УДК 519.612

В. С. Абрамчук*, канд. фіз.-мат. наук,
І. В. Абрамчук,** старший викладач

*Вінницький державний педагогічний університет
імені М. Коцюбинського, м. Вінниця,
**Вінницький національний технічний університет, м. Вінниця

ОПТИМІЗАЦІЙНІ МЕТОДИ РОЗВ'ЯЗУВАННЯ СИСТЕМ $A\vec{x} = \vec{b}$ З ПОГАНО ЗУМОВЛЕНИМИ МАТРИЦЯМИ

Досліджені проблемні задачі, пов'язані з оцінкою похибок у розв'язках зашумлених систем, мінімізацією похибок обчислень при гауссових перетвореннях, прискоренням швидкості збіжності ітераційних методів.

Ключові слова: *оцінка похибок розв'язку, багатошаровий метод гауссового виключення, локальний базис методу напрямленого пошуку, відношення Релея.*

Вступ. СЛАР з погано зумовленими матрицями необхідно розв'язувати у багатьох задачах практики: в задачах математичної обробки і прогнозування експериментів, в задачах оптимального управління і оптимального прогнозування технічних систем, при дослідженні природних явищ і катаклізмів [1–10, 13]. Одним з підходів до розв'язування таких задач є різницеві рівняння і метод скінченних елементів. При цьому матриці таких систем є розрідженими, великих порядків і погано зумовленими. Для створення ефективних методів розв'язування систем з погано зумовленими матрицями необхідно вияснити причини, що перешкоджають збіжності ітераційних методів та розробити нові принципи прискорення їх збіжності та мінімізації похибок обчислень при гауссових перетвореннях матриці. Основою таких досліджень є роботи [11, 12].

Постановка проблеми. Нехай необхідно розв'язати систему лінійних алгебричних рівнянь (СЛАР), яка має вид $A\vec{x} = \vec{b}$, $A \in M_{n\times n}(\mathbb{R})$, $\vec{b} \in \mathbb{R}^n$,з матрицею A довільної структури, погано зумовленою, великого порядку. Нерозв'язаними або недостатньо дослідженими є проблемні задачі: 1) вплив похибок на розв'язок СЛАР $\overrightarrow{\Delta x}\left(A, \vec{b}, E, \overrightarrow{\delta b}\right)$: $\left(A+E\right)\left(\vec{x}+\overrightarrow{\Delta x}\right) = \vec{b} + \overrightarrow{\delta b}$, E, $\overrightarrow{\delta b}$ — похибки даних і обчислень; 2) причини, що не дозволяють існуючим методам бути ефективними при розв'язуванні таких СЛАР; 3) стратегія розробки стійких алгоритмів розв'язування СЛАР в умовах похибок обчислень.

Аналіз актуальних досліджень. СЛАР з погано зумовленими матрицями необхідно розв'язувати у багатьох задачах практики: в задачах

математичної обробки і прогнозування експериментів, в задачах оптимального управління і оптимального прогнозування технічних систем, при дослідженні природних явищ і катаклізмів [1–10]. При розв'язувані СЛАР з погано зумовленими матрицями, в літературі не існує єдиного підходу: з однієї сторони, обгрунтовується, що необхідно відмовитися від розв'язання систем з матрицями типу матриці Гільберта, оскільки число зумовленості таких матриць наближається до $e^{3.5n}$ [3]; з другої сторони матриці Гільберта широко застосовуються у регресійному аналізі, тощо, тому необхідно розробляти методи розв'язування таких систем [2; 8]; по-третє, існує цілий клас методів спряжених напрямів, заснованих на погано зумовленому базисі Крилова, які ефективно використовуються для розв'язування різницевих рівнянь [2; 4]. Причиною таких протиріч є недостатній аналіз факторів, які впливають на ефективність методів розв'язування систем з погано зумовленими матрицями.

Мета статті. 1. Виконати аналіз впливу похибок обчислень на похибку розв'язку СЛАР. 2. Розробити багатошаровий метод гауссових перетворень, що мінімізує похибку обчислень. 3. Розробити стратегію побудови ітераційних методів розв'язування СЛАР на основі системи повних базисів Крилова.

Основна частина. 1. Аналіз похибок у розв'язках зашумлених систем. Однією з основних проблем теорії лінійних алгебраїчних систем є аналіз впливу похибок заокруглення і похибок початкових даних (похибок невизначеності) на розв'язок, обчислений на ЕОМ у скінченорозрядній арифметиці [2–4; 10]. На основі оберненого аналізу, на місце розв'язку системи $A\vec{x} = \vec{b}$, $A \in M_{n\times n}(\mathbb{R})$, $rank\, A = n$, $\vec{b} \in im\, A$, досліджується точний розв'язок зашумленої системи $(A+E)(\vec{x}+\overrightarrow{\delta x}) = \vec{b} + \overrightarrow{\delta b}$, де $E, \overrightarrow{\delta b}$ — похибки. Якщо $\rho(A^{-1}E) < 1$, то матриця $A+E$ має обернену і можна оцінити відносну похибку розв'язку зашумленої системи через число зумовленості матриці [2; 3; 10]. Ця оцінка є апріорною, оскільки в неї не входить ні обчислений розв'язок $\tilde{\vec{x}} = \vec{x} + \overrightarrow{\delta x}$, ні розв'язок системи $A\vec{x} = \vec{b}$ [10]. Вона не дає способу, як керувати обчислювальним процесом, щоб зменшити похибку розв'язку.

Допустимо, що похибки коефіцієнтів матриці є неперервними параметрами, тоді квадратну матрицю $A+E$ можна вважати лінійним оператором деякої системи лінійних диференціальних рівнянь. Задача дослідження поведінки матриці і розв'язків системи рівнянь зведеться до структурної стійкості матриці при збуренні параметрів (теорія канонічних форм Жордана-Арнольді [9]). Проте цей підхід також не дає способу мінімізації похибки розв'язку.

У роботі [12] пропонується розв'язок «зашумленої» системи досліджувати як дробово-раціональну функцію похибок на основі формул Крамера. Скористаємось тим, що визначник матриці є лінійною функцією вектор-стовпців (рядків) матриці [4].

Запишемо розв'язки систем $A\vec{x}=\vec{b}$, $(A+E)\left(\vec{x}+\overrightarrow{\delta x}\right)=\vec{b}+\overrightarrow{\delta b}$ через визначники: $x_i = D_i/D_0$, $x_i+\delta x_i=\left(D_i+\delta D_i\right)/\left(D_0+\delta D_0\right)$. Тоді похибки компонент розв'язку $\delta x_i=\left(\delta D_i - x_i\cdot\delta D_0\right)/\left(D_0+\delta D_0\right)$.

Не втрачаючи загальності, запишемо зашумлену систему другого порядку у матрично-векторній формі $\left[A_{1.}+E_{1.}\quad A_{2.}+E_{2.}\right]\left(\vec{x}+\overrightarrow{\delta x}\right)=\vec{b}+\overrightarrow{\delta b}.$ Тоді в силу лінійності визначників відносно вектор-стовпців матриці, матимемо

$$\begin{aligned}
&D_0+\delta D_0=\det\left[A_{1.}+E_{1.}\ A_{2.}+E_{2.}\right]=\det\left[A_{1.}\ A_{2.}\right]+\left(\det\left[A_{1.}\ E_{2.}\right]+\right.\\
&\left.+\det\left[E_{1.}\ A_{2.}\right]\right)+\det\left[E_{1.}\ E_{2.}\right]=D_0+L_0\left(E\right)+{K_0}^2\left(E\right),\\
&D_1+\delta D_1=\det\left[\vec{b}+\overrightarrow{\delta b}\ A_{2.}+E_{2.}\right]=\det\left[\vec{b}\ A_{2.}\right]+\left(\det\left[\overrightarrow{\delta b}\ A_{2.}\right]+\right.\\
&\left.+\det\left[\vec{b}\ E_{2.}\right]\right)+\det\left[\overrightarrow{\delta b}\ E_{2.}\right]=D_1+L_1\left(\overrightarrow{\delta b},E_{2.}\right)+{K_1}^2\left(\overrightarrow{\delta b},E_{2.}\right),
\end{aligned}\qquad(1)$$

аналогічно $D_2+\delta D_2=D_2+L_2\left(\overrightarrow{\delta b},E_{1.}\right)+{K_2}^2\left(\overrightarrow{\delta b},E_{1.}\right)$.

