МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ
ОДЕССКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ
ИМЕНИ И. И. МЕЧНИКОВА
ИНСТИТУТ МАТЕМАТИКИ, ЭКОНОМИКИ И МЕХАНИКИ

В. В. Вербицкий, В. В. Реут

ВВЕДЕНИЕ В ЧИСЛЕННЫЕ МЕТОДЫ АЛГЕБРЫ

Учебное пособие
для студентов высших учебных заведений,
обучающихся по специальности "Прикладная математика"

ОДЕССА
ОНУ
2015

УДК 519.61 (075.8)
ББК 22.192я73
В31

Рекомендовано к печати решением Ученого совета Одесского национального университета имени И. И. Мечникова.
Протокол № 1 от 30.09.2014 г.

Рецензенты:

Г. А. Шынкаренко, доктор физико-математических наук, профессор, заведующий кафедрой информационных систем Львовского национального университета имени Ивана Франка;

А. В. Плотников, доктор физико-математических наук, профессор, заведующий кафедрой прикладной вычислительной математики и САПР Одесской государственной академии строительства и архитектуры;

В. И. Острик, доктор физико-математических наук, доцент, заведующий НИЛ механико-математического факультета Киевского национального университета имени Тараса Шевченко.

Вербицкий В. В., Реут В. В.

В31 Введение в численные методы алгебры: учебное пособие / В. В. Вербицкий, В. В. Реут. — Одесса: Одесский национальный университет имени И. И. Мечникова, 2015. — 165 с.
ISBN 978-617-689-104-8

В учебном пособии излагаются основные численные методы решения линейных и нелинейных систем уравнений, полной и частичной проблем собственных значений, линейной задачи наименьших квадратов.

Для студентов высших учебных заведений, обучающихся по специальности "Прикладная математика".

УДК 519.61 (075.8)
ББК 22.192я73

ISBN 978-617-689-104-8 © В. В. Вербицкий, В. В. Реут, 2015
© Одесский национальный университет имени И. И. Мечникова, 2015

Оглавление

Введение

Предлагаемое учебное пособие является конспектом лекций по первой части курса „Численные методы“, читаемого в Одесском национальном университете имени И. И. Мечникова для студентов специальности „Прикладная математика“. Пособие может быть использовано для первоначального ознакомления с основными современными численными методами линейной алгебры и нелинейных уравнений.

В разделе 1 пособия приведены некоторые сведения из линейной алгебры и теории матриц, необходимые для понимания излагаемого в дальнейшем материала. В разделе 2 рассмотрена арифметика с плавающей точкой, введены понятия устойчивого (неустойчивого) алгоритма и хорошо (плохо) обусловленной задачи. Методам решения систем линейных алгебраических уравнений посвящены разделы 3 и 5. В разделе 3 рассматриваются прямые методы, в разделе 5 — итерационные. Методы решения линейной задачи наименьших квадратов обсуждаются в разделе 4. Методы решения задач на собственные значения для симметричных матриц приведены в разделе 6. Основные методы решения нелинейных скалярных уравнений и нелинейных систем уравнений рассмотрены в разделе 7.

Продолжить знакомство с численными методами алгебры можно по многочисленным учебникам и монографиям [1, 7, 9, 26, 27, 31, 33, 34, 35, 36].

Алгоритмы численных методов, приведенные в пособии, должны помочь студентам составлять учебные программы (рекомендуется язык программирования пакета MATLAB) с целью лучшего усвоения материала.

1. Предварительные сведения из линейной алгебры и теории матриц

Построение и анализ алгоритмов вычислительной линейной алгебры требуют хороших знаний в области линейной алгебры и теории матриц. В этом разделе приведены основные сведения, необходимые для понимания излагаемого в дальнейшем материала.

1.1. Матрицы и подпространства

Множество векторов $\{y_1, \ldots, y_m\}$ из R^n $(m \leq n)$ называется *линейно независимым*, если из

$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

следует $\alpha_i = 0,\ i = \overline{1,m}$. В противном случае множество векторов *линейно зависимо*.

