ОСНОВИ МЕТОДІВ ОБЧИСЛЕНЬ

У ваших руках конспект лекцій з нормативного курсу “Основи методів обчислень”, прочитаного доц., к.ф.-м.н. Риженком Андрієм Івановичем на третьому курсі спеціальності прикладна математика факультету комп’ютерних наук та кібернетики Київського національного університету імені Тараса Шевченка восени 2018-го року.

Конспект у компактній формі відображає матеріал курсу, допомагає сформувати загальне уявлення про предмет вивчення, правильно зорієнтуватися в даній галузі знань. Конспект лекцій з названої дисципліни сприятиме більш успішному вивченню дисципліни, причому більшою мірою для студентів заочної форми, екстернату, дистанційного та індивідуального навчання.

Комп’ютерний набір та верстка – Скибицький Нікіта Максимович.

Зміст

1 Аналіз похибок заокруглення

1.1 Види похибок

Нехай необхідно розв'язати рівняння

$$Au = f. \tag{1.1.1}$$

За рахунок неточно заданих вхідних даних насправді ми маємо рівняння

$$\tilde{A}\tilde{u} = \tilde{f}. \tag{1.1.2}$$

Назвемо $\delta_1 = u - \tilde{u}$ – *неусувною похибкою.*

Застосування методу розв'язання (1.1.2) приводить до рівняння

$$\tilde{A}_h\tilde{u}_h = \tilde{f}_h, \tag{1.1.3}$$

де $h > 0$ – малий параметр. Назвемо $\delta_2\tilde{u} - \tilde{u}_h$ – *похибкою методу.*

Реалізація методу на ЕОМ приводить до рівняння

$$\tilde{A}_h^*\tilde{u}_h^* = \tilde{f}_h^*. \tag{1.1.4}$$

Назвемо $\delta_3 = \tilde{u}_h^* - \tilde{u}_h$ – *похибкою заокруглення.*

Тоді *повна похибка* $\delta = u - \tilde{u}_h^* = \delta_1 + \delta_2 + \delta_3$.

Визначення 1. *Кажуть, що задача (1.1.1) коректна, якщо*

1. $\forall f \in F\ \exists! u \in U$;
2. *задача (1.1.1) стійка, тобто*

$$\forall\varepsilon > 0 \quad \exists\delta > 0 : \|A - \tilde{A}\| < \delta, \|f - \tilde{f}\| < \delta \Rightarrow \|u - \tilde{u}\| < \varepsilon.$$

Якщо задача (1.1.1) *некоректна*, то або розв'язок її не існує, або він неєдиний, або він нестійкий, тобто

$$\exists\varepsilon > 0 : \forall\delta > 0 : \|A - \tilde{A}\| < \delta, \|f - \tilde{f}\| < \delta \Rightarrow \|u - \tilde{u}\| > \varepsilon.$$

Абсолютна похибка $\Delta x \leq |x - x^*|$.

Відносна похибка $\delta x \le \dfrac{\Delta x}{|x|}$, або $\dfrac{\Delta x}{|x^*|}$.

Значущими цифрами називаються всі цифри, починаючи з першої ненульової зліва.

Вірна цифра – це значуща, якщо абсолютна похибка за рахунок відкидання всіх молодших розрядів не перевищує одиниці розряду цієї цифри. Тобто, якщо

$$x^* = \overline{\alpha_n \dots \alpha_0.\alpha_{-1} \dots \alpha_{-p} \dots},$$

то α_{-p} – вірна, якщо $\Delta x \le 10^{-p}$ (інколи $\Delta x \le w \cdot 10^{-p}$, $\frac{1}{2} \le w < 1$, наприклад, $w = 0.55$).

1.2 Підрахунок похибок в ЕОМ

Підрахуємо відносну похибку заокруглення числа x на ЕОМ з плаваючою комою. В β-ічній системі числення число представляється у вигляді

$$x = \pm(\alpha_1\beta^{-1} + \alpha_2\beta^{-2} + \dots + \alpha_t\beta^{-t} + \dots)\beta^p, \tag{1.2.1}$$

де $0 \le \alpha_k < \beta$, $\alpha_1 \ne 0$, $k = 1, 2, \dots$.

Якщо в ЕОМ t розрядів, то при відкиданні молодших розрядів ми оперуємо з наближеним значенням

$$x^* = \pm(\alpha_1\beta^{-1} + \alpha_2\beta^{-2} + \dots + \alpha_t\beta^{-t})\beta^p,$$

і відповідно похибка заокруглення $x - x^* = \pm\beta^p(\alpha_{t+1}\beta^{-t-1} + \dots)$. Тоді її можна оцінити так

$$|x - x^*| \le \beta^{p-t-1}(\beta-1)(1 + \beta^{-1} + \dots) \le \beta^{p-t-1}(\beta-1)\frac{1}{1-\beta^{-1}} = \beta^{p-t}.$$

Якщо в представлені (1.2.1) взяти $\alpha_1 = 1$, то $|x| \ge \beta^p\beta^{-1}$. Звідси остаточно

$$\delta x \le \frac{\beta^{p-t}}{\beta^{p-1}} = \beta^{1-t}.$$

При більш точних способах заокруглення можна отримати оцінку $\delta x \le \frac{1}{2}\beta^{1-t} = \varepsilon$. Число ε називається “машинним іпсилон”. Наприклад, для $\beta = 2$, $t = 24$, $\varepsilon = 2^{-24} \approx 10^{-7}$.

1.3 Підрахунок похибок обчислення значення функції

Нехай задана функція $y = f(x_1, \ldots, x_n) \in C^1(\Omega)$.. Необхідно обчислити її значення при наближеному значенні аргументів $\vec{x}^* = (x_1^*, \ldots, x_n^*)$, де $|x_i - x_i^*| \le \Delta x_i$ та оцінити похибку обчислення значення функції $y^* = f(x_1^*, \ldots, x_n^*)$. Маємо

$$|y - y^*| = |f(\vec{x}) - f(\vec{x}^*)| = \left| \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{\xi})(x_i - x_i^*) \right| \le \sum_{i=1}^{n} B_i \cdot \Delta x_i,$$

де $B_i = \max\limits_{\vec{x} \in U} \left| \dfrac{\partial f}{\partial x_i}(\vec{x}) \right|$.

Тут $U = \{\vec{x} = (x_1, \ldots, x_n) : |x_i - x_i^*| \le \Delta x_i\} \in \Omega$, $i = \overline{1, n}$. Отже з точністю до величин першого порядку малості по $\Delta x = \max_i \delta x_i$, $\Delta y = |y - y^*| \prec \sum\limits_{i=1}^{n} n_i \cdot \Delta x_i$,

де $b_i = \left| \dfrac{\partial f}{\partial x_i}(\vec{x}^*) \right|$ та "$\prec$" означає приблизно менше.

Розглянемо похибки арифметичних операцій.

1. Сума: $y = x_1 + x_2$, $x_1, x_2 > 0$,

$$\Delta y \le \Delta x_1 + \Delta x_2, \quad \delta y \le \frac{\Delta x_1 + \Delta x_2}{x_1 + x_2} \le \max(\delta x_1, \delta x_2).$$

2. Різниця: $y = x_1 - x_2$, $x_1 > x_2 > 0$,

$$\Delta y \le \Delta x_1 + \Delta x_2, \quad \delta y \le \frac{x \delta x_1 + x_1 \delta x_2}{x_1 - x_2}.$$

   При близьких x_1, x_2 зростає відносна похибка (за рахунок втрати вірних цифр).