Тут позначено: $\det\left[\vec{c}_{1.}\ \vec{c}_{2.}\right]$ — визначник матриці, складеної з двох вектор-стовпців $\vec{c}_{1.}$, $\vec{c}_{2.}$; L — лінійна форма похибок; K^m — похибки вищих порядків.

Компоненти похибок розв'язку зашумленої системи другого порядку

$$\delta x_i=\left[\left(L_i-x_iL_0\right)+\left(K_i^2-x_i{K_0}^2\right)\right]\Big/\left[D_0+\left(L_0+{K_0}^2\right)\right],\ i=1,2.$$

Для довільної зашумленої системи n-го порядку похибки компонент розв'язку $\forall i=\left[1:n\right]$ матимуть вид:

$$\delta x_i=\left[L_i-x_iL_0+\sum\nolimits_{j=2}^{n-1}\left(K_i^{\,j}-x_iK_0^j\right)\right]\Big/\left[D_0+L_0+\sum\nolimits_{j=2}^{n-1}K_0^j\right],\qquad(2)$$

де L_0 — лінійна форма похибок матриці, L_0 — сума n визначників матриць $\left(A\setminus A_{i.}\right)\cup E_{i.}$, де кожен i-й стовпчик $A_{i.}$ заміщується вектором похибок $E_{i.}$; L_i — лінійні форми похибок матриць $\left[\overrightarrow{\delta b},E\setminus E_{i.}\right]$, в яких на місце вектора $E_{i.}$ ставиться вектор $\overrightarrow{\delta b}$ (є сумами відповідних визначників).

Теорема 1. Для того, щоб зашумлена система $(A+E)\left(\vec{x}+\overrightarrow{\delta x}\right)=$ $=\vec{b}+\overrightarrow{\delta b}$ була стійкою до похибок $\left(E,\overrightarrow{\delta x}\right)$, необхідно, щоб допущені похибки задовольняли нерівність

$$\operatorname{sgn} D_0 \cdot \left|L_0 + \sum\nolimits_{j=2}^{n-1} K_0{}^j\right| \Big/ D_0 < 1, \qquad (3)$$

необхідно і достатньо, щоб виконувалась нерівність (3) і $\forall k=[1:n]$ нерівності $\left|\overrightarrow{\delta x_k}\right|<\varepsilon$, що еквівалентно нерівностям

$$\left|L_k - x_k L_0 + \sum\nolimits_{j=2}^{n-1}\left(K_k{}^j - x_k K_0{}^j\right)\right| < \varepsilon \cdot \left|D_0 + L_0 + \sum\nolimits_{j=2}^{n-1} K_0{}^j\right|, \qquad (4)$$

($|\circ|$ — абсолютна величина).

Доведення теореми 1 випливає з вищенаведених формул (1), (2).

Висновок. Зростом розмірності матриці похибки у розв'язках зростають, оскільки похибки матриці і правої частини належать простору $\mathbb{R}^{n^2+n}$ (у той час, як розв'язок належить простору $\mathbb{R}^n$).

На основі проведеного аналізу побудуємо метод гауссового перетворення матриці, що мінімізує похибку обчислень.

2. Багатошаровий алгоритм гауссового перетворення (виключення). При розв'язуванні систем $A\vec{x}=\vec{b}$, $A\in M_{n\times n}(\mathbb{R})$, $rank\,A=n$, $\vec{b}\in im\,A$, з допомогою машинної арифметики методом гауссового виключення (або іншим прямим методом) необхідно досліджувати дві проблемні задачі, пов'язані з запобіганням: 1) катастрофічної втрати точності (втрати старших розрядів [2]) у компонентах перетвореної матриці або, навпаки, у непомірному зростанні значень компонент [3]; 2) зростання числа зумовленості при перетвореннях матриці (наближення матриці до виродженої). Ці проблеми тісно пов'язані між собою і вимагають оцінки наближених розв'язків [2–4; 10].

Із загальної теорії збурень випливає, що повністю усунути похибки обчислень при лінійних перетвореннях неможливо [2–10], їх можна лише зменшити за рахунок оптимізації методу перетворень системи. У роботі [11] доведено, що метод гауссового виключення на кожному i-му кроці, $i\in[1:n-1]$, є процесом ортогоналізації вектор-рядків $A_{.k}$ $\forall k\in[i+1:n]$ до одиничного вектора $\vec{e}_i$. Тому для розв'язання першої проблемної задачі, необхідно на кожному i-му кроці знаходити розв'язуючий вектор-рядок $A_{.i_0}$, що утворює найменший кут з вектором $\vec{e}_i$ [11].

Похибка обчислень при перетвореннях $\left(\tilde{A}_{.k},\vec{e}_i\right)=0\Rightarrow$ $\Rightarrow\left(A_{.k}+\alpha_k A_{.i},\vec{e}_i\right)=0\Rightarrow A_{k,i}+\alpha_k A_{i,i}=0$ буде найменшою для всіх

$k \in [i+1:n]$, якщо абсолютне значення $|A_{i,i}|$ найбільше (такий елемент $A_{i,i}$ називають головним [2; 3]). Оскільки головний елемент повинен належати розв'язуючому вектор-рядку, то задача його визначення еквівалентна задачі [11]: обчислити індекси i_0, j_0 для яких досягається

$$\max_{j\in[i:n]}\max_{k\in[1:n]}\left|\cos\left\{A_{.k},\vec{e}_i\right\}\right| = \max_{j\in[i:n]}\max_{k\in[1:n]}\left|A_{k,j}\right| \Big/ \left\|A_{.k}\right\|_2 .$$

Отже, задача зменшення похибки обчислень на i-му кроці в методі гауссового виключення (який назвемо одношаровим) полягає у знаходженні як розв'язуючого вектор-рядка $A_{.i_0}$ так і розв'язуючого вектор-стовпця $A_{j_0.}$, на перетині яких знаходиться головний елемент A_{i_0,j_0} (який обмінюється з елементом $A_{i,i}$ шляхом перестановки вектор-рядків, вектор-стовпців). Щоб одночасно розв'язати обидві проблемні задачі, побудуємо багатошаровий метод гауссового перетворення шляхом одночасного обнулення піддіагональних елементів у стовпцях $A_{i.},\ldots,A_{(i+m-1).}$, $m \le n$.

Виконаємо обернений аналіз для $m=2$, тоді для $\forall k \in [i+2:n]$:

$$\left(\tilde{A}_{.k},\vec{e}_i\right)=0,\left(\tilde{A}_{.k},\vec{e}_{i+1}\right)=0, \qquad (5)$$

$$\tilde{A}_{.k} := A_{.k}+\alpha_k A_{.i}+\beta_k A_{.(i+1)} \Rightarrow \begin{cases} A_{k,i}+\alpha_k A_{i,i}+\beta_k A_{i,i+1}=0, \\ A_{k,i+1}+\alpha_k A_{i+1,i}+\beta_k A_{i+1,i+1}=0. \end{cases}$$

Мінімізація похибки обчислень при розв'язуванні сукупності систем (5) $\forall k \in [i+2:n]$ досягається тоді і лише тоді, якщо головний визначник $D_0^{(2)} = A_{i,i}A_{i+1,i+1} - A_{i+1,i}A_{i,i+1}$ приймає найбільше абсолютне значення. Отже необхідно знайти нову пару розв'язуючих векторів $A_{.i_1}$, $A_{j_1.}$ з умови $\max\limits_{k\in[i+1:n]}\max\limits_{j\in[i+1:n]}\left|A_{i,i}A_{k,j}-A_{k,i}A_{i,j}\right|$, що реалізується простою алгоритмічною процедурою.