Если задано множество векторов $\{y_1, \ldots, y_m\}$ из R^n $(m \leq n)$, то множество всевозможных линейных комбинаций этих векторов является линейным подпространством R^n, называемым *линейной оболочкой* $\{y_1, \ldots, y_m\}$:

$$\operatorname{span}\{y_1, \ldots, y_m\} = \left\{ \sum_{i=1}^{m} \alpha_i y_i \ :\ \alpha_i \in R,\, i = \overline{1,m} \right\}.$$

Если $S_1, \ldots, S_k$ — подпространства R^n, то их сумма определяется как подпространство

$$S = \{y_1 + \cdots + y_k \ :\ y_i \in S_i,\, i = \overline{1,k}\}.$$

Подпространство S является *прямой суммой* подпространств $S_1, \ldots, S_k$ и обозначается

$$S = S_1 \oplus \cdots \oplus S_k,$$

если каждый вектор $x \in S$ имеет единственное представление

$$x = y_1 + \cdots + y_k, \quad \text{где } y_i \in S_i,\ i = \overline{1,k}.$$

Подмножество $\{y_{i_1}, \ldots, y_{i_k}\}$ называется *максимальным линейно независимым подмножеством* множества $\{y_1, \ldots, y_m\}$, если оно линейно независимо и не является подмножеством никакого другого линейно независимого подмножества множества $\{y_1, \ldots, y_m\}$.

Если $\{y_{i_1}, \ldots, y_{i_k}\}$ — максимальное линейно независимое подмножество множества $\{y_1, \ldots, y_m\}$, то

$$\text{span}\{y_1, \ldots, y_m\} = \text{span}\{y_{i_1}, \ldots, y_{i_k}\}$$

и $\{y_{i_1}, \ldots, y_{i_k}\}$ является базисом для подпространства $\text{span}\{y_1, \ldots, y_m\}$.

Если S — подпространство R^n, то можно найти базис $\{y_1, \ldots, y_m\}$ так, что

$$S = \text{span}\{y_1, \ldots, y_m\}.$$

Все базисы подпространства S имеют одинаковое количество элементов. Это число называется *размерностью* S и обозначается $\dim(S)$.

Со всякой матрицей $A \in R^{m \times n}$ связаны два важных подпространства. *Область значений* (*образ*) матрицы A, которое определяется так:

$$\text{range}(A) = \{y \in R^m \ : \ y = Ax \text{ для некоторого } x \in R^n\}.$$

Ядро (*нуль–пространство*) матрицы A, которое определяется следующим образом:

$$\text{null}(A) = \{x \in R^n \ : \ Ax = 0\}.$$

Если матрицу A определить через ее столбцы, $A = [a_1, \ldots, a_n]$, то

$$\text{range}(A) = \text{span}\{a_1, \ldots, a_n\}.$$

Ранг матрицы A определяется следующим образом:

$$\text{rank}(A) = \dim(\text{range}(A)).$$

Можно показать, что

$$\text{rank}(A) = \text{rank}(A^T).$$

Таким образом ранг матрицы равен числу линейно независимых строк (столбцов).

Для любой матрицы $A \in R^{m \times n}$ имеем

$$\dim(\text{null}(A)) + \text{rank}(A) = n.$$

Единичная матрица $I \in R^{n \times n}$ определяется следующим столбцовым представлением

$$I = [e_1, \ldots, e_n],$$

где

$$e_k = [\underbrace{0, \ldots, 0}_{k-1}, 1, \underbrace{0, \ldots, 0}_{n-k}]^T.$$

Если матрицы A и X из $R^{n\times n}$ удовлетворяют равенству

$$AX = I,$$

то X называется *обратной* к A и для нее используется обозначение A^{-1}.

Некоторые свойства обратной матрицы играют важную роль в вычислительной линейной алгебре. Матрица, обратная к произведению матриц, является произведением обратных к сомножителям, взятым в обратном порядке:

$$(AB)^{-1} = B^{-1}A^{-1}.$$

Транспонирование обратной матрицы — это то же самое, что обращение транспонированной:

$$(A^{-1})^T = (A^T)^{-1} \equiv A^{-T}.$$

Если $A = (a) \in R^{1\times 1}$, то ее *определитель* (*детерминант*) задается равенством $\det(A) = a$. Определитель $n \times n$–матрицы задается через определители $(n-1) \times (n-1)$–матриц. А именно, для $A \in R^{n\times n}$ полагают:

$$\det(A) = \sum_{j=1}^{n} (-1)^{j+1} a_{1j} \det(A_{1j}),$$

где A_{1j} — это $(n-1)\times(n-1)$–матрица, получаемая из A вычеркиванием первой строки и j–столбца.