3. Добуток: $y = x_1 \cdot x_2$, $x_1, x_2 > 0$,

$$\Delta y \prec x_2 \Delta x_1 + x_1 \Delta x_2, \quad \delta y \le \delta x_1 + \delta x_2.$$

4. Частка: $y = \dfrac{x_1}{x_2}$, $x_1, x_2 > 0$,

$$\Delta y \prec \frac{x_2 \Delta x_1 + x_1 \Delta x_2}{x_2^2}, \quad \delta y \le \delta x_1 + \delta x_2.$$

   При малих x_2 зростає абсолютна похибка (за рахунок зростання результату ділення).

Пряма задача аналізу похибок: обчислення $\Delta y, \delta y$ по заданих Δx_i, $i = \overline{1,n}$.

Обернена задача: знаходження Δx_i, $i = \overline{1,n}$ по заданих $\Delta y, \delta y$. Якщо $n > 1$, маємо одну умову $\sum_{i=1}^{n} b_i \Delta x_i < \varepsilon$ для багатьох невідомих Δx_i. Вибирають їх із однієї з умов

$$\forall i : b_i \Delta x_i < \frac{\varepsilon}{n} \quad \text{або} \quad \Delta x_i < \frac{\varepsilon}{\sum_{i=1}^{n} b_i}.$$

2 Нелінійні рівняння

Постановка задачі. Нехай маємо рівняння $f(x) = 0$, $\overline{x}$ – його розв'язок, тобто $f(\overline{x}) \equiv 0$.

Задача розв'язання цього рівняння розпадається на етапи:

1. Існування та кількість коренів.
2. Відділення коренів, тобто розбиття числової вісі на інтервали, де знаходиться один корінь.
3. Обчислення кореня із заданою точністю ε.

Для розв'язання перших двох задач використовуються методи математичного аналізу та алгебри, а також графічний метод. Далі розглядаються методи розв'язання третього етапу.

2.1 Метод ділення навпіл

Припустимо на $[a,b]$ знаходиться лише один корінь рівняння

$$f(x) = 0, \tag{2.1.1}$$

для $f(x) \in C[a,b]$, який необхідно визначити. Нехай $f(a)f(b) < 0$.

Припустимо, що $f(a) > 0$, $f(b) < 0$. Покладемо $x_1 = \frac{a+b}{2}$ і підрахуємо $f(x_1)$. Якщо $f_1(x) < 0$, тоді шуканий корінь x знаходиться на інтервалі (a, x_1). Якщо ж $f_1(x) > 0$, то $\overline{x} \in (x_1, b)$, тобто з двох інтервалів (a, x_1) і (x_1, b) вибираємо той, на границях якого функція $f(x)$ має різні знаки, знаходимо точку x_2 – середину вибраного інтервалу, підраховуємо $f(x_2$ і повторюємо вказаний процес.

В результаті отримаємо послідовність інтервалів, що містять шуканий корінь $\overline{x}$, причому довжина кожного послідуючого інтервалу вдвічі менше попереднього.

Цей процес продовжується до тих пір, поки довжина отриманого інтервалу (a_n, b_n) не стане меншою за $b_n - a_n < 2\varepsilon$. Тоді x_{n+1}, як середина інтервалу (a_n, b_n) пов'язане з $\overline{x}$ нерівністю

$$|x_n + 1 - \overline{x}| < \varepsilon. \tag{2.1.2}$$

Ця умова для деякого n буде виконуватись за теоремою Больцано-Коші.

Оскільки

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

то

$$|x_{n+1} - \overline{x}| \le \frac{1}{2^{n+1}}(b - a) < \varepsilon. \tag{2.1.3}$$

Звідси отримаємо нерівність для обчислення кількості ітерацій n для виконання умови (2.1.2):

$$n = n(\varepsilon) \ge \left\lfloor \log\left(\frac{b-a}{\varepsilon}\right)\right\rfloor + 1.$$

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

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

2.2 Метод простої ітерації

Спочатку рівняння

$$f(x) = 0 \tag{2.2.1}$$

замінюється еквівалентним

$$x = \varphi(x) \tag{2.2.2}$$

Ітераційний процес має вигляді

$$x_{n+1} = \varphi(x_n), \quad n = 0, 1, \ldots \tag{2.2.3}$$

Початкове наближення x_0 задається.

Для збіжності велике значення має вибір функції $\varphi(x)$. Перший спосіб заміни рівняння полягає в відділенні змінної з якогось члена рівняння. Більш продуктивним є перехід від рівняння (2.2.1) до (2.2.2) з функцією $\varphi(x) = x + \tau(x)f(x)$,

де $\tau(x)$ – знакостала функція на тому відрізку, де шукаємо корінь.

Кажуть, що ітераційний метод *збігається*, якщо $\lim\limits_{k\to\infty} x_k = \overline{x}$.

Далі $U_r = \{x : |x - a| \le r\}$ відрізок довжини $2r$ з серединою в точці a.

З'ясуємо умови , при яких збігається метод простої ітерації.

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

$$|x_n - \overline{x}| \le \frac{q^n}{1-q}|x_0 - x_1| \le \frac{q^n}{1-q}(b-a). \tag{2.2.4}$$

Доведення. Нехай $x_{k+1}, x_k \in U_r$. Тоді має місце допоміжна нерівність:

$$\begin{aligned} |x_{k+1} - x_k| = |\varphi(x_k) - \varphi(x_{k-1})| = |\varphi'(\xi_k)(x_k - x_{k-1}) \le \\ \xi_k = x_k + \theta_k(x_{k+1} - x_k), \quad 0 < \theta_k < 1 \\ \le |\varphi'(\xi_k)| \cdot |x_k - x_{k-1}| \le q|x_k - x_{k-1}| = \ldots = q^k|x_1 - x_0|. \end{aligned}$$

Використаємо її для доведення теореми:

$$|x_{k+p} - x_k| = |x_{k+p} - x_{k+p-1} + \ldots + x_{k+1} - x_k| \le |x_{k+p} - x_{k+p-1}| + \ldots + |x_{k+1} - x_k| \le$$
$$\le (q^{k+p-1} + q^{k+p-2} + \ldots + q^k)|x_1 - x_0| = \frac{q^k - q^{k+p-1}}{1-q}|x_1 - x_0| \xrightarrow[k\to\infty]{} 0. \tag{2.2.5}$$

Бачимо що $\{x_k\}$ – фундаментальна послідовність. Значить вона збіжна. При $p \to \infty$ в (2.2.5) отримуємо (2.2.4). □

Визначимо кількість ітерацій для досягнення точності ε. З оцінки в теоремі отримаємо

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

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

Зауваження. Умова збіжності методу може бути замінена на умову Ліпшиця

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

Переваги методу: простота; при $q < \frac{1}{2}$ – швидше збігається ніж метод ділення навпіл; метод узагальнюється на системи. Недоліки методу:

1. при $q > \frac{1}{2}$ – збігається повільніше ніж метод ділення навпіл,
2. виникають труднощі при зведенні $f(x) = 0$ до $x = \varphi(x)$.

