в Ω_1 нулем, а функція $a^\varepsilon(x)$ має вигляд

$$a^\varepsilon(x) = \begin{cases} a(x), & x\in\Omega,\\ \varepsilon^{-2}, & x\in\Omega_1.\end{cases}$$

Якщо вихідні дані задачі (14) забезпечують її розв'язку належність до класу $W_2^3(\Omega)$, то має місце оптимальна за порядком ε оцінка швидкості збіжності

$$\|u_\varepsilon - u\|_{W_2^1(\Omega)} \leqslant c\varepsilon\|f\|_{W_2^1(\Omega)},$$

де c — незалежна від ε та u додатна стала.

Існують також інші варіанти МФО. Його можна застосувати також до інших задач, наприклад третьої крайової задачі, нестаціонарних задач, до рівнянь вищих порядків та ін.

Таким чином, можна дати таку схему наближеного розв'язування задачі (14) в довільній обмеженій області Ω: 1) доповнюємо область Ω

до найменшого паралелепіпеда Ω_0 і в задачі (15) (або (16)) вибираємо ε так, щоб її розв'язок наближав розв'язок задачі (14) з необхідною точністю; 2) розв'язуємо задачу МФО методом сіток з потрібною точністю. Зауважимо, що на другому етапі здійснення цієї схеми можемо зіткнутися з трудністю, яка пов'язана з великим діапазоном зміни коефіцієнтів задачі (15) або (16), оскільки ε треба вибирати досить малим. В п. 5.4 розглянуто спеціальні варіанти поперемінно-трикутного методу, які враховують можливість сильного коливання величини коефіцієнтів у двовимірному еліптичному рівнянні другого порядку.

7.2. Метод граничних елементів

Метод граничних елементів (МГЕ) грунтується на таких розділах математики, як теорія інтегральних рівнянь та метод скінченних елементів. Давно відомо, що багато крайових задач для багатовимірних рівнянь з частинними похідними можна звести до так званих граничних інтегральних рівнянь, теорія яких грунтується на роботах Г. Гріна (1828) та І. Фредгольма (1903). Фактично це зводить диференціальну задачу, поставлену для деякої функції у n-вимірному евклідовому просторі, до інтегральної задачі для функції в $(n-1)$-вимірному просторі, тобто знижує розмірність задачі на одиницю. Граничне інтегральне рівняння наближено розв'язується за допомогою ідей методу скінченних елементів. Отже, метод граничних елементів для крайових задач для диференціальних рівнянь складається з трьох етапів: 1) зведення вихідної диференціальної задачі до граничного інтегрального рівняння; 2) наближене розв'язування інтегрального рівняння методом граничних елементів, тобто апроксимація границі скінченними елементами, заміна шуканої функції лінійною комбінацією «більш простих» функцій, заданих на цих граничних елементах, і, накінець, розв'язування системи алгебраїчних рівнянь відносно коефіцієнтів лінійної комбінації; 3) відновлення розв'язку вихідної диференціальної задачі всередині області.

Розглянемо це спочатку на прикладі одновимірної задачі

$$Lu \equiv -\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) + q(x)u(x) = f(x), \tag{1}$$

$$x \in \Omega \equiv (0, 1), \quad u(0) = u_0, \quad u'(1) = u_1, \tag{2}$$

де $q(x)$, $f(x)$ — задані функції; u_0, u_1 — задані числа. Помножимо рівняння (1) на деяку функцію $w(x)$ і проінтегруємо від 0 до 1 (іншими словами, помножимо скалярно в $L_2(0, 1)$):

$$\int_0^1 (Lu - f)\, w dx = 0. \tag{3}$$

Проінтегрувавши частинами, дістанемо

$$\int_0^1 \left(k\frac{du}{dx}\frac{dw}{dx} + quw - fw\right)dx - \left[k\frac{du}{dx}w\right]_0^1 = 0, \tag{4}$$

де $[v]_a^b = v(b) - v(a)$. Саме такого вигляду співвідношення ми брали в п. 4.3, п. 2—6 за узагальнену постановку задачі (1), (2) і на їх основі будували сіткові схеми методу скінченних елементів та різницеві схеми.

Вважаючи, що w має другі похідні, проінтегруємо в (4) ще раз частинами:

$$\int_0^1 (Lwu - fw)\, dx - \left[k\frac{du}{dx}w\right]_0^1 + \left[k\frac{dw}{dx}u\right]_0^1 = 0 \tag{5}$$

або, враховуючи умови (2),

$$\int_0^1 (Lwu - fw)\, dx - ku_1 w\,|_{x\in\Gamma_2} + k\frac{du}{dx}w\,|_{x\in\Gamma_1} + \\ + k\frac{dw}{dx}u\,|_{x\in\Gamma_2} - k\frac{dw}{dx}u_0\,|_{x\in\Gamma_1} = 0, \tag{6}$$

де $\Gamma_1 = \{0\}$, $\Gamma_2 = \{1\}$ — частини границі області $\Omega = (0, 1)$. Отже, рівність (6) для будь-якої функції w є наслідком співвідношень (1), (2), тобто розв'язок задачі (1), (2) задовольняє (6).

Розглянемо тепер таку задачу: знайти функцію u, задану в області $\bar{\Omega} = \Omega \cup \Gamma_1 \cup \Gamma_2$, яка задовольняє рівність (6) при будь-яких w, де k, f_1, u_0, u_1 — задані. Інтегруючи частинами двічі, перепишемо (6) у вигляді

$$\int_0^1 (Lu - f)\, w dx + (u - u_0)\, k\frac{dw}{dx}\Big|_{x\in\Gamma_1} + \left(\frac{du}{dx} - u_1\right) kw\,|_{x\in\Gamma_0} = 0. \tag{6'}$$

Це співвідношення показує, що (6) можна взяти за основу при побудові методів наближеного розв'язку задачі (1), (2). Якщо ми зможемо знайти функцію u з (6'), яка перетворює всі три складові в останній рівності на нуль при довільній функції w, то дістанемо точний розв'язок задачі (1), (2). Якщо ж деяка функція дає близькі до нуля значення цих доданків, то слід чекати, що вона близька до розв'язку задачі (1), (2), але це треба довести.

Припустимо, що ми можемо явно знайти розв'язок однорідного рівняння

$$Lw = 0.$$

Нехай це буде функція $\tilde{w}(x)$. Тоді з (5) дістанемо співвідношення

$$-\int_0^1 f\tilde{w}dx - \left[k\frac{du}{dx}\tilde{w}\right]_{x=0}^{x=1} + \left[k\frac{d\tilde{w}}{dx}u\right]_{x=0}^{x=1} = 0, \tag{7}$$

або в силу граничних умов

$$-\int_0^1 f\tilde{w}dx - k(1)u_1\tilde{w}(1) + k(0)\frac{du(0)}{dx}\tilde{w}(0) +$$

$$+ k(1)\frac{d\tilde{w}}{dx}u(1) - k(0)\frac{d\tilde{w}(0)}{dx}u_0. \qquad (8)$$

Це аналог граничного інтегрального рівняння в одновимірному випадку, який можна покласти за основу побудови наближених методів визначення невідомих $q_0 \equiv du(0)/dx$ та $p_1 \equiv u(1)$. Якщо відомі два лінійно незалежних розв'язки $\tilde{w}_i(x)$, $i = 1, 2$ рівняння (7), то замість (9) матимемо систему двох рівнянь

$$k(0)\tilde{w}_i(0)q_0 + k(1)\frac{d\tilde{w}_i(1)}{dx}p_1 = \int_0^1 f\tilde{w}_i dx + k(1)u_1\tilde{w}_i(1) +$$

$$+ k(0)\frac{d\tilde{w}_i(0)}{dx}u_0, \quad i = 1, 2, \qquad (9)$$

з якої точно знайдемо невідомі $q_0 = \frac{du(0)}{dx}$ та $p_1 = u(1)$.

Таким чином, цей метод дає змогу знайти розв'язок задачі (1), (2) та його похідної на границі області Ω в точках $x = 0$ та $x = 1$. Після цього ще треба відновити розв'язок всередині області Ω (третій етап МГЕ). У даному разі це можна зробити, наприклад, розв'язуючи замість (1), (2) задачу Діріхле для рівняння (1). Розглянемо застосування такого самого методу для крайових умов Діріхле для наступної конкретної задачі.

Приклад 1. Знайти похідну від розв'язку задачі

$$-\frac{d^2u}{dx^2} - u = x, \quad x \in \Omega \equiv (0, 1), \quad u(0) = u(1) = 0$$

на границі області Ω, тобто в точках $x = 0$, $x = 1$.

Розв'язання. Аналог співвідношення (5) має вигляд

$$-\int_0^1 \left(\frac{d^2w}{dx^2} + w\right)udx - \int_0^1 xwdx - [qw]_{x=0}^{x=1} + \left[u\frac{dw}{dx}\right]_{x=0}^{x=1} = 0,$$

де $q(x) = \frac{du}{dx}$, $q_0 = q(0)$, $q_1 = q(1)$. Лінійно незалежні розв'язки рівняння

$$\frac{d^2w}{dx^2} + w = 0$$

мають вигляд $w_1(x) = \cos x$, $w_2(x) = \sin x$. Аналогом системи рівнянь (9) є система

$$-q_1\cos 1 + q_0 = \int_0^1 x\cos x dx,$$

$$-q_1\sin 1 = \int_0^1 x\sin x dx,$$

звідки знаходимо $q_1 \equiv \frac{du(1)}{dx} = \frac{\cos 1}{\sin 1} = 1$, $q_0 \equiv \frac{du(0)}{dx} = \frac{1}{\sin 1} - 1$.

У розглянутому методі значна частина зусиль із розв'язання вихідної крайової задачі випадає на третій етап МГЕ, тобто відновлення розв'язку вихідної задачі всередині області. Це затруднення в МГЕ можна подолати, використовуючи граничні інтегральні співвідношення, визначені за допомогою так званих фундаментальних розв'язків.

Розглянемо задачу

$$Lu = f(x), \quad x \in \Omega \equiv (0, 1), \quad u(0) = u_0, \quad u(1) = u_1. \qquad (10)$$

За основу конструювання наближеного розв'язку цієї задачі візьмемо співвідношення вигляду (5), яке в даному випадку набирає вигляду

$$\int_0^1 (Lwu - fw)\,dx - \left[k\frac{du}{dx}w\right]_0^1 + k(1)u_1\frac{dw(1)}{dx} - k(0)u_0\frac{dw(0)}{dx} = 0. \qquad (11)$$

Функцію $w^*(x, x_i)$, яка є розв'язком рівняння

$$Lw^* = \delta(x - x_i),$$

де $\delta(x - x_i)$ — дельта-функція Дірака, зосереджена в точці x_i, назвемо фундаментальним розв'язком. Дельта-функція $\delta(x - x_i)$ є похідною від функції Хевісайда

$$H(x - x_i) = \begin{cases} 1, & x > x_i, \\ 0, & x < x_i. \end{cases}$$

Оскільки з (11) випливає рівність

$$-\frac{d}{dx}\left(k(x)\frac{dw^*(x)}{dx}\right) + q(x)w^*(x) = \frac{dH(x - x_i)}{dx},$$

то звідси формально маємо

$$\frac{dw^*(x, x_i)}{dx} \equiv \frac{dw^*(x)}{dx} = -k^{-1}(x)H(x - x_i) +$$

$$+ k^{-1}(x)\int_{x_i}^{x} q(x)w^*(x)\,dx.$$

Ця формула показує, що похідна від фундаментального розв'язку має стрибок величини $k^{-1}(x_i)$ при переході через точку x_i (якщо рухатися зліва направо по числовій осі). Зокрема,

$$\frac{dw^*(0,0)}{dx} = -k^{-1}(0), \quad \frac{dw^*(1,1)}{dx} = 0.$$

Підставивши в (11) замість w функцію w^*, дістанемо

$$u(x_i) - \int_0^1 fw^*(x, x_i)\,dx - \left[k\frac{du}{dx}w^*(x, x_i)\right]_{x=0}^{x=1} + u_0 = 0, \quad (12)$$

де x_i — довільна точка відрізка $[0, 1]$.