Якщо уже виконане упорядкування матриці $A[i+1:n, i+1:n]$ шляхом перестановки вектор-рідків $A_{.(i+1)}$, $A_{.i_1}$, вектор-стовпців $A_{(i+1).}$, $A_{j_1.}$, то наступна тришарова процедура гауссового виключення: $\forall k \in [i+3:n]$, $\left(\tilde{A}_{.k},\vec{e}_i\right)=0$, $\left(\tilde{A}_{.k},\vec{e}_{i+1}\right)=0$, $\left(\tilde{A}_{.k},\vec{e}_{i+2}\right)=0$,

$$\tilde{A}_{.k} := A_{.k}+\alpha_k A_{.i}+\beta_k A_{.(i+1)}+\gamma_k A_{.(i+2)}, \qquad (6)$$

зведеться до пошуку нової пари розв'язуючих векторів $A_{.i_2}$, $A_{j_2.}$ шляхом максимізації абсолютного значення визначника третього порядку системи (6). Оскільки процедура введення нових розв'язуючих векторів $A_{.i_t}$, $A_{i_t.}$, $t \in [1:m-1]$ є вкладеною (рекурентною) (без перетворень матриці), то алгоритм m-шарового, $m \geq 2$, гауссового перетворення легко здійснюється для довільних щільно заповнених матриць. Параметри m-шарових гауссових перетворень після визначення розв'язуючих векторів можна знайти методом Крамера, оскільки головні визначники обчислені. З використанням, наприклад, тришарового процесу (6) одночасно обнуляються піддіагональні елементи в трьох вектор-стовпцях $A_{i.}$, $A_{(i+1).}$, $A_{(i+2).}$ для $\forall k \in [i+3:n]$ без попереднього використання одношарових і двошарових процедур. Після завершення тришарового процесу для $k \in [i+3:n]$ перетворюється рядок $A_{.(i+2)}$ двошаровою процедурою, і нарешті $A_{.(i+1)}$ — одношаровою.

Теорема 2. m-шаровий процес гауссового виключення коректний: мінімізує похибку обчислень, однозначно визначає параметри перетворень матриці, не приводить до виродження системи.

Доведення теореми 2. Оскільки за умовою матриця A не вироджена, то для неї існують відмінні від нуля визначники всіх порядків $2 \leq m < n$. Їх найбільше абсолютне значення гарантується скінченим числом переборів по $k, j \in [i+t:n]$, $t \in [1:m-1]$ (тобто простою двоциклічною програмною реалізацією). Процес обчислень економний, оскільки використовує лінійні векторні перетворення типу (6). Мінімізація похибки обчислень гарантується вибором розв'язуючих векторів.

3. Модифікований алгоритм гауссового виключення для щільно заповнених погано зумовлених матриць. Для того, щоб процес гауссового виключення був стійким при машинній реалізації, пропонується масштабування матриці $\left(A, \vec{b}\right)$ [2; 3].

Запропонований в п. 2 алгоритм гауссового виключення дозволяє модифікувати обчислювальний процес в залежності від типу матриці. Нехай розв'язується система з матрицею типу матриці Гільберта $H = \left(1/(i+j-1)\right)_{i,j=1}^{n}$, яка виникає у багатьох практичних дослідженнях (поліноміальна регресія [2, 3, 10]) і є погано зумовленою (число зумовленості асимптотично наближається до e^{cn}, $c \approx 3.5$ [10], наприклад, похибки в даних поліноміальної регресії 10-го порядку зростають на множник, більший, ніж $3 \cdot 10^{12}$ [3].

Виникає питання, у чому причина сильного зростання похибок у розв'язках систем з погано зумовленими матрицями. Очевидна відповідь, що матриця типу матриці Гільберта є погано зумовленою — одностороння. Більш глибока причина полягає у тому, що не дивлячись на частковий або повний вибір головного елемента, гауссове перетворення матриці Гільберта приводить до зростання числа зумовленості перетворюваних векторів і не зростання значень визначників у перетворених матрицях $A\left[i:n,i:n\right]$, $i\in\left[1:n-1\right]$ (наближення матриці $A\left[i:n,i:n\right]$ до виродженої — до структурної нестійкості).

Побудуємо алгоритм гауссового виключення для матриць типу матриці Гільберта або матриці Коші [2], що мінімізує похибку обчислень. Нехай на кожному i-му кроці перетворень матрицю можна масштабувати, наприклад, за умовою $A_{k,i}=1$, $k\in\left[i:n\right]$. Тоді, на основі п.2, розв'язуючим вектор-рядком є той, для якого досягається $\max\left|A_{k,i+1}\right|$, $k\in\left[i:n\right]$. Для матриці Гільберта порядку $n\geq 2$, розв'язуючим вектором для $i=1$ є не $A_{\cdot 1}$ (що визначається стандартним методом гауссового виключення [2], а вектор $\tilde{A}_{\cdot n}:=A_{\cdot n}/A_{n,1}$. Всі наступні перетворення з умовою масштабування $A_{k,i}=1$, $k\in\left[i:n\right]$, для матриці Гільберта є стійкими.

Нехай матриця A не може бути задовільно масштабована умовою $A_{k,i}=1$, $k\in\left[i:n\right]$ (частковий вибір), або $A_{k,j}=1$, $k,j\in\left[i:n\right]$ (повний вибір) [3]. На початковому кроці розширимо матрицю $\left(A,\vec{b}\right)$, $A\in M_{n\times n}$ до матриці $\left(A1,\overrightarrow{b1}\right)$, $A1\in M_{(n+1)\times(n+1)}$ введенням першого стовпця з одиничних елементів $A1_{k,1}=1$, $k\in\left[1:n+1\right]$, першого рядка з елементів одиничного вектора $A1_{\cdot 1}=\vec{e}_1$, $\forall i,j\in\left[2:n+1\right]$, $A1_{i,j}=A_{i-1,j-1}$; з елементами правих частин: $b1=1$, $\forall i\in\left[2:n+1\right]$, $b1_i=b_{i-1}+1$.

Алгоритм. Сформувати матрицю $\left(A1,\overrightarrow{b1}\right)$.

Для $i\in\left[1:n+1\right]$ вибрати пару розв'язуючих векторів $\vec{a}$, $\vec{c}$ з вектор-рядків $A_{\cdot k}$, $k\in\left[i:n+1\right]$ таких, що для всіх векторів $\vec{d}\in\left\{A_{\cdot(i+1)},\ldots,A_{\cdot(n+1)}\right\}$, $\vec{d}\neq\vec{a}\neq\vec{c}$ виконуються умови: $\tilde{\vec{d}}:=\vec{d}+\alpha\vec{a}+\beta\vec{c}$, $\left(\tilde{\vec{d}},\vec{e}_i\right)=0$, $\left(\tilde{\vec{d}},\vec{e}_{i+1}\right)=1$, де $\alpha=-\left(1+\beta\right)$, $\beta=\left(1-d_{i+1}+a_{i+1}\right)/\left(c_{i+1}-a_{i+1}\right)$.