2.3 Метод релаксації

Якщо в методі простої ітерації для рівняння $x = x + \tau f(x) \equiv \varphi(x)$ вибрати $\tau(x) = \tau = \text{const}$, то ітераційний процес приймає вигляд

$$x_{n+1} = x_n + \tau f(x_n), \quad n = 0, 1, 2, \ldots \tag{2.3.1}$$

x_0 – задано. Метод можна записати у вигляді $\dfrac{x_{k+1} - x_k}{\tau} = f(x_k)$, $k = 0, 1, \ldots$. Оскільки $\varphi'(x) = 1 + \tau f'(x)$, то метод збігається при умові

$$|\varphi'(x)| = |1 + \tau f'(x)| \le q < 1.$$

Нехай $f'(x) < 0$, тоді (2.2.4) запишеться у вигляді: $-q \le 1 + \tau f'(x) \le q < 1$. Звідси

$$\tau |f'(x_k)| \le 1 + q < 2, \quad \text{і} \quad 0 < \tau < \frac{2}{|f'(x)|}.$$

Поставимо задачу знаходження τ, для якого $q = q(\tau) \to \min$. Для того, щоб вибрати оптимальний параметр τ, розглянемо рівняння для похибки $z_k = x_j - \overline{x}$.

Підставивши $x_k = x + z_k$ в (2.3.1), отримаємо

$$z_{k+1} = z_k + \tau f(\overline{x} + z_k).$$

В припущені $f(x) \in C^1[a, b]$ з теореми про середнє маємо

$$f(\overline{x} + z_k) = f(\overline{x}) + z_k f'(\overline{x} + \theta z_k) = z_k f'(\overline{x} + \theta z_k) = z_k f'(\xi_k),$$

$$z_{k+1} = z_k + \tau f'(\xi_k) z_k,$$

$$|z_{k+1}| \le |1 + \tau f'(\xi_k)| \cdot |z_k| \le \max_U |1 + \tau f'(\xi_k)| \cdot |z_k|,$$

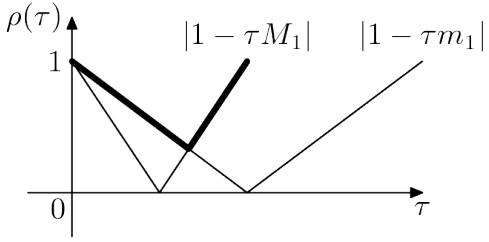
$$|z_{k+1}| \le \max\{|1 - \tau M_1|, |1 - \tau m_1|\} |z_k|,$$

$$m_1 = \min_{[a,b]} |f'(x)|, \quad M_1 = \max_{[a,b]} |f'(x)|.$$

Таким чином, задача вибору оптимального параметра зводиться до знаходження τ, для якого функція

$$q(\tau) = \max\{|1 - \tau M_1|, |1 - \tau m_1|\}$$

приймає мінімальне значення: $q(\tau) \to \min$.



З графіка видно, що точка мінімуму визначається умовою $|1-\tau M_1| = |1-\tau m_1|$.

Тому

$$1 - \tau_0 m_1 = \tau_0 M_1 - 1 \Rightarrow \tau_0 = \frac{2}{M_1 + m_1} < \frac{2}{|f'(x)|}.$$

При цьому значенні τ маємо

$$q(\tau_0) = \rho_0 = \frac{M_1 - m_1}{M_1 + m_1}.$$

Тоді для похибки вірна оцінка

$$|x_n - \overline{x}| \le \frac{\rho_0^n}{1 - \rho_0(b-a) < \varepsilon}.$$

Кількість ітерацій

$$n = n(\varepsilon) \ge \left\lfloor \frac{\ln\left(\dfrac{\varepsilon(1-\rho_0)}{b-a}\right)}{\ln \rho_0} \right\rfloor + 1.$$

Задача 1. Дати геометричну інтерпретацію методу простої ітерації для випадків:

$$0 < \varphi'(x) < 1; \quad -1 < \varphi'(x) < 0; \quad \varphi'(x) < -1; \quad \varphi'(x) > 1,$$

Задача 2. Знайти оптимальне $\tau = \tau_0$ для методу релаксації при $f'(x) > 0$.

2.4 Метод Ньютона (метод дотичних)

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

$$0 = f(\overline{x}) = f(x_k + \overline{x} - x_k) = f(x_k) + f'(\xi_k)(\overline{x} - x_k),$$

де $\xi_k = x_k + \theta_k(\overline{x} - x_k)$, $0 < \theta_k < 1$, $\xi_k \approx x_k$. Тому наступне наближення виберемо з рівняння

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

Звідси маємо ітераційний процес

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

Метод Ньютона ще називають методом лінеаризації або методом дотичних.

Задача 3. Дати геометричну інтерпретацію методу Ньютона.

Метод Ньютона можна інтерпретувати як метод простої ітерації з

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

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

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

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

Недолік методу Ньютона: необхідність обчислювати на кожній ітерації не тільки значення функції, а й похідної.

Модифікований метод Ньютона позбавлений цього недоліку і має вигляд:

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

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

Задача 4. Дати геометричну інтерпретацію модифікованого методу Ньютона.

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

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

Задача 5. Дати геометричну інтерпретацію методу січних.

2.5 Збіжність методу Ньютона

Теорема 2. *Нехай $f(x)\in C^2[a,b]$; $\overline{x}$ простий дійсний корінь рівняння*

$$f(x)=0 \tag{2.5.1}$$

і $f'(x)\neq 0$ при $x\in U_r=\{x:|x-\overline{x}|<r\}$. Якщо

$$q=\frac{M_2|x_0-\overline{x}|}{2m_1}<1 \tag{2.5.2}$$

де $m_1=\min\limits_{U_r}|f'(x)|$, $M_2=\max\limits_{U_r}|f''(x)|$, то для $x_0\in U_r$ метод Ньютона

$$x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)} \tag{2.5.3}$$

збігається і має місце оцінка

$$|x_n-\overline{x}\leq q^{2^n-1}|x_0-\overline{x}|. \tag{2.5.4}$$

Доведення. З (2.5.3) маємо

$$x_{k+1}-\overline{x}=x_k-\frac{f(x_k)}{f'(x_k)}-\overline{x}=\frac{(x_k-\overline{x})f'(x_k)-f(x_k)}{f'(x_k)}=\frac{F(x_k)}{f'(x_k)}, \tag{2.5.5}$$

де $F(x)=(x-\overline{x})f'(x)-f(x)$, така, що

1. $F(x)=0$;
2. $F'(x)=(x-\overline{x})f''(x)$;

Тоді

$$F(x_k)=F(\overline{x})+\int\limits_{\overline{x}}^{x_k}F'(t)\,\mathrm{d}t=\int\limits_{\overline{x}}^{x_k}(t-\overline{x})f''(t)\,\mathrm{d}t.$$

Так як $(t-x)$ не міняє знак на відрізку інтегрування, то скористаємося теоремою про середнє значення:

$$F(x_k) = f''(\xi_k)\int\limits_{\overline{x}}^{x_k}(t-\overline{x})\,\mathrm{d}t = \frac{(x_k-\overline{x})^2}{2}f''(\xi_k), \tag{2.5.6}$$

де $\xi_k = \overline{x} + \theta_k(x_k-\overline{x})$, $0<\theta_k<1$. З (2.5.5), (2.5.6) маємо

$$x_{k+1}-\overline{x} = \frac{(x_k-\overline{x})^2}{2f'(x_k)}f''(\xi_k). \tag{2.5.7}$$

Доведемо оцінку (2.5.3) за індукцією. Так як $x_0 \in U_r$, то

$$|\xi_0-\overline{x}| = |\theta_0(x_0-\overline{x})| < |\theta_0|\cdot|x_0-\overline{x}| < r \Rightarrow \xi_0 \in U_r.$$

Тоді $f''(\xi_0) \le M_2$, тому

$$|x_1-\overline{x}| \le \frac{(x_0-\overline{x})^2M_2}{2m_1} = \frac{M_2|x_0-\overline{x}|}{2m_1}|x_0-\overline{x}| = q|x_0-\overline{x}| = q|x_0-\overline{x}| < r, \quad x_1 \in U_r.$$

Ми довели твердження (2.5.4) при $n=1$. Нехай воно справджується при $n=k$:

$$|x_k-\overline{x}| \le q^{2^k-1}|x_0-\overline{x}| < r, \quad |\xi_k-\overline{x}| = |\theta_k(x_k-\overline{x})| < r.$$

Тоді $x_k, \xi_k \in U_r$.