Покладаючи в (12) $x_i = 0$ та $x_i = 1$ і враховуючи, що $u(0) = u_0$, $u(1) = u_1$, дістаємо відносно невідомих $du(0)/dx$ та $du(1)/dx$ систему двох рівнянь (або граничні співвідношення)

$$k(0)w^*(0,0)\frac{du(0)}{dx} - k(1)w^*(1,0)\frac{du(1)}{dx} = -2u_0 + \int_0^1 fw^*(x,0)\,dx,$$

$$k(0)w^*(0,1)\frac{du(0)}{dx} - k(1)w^*(1,1)\frac{du(1)}{dx} = -u_1 + \int_0^1 fw^*(x,1)\,dx. \quad (13)$$

Знайшовши з (13) $\frac{du(0)}{dx}$, $\frac{du(1)}{dx}$, за формулою (12) можемо відновити розв'язок вихідної задачі (10) в будь-якій внутрішній точці x_i області Ω за формулою (12). При цьому залишається обгрунтувати, що знайдений нами розв'язок дійсно наближає розв'язок задачі (10).

Приклад 2. За допомогою фундаментального розв'язку знайти функцію u, що задовольняє умови

$$-\frac{d^2u}{dx^2} - u = x, \quad x \in (0, 1), \quad u(0) = 0, \quad u(1) = 0. \quad (14)$$

Розв'язання. Функція w^*, яка є розв'язком рівняння

$$-\frac{d^2w^*}{dx^2} - w^* = \delta(x - x_i),$$

має вигляд $w^*(x, x_i) = -\frac{1}{2}\sin|x - x_i|$. Оскільки

$$u_0 = u_1 = 0,$$

$$\int_0^1 xw^*(x,0)\,dx = -\frac{1}{2}\int_0^1 x\sin x\,dx = \frac{1}{2}x\cos x\Big|_0^1 - \frac{1}{2}\int_0^1 \cos x\,dx =$$

$$= \frac{1}{2}\cos 1 - \frac{1}{2}\sin x\Big|_0^1 = \frac{1}{2}\cos 1 - \frac{1}{2}\sin 1,$$

$$\int_0^1 xw^*(x,1)\,dx = -\frac{1}{2}\int_0^1 x\sin(1-x)\,dx = -\frac{1}{2}x\cos(1-x)\Big|_0^1 +$$

$$+\frac{1}{2}\int_0^1 \cos(1-x)\,dx = -\frac{1}{2} - \sin(1-x)\Big|_2^1 = \frac{1}{2}\sin 1 - \frac{1}{2},$$

то система (13) має вигляд

$$\begin{cases} \frac{1}{2}\sin 1\,\frac{du(1)}{dx} = \frac{1}{2}\cos 1 - \frac{1}{2}\sin 1, \\ -\frac{1}{2}\sin 1\,\frac{du(0)}{dx} = \frac{1}{2}\sin 1 - \frac{1}{2}, \end{cases}$$

звідки $\frac{du(1)}{dx} = \frac{\cos 1}{\sin 1} - 1$, $\frac{du(0)}{dx} = \frac{1}{\sin x} - 1$. За формулою (12) можна знайти, наприклад, значення розв'язку в точці $x_i = 1/2$:

$$u\left(\frac{1}{2}\right) = -\frac{1}{2}\int_0^{1/2} x\sin\left(\frac{1}{2} - x\right)dx - \frac{1}{2}\int_{1/2}^1 x\sin\left(x - \frac{1}{2}\right)dx -$$

$$-\frac{du(1)}{dx}\sin\frac{1}{2} + \frac{du(0)}{dx}\sin\frac{1}{2} = -\frac{1}{2}\frac{\cos 1 - 1}{\sin 1}\sin\frac{1}{2} -$$

$$-\frac{1}{2}\left(\frac{1}{2} - \sin\frac{1}{2}\right) - \frac{1}{2}\left(\frac{1}{2} + \sin\frac{1}{2} - \cos\frac{1}{2}\right) = 0{,}069746964,$$

що збігається з точним розв'язком, який неважко знайти аналітично.

Розглянемо далі задачу в частинних похідних

$$\Delta u = 0, \quad x \in \Omega, \quad u(x) = u_1(x), \quad x \in \Gamma_1, \quad (15)$$

$$\frac{\partial u(x)}{\partial n} = u_2(x), \quad x \in \Gamma_2,$$

де Ω — деяка область на площині (x_1, x_2) або (x_1, x_2, x_3) з границею $\Gamma = \Gamma_1 \cup \Gamma_2$; u_1, u_2 — задані функції; $\Delta = \partial^2/\partial x_1^2 + \partial^2/\partial x_2^2$, або $\Delta = \sum_{\alpha=1}^{3} \partial^2/\partial x_\alpha^2$ — оператор Лапласа; n — зовнішня нормаль до границі Γ.

Зведемо задачу (15) до граничного інтегрального рівняння, що, як і в одновимірному випадку, можна зробити не єдиним способом. Наводимо один із них.

Інтегруючи частинами аналогічно тому, як це робилося при виведенні співвідношення (6) або (6′), дістанемо з (15)

$$\int_{\Omega} \Delta u(x)\, u^*(\xi, x)\, dx = \int_{\Gamma_2} \left[\frac{\partial u(x)}{\partial n} - u_2(x)\right] u^*(\xi, x)\, d\Gamma - \int_{\Gamma_1} [u(x) - u_1(x)]\, u^*(\xi, x)\, d\Gamma, \tag{16}$$

де $u^*(\xi, x)$ — деяка вагова функція; $\xi = (\xi_1, \xi_2)$ — деяка змінна.

Розглянемо тепер (16) як самостійний об'єкт, де $u(x)$ — функція, яку потрібно знайти (поки що ми не знаємо, чи $u(x)$ в (16) і $u(x)$ в (15) мають щось спільне, чи ні). Інтегруючи частинами, з (16) дістаємо

$$-\int_{\Omega} \sum_{i=1}^{2} \frac{\partial u(x)}{\partial x_i} \frac{\partial u^*(\xi, x)}{\partial x_i}\, dx = -\int_{\Gamma_1} \frac{\partial u(x)}{\partial n} u^*(\xi, x)\, d\Gamma - \int_{\Gamma_2} u_2(x)\, u^*(\xi, x)\, d\Gamma - \int_{\Gamma_1} [u(x) - u_1(x)] \frac{\partial u^*(\xi, x)}{\partial n}\, d\Gamma,$$

а після повторного інтегрування частинами матимемо

$$\int_{\Omega} \Delta_x u^*(\xi, x)\, u(x)\, dx = -\int_{\Gamma_1} \frac{\partial u(x)}{\partial n} u^*(\xi, x)\, d\Gamma - \int_{\Gamma_2} u_2(x)\, u^*(\xi, x)\, d\Gamma + \int_{\Gamma_2} u(x) \frac{\partial u^*(\xi, x)}{\partial n}\, d\Gamma + \int_{\Gamma_1} u_1(x) \frac{\partial u^*(\xi, x)}{\partial n}\, d\Gamma$$

або

$$\int_{\Omega} \Delta_x u^*(\xi, x)\, u(x)\, dx = -\int_{\Gamma} \frac{\partial u(x)}{\partial n} u^*(\xi, x)\, d\Gamma + \int_{\Gamma} u(x) \frac{\partial u^*(\xi, x)}{\partial n}\, d\Gamma, \tag{17}$$

де Δ_x — оператор Лапласа по змінній x. Виберемо далі $u^*(\xi, x)$ як розв'язок рівняння

$$\Delta_x u^*(\xi, x) = -2\alpha\pi\delta(x, \xi), \tag{18}$$

де $\alpha = 1$ для двовимірної задачі і $\alpha = 2$ — для тривимірної, $\delta(x, \xi)$ — відповідна двовимірна або тривимірна δ-функція Дірака (нестрого її можна визначити властивостями $\delta(x, \xi) = 0$ при $\xi \neq x$, $\delta(x, \xi) = \infty$ при $x = \xi$ і $\int_{\Omega} u(x)\, \delta(x, \xi)\, dx = u(\xi)$). Нагадаємо, що розв'язок рівняння (18) називається *фундаментальним розв'язком рівняння Лапласа*. Неважко перевірити, що для двовимірної задачі (18) $u^*(\xi, x) = \ln \frac{1}{r}$, $r = \sqrt{(\xi_1 - x_1)^2 + (\xi - x_2)^2}$, а для тривимірної $u^*(\xi, x) = \frac{1}{r}$, $r = \sqrt{\sum_{\alpha=1}^{3} (\xi_\alpha - x_\alpha)^2}$. Підставляючи цей розв'язок в (17), дістаємо граничне інтегральне рівняння

$$2\alpha\pi u(\xi) + \int_{\Gamma} u(x) \frac{\partial u^*(\xi, x)}{\partial n}\, d\Gamma = \int_{\Gamma} \frac{\partial u(x)}{\partial n} u^*(\xi, x)\, d\Gamma, \tag{19}$$

де ξ — будь-яка точка площини (x_1, x_2) при $\alpha = 1$; ξ — будь-яка точка простору (x_1, x_2, x_3) при $\alpha = 2$; $u(x)$ — шукана функція із задачі (16).

Розглянемо далі двовимірну задачу (15). Якщо взяти точку ξ на границі Γ, то можна довести, що (19) матиме вигляд

$$c(\xi)\, u(\xi) + \int_{\Gamma} u(x) \frac{\partial u^*(\xi, x)}{\partial n}\, d\Gamma = \int_{\Gamma} \frac{\partial u(x)}{\partial n} u^*(\xi, x)\, d\Gamma, \tag{20}$$

де $c(\xi)$ — величина внутрішнього кута границі в точці ξ ($c(\xi) = \pi$, якщо в точці ξ границя є гладкою), ξ в інтегралах по границі є параметром. Зауважимо, що задача (20) має розмірність на одиницю меншу, ніж (16), і далі саме вона буде покладена в основу наближеного методу граничних елементів. Нагадаємо також, що для обгрунтування цього методу потрібно доводити, що його розв'язок наближує розв'язок вихідної задачі (15).