Приведем несколько важных свойств определителей:

$$\begin{array}{ll}
\det(AB) = \det(A)\det(B), & A, B \in R^{n\times n}, \\
\det(A^T) = \det(A), & A \in R^{n\times n}, \\
\det(\alpha A^T) = \alpha^n \det(A), & A \in R^{n\times n}, \alpha \in R, \\
\det(A) \neq 0 \Leftrightarrow A \text{ невырождена}, & A \in R^{n\times n}.
\end{array}$$

Доказательства этих свойств и свойств обратной матрицы можно найти в любом учебнике по линейной алгебре, например в [11, 4, 12].

1.2. Векторные и матричные нормы

Пусть X — вещественное линейное пространство. Функция $||\cdot|| : X \to R$ со свойствами:

1. $||x|| \geq 0$,
2. $||x|| = 0 \;\Leftrightarrow\; x = 0$,
3. $||\alpha x|| = |\alpha|\,||x||$,

4. $||x + y|| \leq ||x|| + ||y||$

для всех элементов $x, y \in X$ и для любых $\alpha \in R$, называется нормой на X.

Две нормы $||\cdot||_\alpha$ и $||\cdot||_\beta$ в X называются эквивалентными, если существуют такие положительные константы c_1 и c_2, что

$$c_1||x||_\alpha \leq ||x||_\beta \leq c_2||x||_\alpha$$

для всех $x \in X$.

В конечномерном пространстве все нормы эквивалентны.

Если X — это пространство векторов R^n, норму будем называть векторной.

Полезный класс векторных норм — это p-нормы, определяемые как

$$||x||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}.$$

Наиболее важными из векторных p-норм являются 1, 2 и ∞ нормы:

$$||x||_1 = \sum_{i=1}^{n} |x_i|,$$
$$||x||_2^2 = \sum_{i=1}^{n} x_i^2 = x^T x,$$
$$||x||_\infty = \max_{1 \leq i \leq n} |x_i|.$$

Векторная 2-норма называется евклидовой нормой. Эти векторные нормы эквивалентны, и имеют место неравенства:

$$||x||_2 \leq ||x||_1 \leq \sqrt{n}||x||_2, \tag{1.1}$$
$$||x||_\infty \leq ||x||_2 \leq \sqrt{n}||x||_\infty, \tag{1.2}$$
$$||x||_\infty \leq ||x||_1 \leq n||x||_\infty. \tag{1.3}$$

Если в качестве пространства X взять пространство матриц $R^{m \times n}$, то норма называется матричной.

В дальнейшем будем использовать только мультипликативные матричные нормы, т. е., нормы которые удовлетворяют свойству:

$$||AB|| \leq ||A||\,||B|| \quad \forall A \in R^{m \times p},\ \forall B \in R^{p \times n}.$$

Наиболее часто используются следующие матричные нормы:

$$||A||_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|,$$
$$||A||_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}|,$$
$$||A||_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \text{ (норма Фробениуса)}.$$

Эти матричные нормы эквивалентны, и имеют место неравенства:

$$||A||_2 \le ||A||_F \le \sqrt{n}||A||_2, \tag{1.4}$$

$$\frac{1}{\sqrt{n}}||A||_\infty \le ||A||_2 \le \sqrt{m}||A||_\infty, \tag{1.5}$$

$$\frac{1}{\sqrt{m}}||A||_1 \le ||A||_2 \le \sqrt{n}||A||_1. \tag{1.6}$$

Матричная норма $||\cdot||_\beta$ называется согласованной с векторной нормой $||\cdot||_\alpha$, если

$$||Ax||_\alpha \le ||A||_\beta ||x||_\alpha \quad \forall x \in R^n.$$

Например, матричная ∞-норма согласована с векторной ∞-нормой. Действительно,

$$||Ax||_\infty = \max_{1\le i\le m} |\sum_{j=1}^{n} a_{ij}x_j| \le \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|\,|x_j| \le$$

$$\le (\max_{1\le j\le n} |x_j|)(\max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|) = ||A||_\infty ||x||_\infty.$$