Доведемо (2.5.4) для $n=k+1$. З (2.5.7) маємо

$$|x_{k+1}-\overline{x}| \le \frac{|x_k-\overline{x}|^2M_2}{2m_1} \le (q^{2^k-1})^2\frac{|x_0-\overline{x}|^2M_2}{2m_1} = q^{2^{k+1}-2}\frac{|x_0-\overline{x}|M_2}{2m_1}|x_0-\overline{x}| = q^{2^{k+1}-1}|x_0-\overline{x}|.$$

Таким чином (2.5.4) справджується для $n=k+1$. Значить (2.5.4) виконується і для довільного n. Таким чином $x_n \xrightarrow[n\to\infty]{} \overline{x}$. □

З (2.5.4) маємо оцінку кількості ітерацій для досягнення точності ε:

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

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

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

Для методу Ньютона $|x_{k+1} - \overline{x}| = \dfrac{|x_k - \overline{x}|^2 \cdot |f''(\xi_k)|}{2|f'(x_k)|} | \Rightarrow |x_{k+1} - \overline{x}| = O(|x_k - \overline{x}|^2)$.

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

Теорема 3. *Нехай $f(x) \in C^2[a,b]$ та $\overline{x}$ простий корінь рівняння $f(x) = 0$ а також $\forall x \in [a,b]$: $f'(x) \neq 0$. Якщо $f'(x)f''(x) > 0$ ($f'(x)f''(x) < 0$) то для методу Ньютона при $x_0 = b$ послідовність наближень $\{x_k\}$ монотонно спадає (монотонно зростає при $x_0 = a$).*

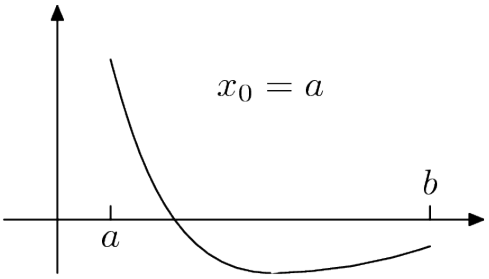
Задача 6. Довести теорему при

1. $f'(x)f''(x) > 0$;
2. $f'(x)f''(x) < 0$.

Задача 7. Знайти степінь збіжності методу січних.

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

Якщо $f(b)f''(b) > 0$, то $x_0 = b$; маємо $\{x_k\} \downarrow \overline{x}$. Пояснення на рисунку:



Зауваження 1. Якщо $\overline{x}$ p-кратний корінь тобто $f^{(m)}(x) = 0$, $m = \overline{0, p-1}$, $f^{(p)}(x) \neq 0$, то в методі Ньютона необхідна наступна модифікація

$$x_{k+1} = x_k - p\frac{f(x_k)}{f'(x_k)} \quad \text{і} \quad q = \frac{M_{p+1}|x_0 - \overline{x}|}{m_p p(p+1)} < 1.$$

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

Переваги методу Ньютона:

1. висока швидкість збіжності;
2. узагальнюється на системи рівнянь;
3. узагальнюється на комплексні корені.

Недоліки методу Ньютона:

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

3 Методи розв'язання СЛАР

Методи розв'язування СЛАР поділяються на прямі та ітераційні. При умові точного виконання обчислень прямі методи за скінчену кількість операцій в результаті дають точний розв'язок. Використовуються вони для невеликих та середніх СЛАР $n = 10^2..10^4$. Ітераційні методи використовуються для великих СЛАР $n > 10^5$, як правило розріджених. В результаті отримуємо послідовність наближень, яка збігається до розв'язку.

3.1 Метод Гаусса

Розглянемо задачу розв'язання СЛАР

$$A\vec{x} = \vec{b}, \tag{3.1.1}$$

причому $A = (a_{i,j})_{i,j=1}^n$, $\det A \neq 0$, $\vec{x} = (x_i)_{i=1}^n$, $\vec{b} = (b_j)_{j=1}^n$. Метод Крамера з обчисленням визначників для такої системи має складність $Q = O(n!n)$.

Запишемо СЛАР у вигляді

$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \ldots + a_{1,n}x_n = b_1 \equiv a_{1,n+1} \\ a_{2,1}x_1 + a_{2,2}x_2 + \ldots + a_{2,n}x_n = b_1 \equiv a_{2,n+1} \\ \ldots \\ a_{n,1}x_1 + a_{n,2}x_2 + \ldots + a_{n,n}x_n = b_1 \equiv a_{n,n+1} \end{cases} \tag{3.1.2}$$

Якщо $a_{1,1} \neq 0$, то ділимо перше рівняння на нього і виключаємо x_1 з інших рівнянь:

$$\begin{cases} x_1 + a^{(1)}_{1,2}x_2 + \ldots + a^{(1)}_{1,n}x_n = a^{(1)}_{1,n+1} \\ a^{(1)}_{2,2}x_2 + \ldots + a^{(1)}_{2,n}x_n = a^{(1)}_{2,n+1} \\ \ldots \\ a^{(1)}_{n,2}x_2 + \ldots + a^{(1)}_{n,n}x_n = a^{(1)}_{n,n+1} \end{cases}$$

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

$$\begin{cases} x_1 + a^{(1)}_{1,2}x_2 + \ldots + a^{(1)}_{1,n}x_n = a^{(1)}_{1,n+1} \\ x_2 + \ldots + a^{(2)}_{2,n}x_n = a^{(2)}_{2,n+1} \\ \ldots \\ x_n = a^{(n)}_{n,n+1} \end{cases} \tag{3.1.3}$$

Це прямий хід методу Гаусса. Формули прямого ходу

$$\begin{cases} k = \overline{1, n-1}: \\ a^{(k)}_{k,j} = \dfrac{a^{(k-1)}_{k,j}}{a^{(k-1)}_{k,k}}, \quad j = \overline{k+1, n+1}, \\ a^{(k)}_{i,j} = a^{(k-1)}_{i,j} - a^{(k-1)}_{i,k}a^{(k)}_{k,j}, \\ i = \overline{k+1, n}, \quad j = \overline{k+1, n+1}. \end{cases}$$

Звідси

$$x_n = a^{(n)}_{n,n+1}, \quad x_i = a^{(i)}_{i,n+1} - \sum_{j=i+1}^{n} a^{(i)}_{i,j}x_j, \quad i = \overline{n-1, 1}. \tag{3.1.4}$$

Це формули оберненого ходу.

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

ходу.

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

Введемо матриці

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

елементи якої обчислюється так $m_{k,k} = \dfrac{1}{a_{k,k}^{(k-1)}}$, $m_{k,k} = -\dfrac{a_{i,k}^{(k-1)}}{a_{k,k}^{(k-1)}}$.