Розіб'ємо границю Γ на сегменти Γ_j так, що $\Gamma = \bigcup_{j=1}^{N} \Gamma_j$. Вважатимемо, що $N = N_1 + N_2$, де N_1 елементів покриває частину границі Γ_1, N_2 — частину границі Γ_2. Сегменти Γ_j називають граничними елементами. Тоді (20) можна записати у вигляді

$$c(\xi)\, u(\xi) + \sum_{j=1}^{N} \int_{\Gamma_j} u(x) \frac{\partial u^*(\xi, x)}{\partial n}\, d\Gamma = \sum_{j=1}^{N} \int_{\Gamma_j} \frac{\partial u(x)}{\partial n} u^*(\xi, x)\, d\Gamma. \tag{21}$$

Нехай відомі параметричні рівняння елемента Γ_j

$$x_1 \equiv x_1(\eta) = \psi_1(\eta), \quad x_2 \equiv x_2(\eta) = \psi_2(\eta), \quad \eta \in [-1, 1]. \tag{22}$$

Тоді

$$\int_{\Gamma_j} u(x) \frac{\partial u^*(\xi, x)}{\partial n}\, d\Gamma = \int_{-1}^{1} \tilde{u}(\eta) \frac{\partial \tilde{u}^*(\xi, \eta)}{\partial n} |G(\eta)|\, d\eta,$$

$$\int_{\Gamma_j} \frac{\partial u(x)}{\partial n} u(\xi, x)\, d\Gamma = \int_{-1}^{1} \frac{\partial \tilde{u}(\eta)}{\partial n} \tilde{u}^*(\xi, \eta) |G(\eta)|\, d\eta,$$

де $\tilde{v}(\eta) = v(x(\eta)) = v(x_1(\eta), x_2(\eta))$; G — якобіан переходу

$$|G| = |G(\eta)| = \sqrt{\left(\frac{dx_1}{d\eta}\right)^2 + \left(\frac{dx_2}{d\eta}\right)^2} = \frac{d\Gamma}{d\eta}, \quad d\Gamma = |G|\, d\eta.$$

Виберемо на кожному елементі k точок, які називатимемо вузлами $x_i = (x_{1i}, x_{2i}) = (x_1(\eta_i), x_2(\eta_i))$, $i = \overline{1, k}$ (точки x_i та η_i будемо ототожнювати), і побудуємо інтерполяційні многочлени

$$p(\eta; \tilde{u}) = \sum_{l=1}^{k} u_l \varphi_l(\eta) = \tilde{u}(\eta) - R_u(\eta),$$
$$p(\eta, \tilde{q}) = \sum_{l=1}^{k} q_l \varphi_l(\eta) = \tilde{q}(\eta) - R_q(\eta), \tag{24}$$

де $u_l = u(x_l) \equiv \tilde{u}(\eta_l)$, $q_l = \dfrac{\partial u(x_l)}{\partial n} = \dfrac{\partial \tilde{u}(\eta_l)}{\partial n} \equiv \tilde{q}(\eta_l)$, $q(x) = \dfrac{\partial u(x)}{\partial n}$, $\varphi_l(\eta)$ — фундаментальні многочлени інтерполяції, тобто многочлени не вище $(k-1)$-го степеня і такі, що $\varphi_l(\eta_j) = \delta_{lj}$; δ_{lj} — символ Кронекера; R_u, R_q — залишкові члени інтерполяції. Далі виконаємо в (21) заміну (22) з урахуванням (23), покладемо $\xi = x_i$ і підставимо замість $\tilde{u}$ та $\tilde{q}$ їх вирази з (24):

$$c_i u_i + \sum_{j=1}^{N} \int_{-1}^{1} p(\eta; u) \frac{\partial \tilde{u}^*(\eta_i, \eta)}{\partial n} |G(\eta)| d\eta =$$
$$= \sum_{j=1}^{N} \int_{-1}^{1} p(\eta; q) \tilde{u}^*(\eta_i; \eta) |G(\eta)| d\eta + R, \quad i = 1, \overline{N \times k}, \tag{25}$$

де

$$R = \sum_{j=1}^{N} \int_{-1}^{1} \left[-R_u(\eta) \frac{\partial \tilde{u}^*(\eta_i, \eta)}{\partial n} + R_q(\eta) \tilde{u}^*(\eta_i; \eta) \right] |G(\eta)| d\eta. \tag{26}$$

Відкидаючи в (25) залишковий член R і зберігаючи знак рівності, дістаємо систему лінійних алгебраїчних рівнянь відносно y_l та v_l:

$$c_i y_i + \sum_{j=1}^{N} \int_{-1}^{1} p(\eta; y) \frac{\partial \tilde{u}^*(\eta_i; \eta)}{\partial n_h} |G(\eta)| d\eta =$$
$$= \sum_{j=1}^{N} \int_{-1}^{1} p(\eta; v) \tilde{u}^*(\eta_i; \eta) |G(\eta)| d\eta, \quad i = 1, \overline{N \times k}, \tag{27}$$

де

$$p(\eta; y) = \sum_{l=1}^{k} y_l \varphi_l(\eta), \quad p(\eta; v) = \sum_{l=1}^{k} v_l \varphi_l(\eta),$$

y_l — деякі наближення для $u(x_l)$, v_l — наближення для $q_l = \dfrac{\partial u(x_l)}{\partial n}$. Оскільки на N_1 сегментах нам відомі значення $u(x)$, а на N_2 сегментах — значення $\dfrac{\partial u(x)}{\partial n}$, то покладемо у вузлах цих сегментів відповідно $y_l = u_1(x_l)$, $v_l = u_2(x_l)$. Тоді в системі (27) кількість невідомих і рівнянь збігатиметься і її можна записати у вигляді $AY = F$, де вектор Y складається з невідомих значень y_l та v_l.

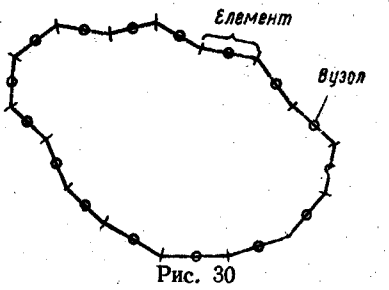


Рис. 30

Щоб спростити обчислення інтегралів у (27), можна елемент границі Γ_j апроксимувати за допомогою тих самих функцій $\varphi_l(\eta)$, тобто замість (22) наближено покласти

$$x_1 \approx \tilde{x}_1 = \tilde{x}_1(\eta) = \sum_{l=1}^{k} x_1^l \varphi_l(\eta),$$
$$x_2 \approx \tilde{x}_2 = \tilde{x}_2(\eta) = \sum_{l=1}^{k} x_2^l \varphi_l(\eta), \tag{28}$$

а коефіцієнти x_1^l, x_2^l вибрати з умови інтерполяції кривої Γ_j кривою Γ_{hj}, яка задається рівняннями (28):

$$\sum_{l=1}^{k} x_1^l \varphi_l(\eta_i) = x_{1i}, \quad \sum_{l=1}^{k} x_2^l \varphi_l(\eta_i) = x_{2i}, \quad i = \overline{1, k}.$$

Звідси маємо $x_1^l = x_{1l}$, $x_2^l = x_{2l}$, $l = \overline{1, k}$. У цьому разі Γ_{hj} називається *граничним елементом* $(k-1)$*-го порядку* і спрощується задача обчислення якобіана $G(\eta)$, отже, і інтегралів у (27). Крім того, можна наближено покласти $\partial \tilde{u}^*/\partial n \approx \partial \tilde{u}^*/\partial n_h$, де n_h — зовнішня нормаль до елемента Γ_{hj}.

Розв'язавши (27), розв'язок задачі (15) у будь-якій внутрішній точці ξ області Ω можна відновити за формулою (19), точніше за формулою

$$y(\xi) = \sum_{j=1}^{N} G(\xi, x_j) v_j - \sum_{j=1}^{N} \hat{H}(\xi, x_j) y_j,$$

де $G(\xi, x_j) = \int_{\Gamma_j} u^*(\xi, x) d\Gamma$; $\hat{H}(\xi, x_j) = \int_{\Gamma_j} \dfrac{\partial u^*(\xi, x)}{\partial n} d\Gamma$; $u^* = \dfrac{1}{2\pi} \ln \dfrac{1}{r}$; $y(\xi)$ — наближення для $u(\xi)$.

Розглянемо тепер найпростіший приклад граничних елементів нульового порядку або сталих граничних елементів для задачі (15), коли на кожному елементі обирається лише один вузол x_j. Оберемо його посередині кожного елемента (рис. 30) і апроксимуємо границю Γ об'єднанням відрізків прямих $\Gamma_h = \bigcup_{i=1}^{N} \Gamma_{hj}$. Якщо обрати $u^* =$

$= \frac{1}{2\pi} \ln \frac{1}{r}$ (у тривимірному випадку $u^* = \frac{1}{4\pi} \frac{1}{r}$), то у випадку гладкої границі дістаємо рівність (20) з $c(\xi) = \frac{1}{2}$ і, заміняючи в (21) Γ_j на Γ_{hj}, $u(x)$ та $\frac{\partial u(x)}{\partial n}$ на Γ_{hj} їхніми інтерполяційними многочленами нульового степеня, дістанемо

$$\frac{1}{2} y_i + \sum_{j=1}^{N} \hat{H}_{ij} y_j = \sum_{j=1}^{N} G_{ij} v_j, \quad i = \overline{1, N}. \tag{29}$$

Ввівши позначення

$$H_{ij} = \begin{cases} \hat{H}_{ij}, & i \neq j, \\ \hat{H}_{ij} + \frac{1}{2}, & i = j, \end{cases}$$

систему (29) можна записати у вигляді

$$\sum_{j=1}^{N} H_{ij} y_j = \sum_{j=1}^{N} G_{ij} v_j, \quad i = \overline{1, N} \tag{30}$$

або

$$Hy = Gv, \tag{31}$$

де

$$H = (H_{ij})_{i,j=1}^{N}, \quad G = (G_{ij})_{i,j=1}^{N}, \quad y = (y_i)_{i=1}^{N}, \quad v = (v_j)_{j=1}^{N},$$

$$\hat{H}_{ij} = \int_{\Gamma_{hj}} \frac{\partial u^*(x_i, x)}{\partial n} d\Gamma, \quad G_{ij} = \int_{\Gamma_{hj}} u^*(x_i, x)\, d\Gamma, \quad u^*(\xi, x) = \frac{1}{2\pi} \ln r,$$

$$r = \sqrt{(x_1 - \xi_1)^2 + (x_2 - \xi_2)^2}.$$

Поклавши $y_i = u_1(x_i)$, $v_j = u_2(x_j)$ для тих вузлів, які лежать на Γ_1 та Γ_2 відповідно, запишемо (31) у вигляді системи лінійних алгебраїчних рівнянь

$$AY = F,$$

де в Y входять лише невідомі y_i, v_j.