Розв'язуючі вектори $\vec{a}$, $\vec{c}$ вибираються з умови $\max\limits_{j\in[i:n]}\left|c_{i+1}-a_{i+1}\right| =$ $= \max\limits_{k,j\in[i:n+1]}\left|A1_{k,i+1}-A1_{j,i+1}\right|$ (частковий вибір), або $\max\limits_{k,j\in[i:n+1]}\left|A1_{k,t}-A1_{j,t}\right|$ (повний вибір для всіх стовпчиків $t\in[i+1:n+1]$), $\max\limits_{t\in[i+1:n+1]}\left|c_t-a_t\right|>0$ існує (у протилежному випадку матриця $A1$ вироджена).

4. Конус $K_{\min}$**.** При розв'язуванні систем $A\vec{x}=\vec{b}$ ітераційними методами основна увага приділена асимптотичній швидкості збіжності, з розрахунку що буде досягнута на деякому кроці задана точність $\|\vec{r}\|<\varepsilon$. Ця теоретична оцінка лежить, як правило, в основі порівняння методів на ефективність. Але в дійсності оцінка не виконується, особливо при розв'язуванні систем з погано зумовленими матрицями. На перешкоді стає конус $K_{\min}$. Простір $\mathbb{R}^n$ розбивається гіперплощинами $A_{.i}\vec{x}=\vec{b_i}$, $i=\{1,...,n\}$, на 2^n конусів з вершиною у точці $\vec{x}^*=A^{-1}\vec{b}$. Пару конусів, що містять сингулярну пряму $\vec{x}=\vec{x}^*+t\cdot\vec{\varepsilon}_{\min}$, $\vec{\varepsilon}_{\min}$ — власний вектор матриці A^TA, що відповідає $\lambda_{\min}\left(A^TA\right)$, назвемо конусом $K_{\min}$ (конусом мінімальних нев'язок). Щоб пояснити деталі, розглянемо приклад у просторі $\mathbb{R}^2$.

Приклад. Проаналізувати систему

$$A\vec{x}=\vec{b},\ A=\begin{bmatrix}1 & 1\\ 1 & 0.9999\end{bmatrix},\ \vec{b}=\begin{bmatrix}0\\ 0.01\end{bmatrix},\ \vec{x}^*=\begin{bmatrix}100\\ -100\end{bmatrix}.$$

Обчислимо основні характеристики:

$$\textbf{\textit{cond}}\,A=\sqrt{\lambda_{\max}\left(A^TA\right)\Big/\lambda_{\min}\left(A^TA\right)}=12648.79453,\ \det A=-0.0001,$$

$$\lambda_{\max}\left(A^TA\right)=3.99980075,\ \lambda_{\min}\left(A^TA\right)=0.000000025,$$

$$S_{\min}:\begin{cases}x_1=100-0.999950012t,\\ x_2=-100+t;\end{cases}\quad S_{\max}:\begin{cases}x_1=100+1.000967t,\\ x_2=-100+t;\end{cases}$$

$t\in\mathbb{R}$, $S_{\min}$, $S_{\max}$, — сингулярні прямі, що визначаються власними векторами $\vec{\varepsilon}_{\min}$, $\vec{\varepsilon}_{\max}$ матриці A^TA.

Малість нев'язки в області $K_{\min}$ не дозволяє оцінити реальний розв'язок, оскільки норма нев'язки може стати малою величиною, а норма похибки залишатись великою величиною. Будь-який ітераційний процес, що заснований на мінімізації норми вектора нев'язки або

норми вектора похибки, швидко приводить наближення в область $K_{\min}$. Дійсно, виберемо дві початкові точки $\vec{x}^{(1)} = [150;50]^T \notin K_{\min}$, $\left\|\vec{x}^{(1)} - \vec{x}^*\right\|_2 \approx 158$, $\vec{x}^{(2)} = [0;0]^T \in K_{\min}$, $\left\|\vec{x}^{(2)} - \vec{x}^*\right\|_2 \approx 14$.

Застосуємо один і той же ітераційний процес $\vec{\tilde{x}}^{(k)} = \vec{x}^{(k)} + \alpha_k A^T \vec{r}^{(k)}$, $\alpha_k = -\left\|\vec{r}^{(k)}\right\|_2^2 \Big/ \left\|A^T \vec{r}^{(k)}\right\|_2^2$, що мінімізує норму вектора похибки, дістанемо $\vec{\tilde{x}}^{(1)} \approx [49.999 - 49.996]^T$; $\vec{\tilde{x}}^{(2)} \approx [0.0050005\ 0.005]^T$; $\vec{\tilde{x}}^{(1)}, \vec{\tilde{x}}^{(2)} \in K_{\min}$, (обчислення проведені на 10-розрядному калькуляторі). Незалежно від того, що $\vec{x}^{(1)} \notin K_{\min}$, $\vec{x}^{(2)} \in K_{\min}$, ітераційний процес за один крок привів нові наближення в область $K_{\min}$ (точніше, точка $\vec{x}^{(1)}$, яка не належить $K_{\min}$ і знаходилась на великій відстані від розв'язку, перейшла в ближчу точку $\vec{\tilde{x}}^{(1)} \in K_{\min}$, ніж точка $\vec{x}^{(2)}$, що знаходилась в $K_{\min}$ на меншій відстані від розв'язку, яка майже не змінила своє положення). Обчислені нев'язки $\vec{\tilde{r}}^{(1)}$, $\vec{\tilde{r}}^{(2)}$ не вносять ніякого роз'яснення в дану ситуацію.

5. Ітераційні методи розв'язування систем $A\vec{x} = \vec{b}$ великих порядків. Щоб обгрунтувати можливу ефективність довільного методу розв'язування системи лінійних алгебричних рівнянь з невиродженими погано зумовленими матрицями великих порядків, необхідно виконати аналіз впливу похибок заокруглення на результат обчислювального процесу реалізації алгоритму методу у машинній арифметиці (в умовах скінчено-розрядної сітки). Оскільки прямі методи пов'язані з накопиченням похибок при перетвореннях матриці, то вони відносяться до неефективних методів розв'язування систем з погано зумовленими матрицями великих порядків.

Нерозв'язаною проблемою є: «чи існує стратегія побудови ітераційних методів розв'язування таких систем, яка забезпечує як теоретичну збіжність, так і її практичну алгоритмічну реалізацію у скінчено-розрядній арифметиці при економній кількості обчислень». Існують дві принципово різні стратегії побудови ітераційних методів.

Перша — ітераційний процес задається у формі $\forall k = 0,1,...$ $\vec{x}^{(k+1)} = \vec{x}^{(k)} + \alpha_k \vec{c}^{(k)}$ (або у більш загальній формі $\vec{x}^{(k+1)} = B^{(k)}\vec{x}^{(k)} + \beta_k \vec{w}^{(k)}$), де $B^{(k)}$, $\vec{c}^{(k)}$, $\vec{w}^{(k)}$, задаються в залежності від класу методів, як функції матриці A, правої частини $\vec{b}$, наближення до A^{-1}, можуть враховувати структуру матриці та попередні наближення

за t кроків [2; 4]. Збіжність цих процесів заснована на мінімізації вектора похибки $\vec{\varepsilon}^{(k+1)} = \vec{x}^{(k+1)} - \vec{x}^*$ або вектора нев'язки $\vec{r}^{(k+1)} = A\vec{x}^{(k+1)} - \vec{b}$, або на іншій процедурі, наприклад, на побудові спряжених, біспряжених напрямів, тощо. Проте всі ці процедури, за умови збіжності, приводять в область $K_{\min}$, де процес збіжності сповільнюється і може стати, в силу похибок обчислень, неконтрольованим відносно монотонності процесу. Більш глибокі результати у розв'язуванні систем з погано зумовленими матрицями засновані на принципі регуляризації — побудові псевдорозв'язку з мінімальною нормою $\|\circ\|_2$ [8]. Але, якщо шукати нормальний розв'язок відносно $\vec{x}^* = A^{-1}\vec{b}$, то знову виникне проблемна задача впливу похибок заокруглення на збіжність методу (оскільки абсолютний нуль у машинній арифметиці відсутній).