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

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

Позначимо останню систему, яка співпадає з (3.1.2), так

$$U\vec{x} = \vec{c}, \quad U = (u_{i,j})_{i,j=1}^n, \tag{3.1.5}$$

причому

$$\begin{cases} u_{i,i} = 1 \\ u_{i,j} = 0, \quad i > j \end{cases}$$

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

$$L_k = \begin{pmatrix} 1 & & 0 & & 0 \\ & \ddots & & & \\ 0 & & a_{k,k}^{(k-1)} & & 0 \\ & & \vdots & \ddots & \\ 0 & & a_{n,k}^{(k-1)} & & 1 \end{pmatrix},$$

Тоді

$$A = L_1 \dots L_n U = LU; \quad L = L_1 \dots L_n,$$

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

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

За допомогою цих матриць перехід до трикутної системи (**??**) тепер має вигляд:

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

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

Висновки про переваги трикутного розкладу:

1. Розділення прямого і оберненого ходів дає змогу економно розв'язувати декілька систем з одноковою матрицею та різними правими частинами.
2. Зберігання M, або L та U на місці A.
3. Обчислюючи l – кількість перестановок, можна встановити знак визначника.

3.2 Метод квадратних коренів

Цей метод призначений для розв'язання систем рівнянь із симетричною матрицею

$$A\vec{x} = \vec{b}, \quad A^T = A. \tag{3.2.1}$$

Він оснований на розкладі матриці A в добуток:

$$A = S^T D S, \tag{3.2.2}$$

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

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

$$\begin{aligned}(DS)_{i,j} &= \begin{cases} d_{i,i}s_{i,j}, & i \le j \\ 0, & i > j \end{cases} \\ (S^T DS)_{i,j} &= \sum_{l=1}^{n} s_{i,l}^T d_{l,l} s_{l,j} = \sum_{l=1}^{i-1} s_{l,i}s_{l,j}d_{l,l} + s_{i,i}s_{i,j}d_{i,i} + \\ &+ \underbrace{\sum_{l=i+1}^{n} s_{l,i}s_{l,j}s_{l,l}}_{=0} = a_{i,j}, \quad i, i = \overline{1,n}.\end{aligned} \tag{3.2.3}$$

Якщо $i = j$, то

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

Тому

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

Якщо $i < j$, то

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

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

Знайдемо розв'язок рівняння (**??**). Враховуючи (**??**), маємо:

$$S^T D\vec{y} = \vec{b} \tag{3.2.4}$$

$$S\vec{x} = \vec{y} \tag{3.2.5}$$

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

$$y_i = \frac{b_i - \sum_{j=1}^{i-1} s_{j,i} s_{j,j} y_j}{s_{i,i} d_{i,i}}, \quad i = \overline{1,n} \tag{3.2.6}$$

$$x_n = \frac{y_n}{s_{n,n}}, \quad x_i = \frac{y_i - \sum_{j=1}^{i-1} s_{i,j} x_j}{s_{i,i}}, \quad i = \overline{n-1,1}. \tag{3.2.7}$$

Метод застосовується лише для симетричних матриць. Його складність – $Q = \frac{1}{3}n^3 + O(n^2)$.

Переваги цього методу:

1. він витрачає в 2 рази менше пам'яті ніж метод Гаусса для зберігання $A^T = A$ (необхідний об'єм пам'яті $\frac{n(n+1)}{2} \sim \frac{n^2}{2}$;
2. метод однорідний, без перестановок;
3. якщо матриця A має багато нульових елементів, то і матриця S також.

Остання властивість дає економію в пам'яті та кількості арифметичних операцій. Наприклад, якщо A має m ненульових стрічок по діагоналі, то $Q = O(m^2 n)$.

3.3 Обчислення визначника та оберненої матриці

Кількість операцій обчислення детермінанту за означенням – $Q_{\det} = n!$. В методі Гаусса – $PA = LU$. Тому

$$\det P \det A = \det L \det U \Rightarrow \det A = (-1)^l \det L \det U = (-1)^l \prod_{k=1}^{n} a_{k,k}^{(k)}, \quad (3.3.1)$$

де l – кількість перестановок. Ясно, що за методом Гаусса

$$Q_{\det} = \frac{2}{3}n^3 + O(n^2).$$

В методі квадратного кореня $A = S^T D S$. Тому

$$\det A = \det S^T \det D \det S = \prod_{k=1}^{n} d_{k,k} \prod_{k=1}^{n} s_{k,k}^2. \quad (3.3.2)$$

Тепер $Q_{\det} = \frac{1}{3}n^3 + O(n^2)$.

За означенням

$$AA^{-1} = E \quad (3.3.3)$$

де A^{-1} обернена до матриці A. Позначимо

$$A^{-1} = (\alpha_{i,j})_{i,j=1}^{n}.$$

Тоді $\vec{\alpha}_j = (\alpha_{i,j})_{i=1}^n$ – вектор-стовпчик оберненої матриці. З (3.3.3) маємо

$$A\vec{\alpha}_j = \vec{e}_j, \quad j = \overline{1,n}, \tag{3.3.4}$$

$\vec{e}_j$ – стовпчики одиничної матриці: $\vec{e}_j = (\delta_{i,j})_{i=1}^n$, $\delta_{i,j} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$.

Для знаходження $^{-1}$ необхідно розв'язати n систем. Для знаходження A^{-1} методом Гаусса необхідна кількість операцій $Q = 2n^3 + O(n^2)$.

3.4 Метод прогонки

Це економний метод для розв'язання СЛАР з три діагональною матрицею:

$$-c_0 y_0 + b_0 y_1 = -f_0, \tag{3.4.1}$$

$$a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \tag{3.4.2}$$

$$a_N y_{N-1} - c_N y_N = -f_N. \tag{3.4.3}$$

Матриця системи

$$A = \begin{pmatrix} -c_0 & b_1 & & 0 \\ a_0 & -c_1 1 & \ddots & \\ & \ddots & \ddots & b_N \\ 0 & & a_N & -c_N \end{pmatrix}$$

тридіагональна.

Розв'язок представимо у вигляді

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}, \quad i = \overline{0, N-1}. \tag{3.4.4}$$

Замінимо в (3.4.4) $i \mapsto i-1$ і підставимо в (3.4.2), тоді

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

Звідси

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

Тому з (3.4.2)

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

Умова розв'язності (3.4.1) $c_i - a_i\alpha_i \neq 0$.

Щоб знайти всі α_i, β_i, треба задати перші значення. З (3.4.1):

$$\alpha_1 = \frac{b_0}{c_0}, \quad \beta_1 = \frac{f_0}{c_0}. \tag{3.4.5}$$

Після знаходження всіх α_i, β_i обчислюємо y_N з системи

$$\begin{cases} a_N y_N - c_n y_N = -f_N, \\ y_{N-1} = \alpha_N y_N + \beta_N \end{cases}$$

Звідси

$$y_N = \frac{f_N + a_n\beta_N}{c_N - a_N\alpha_N}. \tag{3.4.6}$$

Алгоритм

1. Покладемо $\alpha_1 = \frac{b_0}{c_0}$, $\beta_1 = \frac{f_0}{c_0}$.
2. Позначимо $z_i = c_i - a_i\alpha_i$, обчислимо $\alpha_{i+1} = \frac{b_i}{z_i}$, $\beta_{i+1} = \frac{f_i + a_i\beta_i}{z_i}$, для $i = \overline{1, N-1}$.
3. Знайдемо $y_N = \frac{f_N + a_N\beta_N}{c_N - a_N\alpha_N}$.
4. Обчислюємо $y_i = \alpha_{i+1}y_{i+1} + \beta_{i+1}$, $i = \overline{N-1, 0}$.