З формул (23) бачимо, що коли вузол x_i належить елементу Γ_{hi} (такий елемент називається сингулярним), то інтеграли у виразах для $\hat{H}_{ii}$ та G_{ii} мають особливості, а в іншому разі є інтегралами від гладких функцій. Тому при $i \neq j$ для обчислення $\hat{H}_{ij}$, G_{ij} використовуються квадратурні формули Гаусса

$$\hat{H}_{ij} = \int_{\Gamma_{hj}} \frac{\partial u^*(x_i, x)}{\partial n_h} d\Gamma = \int_{-1}^{1} \frac{\partial \tilde{u}^*(\eta_i, \eta)}{\partial n_h} |G(\eta)|\, d\eta \approx$$

$$\approx \frac{l_j}{2} \sum_{k=1}^{M} \frac{\partial \tilde{u}^*(\eta_i, \eta_k)}{\partial n} w_k,$$

$$G_{ij} = \int_{\Gamma_{hj}} u^* d\Gamma = \int_{-1}^{1} \tilde{u}^*(\eta_i, \eta) |G(\eta)|\, d\eta \approx \frac{l_j}{2} \sum_{k=1}^{M} \tilde{u}^*(\eta_i, \eta_k) w_k,$$

де l_j — довжина елемента; w_k — вагові коефіцієнти формул Гаусса; η_k — абсциси формули Гаусса (як правило, $M \leqslant 4$). Неважко помітити, що

$$\hat{H}_{ii} = \int_{\Gamma_{hi}} \frac{\partial u^*}{\partial n_h} d\Gamma = 0,$$

бо нормаль і елемент ортогональні. Для G_{ii} маємо

$$G_{ii} = \int_{\Gamma_{hi}} u^* d\Gamma = \frac{1}{2\pi} \int_{\Gamma_{hi}} \ln \frac{1}{r} d\Gamma.$$

Вводячи параметр η за формулою $r = \eta |r_1|$, $|r_1| = |r_2|$, дістаємо (рис. 31)

$$G_{ii} = \frac{1}{2\pi} \int_{(1)}^{(2)} \ln \frac{1}{r} d\Gamma = \frac{1}{\pi} \int_{(0)}^{(2)} \ln \frac{1}{r} d\Gamma = \frac{|r_1|}{\pi} \left[\ln \frac{1}{|r_1|} + \right.$$

$$\left. + \int_{0}^{1} \ln \frac{1}{\eta} d\eta \right] = \frac{1}{\pi} |r_1| \left(\ln \frac{1}{|r_1|} + 1 \right).$$

Розглянемо граничні елементи першого порядку або лінійні граничні елементи. У цьому разі на кожному сегменті Γ_j вибираємо два вузли на його кінцях і апроксимуємо Γ_j хордою Γ_{hj}, яка стягує ці два вузли (рис. 32). Множину вузлів позначимо $\omega_h = \{x_i \in \Gamma : i = \overline{1, N}\}$, $\Gamma = \bigcup_{i=1}^{N} \Gamma_j$.

На елементі Γ_j, кінці якого позначимо x_1 та x_2, зобразимо

$$u(x) = \tilde{u}(\eta) = p_1(\eta;\ \tilde{u}) + R = u_1 \varphi_1(\eta) + u_2 \varphi_2(\eta) + R =$$

$$= [\varphi_1 \varphi_2] \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + R = \varphi^T u^n + R, \tag{32}$$

$$q(x) \equiv \frac{\partial u(x)}{\partial n} = \tilde{q}(\eta) = p_1(\eta;\ \tilde{q}) + R_q = q_1 \varphi_1(\eta) +$$

$$+ q_2 \varphi_2(\eta) + R_q = [\varphi_1\ \varphi_2] \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} + R_q = \varphi^T q^n + R_q,$$

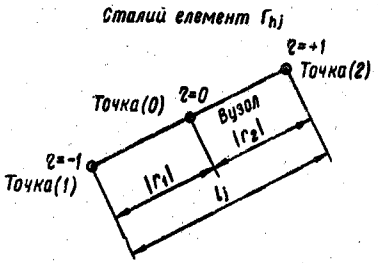


Рис. 31

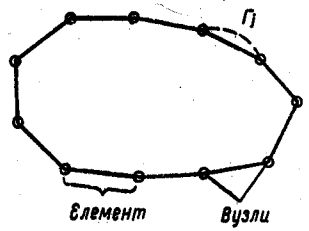


Рис. 32

де $u_i = u(x_i)$; $q_i = q(x_i)$; $\varphi_1 = \frac{1}{2}(1-\eta)$; $\varphi_2 = \frac{1}{2}(1+\eta)$; R_u, R_q — залишкові члени. Тоді (21) для гладкої границі $i\xi \in \omega_h$ набере вигляду

$$\frac{1}{2}u_i + \sum_{j=1}^{N}\int_{\Gamma_j} u(x)\frac{\partial u^*(x_i, x)}{\partial n}d\Gamma = \sum_{j=1}^{N}\int_{\Gamma_j}\frac{\partial u(x)}{\partial n}u^*(x_i, \xi)\,d\Gamma. \quad (33)$$

Нехай y_i, v_i — наближення відповідно для $u(x_i)$, $\frac{\partial u(x_i)}{\partial n}$. Замінимо в (33) $u(x)$, $\frac{\partial u(x)}{\partial n}$ на інтерполяційні многочлени $p_1(\eta; y)$ та $p_1(\eta; v)$ відповідно. Тоді

$$\int_{\Gamma_j} u\frac{\partial u^*(x_i, x)}{\partial n}d\Gamma \approx \int_{\Gamma_j}[\varphi_1\varphi_2]\frac{\partial u^*(x_i, x)}{\partial n}d\Gamma\begin{pmatrix}y_1\\y_2\end{pmatrix} \equiv [h_{ij}^{(1)}, h_{ij}^{(2)}]\begin{pmatrix}y_1\\y_2\end{pmatrix}, \quad (34)$$

де $h_{ij}^{(1)} = \int_{\Gamma_j}\varphi_1\frac{\partial u^*(x_i, x)}{\partial n}d\Gamma$, $h_{ij}^{(2)} = \int_{\Gamma_j}\varphi_2\frac{\partial u^*(x_i; x)}{\partial n}d\Gamma$. Величини $h_{ij}^{(k)}$ називаються коефіцієнтами впливу і характеризують зв'язок між вузлами x_i і k на елементі j.

Далі покладемо

$$\int_{\Gamma_j}\frac{\partial u(x)}{\partial n}u^*(x_i; x)\,d\Gamma \approx \int_{\Gamma_j}[\varphi_1\varphi_2]\,u^*(x_i, x)\,d\Gamma\begin{pmatrix}v_1\\v_2\end{pmatrix} \equiv [g_{ij}^{(1)}g_{ij}^{(2)}]\begin{pmatrix}v_1\\v_2\end{pmatrix}, \quad (35)$$

де

$$g_{ij}^{(1)} = \int_{\Gamma_j}\varphi_1 u^* d\Gamma, \quad g_{ij}^{(2)} = \int_{\Gamma_j}\varphi_2 u^* d\Gamma.$$

Враховуючи, що кожен вузол належить двом сусіднім елементам, дістанемо таке наближення для (33):

$$\frac{1}{2}y_i + [\hat{H}_{i1}\hat{H}_{i2}\ \dots\ \hat{H}_{iN}]\begin{pmatrix}y_1\\\dots\\y_N\end{pmatrix} = [G_{i1}\ \dots\ G_{iN}]\begin{pmatrix}v_1\\\dots\\v_N\end{pmatrix}, \quad i = \overline{1, N}, \quad (36)$$

де $\hat{H}_{ij}$ дорівнює сумі компоненти $h^{(2)}$ $(j-1)$-го елемента і компоненти $h^{(1)}$ j-го елемента, якщо використовується нумерація елементів у напрямі проти годинникової стрілки. Покладаючи $y_i = u_1(x_i)$ у вузлах x_i, які попадають на Γ_1 і $v_i = u_2(x_i)$ у вузлах, які попадають на Γ_2, з виразу (36) дістанемо систему лінійних алгебраїчних рівнянь виду

$$AY = F, \quad (37)$$

де A — повністю заповнена матриця порядку N; Y — вектор розмірності N, до якого входять невідомі значення y_k та v_k. Зауважимо, що коли обчислення інтегралів по Γ_j в (34), (35) викликає затруднення, можна їх замінити інтегралами по Γ_{h_j} і покласти $\partial u^*/\partial n \approx \partial u^*/\partial n_h$, де n_h — зовнішня нормаль до Γ_{h_j}.

Елементи вищого порядку дають змогу розширити кількість можливих апроксимацій. Наприклад, для кубічного елемента, який задається чотирма вузлами, можна використовувати апроксимацію

$$u = \sum_{i=1}^{4} u_i\varphi_i(\eta) + R_u,$$

$$\varphi_1 = \frac{1}{16}(1-\eta)[-10 + 9(\eta^2+1)],$$

$$\varphi_2 = \frac{1}{16}(1+\eta)[-10 + 9(\eta^2+1)],$$

$$\varphi_3 = \frac{9}{16}(1-\eta^2)(1-3\eta), \quad \varphi_4 = \frac{9}{16}(1-\eta^2)(1+3\eta)$$

або

$$u = u_1\varphi_1 + \left(\frac{\partial u}{\partial\eta}\right)_1\varphi_2 + u_2\varphi_3 + \left(\frac{\partial u}{\partial\eta}\right)_2\varphi_4,$$

$$\varphi_1 = \frac{1}{4}(\eta-1)^2(\eta+2), \quad \varphi_2 = \frac{1}{4}(\eta-1)^2(\eta+1),$$

$$\varphi_3 = \frac{1}{4}(\eta+1)^2(\eta-2), \quad \varphi_4 = \frac{1}{4}(\eta+1)^2(\eta-1),$$

які приведуть до різних систем виду (37).

На цьому закінчимо короткий виклад методу граничних елементів для задачі (15), залишаючи за рамками цієї книги питання його збіжності та порядку точності.

Розглянемо неоднорідну задачу:

$$\Delta u = f(x), \quad x \in \Omega, \quad u(x) = u_1(x), \quad x \in \Gamma_1, \quad (37')$$
$$\partial u(x)/\partial n = u_2(x), \quad x \in \Gamma_2.$$

Поступаючи аналогічно попередньому, замість (20) дістаємо

$$c(\xi)u(\xi)+\int_\Gamma u(x)\frac{\partial u^*(\xi,x)}{\partial n}d\Gamma+\int_\Omega f(x)u^*(\xi,x)dx=$$

$$=\int_\Gamma \frac{\partial u(x)}{\partial n}u^*(\xi,x)d\Gamma, \qquad (38)$$

і метод граничних елементів для (37′) відрізняється від методу граничних елементів для (15) лише тим, що потрібно додатково обчислювати інтеграли

$$B_i=\int_\Omega f(x)u^*(x_i,x)dx$$

одним з чисельних методів.