Запропонуємо двоциклічну стратегію розв'язування цієї проблеми на основі методу напрямленого пошуку [12]:

$$\vec{x}^{(k,j)} = \vec{x}^{(j)} + \alpha_{k,j}\vec{c}^{(k,j)},\ k = 0,1,...,m,\ j = 1,...,m_2, \qquad (7)$$

де у внутрішньому циклі по $k \in \{0,1,...,m\}$ наближення $\vec{x}^{(j)}$, нев'язки $\vec{r}^{(j)} = A\vec{x}^{(j)} - \vec{b}$ не змінюються і а) наближення $\vec{x}^{(j+1)}$, $j \in \{0,1,...,m_2\}$ (зовнішній цикл) оптимізується у внутрішньому циклі за рахунок напрямного вектора $\vec{c}^{(k,j)} \in K_m$, $\left(\vec{c}^{(k,j)}, \vec{r}^{(j)}\right) \neq 0$, шляхом мінімізації норми $\|\circ\|_2$ вектора похибки $\vec{\varepsilon}^{(k,j)} = \vec{x}^{(k,j)} - \vec{x}^*$, але вектор $\vec{x}^{(k,j)}$ не обчислюється); б) нев'язка $\vec{r}^{(k+1)}$ оптимізується у внутрішньому циклі за рахунок вектора-поправки $\vec{c}^{(k,j)} \in K_m$, $\left(A\vec{c}^{(k,j)}, \vec{r}^{(j)}\right) \neq 0$ шляхом мінімізації норми $\|\circ\|_2$ вектора нев'язки $\vec{r}^{(k,j)} = A\vec{x}^{(k,j)} - \vec{b}$, але вектори $\vec{x}^{(k,j)}$, $\vec{r}^{(k,j)}$ не обчислюються. Ітераційні процеси а), б) формально ідентичні з точки зору організації обчислень, проте мають різний геометричний зміст.

Побудуємо алгоритм процедури б) (для п. а будується аналогічно), який складається з двох етапів: 1) обчислення початкового вектора $\vec{x}^{(0)}$; 2) уточнення розкладу вектора $\vec{b}$.

Алгоритм обчислення початкового вектора $\vec{x}^{(0)}$. Вектор $\vec{x}$, розв'язок системи $A\vec{x} = \vec{b}$, визначає розклад вектора $\vec{b}$ по неортогональному базису $\left\{A_{i\cdot} = A\vec{e}_i\right\}_{i=1}^{n}$. На місце фактичного розкладу, виконаємо ортогональні проекції $\vec{b}$ на вектори $A_{i\cdot}$, дістанемо наближені

значення компонент $x_i^{(0)} = \left(\vec{b}, A\vec{e}_i\right) \Big/ \left\|A\vec{e}_i\right\|_2^2$ розкладу. Вектор $\vec{b}' = \sum_{i=1}^{n} A_i . x_i^{(0)}$ за напрямом і евклідовою довжиною не збігається з вектором $\vec{b}$. Масштабуємо вектор $\vec{b}'$ за умови $\left\|\alpha \vec{b}'\right\|_2 = \left\|\vec{b}\right\|_2 \Rightarrow \alpha = \left\|\vec{b}\right\|_2 \Big/ \left\|\vec{b}'\right\|_2$. Масштабований вектор $\vec{y}^{(0)} = \alpha \vec{x}^{(0)}$ приймемо за початковий розв'язок системи $A\vec{x} = \vec{b}$.

Побудуємо алгоритм уточнення розв'язку системи $A\vec{x} = \vec{b}$ (з матрицею – чорний ящик) на основі базисних векторів $Kr_m = \left\{\vec{r}^{(0)}, A\vec{r}^{(0)}, ..., A^{m-1}\vec{r}^{(0)}\right\} = \left\{\vec{u}_0, \vec{u}_1, ..., \vec{u}_{m-1}\right\}$ підпростору Крилова і системи повних базисів $\left\{Ke_i\right\}_{i=1}^{n}$. Базисні вектори підпростору Крилова є сильнозумовленими, тому їх ортогоналізація методом Грама-Шмідта (Арнольді або Ланцоша) приводить до швидкої розортогоналізації в силу похибок обчислень [2; 4], отже $m \ll n$. Систему повних базисів найпростіше формувати на основі одиничного базису $\left\{\vec{e}_i\right\}_{i=1}^{n}$ у формі криловських базисів $Ke_i = \left\{\vec{e}_i, A\vec{e}_i, ..., A^{m-1}\vec{e}_i\right\}$, $i \in [1:n]$. Розмірність m криловських базисів $K_m = \left\{\vec{p}, A\vec{p}, ..., A^{m-1}\vec{p}\right\}$ вибиратимемо в залежності від зумовленості матриці A, за правилом: $\forall j = 0,1,...,t \le m-2$ виконується $\left|\cos\left\{A^j\vec{p}, A^{j+1}\vec{p}\right\}\right| < 1 - 10^{-s}$ і $\left|\cos\left\{A^{m-1}\vec{p}, A^m\vec{p}\right\}\right| \ge 1 - 10^{-s}$ ($s > 1$ — ціле додатне число, що вибирається в залежності від машинної точності.

Базис $Kr_m = \left\{\vec{r}^{(0)}, A\vec{r}^{(0)}, ..., A^{m-1}\vec{r}^{(0)}\right\} = \left\{\vec{u}_0, \vec{u}_1, ..., \vec{u}_{m-1}\right\}$, $m \ll n$ як не повний базис, використаємо для побудови початкового вектора $\vec{c}^{(0)}$. Методом Грама-Шмідта з векторів системи $\left\{\vec{u}_i\right\}_{i=0}^{m-1}$ побудуємо A-ортогональну систему $\left\{\tilde{\vec{u}}_i\right\}_{i=0}^{m-1}$, $\left(A\tilde{\vec{u}}_i, A\tilde{\vec{u}}_j\right) = 0$, $i \ne j$. Вектор $\vec{c}^{(0)}$, що оптимізує норму $\|\circ\|_2$ вектора нев'язки $\tilde{\vec{r}} = \vec{r}_0 + \sum_{i=0}^{m-1} \alpha_i A\tilde{\vec{u}}_i$ запишемо у формі

$$\vec{c}_0 = \sum_{i=0}^{m-1} \alpha_i \tilde{\vec{u}}_i \,, \alpha_i = -\left(\vec{r}^{(0)}, A\tilde{\vec{u}}_i\right) \Big/ \left\|A\tilde{\vec{u}}_i\right\|_2^2 . \qquad (8)$$

Результатом є вектор $\vec{c}^{(0)}$, що забезпечує монотонність процесу: $\tilde{\vec{x}} = \vec{x}^{(0)} + \gamma \vec{c}^{(0)}$, $\gamma = -\left(\vec{r}^{(0)}, \vec{c}^{(0)}\right) \Big/ \left\|A\vec{c}^{(0)}\right\|_2^2$. Дійсно $\left\|\tilde{\vec{r}}\right\|_2^2 = \left\|\vec{r}^{(0)}\right\|_2^2 -$

$-\left(\vec{r}^{(0)}, A\vec{c}^{(0)}\right)^2 \Big/ \left\|A\vec{c}^{(0)}\right\|_2^2 < \left\|\vec{r}^{(0)}\right\|_2^2$ Оптимальні параметри α_i розраховані теоретично в абсолютній арифметиці. Чим менше похибка розортогоналізації, тим ближче реально обчислений вектор $\vec{c}^{(0)}$ до оптимального.