Складність алгоритму $Q = 8N - 2$.

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

Теорема 4 (про достатні умови стійкості метода прогонки)**.** *Нехай*

$$\forall i : a_i, b_i \neq 0, \quad ma \quad |c_i| \geq |a_i| + |b_i| \quad (a_0 = b_N = 0),$$

та хоча би одна нерівність строга. Тоді

$$|\alpha_i| \leq 1, \quad ma \quad z_i = c_i - a_i\alpha_i \neq 0, \quad i = \overline{1, N}.$$

Задача 8. Довести теорему про стійкість методу прогонки.

3.5 Обумовленість систем лінійних алгебраїчних рівнянь

Нехай задано СЛАР

$$A\vec{x} = \vec{b}. \tag{3.5.1}$$

Припустимо, що матриця і права частина системи задані неточно і фактично розв'язуємо систему

$$B\vec{y} = \vec{h}. \tag{3.5.2}$$

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

Малість детермінанту $\det A \ll 1$ не є необхідною умовою різкого збільшення похибки. Це ілюструє наступний приклад:

$$A = diag(\varepsilon), \quad a_{i,j} = \varepsilon\Delta_{i,j}.$$

Тоді $\det A = \varepsilon^n \ll 1$, але $x_i = \frac{b_i}{\varepsilon}$. Тому $\Delta x_i = \frac{\Delta b_i}{\varepsilon}$.

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

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

Віднімемо від цієї рівності (**??**) $A\vec{z} + C\vec{x} + C\vec{z} = \vec{\eta}$. Тоді

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

Введемо норми векторів: $\|\vec{z}\|$:

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

Норми матриці, що відповідають нормам вектора, тобто

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

такі:

$$\|A\|_1 = \max_j \sum_{i=1}^{n} |a_{i,j}|, \quad \|A\|_2 = \max_i \sqrt{\lambda_i(A^T A)}, \quad \|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{i,j}|,$$

де $\lambda_i(B)$ – власні значення матриці B.

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

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

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

Теорема 5. *Якщо* $\exists A^{-1}$ *та* $\|A^{-1}\|\|C\| < 1$, *то*

$$\delta(\vec{x}) \le \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\delta(A)}(\delta(A) + \delta(\vec{b})), \tag{3.5.3}$$

де $\mathrm{cond}(A)$ – *число обумовленості.*

Доведення.

$$A\vec{z} = \vec{\eta} - C\vec{x} - C\vec{z}, \quad \vec{z} = A^{-1}\vec{\eta} - A^{-1}C\vec{x} - A^{-1}C\vec{z}$$

$$\|\vec{z}\| \le \|A^{-1}\vec{\eta}\| + \|A^{-1}C\vec{x}\| + \|A^{-1}C\vec{z}\| \le \|A^{-1}\|\cdot\|\vec{\eta}\| + \|A^{-1}\|\cdot\|C\|\cdot\|\vec{x}\| + \|A^{-1}\|\cdot\|C\|\cdot\|\vec{z}\|.$$

$$\|\vec{z}\| \le \frac{\|A^{-1}\|(\|\vec{\eta}\| + \|C\|\cdot\|\vec{x}\|)}{1 - \|A^{-1}\|\cdot\|C\|}$$

Оцінка похибки

$$\delta(\vec{x}) \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\cdot\|C\|}\left(\frac{\|\vec{\eta}\|}{\|\vec{x}\|} + \|C\|\right) =$$

$$= \frac{\|A^{-1}\|\cdot\|A\|}{1 - \|A^{-1}\|\cdot\|A\|\cdot\dfrac{\|C\|}{\|A\|}}\left(\frac{\|\vec{\eta}\|}{\|A\|\cdot\|\vec{x}\|} + \delta(A)\right) \le$$

$$\le \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\delta(A)}\left(\frac{\|\vec{\eta}\|}{\|\vec{x}\|} + \delta(A)\right).$$

□

Наслідок. Якщо $C \equiv 0$, то $\delta(\vec{x}) \le \mathrm{cond}(A)\delta(\vec{b})$.

Властивості $\mathrm{cond}(A)$:

1. $\mathrm{cond}(A) \ge 1$;
2. $\mathrm{cond}(A) \ge \dfrac{\max\limits_i |\lambda_i(A)|}{\min\limits_i |\lambda_i(A)|}$;

3. $\mathrm{cond}(AB) \le \mathrm{cond}(A) \cdot \mathrm{cond}(B)$;

4. $A^T = A^{-1} \Rightarrow \mathrm{cond}(A) = 1$.

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

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

Якщо $1 \ll \mathrm{cond}(A)$, то система називається *погано обумовленою*.

Оцінка впливу похибок заокруглення при обчисленні розв'язку СЛАР така (Дж. Уілкінсон): $\delta(A) = O(n\beta^{-t})$, $\delta(\vec{b}) = O(\beta^{-t})$, де β – розрядність ЕОМ, t – кількість розрядів, що відводиться під мантису числа. З оцінки (??) витікає: $\delta(\vec{x}) = \mathrm{cond}(A) \times O(n\beta^{-t})$. Висновок: найпростіший спосіб підвищити точність обчислення розв'язку погано обумовленої СЛАР – збільшити розрядність ЕОМ при обчисленнях. Інші способи пов'язані з розглядом цієї СЛАР як некоректної задачі із застосуванням відповідних методів її розв'язання.

Приклад погано обумовленої системи – системи з матрицею Гільберта

$$H_n = \left(\frac{1}{i+j-1}\right)_{i,j=1}^{n}, \quad \text{наприклад } \mathrm{cond}(H_8) \approx 10^9.$$

4 Ітераційні методи для систем

4.1 Ітераційні методи розв'язання СЛАР

Систему

$$A\vec{x} = \vec{b} \tag{4.1.1}$$

зводимо до вигляду

$$\vec{x} = B\vec{x} + \vec{f}. \tag{4.1.2}$$

Будь яка система

$$\vec{x} = \vec{x} - C(A\vec{x} - \vec{b}) \tag{4.1.3}$$

має вигляд (4.1.2) і при $\det C \ne 0$ еквівалентна системі (4.1.1). Наприклад, для $C = \tau E$:

$$\vec{x} = \vec{x} - \tau(A\vec{x} - \vec{b}). \tag{4.1.4}$$

4.1.1 Метод простої ітерації

Цей метод застосовується до рівняння (4.1.2)

$$\vec{x}^{(k+1)} = B\vec{x}^{(k)} + \vec{f}. \tag{4.1.5}$$

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

Ітераційний процес збігається, тобто $\|\vec{x}^{(k)} - \vec{x}\| \to 0$, $k \to \infty$, якщо

$$\|B\| \le q < 1 \tag{4.1.6}$$

При цьому має місце оцінка

$$\|\vec{x}^{(n)} - \vec{x}\| \le \frac{q^n}{1-q}\|\vec{x}^{(1)} - \vec{x}^{(0)}\|. \tag{4.1.7}$$

4.1.2 Метод Якобі

Припустимо $\forall i a_{i,i} \neq 0$. Зведемо систему (4.1.1) до вигляду

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

Ітераційний процес запишемо у вигляді

$$x_i^{(k+1)} = -\sum_{j=1}^{i-1} \frac{a_{i,j}}{a_{i,i}} \cdot x_j^{(k)} - \sum_{j=i+1}^{n} \frac{a_{i,j}}{a_{i,i}} \cdot x_j^{(k)} + \frac{b_i}{a_{i,i}}, \quad k = 0,1,\ldots, \quad i = \overline{1,n}. \tag{4.1.8}$$