Метод граничних елементів можна застосувати і для більш загальних еліптичних задач виду

$$Lu=f(x),\quad x\in\Omega,\quad B_iu=0,\quad x\in\partial\Omega, \qquad (39)$$

де

$$Lu=\sum_{\mu=0}^{m}\sum_{p_1+p_2+\cdots+p_n=\mu}a_{p_1\ldots p_n}(x)\frac{\partial^\mu u}{\partial x_1^{p_1}\ldots\partial x_n^{p_n}}\equiv\sum_{|p|\leqslant m}a_p(x)D^pu,$$

$$p=(p_1,\ldots,p_n),\quad |p|=p_1+\cdots+p_n,$$

$$D=(D_1,\ldots,D_n),\quad D_j=\frac{\partial}{\partial x_j},$$

B_j — деякий граничний оператор.

Оператор

$$L^*u=\sum_{|p|\leqslant m}(-1)^{(p)}D^p(a_p(x)u)$$

називається *формально спряженим до* L. Помноживши (39) на деяку функцію w і виконавши інтегрування частинами, аналогічно (17) прийдемо до рівності

$$\int_\Omega L^*wudx=Gu,$$

де G — деякий граничний інтегральний оператор. Якщо $u^*(\xi,x)$ — фундаментальний розв'язок рівняння $L^*w=\delta(x-\xi)$, то аналогічно (19) дістанемо деяке граничне рівняння, до якого можна застосувати розглянуту обчислювальну схему.

Використання фундаментальних розв'язків — це не єдиний шлях виведення граничних інтегральних рівнянь. Відомо [32], наприклад,

що у двовимірному випадку функція

$$u(x)=\int_\Gamma \mu(\xi)\frac{\partial}{\partial n(\xi)}\left(\ln\frac{1}{r(\xi,x)}\right)d\Gamma(\xi), \qquad (40)$$

яка називається потенціалом подвійного шару ($n(\xi)$ — зовнішня нормаль до Γ у точці ξ), задовольняє рівняння Лапласа

$$\Delta^2u(x)=0,\quad x\in\Omega, \qquad (41)$$

при будь-якій функції $\mu(\xi)$, причому

$$u^+(x)-u^-(x)=-2\pi\mu(x), \qquad (42)$$

де $u^\mp(x)$ — значення потенціалу при прямуванні x до Γ відповідно зовні і з середини області Ω. Тому розв'язок задачі Діріхле для рівняння (41) при $u(x)=u_0(x)$, $x\in\Gamma$, u_0 — задана, можна шукати у вигляді (40) з невідомою функцією $\mu(\xi)$, яка називається *щільністю потенціалу подвійного шару*. Враховуючи (42), дістаємо для $\mu(\xi)$ граничне інтегральне рівняння Фредгольма другого роду

$$u_0(x)=-2\pi\mu(x)+\int_\Gamma \mu(\xi)\frac{\partial}{\partial n(\xi)}\left(\ln\frac{1}{r(\xi,x)}\right)d\Gamma(\xi), \qquad (43)$$

до якого можна застосувати обчислювальну схему методу граничних елементів. Теорію потенціалів можна використати і для розв'язування інших задач.

На закінчення підкреслимо ще раз переваги методу граничних елементів: 1) зменшення розмірності задачі на 1 і, як наслідок, на порядок зменшується кількість невідомих у дискретній задачі; 2) можливість застосування для областей довільної форми; 3) можливість розпаралелювання обчислення значень невідомої функції в середині області.

Недоліками цього методу є: 1) обмеженість класу задач, для яких можлива еквівалентна постановка у вигляді граничного інтегрального рівняння; 2) цілковита заповненість матриці дискретної задачі (37). Слід зауважити, що останній недолік часто повністю компенсується першою перевагою.

7.3. Багатопроцесорні ЕОМ та паралельні алгоритми обчислень

7.3.1. Суперкомп'ютери і паралельні обчислення. Більшість сучасних ЕОМ виконують обчислення послідовно. Це частково пов'язано з історією розвитку ЕОМ, в ході якої склалась їхня структура, а частково — з традиційним послідовним способом мислення людини.

Якщо не брати до уваги різні периферійні пристрої, наприклад пристрій вводу — виводу (ПВВ), то така ЕОМ складається з трьох

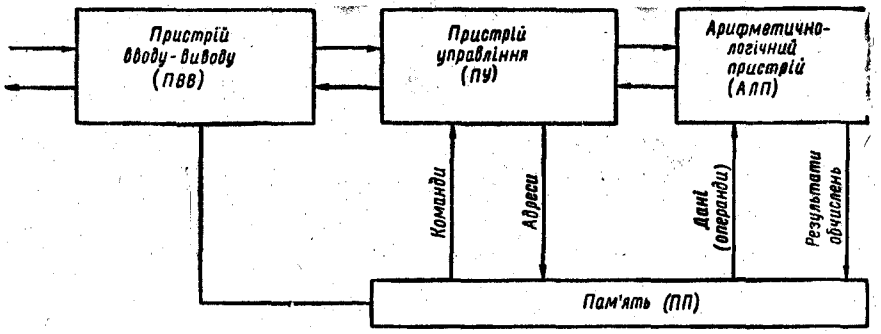


Рис. 33

основних блоків (рис. 33): пам'яті (ПП), пристрою управління (ПУ) та арифметично-логічного пристрою (АЛП). Склад ЕОМ та зв'язки між її складовими називають ще *архітектурою* ЕОМ. Блоки ПУ та АЛП називають також *центральним процесором* (ЦП) і оскільки ЦП всього один, то такі ЕОМ називають ще *однопроцесорними*. На регістрах ЦП виконуються всі операції ЕОМ. У комірках пам'яті зберігаються команди (програми), які виконуються пристроєм управління, і дані, які обробляються на АЛП. Це ЕОМ так званої класичної архітектури, авторами якої вважаються А. Бйоркс, Г. Голдстайн і Дж. фон Нейман.

Цифрові ЕОМ є пристроями, які працюють дискретно за часом, і крок цієї дискретизації (такт ЕОМ або час циклу) строго фіксований. За час циклу встановлюються всі перехідні процеси в електронних схемах ЕОМ, отже, стан ЕОМ, під яким розуміють вміст комірок пам'яті і регістрів ЦП, визначений однозначно в кожний момент дискретного часу. Виконання команди зводиться до її пошуку, тобто читання, розшифровки і формування адрес (у пам'яті) операндів пристроєм управління і виконання в АЛП передбаченої операції над операндами. Лише деякі елементарні команди виконуються за один такт ЕОМ, а інші реалізуються за декілька тактів. Як правило, АЛП містить декілька функціональних пристроїв (ФП), які спеціалізуються на виконанні окремих фаз операцій і умовно можна вважати, що кожен з них спрацьовує за один такт.

Щоб підвищити продуктивність ЕОМ, можна або збільшити її швидкість дії, тобто скоротити час циклу (такту), або ж збільшити кількість операцій, які виконує ЕОМ за одиницю часу.

У зв'язку з фізичними обмеженнями на час спрацювання електронних схем і на час поширення електричних сигналів перший із вказаних шляхів підвищення продуктивності вже зараз майже вичерпав свої можливості. Другий шлях пов'язаний зі створенням ЕОМ, які можуть одночасно (паралельно) виконувати різні операції. Такі ЕОМ називаються *паралельними* і за їхню високу продуктивність іноді називають також *суперкомп'ютерами*. Природно, що такі ЕОМ повинні мати багато одночасно взаємодіючих процесорів та інших пристроїв. За видом цієї взаємодії вони належать до різних класів. Паралельні ЕОМ можна класифікувати таким чином: 1) ЕОМ типу IMD (один потік команд, багато потоків даних), до яких належать Illioc IV, Staran, Cyber 203/205, Gray = 1 і т. д.; 2) ЕОМ типу MIMD (декілька потоків команд, декілька потоків даних), до яких належать DDA, C. mmp, SMS 201 та ін.

Одним із таких класів є паралельні ЕОМ, які працюють за конвеєрним принципом. Розглянемо її роботу на прикладі додавання двох N-вимірних векторів $A = (a_i)$ та $B = (b_i)$. Вважатимемо, що операція додавання двох чисел на деякій ЕОМ може реалізовуватися на p функціональних пристроях (ФП), причому кожний ФП спрацьовує за один такт. Якщо операнди подавати в АЛП з інтервалом у p тактів, чекаючи кінця обробки попередньої пари, то всі N операцій додавання $A + B$ будуть виконані за $T_1(N) = N_p$ тактів. Такий спосіб обробки даних називається *послідовним*. Його недоліком є недостатня завантаженість ФП, бо на кожному такті працює лише один ФП (рис. 34, *а*). Арифметичні операції в послідовних ЕОМ виконуються значно повільніше, ніж передача даних, тому головна вимога до чисельних алгоритмів для них — мінімізація кількості арифметичних операцій.

Якщо ж подавати набори операндів з інтервалом в один такт, то зразу над p наборами операндів виконуватимуться p фаз операції одночасно (рис. 34, *б*): подібно обробці на конвеєрі операнди передаються від пристрою до пристрою з інтервалом в один такт. Тому перший результат дістанемо через p тактів, а останній — через $T_p(N) = N - 1 + p$ тактів після початку обробки. Це і є конвеєрний спосіб обробки даних, а відповідний ланцюжок з p ФП називається конвеєром з p ступенями. Іншими прикладами векторних операцій можуть бути знаходження лінійної комбінації N-вимірних векторів $\alpha A + \beta B + \gamma C$ або обчислення N значень многочлена $y_i = p_k(x_i)$ у N вуз-

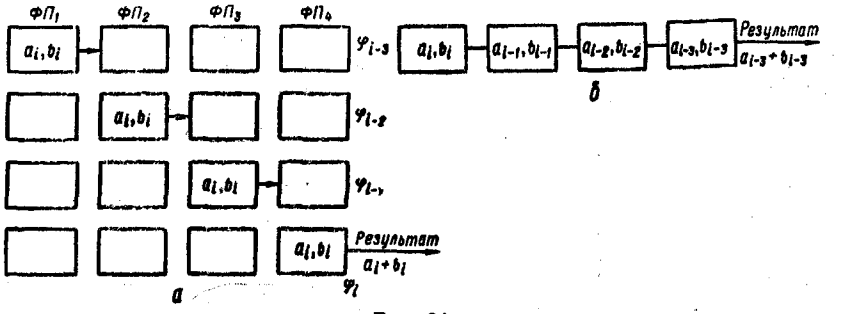


Рис. 34

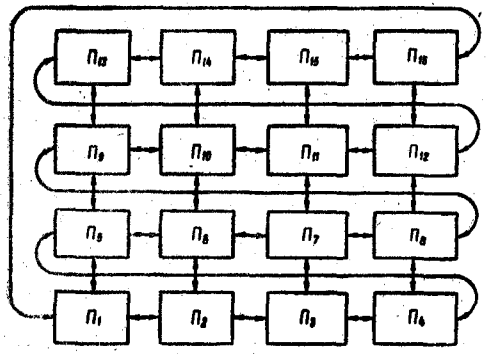


Рис. 35

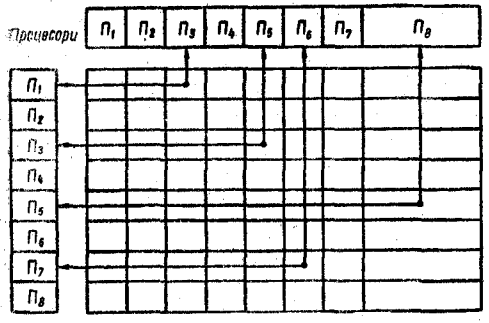


Рис. 36

лах сітки x_i, $i = \overline{1, N}$, і т. п. На p-ступеневому конвеєрі векторна операція може бути прискорена в

$$s_p = \frac{T_1(N)}{T_p(N)} = \frac{1}{1 + \frac{p-1}{N}} p$$

разів. У сучасних конвеєрних або векторних ЕОМ число ступенів коливається від 2 до 50. Внаслідок великого числа ступенів та зменшення часу циклу швидкість дії таких конвеєрних суперкомп'ютерів досягає (0,1 ÷ 1,5) млрд операцій за секунду.