Уточнимо вектор $\vec{c}^{(0)}$, якщо $\varphi = \cos^2\left\{\vec{r}^{(0)}, A\vec{c}^{(0)}\right\} < 1-10^{-s}$ на основі системи повних базисів Ke_i у формі $\vec{c}^{(i)} := \alpha\vec{c}^{(i)} + \beta_{k,i}\vec{s}^{(k,i)}$, $i \in I\left(\vec{r}^{(0)}\right)$, $k \in \{0,1,...,m-1\}$, де $I\left(\vec{r}^{(0)}\right)$ — множина індексів $i \in \{1,...,n\}$ упорядкована за неспаданням послідовності $\left|\vec{r}^{(0)}, A\vec{e}_i\right| \Big/ \left\|A\vec{e}_i\right\|_2^2$. Вектори $\vec{s}^{(k,i)}$ формуються з векторів $Ke_i = \left\{\vec{w}_0^{(i)}, ..., \vec{w}_{m-1}^{(i)}\right\}$ послідовно за правилом: $\vec{s}^{(0,i)} = \gamma\vec{c}^{(i)} + \vec{w}_0^{(i)}$, $\left(A\vec{s}^{(0,i)}, A\vec{c}^{(i)}\right) = 0 \Rightarrow \gamma = -\left(A\vec{w}_0^{(i)}, A\vec{c}^{(i)}\right) \Big/ \left\|A\vec{c}^{(i)}\right\|_2^2$;

$\forall k, j \in [1:m-1]$, $\vec{s}^{(k,i)} = \gamma\vec{c}^{(i)} + \sum_{j=0}^{k-1}\eta_j\vec{s}^{(j,i)} + \vec{w}^{(k)}$, $\left(A\vec{s}^{(k,i)}, A\vec{c}^{(i)}\right) = 0$,

$$\forall j \in [1:m-1] \quad \left(A\vec{s}^{(k,i)}, A\vec{s}^{(j,i)}\right) = 0,\ k \neq j, \tag{9}$$

$$\Rightarrow \gamma = -\left(A\vec{w}^{(k)}, A\vec{c}^{(i)}\right) \Big/ \left\|A\vec{c}^{(i)}\right\|_2^2,\ \eta_j = -\left(A\vec{w}^{(k)}, A\vec{s}^{(j,i)}\right) \Big/ \left\|A\vec{s}^{(j,i)}\right\|_2^2.$$

Оскільки $\forall k \in [0,1,...,m-1]$ система векторів $\left\{A\vec{s}^{(k,i)}\right\}$ ортогональна і $\left(A\vec{s}^{(k,i)}, A\vec{c}^{(i)}\right) = 0$, то $\vec{s}^{(i)} = \sum_{k=0}^{m-1}\delta_k\vec{s}^{(k,i)}$ при $\delta_k = -\left(\vec{r}^{(0)}, A\vec{s}^{(k,i)}\right) \Big/ \left\|A\vec{s}^{(k,i)}\right\|_2^2$ мінімізує норму $\|\circ\|_2$ вектора нев'язки. Внутрішній цикл по k завершено. Оскільки вектор $\vec{s}^{(i)}$ A-ортогональний до вектора $\vec{c}^{(i)}$, тому вектор $\vec{c}^{(i)} := \alpha\vec{c}^{(i)} + \beta\vec{s}^{(i)}$ на i-му кроці зовнішнього циклу мінімізує норму вектора нев'язки: $\vec{r}^{(i)} = \vec{r}^{(0)} + \alpha_i A\vec{c}^{(i)} + \beta_i A\vec{s}^{(i)}$,

$$\alpha_i = -\left(\vec{r}^{(0)}, A\vec{c}^{(i)}\right) \Big/ \left\|A\vec{c}^{(i)}\right\|_2^2,\ \beta_i = -\left(\vec{r}^{(0)}, A\vec{s}^{(i)}\right) \Big/ \left\|A\vec{s}^{(i)}\right\|_2^2 \tag{10}$$

(це забезпечує строгу монотонність процесу обчислень згладжує похибку обчислень).

Якщо $\cos^2\left\{\vec{r}^{(0)}, A\vec{c}^{(i)}\right\} < 1-10^{-s}$, то продовжити процес для $i \in [1:n]$. Після завершення циклу по i вектор $\vec{x}^{(1)}$ задаємо у формі

$$\vec{x}^{(1)} = \vec{x}^{(0)} + \xi\vec{c}^{(t)},\ \vec{r}^{(1)} = \vec{r}^{(0)} + \xi A\vec{c}^{(t)}, \tag{11}$$

де t — значення з множини $[1:n]$ після завершення процесу. Параметр ξ знайдемо з умови мінімізації вектора нев'язки: $\xi = -\left(\vec{r}^{(0)}, A\vec{c}^{(t)}\right) \Big/ \left\|A\vec{c}^{(t)}\right\|_2^2$. Якщо $\cos^2\left\{\vec{b}^{(1)}, \vec{b}\right\} < 1-10^{-s}$, $\vec{b}^{(1)} = \sum_{i=1}^{n} A_{i.}\vec{x}_i^{(1)}$, то замінивши $\vec{x}^{(0)} := \vec{x}^{(1)}$ процес повторити для нового наближення $\vec{x}^{(1)}$, нев'язки $\vec{r}^{(1)} = A\vec{x}^{(1)} - \vec{b}$.

Алгоритм побудований для матриць типу «чорний ящик». Якщо змінні системи упорядковані так, що матриця розбивається на укрупнені підсистеми ортогональних векторів-рядків (векторів-стовпців), то метод напрямленого пошуку стає економним і прискорюється швидкість збіжності та зменшується вплив похибок заокруглення на швидкість збіжності. Такими системами є різницеві еліптичні рівняння [5; 6; 12].

Теорема 3. Алгоритм оптимізації напрямного вектора (методом мінімізації норми вектора похибок або нев'язок) коректний: збігається, визначає параметри однозначно, мінімізує похибку обчислень (похибку розортогоналізації базисних векторів).

Доведення теореми випливає з формул (7)–(11).

6. Прискорення збіжності ітераційних процесів розв'язування систем $A\vec{x} = \vec{b}$ на основі мінімізації відношення Релея $\rho\left(A^T A, \vec{x} - \vec{x}^*\right)$. Нехай $\vec{x} \in K_{\min}$ — наближений розв'язок системи $A\vec{x} = \vec{b}$. Якщо вектор похибки $\vec{\varepsilon} = \vec{y} - \vec{x}^*$ є власним вектором матриці $A^T A$, то обчислений вектор $\vec{p} = A^T\vec{r} = A^T A\left(\vec{y} - \vec{x}^*\right) = \lambda\left(\vec{y} - \vec{x}^*\right)$ є напрямним вектором і розв'язок системи здійсниться за один крок.