Ітераційний процес збігається до розв'язку, якщо виконується умова

$$\forall i : \sum_{\substack{j=1 \\ i \neq j}}^{n} |a_{i,j}| \le |a_{i,i}|.$$

Це умова діагональної переваги матриці A. Якщо ж

$$\forall i : \sum_{\substack{j=1 \\ i \neq j}}^{n} |a_{i,j}| \le q \cdot |a_{i,i}|, \quad 0 \le q < 1. \tag{4.1.9}$$

то має місце оцінка точності:

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

4.1.3 Метод Зейделя

В компонентному вигляді ітераційний метод Зейделя записується так:

$$x_i^{(k+1)} = -\sum_{j=1}^{i-1} \frac{a_{i,j}}{a_{i,i}} \cdot x_j^{(k+1)} - \sum_{j=i+1}^{n} \frac{a_{i,j}}{a_{i,i}} \cdot x_j^{(k)} + \frac{b_i}{a_{i,i}}, \quad k = 0, 1, \ldots, \quad i = \overline{1, n}. \tag{4.1.10}$$

На відміну від методу Якобі на k-му-кроці попередні компоненти розв'язку беруться з $k+1$-ої ітерації.

Достатня умова збіжності методу Зейделя – $A^T = A > 0$.

4.1.4 Матрична інтерпретація методів Якобі і Зейделя

Подамо матрицю A у вигляді

$$A = A_1 + D + A_2,$$

де A_1 – нижній трикутник матриці A, A_2 – верхній трикутник матриці A, D – її діагональ. Тоді систему (4.1.1) запишемо у вигляді

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

або

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

Матричний запис методу Якобі:

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

методу Зейделя:

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

Необхідна і достатня умова збіжності методу Якобі: всі корені рівняння $\det(D + \lambda(A_1 + A_2)) = 0$ по модулю більше 1.

Необхідна і достатня умова збіжності методу метода Зейделя: всі корені рівняння $\det(A_1 + D + \lambda A_2) = 0$ по модулю більше 1.

4.1.5 Однокрокові (двошарові) ітераційні методи

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

$$B_k \frac{\vec{x}^{(k+1)} - \vec{x}^{(k)}}{\tau_{k+1}} + A\vec{x}^{(k)} = \vec{b}, \tag{4.1.11}$$

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

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

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

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

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

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

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

Методам, що розглянуті вище відповідають:

- методу простої ітерації: $B_k = E$, $\tau_{k+1} = \tau$;
- методу Якобі: $B_k = D$, $\tau_{k+1} = 1.$;
- методу Зейделя: $B_k = D + A_1$, $\tau_{k+1} = 1$.

4.1.6 Збіжності стаціонарних ітераційних процесів у випадку симетричних матриць

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

Нехай для A справедливі нерівності

$$\gamma_1 E \leq A \leq \gamma_2 E, \quad \gamma_1, \gamma_2 > 0. \tag{4.1.12}$$

Тоді при виборі $\tau = \tau_0 = \dfrac{2}{\gamma_1 + \gamma_2}$ ітераційний процес збігається. Найбільш точним значенням γ_1, γ_2 при яких виконуються обмеження (4.1.12) є $\gamma_1 =$

$\min\limits_i \lambda_i(A)$, $\gamma_2 = \max\limits_i \lambda_i(A)$. Тоді $q = q_0 = \frac{\gamma_2-\gamma_1}{\gamma_2+\gamma_1} = \frac{1-\xi}{1+\xi}$, $\xi = \frac{\gamma_1}{\gamma_2}$. (Зауважимо, що аналогічно обчислюється q і для методу релаксації розв'язання нелінійних рівнянь, де $\gamma_1 = m_= \min|f'(x)|$, $\gamma_2 = M_1 = \max|f'(x)|$) і справедлива оцінка

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

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

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

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

4.1.7 Метод верхньої релаксації

Узагальненням методу Зейделя є метод верхньої релаксації:

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

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

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

4.1.8 Методи варіаційного типу

До цих методів відносяться: метод мінімальних нев'язок, метод мінімальних поправок, метод найшвидшого спуску, метод спряжених градієнтів. Вони дозволяють обчислювати наближення без використання апріорної інформації про γ_1, γ_2 в (4.1.12).

Нехай $B = E$. Для методу мінімальних нев'язок параметри τ_{k+1} обчислюються з умови

$$\|\vec{r}^{(k+1)}\|^2 = \|\vec{r}^{(k)}\|^2 - 2\tau_{k+1}(\vec{r}^{(k)}, A\vec{r}^{(k)}) + \tau_{k+1}^2\|A\vec{r}^{(k)}\|^2 \to \min.$$

Тому

$$\tau_{k+1} = \frac{(A\vec{r}^{(k)}, \vec{r}^{(k)})}{\|\vec{r}^{(k)}\|^2}, \quad \text{де } \vec{r}^{(k)} = A\vec{x}^{(k)} - \vec{b} \text{ – нев'язка.}$$

Умова для завершення ітераційного процесу:

$$\|\vec{r}^{(n)} < \varepsilon.$$

Швидкість збіжності цього методу співпадає із швидкістю методу простої ітерації з одним оптимальним параметром $\tau_0 = \dfrac{2}{\gamma_1 + \gamma_2}$.
Аналогічно будуються методи з $B \neq E$. Матриця B називається переобумовлювачем і дозволяє підвищити швидкість збіжності ітераційного процесу. Його вибирають з умов

1. легко розв'язувати СЛАР $B\vec{x}^{(k)} = F_k$ (діагональний, трикутній, добуток трикутніх та інше);
2. зменшення числа обумовленості матриці $B^{-1}A$ у порівнянні з A.

4.2 Методи розв'язання нелінійних систем

Розглянемо систему рівнянь

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

Перепишемо її у векторному вигляді:

$$\vec{f}(\vec{x}) = 0. \tag{4.2.1}$$

4.2.1 Метод простої ітерації

В цьому методі рівняння (4.2.1) зводиться до еквівалентного вигляду

$$\vec{x} = \vec{\Phi}(\vec{x}). \tag{4.2.2}$$

Ітераційний процес представимо у вигляді:

$$\vec{x}^{(k+1)} = \vec{\Phi}(\vec{x}^{(k)}). \tag{4.2.3}$$

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

Нехай оператор $\vec{\Phi}$ визначений на множині H. За теоремою про стискуючі відображення ітераційний процес (4.2.3) сходиться, якщо виконується умова

$$\|\vec{\Phi}(\vec{x}) - \vec{\Phi}(\vec{y})\| \leq q\|\vec{x} - \vec{y}\|, \quad 0 < q < 1, \tag{4.2.4}$$

або

$$\|\vec{\Phi}'(\vec{x})\| \le q < 1, \tag{4.2.5}$$

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

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

Частинним випадком методу простої ітерації є метод релаксації для рівняння 4.2.1:

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

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

4.2.2 Метод Ньютона

Розглянемо рівняння

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

Представимо його у вигляді

$$\vec{F}(\vec{x}^{(k)}) + \vec{F}'(\vec{\xi}^{(k)})(\vec{x} - \vec{x}^{(k)}) = 0, \tag{4.2.6}$$

де $\vec{\xi}^{(k)} = \vec{x}^{(k)} + \theta_k(\vec{x}^{(k)} - \vec{x})$, $0 < \theta_k < 1$. Тут $\vec{F}'(\vec{x}) = \left(\frac{\partial f_i}{\partial x_j}\right)_{i,j=1}^{n}$ – матриця Якобі для $\vec{F}(\vec{x})$. Можемо наближено вважати $\vec{\xi}^{(k)} \approx \vec{x}^{(k)}$. Тоді з (4.2.6) матимемо

$$\vec{F}(\vec{x}^{(k)}) + \vec{F}'(\vec{x}^{(k)})(\vec{x}^{(k+1)} - \vec{x}^{(k)}) = 0. \tag{4.2.7}$$

Ітераційний процес представимо у вигляді:

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \vec{F}'(\vec{x}^{(k)})^{-1}\vec{F}(\vec{x}^{(k)}). \tag{4.2.8}$$

Для реалізації методу Ньютона потрібно, щоб існувала обернена матриця $\vec{F}'(\vec{x}^{(k)})^{-1}$.