Іншим прикладом паралельної ЕОМ є обчислювальна система з p процесорами $П_q$, $q = \overline{1, p}$ одного типу, причому процесор $П_q$ включає блоки $ПУ_q$, $АЛП_q$ та пам'ять p_q. Між процесорами можна встановити $\frac{1}{2} p (p-1)$ попарних зв'язків для обміну інформацією. Але оскільки такі прямі зв'язки важко реалізувати апаратно, то використовується обмежене число регулярних зв'язків, а процесори, які безпосередньо між собою не зв'язані, обмінюються інформацією через інші (рис. 35). Наприклад, обмін $П_1 \to П_8$ реалізується як $П_1 \to П_{16} \to П_{12} \to П_8$. Можлива ще одна схема зв'язків, коли вони здійснюються через програмований комутатор, який дає змогу реалізувати будь-які пари зв'язків (рис. 36). Можливі також інші структури паралельних ЕОМ, які класифікуються за складом і кількістю ФП, ПУ, АЛП і пам'яті, а також за характером зв'язків між пристроями. Структура паралельної ЕОМ має бути максимально пристосована до ефективного розв'язування певного класу задач, і кожна структура ЕОМ потребує пристосованих до неї алгоритмів. Так, конвеєрні ЕОМ ефективні для алгоритмів, які містять багато векторних операцій великого розміру, а в алгоритмах для багатопроцесорних ЕОМ потрібно мінімізувати кількість і тривалість обмінів між процесорами при максимальному завантаженні їх.

Отже, в суперкомп'ютерах однаково важливу роль відіграють такі фактори: елементна база ЕОМ, архітектура ЕОМ, алгоритм, програма,

математичне забезпечення. Основною задачею чисельних методів для паралельних ЕОМ, що зрозуміло із сказаного, є не мінімізація кількості арифметичних операцій (як на ЕОМ послідовної дії), а мінімізація числа тактів при реалізації обчислювального алгоритму, що можна забезпечити його максимальним розпаралелюванням, тобто виділенням у ньому значних частин, які можна виконувати паралельно. Мірою ефективності паралельного алгоритму на конкретній паралельній ЕОМ може бути прискорення

$$s_p = T_1/T_p,$$

яке визначається відношенням часу (в тактах) виконання алгоритму на послідовній (однопроцесорній) T_1 і паралельній T_p ЕОМ. Ще однією характеристикою паралельного алгоритму може бути число

$$E_p = s_p/p \leqslant 1,$$

яке називається *ефективністю паралельного алгоритму*.

Зауважимо, що розробка паралельних алгоритмів є комплексною проблемою, яка пов'язує математичне моделювання, обчислювальну математику, теорію машин, програмування і, нарешті, стиль мислення людини. При розв'язуванні практичних задач паралелізму слід прагнути на всіх етапах: постановці задачі, виборі фізичної та математичних моделей, розробці чисельних методів, програмуванні і т. д.

Щоб оцінити прискорення паралельних алгоритмів, які ми розглянемо далі, введемо такі позначення: τ_a — число тактів, необхідне для виконання однієї арифметичної операції; τ_s — число тактів, необхідне для підготовки двох процесорів до обміну даних між собою; τ_c — число тактів, необхідне для передачі одного числа від процесора до процесора. Тоді параметри $\varkappa_s = \tau_s/\tau_a$, $\varkappa_c = \tau_c/\tau_a$ характеризуватимуть швидкість дії каналу зв'язку відносно швидкості дії АЛП, а передача l чисел потребує $\tau_s + l\tau_c$ тактів.

7.3.2. Паралельний алгоритм підсумування p чисел. Нехай на p-процесорній ЕОМ потрібно обчислити суму

$$z(p) = \sum_{q=1}^{p} z_q,$$

причому в пам'яті кожного процесора $П_q$ зберігається лише одне число z_q, $q = \overline{1, p}$. Із постановки задачі зрозуміло, що задача не може бути розв'язана без обміну інформації між процесорами. Але цей обмін можна організувати по-різному. Наприклад, при послідовній схемі обмінів (рис. 37) і одночасному підсумуванні для обчислення

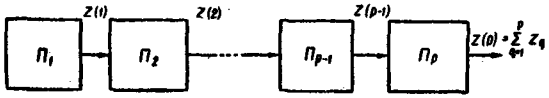


Рис. 37

результату треба

$$T_p^* = (p-1)(\tau_a + \tau_s + \tau_c)$$

тактів.

На однопроцесорній ЕОМ обчислення займуть лише $T_1 = (p-1)\tau_a$ тактів, тобто $T_p^* > T_1$, і таке використання багатопроцесорної ЕОМ не є ефективним.

Розглянемо алгоритм логарифмічного підсумування або підсумування за схемою «спарення». Припустимо для простоти, що $p = 2^m$, m — ціле число. Схема обмінів між процесорами для цього алгоритму зображена на рис. 38.

Неважко помітити, що на k-му кроці такого алгоритму паралельно виконуються $p/2^{k-1}$ обмінів і $p/2^k$ додавань на $p/2^k$ процесорах, тому

$$T_p = (\tau_a + \tau_s + \tau_c)\, m = (\tau_a + \tau_s + \tau_c) \log_2 p.$$

Величина прискорення є

$$s_p = \frac{p-1}{\log_2 p\,(1 + \varkappa_s + \varkappa_c)}.$$

Як видно, прискорення $s_p > 1$, якщо $1 + \varkappa_s + \varkappa_c < (p-1)/m$ $p = 2^m$. Наприклад, при $\varkappa_s + \varkappa_c = 3$ виграш порівняно з однопроцесорною ЕОМ досягається лише для ЕОМ з числом процесорів $p \geqslant 22$.

Вправа. Послідовний алгоритм обчислення суми $z(p) = \sum_{q=1}^{p} z_q$ можна подати у вигляді рекурентного співвідношення першого порядку $z(k) = z_k + z(k-1)$, $z(0) = 0$, $k = \overline{1, p}$. Побудувати паралельний алгоритм логарифмічного підсумуван-

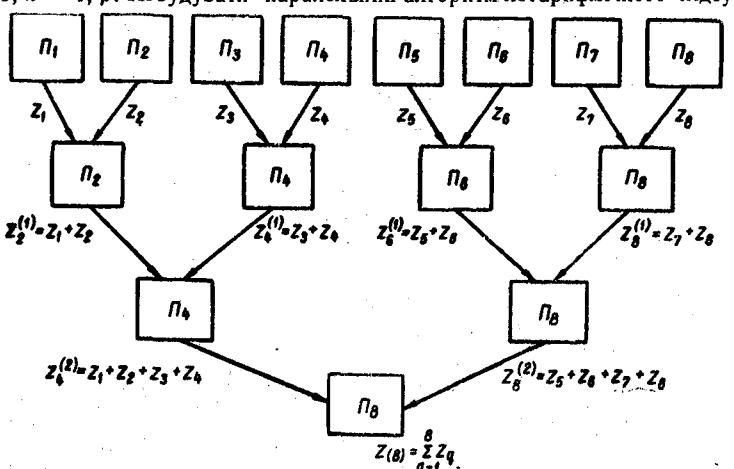


Рис. 38

ня для n рекурентних рівнянь m-го порядку

$$x_k = \begin{cases} 0, & k \leqslant 0, \\ c_k + \sum\limits_{j=k-m}^{k-1} a_{kj} x_j, \end{cases} \qquad (1)$$

де $k = \overline{1, n}$, c_k, a_{kj} — задані числа. У векторному вигляді (1) можна записати так:

$$x = Ax + c,$$

де $x = (x_1, \ldots, x_n)$, $c = (c_1, \ldots, c_n)$, $A = (a_{ij})$ — нижня трикутна матриця.

7.3.3. Паралельний алгоритм розв'язування систем лінійних алгебраїчних рівнянь (СЛАР). Як відомо, одним з найпопулярніших методів розв'язування СЛАР

$$Ax = b$$

на послідовних ЕОМ є так зване LU-зображення матриці A, тобто $A = LU$, де L — нижня трикутна матриця; U — верхня трикутна матриця. Природно виникає ідея знайти таке зображення матриці A, яке було б зручним для реалізації на паралельних ЕОМ.

Розглянемо QIF-зображення

$$A = WZ,$$

де W, Z — матриці вигляду

$$W = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ w_{2,1} & 1 & w_{2,3} & \ldots & w_{2,n} \\ w_{3,1} & w_{3,2} & 1 & \ldots & w_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ w_{n-1,1} & w_{n-1,2} & w_{n-1,3} & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix} \begin{matrix} \\ \\ \\ \\ w_{n-1,n} \\ 1 \end{matrix},$$

$$Z = \begin{pmatrix} z_{1,1} & z_{1,2} & \ldots & z_{1,n} \\ z_{2,1} & z_{2,2} & \ldots & z_{2,n-1} \\ \ldots & \ldots & \ldots & \ldots \\ z_{n,1} & z_{n2} & \ldots & z_{n,n} \end{pmatrix}.$$

Порівнюючи елементи матриці A та WZ, для визначення $w_{i,j}$ та $z_{i,j}$ дістаємо такі рівності. Для елементів першого та останнього рядків матриці Z: а) $z_{1,i} = a_{1,i}$, $z_{n,i} = a_{n,i}$, $i = \overline{1, n}$; б) елементи $w_{i,1}$, $w_{i,n}$, $i = \overline{2, n-1}$, визначаються із систем лінійних рівнянь другого порядку:

$$w_{i,1} z_{1,1} + w_{i,n} z_{n,1} = a_{i,1},$$
$$w_{i,1} z_{1,n} + w_{i,n} z_{n,n} = a_{i,n}.$$

У результаті цього першого етапу дістаємо перший та останній стовпчики матриці W та перший і останній рядки матриці Z. Аналогічно визначаються інші елементи матриць W, Z, для чого потрібно виконати $(n-1)/2$ етапів.

Після цього розв'язання системи $Ax = b$ зводиться до розв'язання двох систем:

$$Zx = y, \quad Wy = b,$$

де y — допоміжний вектор. Система $Wy = b$ має вигляд

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ w_{2,1} & 1 & \dots & w_{2,n} \\ w_{3,1} & w_{3,2} & \dots & w_{3n} \\ \dots & \dots & \dots & \dots \\ w_{n-1,1} & w_{n-1,2} & \dots & w_{n-1,n} \\ 0 & 0 & \dots & 1 \end{pmatrix} \times \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{pmatrix}.$$

З цієї системи можна одночасно знайти y_1, y_n, потім y_2 та y_{n-1} і т. д. Розв'язок системи $Zx = y$ знаходять аналогічно. Ми не будемо зараз зупинятися на питаннях розміщення даних на процесорах та взаємодії цих процесорів, обмежившись принциповою можливістю розпаралелювання.