Побудуємо алгоритм мінімізації відношення Релея $\rho\left(A^T A, \vec{x} - \vec{x}^*\right) = \|\vec{r}\|_2^2 \Big/ \left\|\vec{x} - \vec{x}^*\right\|_2^2$ на сфері $\left\|\vec{x} - \vec{x}^*\right\|_2^2 = \left\|\tilde{\vec{x}} - \vec{x}^*\right\|_2^2$ на основі підпростору Крилова $Kr_4 = \boldsymbol{span}\left\{\tilde{\vec{r}}, A\tilde{\vec{r}}, A^2\tilde{\vec{r}}, A^3\tilde{\vec{r}}\right\} = \boldsymbol{span}\left\{\vec{u}_1, \vec{u}_2, \vec{u}_3, \vec{u}_4\right\}$, $\tilde{\vec{r}} = A\tilde{\vec{x}} - \vec{b}$. Вектор $\vec{x}$, що мінімізує відношення Релея, шукатимемо у формі $\vec{x} = \tilde{\vec{x}} + \alpha A^T\vec{p} + \beta A^T\vec{s} \Rightarrow \vec{r} = \tilde{\vec{r}} + \alpha AA^T\vec{p} + \beta AA^T\vec{s}$, де $\vec{s} = \gamma\vec{u}_1 + \delta\vec{u}_2 + t\vec{u}_3 + \vec{u}_4$, $\vec{p} = \vec{u}_1$, параметри γ, δ знайдемо з умов

$$\left(A^T\vec{s}, A^T\vec{p}\right) = 0,\ \left(AA^T\vec{s}, AA^T\vec{p}\right) = 0. \qquad (12)$$

Розкриємо вирази $\left\|\vec{x} - \vec{x}^*\right\|_2^2$, $\|\vec{r}\|_2^2$, матимемо

$$\varphi = \|\vec{r}\|_2^2 = \left\|\tilde{\vec{r}}\right\|_2^2 + 2\alpha\left(\tilde{\vec{r}}, AA^T\vec{p}\right) + 2\beta\left(\tilde{\vec{r}}, AA^T\vec{s}\right) + \alpha^2\left\|AA^T\vec{p}\right\|_2^2 + \beta^2\left\|AA^T\vec{s}\right\|_2^2.$$

Рівняння сфери $\left\|\vec{x}-\vec{x}^*\right\|_2^2=\left\|\vec{\tilde{x}}-\vec{x}^*\right\|_2^2$ еквівалентне заданню еліпса відносно параметрів α,β:

$$L_1:\quad 2\alpha\left(\vec{\tilde{r}},\vec{p}\right)+2\beta\left(\vec{\tilde{r}},\vec{s}\right)+\alpha^2\left\|A^T\vec{p}\right\|_2^2+\beta^2\left\|A^T\vec{\tilde{s}}\right\|_2^2=0. \qquad (13)$$

Оскільки визначник системи (12) відносно параметрів γ,δ відмінний від нуля: $D_0=\left\|A^T\vec{u}_1\right\|_2^2\left(AA^T\vec{u}_1,AA^T\vec{u}_2\right)-\left\|AA^T\vec{u}_1\right\|_2^2\left(A^T\vec{u}_1,A^T\vec{u}_2\right)\neq 0$ (для векторів підпростору Крилова), то вектор визначиться з системи (12) у формі $\vec{s}=\vec{s}_0t+\vec{s}_1$.

Дві неперервні замкнуті опуклі лінії — лінія рівня L_2 функції $\varphi=\left\|\vec{\tilde{r}}\right\|_2^2$

$$L_2:2\alpha\left(\vec{\tilde{r}},AA^T\vec{p}\right)+2\beta\left(\vec{\tilde{r}},AA^T\vec{s}\right)+\alpha^2\left\|AA^T\vec{p}\right\|_2^2+\beta^2\left\|AA^T\vec{\tilde{s}}\right\|_2^2=0$$

і еліпс L_1, що мають спільну точку $\left(\alpha_0,\beta_0\right)=\left(0,0\right)$, де $\vec{s}=\vec{s}_0t+\vec{s}_1$, $t\in\mathbb{R}$, або перетинаються, або дотикаються. Якщо лінії перетинаються, то існує найближча точка $\left(\alpha_1,\beta_1\right)\in L_1$, в якій опукла функція φ приймає найменше значення: $\varphi\left(\alpha_1,\beta_1\right)<\varphi\left(\alpha_0,\beta_0\right)$. Щоб довести перетин ліній L_1, L_2, допустимо протилежне. Нехай $\left(0,0\right)$ є точкою дотику, тоді вектори $\vec{u}$, $\vec{v}$ з спільним початком у точці $\left(0,0\right)$ і кінцями у центрах ліній L_1, L_2 повинні бути ортогональними: $\cos^2\left(\vec{u},\vec{v}\right)=0$.

Дослідимо функцію $w\left(t\right)=\cos^2\left\{\vec{u}\left(t\right),\vec{v}\left(t\right)\right\}$ по параметру $t\in\mathbb{R}$, де $\vec{u}=\left[-\left(\vec{\tilde{r}},\vec{p}\right)\Big/\left\|A^T\vec{p}\right\|_2^2-\left(\vec{\tilde{r}},\vec{s}\left(t\right)\right)\Big/\left\|A^T\vec{s}\left(t\right)\right\|_2^2\right]^T$,

$$\vec{v}=\left[-\left(\vec{\tilde{r}},AA^T\vec{p}\right)\Big/\left\|AA^T\vec{p}\right\|_2^2-\left(\vec{\tilde{r}},AA^T\vec{s}\left(t\right)\right)\Big/\left\|AA^T\vec{s}\left(t\right)\right\|_2^2\right]^T.$$

Оскільки функція $w\left(t\right)$, визначена, неперервна і не є сталою, то існує принаймні одна точка t_0, для якої $w(t_0)\neq 0$ (отже лінії L_1, L_2 перетинаються).

Пошук точок на еліпсі з найменшим (найбільшим) значенням опуклої функції легко здійснити методами умовного екстремуму.

Існування мінімального значення відношення Релея в області, що є перетином конуса $K_{\min}$ і сфери $\left\|\vec{x}-\vec{x}^*\right\|_2=\left\|\vec{\tilde{x}}-\vec{x}^*\right\|_2$ випливає з умови, що область $K_{\min}$ містить сингулярну пряму $S_{\min}$.

7. Критерій зупинки ітерацій і прийняття наближеного рішення системи $A\vec{x}=\vec{b}$**.** При розв'язуванні систем $A\vec{x}=\vec{b}$ з погано зумовленими матрицями в умовах машинної арифметики критерії зупинки ітерацій: $\left\|\vec{x}^{(k+1)}-\vec{x}^{(k)}\right\|<\varepsilon$ або $\left\|\vec{r}^{(k+1)}\right\|<\varepsilon$, або на основі теоретичного розрахунку числа ітерацій, що забезпечують асимптотичну збіжність з похибкою $\left\|\vec{r}^{(k)}\right\|<\varepsilon$ втрачають зміст. Область $K_{\min}$ і похибки обчислень стають на перешкоді застосуванню критерія прийняття нормального рішення [8]: $\min\limits_{\vec{x}\in\mathbb{R}^n}\left(\left\|A\vec{x}-\vec{b}\right\|_2^2+\alpha\left\|\vec{x}-\vec{x}^*\right\|_2^2\right)$.

За критерій зупинки ітерацій і прийняття рішення приймемо умову: якщо для деякого наближення $\vec{x}^{(k)}\in\mathbb{R}^n$ виконується нерівність $\cos^2\left\{\vec{c}\left(\vec{x}^{(k)}\right),\vec{b}\right\}\geq 1-10^{-t}$, $t>1$ — пов'язане з машинною точністю, $\vec{c}\left(\vec{x}^{(k)}\right)=\sum_{i=1}^{n}A_{i\cdot}\,\vec{x}_i^{(k)}$, то ітераційний процес зупинити і $\vec{x}^{(k)}$ — наближене рішення системи.

Теорема 4. Функціонал $\cos^2\left\{\vec{c}\left(\vec{x}\right),\vec{b}\right\}$ стійкий до похибок збурень.