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

$$\left\{ A_k \vec{z}^{(k)} = \vec{F}(\vec{x}^{(k)}), \quad \vec{x}^{(k+1)} = \vec{x}^{(k)} - \vec{z}^{(k)}, \quad k = 0, 1, 2, \dots \right. \tag{4.2.9}$$

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

Метод має квадратичну збіжність, якщо добре вибрано початкове наближення. Складність методу (при умові використання методу Гаусса розв'язання СЛАР (4.2.9) на кожній ітерації

$$Q_n = \frac{2}{3}n^3 + O(n^2),$$

де n – розмірність системи (4.2.1).

4.2.3 Модифікований метод Ньютона

Ітераційний процес має вигляд :

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

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

Запишемо метод у вигляді системи лінійних рівнянь (аналог (4.2.9))

$$\left\{ A_0\vec{z}^{(k)} = \vec{F}(\vec{x}^{(k)}), \quad \vec{x}^{(k+1)} = \vec{x}^{(k)} - \vec{z}^{(k)}, \quad k = 0, 1, 2, \dots \right. \tag{4.2.10}$$

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

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

Задача 9. Побудувати аналог методу січних для систем нелінійних рівнянь.

5 Проблема власних значень

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

$$A\vec{x} = \lambda\vec{x}. \tag{5.0.1}$$

λ називається власним значенням A, а $\vec{x}$ – власним вектором. З (5.0.1)

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

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

Для розв'язання багатьох задач механіки, фізики, хімії потрібне знаходження всіх власних значень λ_i, $i = \overline{1,n}$, а іноді й всіх власних векторів $\vec{x}_i$, що відповідають λ_i. Цю задачу називають повною проблемою власних значень.

В багатьох випадках потрібно знайти лише максимальне або мінімальне за модулем власне значення матриці. При дослідженні стійкості коливальних процесів іноді потрібно знайти два максимальних за модулем власних значення матриці.

Останні дві задачі називають *частковими проблемами власних значень*.

5.1 Степеневий метод

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

   Нехай $\vec{x}^{(0)}$ – заданий вектор, будемо послідовно обчислювати вектори

   $$\vec{x}^{(k+1)} = A\vec{x}^{(k)}, \quad k = 0, 1, \dots \tag{5.1.1}$$

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

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

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

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

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

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

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

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

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

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

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

(а) $\vec{x}^{(0)} \to \vec{e}_0 = \frac{\vec{x}^{(0)}}{\|\vec{x}^{(0)}\|}$;

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

(в) $|\mu_1^{(k+1)} - \mu_1^{(k)}| \geq \varepsilon$ `goto 2`;

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

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

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

Задача 10. Довести, що якщо $|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \dots$, то

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

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

Задача 11. Побудувати алгоритм обчислення λ_2, $\vec{e}_2$, використовуючи нормування векторів та скалярні добутки для обчислення $\mu_2^{(k)}$).

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

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

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

Тому $\max\limits_i \lambda_i(B) = \lambda_{\max} - \min\limits_i \lambda_i(A)$. Звідси $\lambda_{\min}(A) = \lambda_{\max}(A) - \lambda_{\max}(B)$.

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

Ще один спосіб обчислення мінімального власного значення полягає в використання обернених ітерацій:

$$A\vec{x}^{(k+1)} = \vec{x}^k, \quad k = 0, 1, \ldots \tag{5.1.2}$$

Але цей метод вимагає більшої кількості арифметичних операцій: складність методу на основі формули (5.1.1) $Q = O(n^2)$, а на основі (5.1.2) – $Q = O(n^3)$, оскільки треба розв'язувати СЛАР, але збігається метод (5.1.2) швидче.

5.2 Ітераційний метод обертання

Це метод розв'язання повної проблеми власних значень для симетричних матриць $A^T = A$. Існує матриця U, що приводить матрицю A до діагонального виду:

$$A = U\Lambda U^T, \tag{5.2.1}$$

де Λ – діагональна матриця, по діагоналі якої стоять власні значення λ_i; U – унітарна матриця, тобто: $U^{-1} = U^T$.

З (5.2.1) маємо

$$\Lambda = U^T\Lambda U, \tag{5.2.2}$$

Нехай $\exists\tilde{U}$ – матриця, така що $\tilde{\Lambda} = \tilde{U}^T A\tilde{U}$ і $\tilde{\Lambda} = (\tilde{\lambda}_{i,j})_{i,j=1}^n$, $|\tilde{\lambda}_{i,j}| < \delta \ll 1$, $i \neq j$.

Тоді діагональні елементи мало відрізняються від власних значень

$$|\tilde{\lambda}_{i,i} - \lambda_i(A)| < \varepsilon = \varepsilon(\delta).$$

Введемо $t(A) = \sum\limits_{\substack{i,j=1 \\ i\neq j}}^{n} a_{i,j}^2$. З малості величини $t(A)$ витікає, що діагональні елементи малі. По $A = A_0$ за допомогою матриць обертання що повертають систему векторів на кут φ, побудуємо послідовність $\{A_k\}$ таку, що $A_k \to \Lambda$ при $k \to \infty$.

Задача 12. Показати, що матриця обертання U_k є унітарною: $U_k^{-1} = U_k^T$.

Послідовно будуємо:

$$A_{k+1} = U_K^T A_k U_k, \tag{5.2.3}$$

Процес (5.2.3) називається *монотонним*, якщо: $t(A_{k+1}) < t(A_k)$.

Задача 13. Довести, що для матриці (5.2.3) виконується:

$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} \cos 2\varphi + \frac{1}{2}(a_{j,j}^{(k)} - a_{i,i}^{(k)}) \sin 2\varphi, \tag{5.2.4}$$

Показати, що $t(A_{k+1}) = t(A_k) - 2(a_{i,j}^{(k)})^2$, якщо вибирати φ з умови $a_{i,j}^{(k+1)} = 0$.

Звідси $\varphi = \varphi_k = \frac{1}{2}\arctan(p^{(k)})$, $p^{(k)} = \frac{2a_{i,j}^{(k)}}{a_{i,i}^{(k)} - a_{j,j}^{(k)}}$, де $|a_{i,j}^{(k)}| = \max\limits_{\substack{m,l \\ m \neq l}} |a_{m,l}^{(k)}|$. Тоді $t(A_k) \to 0$, $\to \infty$. Чим більше n тим більше ітерацій необхідно для зведення A до Λ.

Якщо матриця несиметрична, то застосовують QR, QL методи.