7.3.4. Паралельний алгоритм чисельного інтегрування. Інтеграл $I = \int_a^b \rho(x) f(x)\, dx$, як ми бачили в гл. 5, ч. 1, наближено обчислюється як сума вигляду

$$J(N) = \sum_{i=1}^{N} c_i f(x_i), \tag{2}$$

де x_i — задані вузли інтегрування; c_i — відомі коефіцієнти квадратурної формули. Вважатимемо, що значення функції $f(x_i)$ в одному вузлі x_i обчислюються за γ_f тактів або обчислені заздалегідь (тоді $\gamma_f = 0$). Алгоритм обчислення суми (2) на послідовній (однопроцесорній) ЕОМ має вигляд

$$J(0) \leftarrow 0; \tag{3}$$

$$J(i) \leftarrow J(i-1) + c_i f_i; \quad i = \overline{1, N},$$

де $\leftarrow$ — знак операції присвоєння. Цей алгоритм виконується за

$$T_1(N) = (2 + \gamma_f) N \tau_a$$

тактів. Неважко помітити, що

$$J(N) = \sum_{q=1}^{p} z_q, \quad z_q = \sum_{i=n(q-1)+1}^{nq} c_i f_i, \quad q = \overline{1, p}.$$

Тепер обчислення на p-процесорній ЕОМ можна організувати так: 1) кожну суму z_q обчислити на процесорі Π_q, розмістивши в його пам'яті всі необхідні значення функції f_i та вагових коефіцієнтів c_i, $i = n(q-1)+1, \dots, nq$, і застосовуючи алгоритм типу (3), тобто $z \leftarrow 0$; $z_q \leftarrow z_q + c_i f_i$; $i = \overline{n(q-1)+1, \; nq}$, $q = \overline{1, p}$ (для цього потрібно $T_{1,q} = (2 + \gamma_f) n \tau_a$ тактів для всіх q разом), 2) суму $J(N) = \sum_{q=1}^{p} z_q$ обчислити за алгоритмом логарифмічного підсумовування. Вся задача на p-процесорній ЕОМ буде розв'язана за

$$T_p(N) = (2 + \gamma_f) n \tau_a + (\tau_a + \tau_s + \tau_c) \log_2 p$$

тактів, і тому прискорення

$$s_p = \frac{p}{1 + \dfrac{p \log_2 p}{N(2 + \gamma_2)} (1 + \varkappa_s + \varkappa_c)}. \tag{4}$$

Якщо $\varkappa_s + \varkappa_c$ та $\frac{p \log_2 p}{N(2+\gamma_f)}$ малі (це можна трактувати відповідно так, що ЕОМ має швидкі канали зв'язку та велике завантаження кожного процесора), то ефективність p-процесорної ЕОМ на даній задачі висока, бо тоді $s_p \sim p$. Менша ефективність буде для ЕОМ з повільними каналами зв'язку і з малим завантаженням процесорів.

Розглянутий паралельний алгоритм обчислення суми (2) можна застосувати і для розв'язування інших задач.

Нехай, наприклад, на p-процесорній ЕОМ потрібно обчислити квадрат норми сіткової функції $r(x)$, $x \in \omega = \{x = (x_1, x_2) : x_{1i_1} = i_1 h, \; x_{2i_2} = i_2 h, \; i_\alpha = \overline{1, N_1}, \; h = 1/N_1\}$, тобто

$$\|r\|^2 = h^2 \sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_1} r_{i_1 i_2}^2.$$

Подавши $r(x)$, $x \in \omega$ у вигляді вектора $R = (R_i)_{i=\overline{1,N}}$, розмірності $N = N_1^2 = 1/h_1^2$ (це можна зробити, наприклад, поклавши $R_i = r_{i_1 i_2}$, $i = (i_2 - 1) N_1 + i_1$, $i_1 = \overline{1, N_1}$, $i_2 = \overline{1, N_1}$) ми зведемо до задачі вигляду (2):

$$\|r\|^2 = h^2 \sum_{i=1}^{N} R_i^2. \tag{5}$$

Якщо вважати, що всі елементи $r_{i_1 i_2}$ обчислені задані (задача А), то неважко знайти прискорення (див. (4))

$$s_p^{\mathrm{A}} = \frac{p}{1 + \frac{1}{2} h^2 p \log_2 p (1 + \varkappa_s + \varkappa_c)}. \tag{6}$$

Якщо ж кожне значення $r_{i_1 i_2}$ обчислюється за γ_r арифметичних операцій (задача Б), то прискорення

$$s_p^{\text{Б}} = \frac{p}{1 + (2 + \gamma_r)^{-1} h^2 p \log_2 p (1 + \varkappa_s + \varkappa_c)}. \qquad (7)$$

Під задачею Б можна, наприклад, розуміти обчислення норми нев'язки $r = Ay - f$ наближеного розв'язку y рівняння $Au = f$ різницевим оператором на п'ятиточковому шаблоні в сітковому прямокутнику (див. п. 5.1—5.3)

$$(Ay)_{i_1 i_2} = b_{i_1 i_2} y_{i_1 i_2 - 1} + a_{i_1 i_2} y_{i_1 - 1, i_2} - c_{i_1 i_2} y_{i_1 i_2} + \\ + \bar{a}_{i_1 i_2} y_{i_1 + 1, i_2} + \bar{b}_{i_1 i_2} y_{i_1, i_2 + 1},$$

де $a_{i_1 i_2}$, $\bar{a}_{i_1 i_2}$, $b_{i_1 i_2}$, $\bar{b}_{i_1 i_2}$ — коефіцієнти різницевого оператора; $y_{0, i_2} = y_{N_1 + 1, i_2} = y_{i_1, 0} = y_{i_1, N+1} = 0$. У цьому випадку нев'язка обчислюється за $\gamma_r = 10$ арифметичних операцій. Щоб наочно уявити величини прискорень, наводимо таку таблицю:

Крок сітки	$p = 16$, $\varkappa_c + \varkappa_c = 3$		$p = 64$, $\varkappa_s + \varkappa_c = 3$	
	$s_p^{\text{А}}$	$s_p^{\text{Б}}$	$s_p^{\text{А}}$	$s_p^{\text{Б}}$
1/32	14,2	15,6	36,5	56,9
1/64	15,6	15,9	53,9	62,0
1/128	15,8	15,98	61,1	63,5

7.3.5. Паралельний алгоритм розв'язування задачі Коші. Знову обмежимося розглядом принципової можливості розпаралелювання без врахування питань розміщення даних на процесорах і взаємодії процесорів.

Нехай маємо значення y_n розв'язку задачі Коші

$$\frac{du}{dt} = f(t, u), \quad u(0) = u_0 \qquad (8)$$

у точці $t_n = n\tau$ сітки $\omega_\tau = \{t_i = i\tau : i = 0, 1, \ldots\}$. Тоді на k процесорах можна одночасно методом Ейлера знайти блок значень

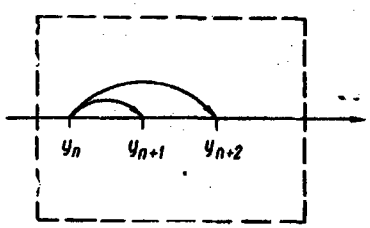


Рис. 39

$$y_{n+i} = y_n + (i\tau)\, y'_n, \quad i = \overline{1, k}, \qquad (9)$$

де $y'_n = f(t_n, y_n)$ (далі обмежимося випадком двохточкового блоку ($k = 2$)). Проте метод (9) має невисоку точність — порядку $O(\tau^2)$ на кроці і $O(\tau)$ на інтервалі інтегрування (див. п. 3.3—3.5). Щоб підвищити точність, побудуємо спочатку триточкові формули методом Адамса (рис. 39).

Із виразу (8) маємо

$$u(t_{n+1}) = u(t_n) + \int_{t_n}^{t_{n+1}} f(t, u(t))\, dt, \qquad (10)$$

$$u(t_{n+2}) = u(t_n) + \int_{t_n}^{t_{n+2}} f(t, u(t))\, dt. \qquad (11)$$

Замінимо в (10), (11) функцію $F(t) \equiv f(t, u(t))$ наближено інтерполяційним поліномом по точках t_n, t_{n+1}, t_{n+2}:

$$p_2(t) = f_n \frac{(t - t_{n+1})(t - t_{n+2})}{2\tau^2} - f_{n+1} \frac{(t - t_n)(t - t_{n+2})}{\tau^2} + \\ + f_{n+2} \frac{(t - t_n)(t - t_{n+1})}{2\tau^2} \equiv \frac{1}{2\tau^2} [y'_n (t - t_{n+1})(t - t_{n+2}) - \\ - 2 y'_{n+1} (t - t_n)(t - t_{n+2}) + y'_{n+2} (t - t_n)(t - t_{n+1})],$$

де $f_n \equiv f(t_n, y_n) \equiv y'_n$.

Інтегруючи в (10), (11) і замінюючи $u(t_n)$ наближенням y_n, дістаємо відповідно

$$y_{n+1} = y_n + \frac{\tau}{12} (5y'_n + 8y'_{n+1} - y'_{n+2}), \qquad (10')$$

$$y_{n+2} = y_n + \frac{\tau}{3} (y'_n + 4y'_{n+1} + y'_{n+2}). \qquad (11')$$

Неважко знайти, що формула (10′) має третій, а формула (11′) — четвертий порядок апроксимації на точному розв'язку, тобто

$$u_{n+1} = u_n + \frac{\tau}{12} (5u'_n + 8u'_{n+1} - u'_{n+2}) + O(\tau^4); \\ u_{n+2} = u_n + \frac{\tau}{3} (u'_n + 4u'_{n+1} + u'_{n+2}) + O(\tau^5). \qquad (12)$$

Візьмемо $y_{n+1,0} = y_n + \tau y'_n$, $y_{n+2,0} = y_n + 2\tau y'_n$ ($y_{n+1,0} - u_{n+1} = O(\tau)$, $y_{n+2,0} - u_{n+2} = O(\tau)$) за початкове наближення і обчислимо

$$y_{n+1,1} = y_n + \frac{\tau}{12} (5y'_n + 8y'_{n+1,0} - y'_{n+2,0}), \\ y_{n+2,1} = y_n + \frac{\tau}{3} (y'_n + 4y'_{n+1,0} + y'_{n+2,0}), \qquad (13)$$

де $y'_{n,0} = f(t_n, y_{n,0})$. При достатній гладкості функції $f(t, u)$ з виразів

(12), (13) легко знаходимо ($y_n - u_n = O(\tau^4)\ \forall\ n$)

$$y_{n+1,1} - u_{n+1} = O(\tau^2), \quad y_{n+2,1} - u_{n+2} = O(\tau^2).$$

Аналогічно дістаємо $y_{n+1,3} - u_{n+1} = O(\tau^4)$, $y_{n+2,3} - u_{n+2} = O(\tau^4)$, де

$$y_{n+1,i+1} = y_n + \frac{\tau}{12}(5y'_n + 8y'_{n+1,i} - y'_{n+2,i}),$$

$$y_{n+2,i+1} = y_n + \frac{\tau}{12}(y'_n + 4y'_{n+1,i} + y'_{n+2,i}), \quad i = 1, 2.$$

Таким чином, знаходимо такий паралельний алгоритм:

а) на двох процесорах обчислити за один такт

$$y_{n+2,0} = y_n + r\tau y'_n, \quad r = 1, 2; \tag{14}$$

б) для $s = 0, 1, 2$ обчислити

$$\begin{aligned} y_{n+1,s+1} &= y_n + \frac{\tau}{12}(5y'_n + 8y'_{n+1,s} - y'_{n+2,s}), \\ y_{n+2,s+1} &= y_n + \frac{\tau}{3}(y'_n + 4y'_{n+1,s} + y'_{n+2,s}), \end{aligned} \tag{15}$$

використовуючи для кожного s два процесори при одночасному обчисленні $y'_{n+1,s}$ та $y'_{n+2,s}$.