Доведення теореми. Нехай

$$\vec{x}=\vec{x}^*+\overrightarrow{\Delta x}\Rightarrow\ \vec{c}\left(\vec{x}\right)=\sum_{i=1}^{n}A_{i\cdot}\left(x_i^*+\Delta x_i\right),$$

$$\vec{c}\left(\vec{x}\right)=\vec{b}+\vec{w}\Rightarrow\ \cos^2\left\{\vec{c},\vec{b}\right\}=\left(\vec{b}+\vec{w},\vec{b}\right)^2\Big/\left(\left\|\vec{b}\right\|_2^2\cdot\left\|\vec{b}+\vec{w}\right\|_2^2\right)=$$

$$=\left(1+\alpha\right)^2\Big/\left(1+2\alpha+\beta^2\right)^2=\varphi,\ \text{де}\ \alpha=\left(\vec{w},\vec{b}\right)\Big/\left\|\vec{b}\right\|_2^2,\ \beta=\left\|\vec{w}\right\|_2\Big/\left\|\vec{b}\right\|_2.$$

Максимальне значення функція φ приймає тоді і лише тоді, якщо $\alpha=\beta=0$. Отже, похибка $\left\|\vec{w}\right\|_2=0$, що можливо лише за умови $\forall i\in[1:n]\ \Delta x_i=0$.

8. Бі-спряжені базиси. Якщо матриця A системи $A\vec{x}=\vec{b}$ близька до виродженої ($\left|\det A\right|$ — мале додатне число, ***cond*** A — велике число), або A, $\vec{b}$ взяті з експерименту (задані з наближенням), то наближений розв'язок (квазірозв'язок) такої системи можна отримати шляхом мінімізації функціоналу $\varphi=\left\|A\vec{x}-\vec{b}\right\|_2^2+\delta\left\|\vec{x}-\vec{y}^{(0)}\right\|_2^2$ (δ — параметр регуляризації [8]). Побудуємо у внутрішньому циклі процесу (7) із базису Ke_i, $i\in[1:n]$ бі-спряжену систему векторів з матрицями A^T та AA^T [13]:

$\left(AA^T\vec{c}^{(t,j)}, AA^T\vec{c}^{(p,j)}\right)=0$, $\left(A^T\vec{c}^{(t,j)}, A^T\vec{c}^{(p,j)}\right)=0$, $\forall t\neq p\in\{0,1,...,m-1\}$.

Тоді мінімум φ існує і єдиний, оптимальні параметри $\alpha_{k,j}$ визначаються за формулами

$$\alpha_{k,j}=-\left(\vec{r}^{(j)}, AA^T\vec{c}^{(k,j)}\right)\Big/\left\|AA^T\vec{c}^{(k,j)}\right\|_2^2-\delta\left(\vec{\rho}^{(j)}, \vec{c}^{(k,j)}\right)\Big/\left\|A^T\vec{c}^{(k,j)}\right\|_2^2,$$

де $\vec{\rho}^{(j)}=A\vec{x}^{(j)}-A\vec{y}^{(0)}$, $\vec{y}^{(0)}\in\mathbb{R}^n$ — вектор відносно якого шукаємо нормальний розв'язок. Якщо $\vec{y}^{(0)}=\vec{x}^*$, то $\vec{\rho}^{(j)}=\vec{r}^{(j)}$.

Висновки. У роботі розв'язані проблемні задачі теорії лінійних систем: 1) встановлена залежність похибок зашумленої системи від похибки матриці, похибки правої частини, від розв'язку, порядку та величини визначника матриці; 2) побудований багатошаровий алгоритм гауссового виключення, що одночасно обнулює піддіагональні елементи в m стовпцях і мінімізує похибку обчислень та модифікований метод для погано зумовлених матриць типу матриці Гільберта та матриць, що не можуть задовільно масштабуватись; 3) показано, що перепоною для відшукання з високою точністю наближених розв'язків у системах з погано зумовленими матрицями стають області $K_{\min}$, що містять сингулярні прямі з найменшими власними значеннями; 4) запропонований двоциклічний ітераційний метод напрямленого пошуку для розв'язування систем з матрицями великих порядків на основі системи повних базисів криловського типу, що мінімізує похибку обчислень за рахунок вибору оптимальних параметрів у підпросторах Крилова малих розмірностей і проведення обчислень поза областю $K_{\min}$; 5) запропонований алгоритм прискорення збіжності на основі мінімізації відношення Релея на сфері.

Список використаних джерел:

1. Хейгеман Л. Прикладные итерационные методы / Л. Хейгеман, Д. Янг ; пер. с англ. — М. : Мир, 1986. — 448 с.
2. Деммель Дж. Вычислительная линейная алгебра / Дж. Деммель. — М. : Мир, 2001. — 429 с.
3. Райс. Дж. Матричные вычисления и математическое обеспечение / Дж. Райс ; пер. с англ. — М. : Мир, 1984. — 264 с.
4. Воеводин В. В. Матрицы и вычисления / В. В. Воеводин, Ю. А Кузнецов. — М. : Наука, 1984. — 320 с.
5. Патанкар С. Численные методы решения задач теплообмена и динамики жидкости / С. Патанкар. — М. : Энергоатомиздат, 1984. — 125 с.
6. Зверев В. Г. Модифицированный полилинейный метод решения разностных эллиптических уравнений / В. Г. Зверев // ЖВМ и МФ. — 1998. — Т. 38. — № 9. — С. 1553–1562.
7. Зенкевич О.Конечные элементы и аппроксимация / О. Зенкевич, К. Морган. — М. : Мир, 1986. — 318 с.

8. Тихонов А. Н. Методы решения некорректных задач / А. Н. Тихонов, В. Я. Арсенин. — М. : Наука, 1974. — 223 с.
9. Гилмор Р. Прикладная теория катастроф / Р. Гилмор. — М. : Мир, 1984. — 350 с. (кн.1), 282 с. (кн.2).
10. Хорн Р. Матричный анализ / Р. Хорн, Ч. Джонсон ; пер. с англ. — М. : Мир, 1989. — 655 с.
11. Абрамчук В. С. Итерационные методы направленного поиска решения систем $Ax = fc$ сингулярно-естественным упорядочением переменных / В. С. Абрамчук // Доп. НАН Украины. — 1996. — № 8. — С. 4–8.
12. Абрамчук В. С. Проблеми, методи, алгоритми розв'язування систем лінійних рівнянь з погано зумовленими матрицями / В. С. Абрамчук, І. В. Абрамчук // Математичне та комп'ютерне моделювання. Серія: Фізико-математичні науки : зб. наук. праць. — Кам'янець-Подільський : Кам'янець-Подільський національний університет імені Івана Огієнка, 2016. — Вип. 13. — С. 5–16.
13. Ильин В. П. Методы би-сопряженных направлений в подпространствах Крылова / В. П. Ильин // СибЖИМ. — 2008. — Т. 11, № 4 (36). — С. 47–60.

The problems, that investigates, was related with estimation of errors in the solutions of noised systems and minimizations errors of calculations that arise in Gaussian transformations and acceleration of convergence of the iterative methods.

Key words: *the estimation of solution's error, multilayer method of Gaussian elimination, the local basis of the direct search method, Rayleigh–Ritz quotient.*

Отримано: 23.03.2017

УДК 517.97

С. М. Бак, канд. фіз.-мат. наук, доцент

Вінницький державний педагогічний університет
імені Михайла Коцюбинського, м. Вінниця

ІСНУВАННЯ СТОЯЧИХ ХВИЛЬ В ДИСКРЕТНОМУ НЕЛІНІЙНОМУ РІВНЯННІ ШРЕДІНГЕРА З КУБІЧНОЮ НЕЛІНІЙНІСТЮ НА ДВОВИМІРНІЙ ҐРАТЦІ

Стаття присвячена вивченню дискретного нелінійного рівняння типу Шредінгера з кубічною нелінійністю на двовимірній ґратці. Одержано результат про існування стоячих хвиль для таких рівнянь.

Ключові слова: *дискретне нелінійне рівняння Шредінгера, двовимірна ґратка, стоячі хвилі, критичні точки, теорема про зачеплення.*

Вступ. Останнім часом значну увагу приділяють моделям, дискретним за просторовою змінною. Серед рівнянь, які описують такі моделі, найбільш відомими є рівняння ланцюгів осциляторів, дискре-