Вправа. Переконатись, що послідовні формули для обчислення y_{n+1} та y_{n+2} (за відомим y_n):

$$\begin{aligned} y_{n+1,0} &= y_n + \tau y'_n, \\ y_{n+1,1} &= y_n + \frac{\tau}{2}(y'_n + y'_{n+1,0}), \\ y_{n+2,1} &= y_n + 2\tau y'_{n+1,1}, \\ y_{n+1,2} &= y_n + \frac{\tau}{12}(5y'_n + 8y'_{n+1,1} - y'_{n+2,1}), \\ y_{n+2,2} &= y_n + \frac{\tau}{3}(y'_n + 4y'_{n+1,2} + y'_{n+2,1}) \end{aligned} \tag{16}$$

мають четвертий порядок апроксимації на розв'язку задачі Коші.

Якщо коефіцієнт прискорення рахувати за кількістю обчислень правої частини, то неважко помітити, що у випадку паралельного алгоритму (14), (15) та послідовного (16) він дорівнює 5/3. Можна побудувати аналогічні формули і для k-точкового блоку з прискоренням $(k + 3)/(2(1 + 1/k))$, а ще більшого прискорення, а саме k, можна досягти за допомогою багатоточкових методів типу предиктор — коректор (див., наприклад, Worland, P. B. Parallel methods for the numerical solution of ordinary differential equations, IEEE Trans. Comp., C—25, 10 (1976) 1045—8; Schampie L. F, Watts H. A. Block implicit one-step methods).

7.3.6. Паралельний алгоритм для явної двох'ярусної схеми. Розглянемо схему

$$\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = \varphi_k, \quad k = 0, 1, \ldots; \quad y_0 \text{ — задано}, \tag{17}$$

де A — $(2l + 1)$ — діагональна матриця розмірності $N \times N$, $N = pn$, τ_{k+1} — задані числа, а кожен елемент векторів φ_k та $Ay_k - \varphi_k$ може бути обчислений відповідно за γ_φ та γ_A арифметичних операцій.

Неважко підрахувати, що на однопроцесорній ЕОМ кожен крок за схемою (17) можна виконати за

$$T_1(N) = (2 + \gamma_A + \gamma_\varphi) N\tau_a$$

тактів.

Паралельний алгоритм виконання одного кроку за схемою (17) побудуємо таким чином. У процесорі Π_q номера q будемо обчислювати n елементів вектора $y_{k+1} = y_k - \tau_{k+1}(Ay_k - \varphi_k)$ з індексами $i = n(q - 1) + 1, \ldots, nq$. Зважаючи на структуру матриці A, зауважимо, що це можливо тоді, коли в пам'яті процесора Π_q зберігаються $n + 2l$ елементів вектора y_k з індексами $i = n(q - 1) + 1 - l, \ldots, nq + l$. Щоб перейти до обчислень на наступному кроці, потрібно переслати l перших і l останніх з обчислених елементів вектора y_{k+1} з пам'яті процесора Π_q у пам'ять процесорів Π_{q-1} і Π_{q+1} відповідно.

Отже, на кожному кроці реалізації схеми (17) процесорами обчислюється вектор y_{k+1} і проводиться обмін по $2l$ чисел. Неважко підрахувати, що

$$T_p(N) = (2 + \gamma_A + \gamma_\varphi) n\tau_a + \tau_s + l\tau_c,$$

$$s_p = \frac{p}{1 + p\dfrac{\varkappa_s + l\varkappa_c}{N(2 + \gamma_A + \gamma_\varphi)}}.$$

Наприклад, при $l = 1$, $\gamma_\varphi = 4$, $\gamma_A = 6$, $N = 64$, $\varkappa_s + \varkappa_c = 3$ дістаємо $s_{16} = 15$, $s_{61} = 51{,}2$.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. *Андреев В. Б., Руховец Л. А.* Проекционные методы // Новое в жизни, науке и технике. Сер. мат., кибернет.— М. : Знание, 1986.— Вып. 11.— С. 32.
2. *Бабенко К. И.* Основы численного анализа.— М. : Наука, 1986.— 744 с.
3. *Бахвалов Н. С.* Численные методы.— М. : Наука, 1973.— 623 с.
4. *Бахвалов Н. С., Жидков Н. П., Кобельков Г. М.* Численные методы.— М. : Наука, 1987. — 598 с.
5. *Березин И. С., Жидков Н. П.* Методы вычислений: В 2 т. М. : ГИФМЛ, 1962.— Т. 1.— 464 с.; Т. 2.— 639 с.
6. *Волков Е. А.* Численные методы.— М. : Наука, 1987.— 256 с.
7. *Гавурин М. К.* Лекции по методам вычислений.— М. : Наука, 1971.— 247 с.
8. *Годунов С. К., Рябенький В. С.* Разностные схемы.— М. : Наука, 1973.— 400 с.
9. *Деккер К., Вервер Я.* Устойчивость методов Рунге — Кутта для жестких нелинейных дифференциальных уравнений: Пер. с англ.— М. : Мир, 1988.— 247 с.
10. *Иванов В. В.* Методы вычислений на ЭВМ: Справ. пособие.— К. : Наук. думка, 1986.— 583 с.
11. *Калиткин Н. Н.* Численные методы.— М. : Наука, 1978.— 512 с.
12. *Крылов В. И., Бобков В. В., Монастырный П. И.* Вычислительные методы.— М. : Наука, 1976—1977.— Т. 1.— 584 с.; Т. 2.— 611 с.
13. *Ляшко И. И., Макаров В. Л., Скоробагатько А. А.* Методы вычислений.— К. : Вища шк. Головное изд-во, 1977.— 406 с.
14. *Марчук Г. И.* Методы вычислительной математики.— М. : Наука, 1980.— 536 с.
15. *Марчук Г. И., Агошков В. И.* Введение в проекционно-сеточные методы.— М. : Наука, 1981.— 416 с.
16. *Марчук Г. И., Шайдуров В. В.* Повышение точности решений разностных схем.— М. : Наука, 1979.— 247 с.
17. *Оганесян Л. А., Руховец Л. А.* Вариационно-разностные методы решения эллиптических уравнений.— Ереван : Изд-во АН АрмССР, 1979.— 235 с.
18. *Ортега Дж., Пул У.* Введение в численные методы решения дифференциальных уравнений.— М. : Наука, 1986.— 288 с.
19. *Самарский А. А.* Введение в численные методы.— М. : Наука, 1982.— 272 с.
20. *Самарский А. А.* Теория разностных схем.— М. : Наука, 1983.— 616 с.
21. *Самарский А. А., Андреев В. Б.* Разностные методы для эллиптических уравнений.— М. : Наука, 1976.— 351 с.
22. *Самарский А. А., Гулин А. В.* Устойчивость разностных схем.—М. : Наука, 1973.— 415 с.
23. *Самарский А. А., Гулин А. В.* Численные методы.— М. : Наука, 1982.— 429 с.
24. *Самарский А. А., Лазаров Р. Д., Макаров В. Л.* Разностные схемы для дифференциальних уравнений с обобщенными решениями.— М. : Высш. шк., 1987.— 296 с.
25. *Самарский А. А., Николаев Е. С.* Методы решения сеточных уравнений.— М. : Наука, 1978.— 591 с.
26. *Форсайт Дж., Малькольм М., Моулер К.* Машинные методы математических вычислений.— М. : Мир, 1980.— 280 с.
27. *Grigorieff R. D.* Numerik gewöhnlicher Differentialgleichungen.— Stuttgart : Teubner, 1972.— Vol. I.
28. *Grigorieff R. D.* Numerik gewöhnlicher Differentialgleichungen.— Stuttgart : Teubner, 1977.— Vol. II.
29. *Hackbusch W.* Multi-grid methods and applications.— New York ; Heidelberg; Berlin : Springer, 1985.
30. *Hackbusch W.* Integralgleichungen. Theorie und Numerik.— Stuttgart : Teubner, 1989.
31. *Isaakson E., Keller H. B.* Analysis of numerical methods. Wiley, New York, 1966.
32. *Maeß G.* Vorlesungen über numerische Mathematik II. Analysis.— Akademie — Verlag Berlin, 1988.
33. *Stoer J.* Extrapolation methods for the solution of initial value problems and their practical realization.— Berlin; Heidelberg; New York : LNM362, Springer, 1972.
34. *Stoer J., Bulirsch R.* Numerische Mathematik.— Berlin : Springer — Verlag, 1990.
35. *Förnig W., Spellucci P.* Numerische Mathematik für Ingeniuere und Physiker. Band 1. Lineare Algebra.— Berlin : Springer — Verlag, 1990.
36. *Canuto C., Hussaini M., Quarteroni A., Zang T.* Spectral Methods in Fluid Dynamics.— Berlin : Springer — Verlag, 1988.
37. *Deuflhard P., Hohmann A.* Numerische Mathematik. Eine algorithmische orientierte Einführung. Walter de Gruyter.— Berlin; New York, 1991.

ЗМІСТ

Навчальне видання

Гаврилюк Іван Петрович

Макаров Володимир Леонідович

МЕТОДИ ОБЧИСЛЕНЬ

У двох частинах

ЧАСТИНА 2

Оправа художника *В. Г. Самсонова*
Художній редактор *С. В. Анненков*
Технічний редактор *Т. Г. Шепновська*
Коректор *Н. І. Кунцевська*

Здано до набору 08.12.94. Підписано до друку 26.06.95. Формат 60×84¹/₁₆. Папір друк. № 2. Гарнітура літературна. Високий друк. Умовн.-друк. арк. 25,11. Умовн. фарбовідб. 25,34. Обл.-вид. арк. 24,32. Вид. № 9817. Замовлення № 4—2174.

Видавництво «Вища школа». 252054, Київ-54, вул. Гоголівська, 7.

Головне підприємство республіканського виробничого об'єднання «Поліграфкнига». 252057, Київ-57, вул. Довженка, 3.