Матричная норма $||\cdot||_\alpha$ называется порождаемой векторной нормой $||\cdot||_\alpha$, если она определяется следующим образом:

$$||A||_\alpha = \max_{x\ne 0} \frac{||Ax||_\alpha}{||x||_\alpha} = \max_{||x||_\alpha=1} ||Ax||_\alpha.$$

Покажем, что матричная ∞-норма порождается векторной ∞-нормой. Пусть $||\cdot||_{\infty,*}$ — матричная норма порождаемая векторной ∞-нормой. Ранее было показано, что матричная ∞-норма согласована с векторной ∞-нормой, поэтому

$$||A||_{\infty,*} = \max_{||x||_\infty=1} ||Ax||_\infty \le ||A||_\infty. \tag{1.7}$$

Пусть

$$||A||_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}| = \sum_{j=1}^{n} |a_{i_0 j}|.$$

Выберем x^* следующим образом, $x_j^* = \frac{|a_{i_0 j}|}{a_{i_0 j}}$, если $a_{i_0 j} \ne 0$, и $x_j^* = 0$ в противном случае ($j = \overline{1,n}$). Очевидно, что $||x^*||_\infty = 1$. Далее

$$||Ax^*||_\infty = \max_{1\le i\le n} |\sum_{j=1}^{n} a_{ij}x_j^*| \ge |\sum_{j=1}^{n} a_{i_0 j}x_j^*| = \sum_{j=1}^{n} |a_{i_0 j}| = ||A||_\infty.$$

Значит,

$$||A||_\infty \le ||Ax^*||_\infty \le \max_{||x||_\infty=1} ||Ax||_\infty = ||A||_{\infty,*}. \tag{1.8}$$

Требуемое следует из (1.7) и (1.9).

Аналогично можно показать, что матричная 1-норма порождается векторной 1-нормой. Векторной 2-нормой порождается матричная 2-норма:

$$|A||_2^2 = \max_{x\neq 0} \frac{||Ax||_2^2}{||x||_2^2} = \max_{x\neq 0} \frac{(Ax)^T(Ax)}{x^T x}. \tag{1.9}$$

Матричную 2-норму часто называют спектральной нормой. К изучению свойств этой матричной нормы вернемся в дальнейшем.

Теорема 1.2.1. *Матричная норма порождаемая векторной нормой является наименьшей из согласованных с этой векторной нормой.*

Доказательство. Согласованность очевидна.

Пусть $||\cdot||_*$ — матричная норма, согласованная с векторной нормой $||\cdot||$, т. е.,

$$||Ax|| \le ||A||_*||x|| \quad \forall x \in R^n.$$

Тогда

$$\frac{||Ax||}{||x||} \le ||A||_* \quad x \neq 0,$$

и

$$||A|| = \max_{x\neq 0} \frac{||Ax||}{||x||} \le ||A||_*.$$

Теорема доказана. □

Теорема 1.2.2. *Каждая норма вектора порождает некоторую матричную норму.*

Доказательство. Функционал $F(x) = ||Ax||$ на сфере единичного радиуса

$$S_1 = \{x \in R^n \,:\, ||x|| = 1\}$$

непрерывен, что следует из свойств нормы, а значит ограничен и достигает наибольшего значения. Таким образом, можно положить

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

□

Отметим, что не всякая норма матрицы порождается нормой вектора. Действительно, с одной стороны порождаемая норма единичной матрицы I равна 1, ибо

$$||I|| = \max_{x \neq 0} \frac{||Ix||}{||x||} = 1.$$

С другой стороны

$$||I||_F = \sqrt{n} > 1 \quad \text{при } n > 1,$$

что противоречит теореме 1.2.1, если предположить, что норма Фробениуса $|| \cdot ||_F$ порождается какой-либо векторной нормой.

1.3. Ортогональные матрицы

Матрица $Q \in R^{n \times n}$ называется *ортогональной*, если

$$Q^T Q = Q Q^T = I.$$

Установим некоторые свойства ортогональных матриц.

Свойство 1.3.1. *Столбцы (строки) ортогональной матрицы ортонормированы.*

Доказательство. Представим матрицу Q ее столбцами,

$$Q = [q_1, q_2, \ldots, q_n].$$

Тогда

$$I = Q^T Q = \begin{bmatrix} q_1^T \\ q_2^T \\ \vdots \\ q_n^T \end{bmatrix} [q_1, q_2, \ldots, q_n] = \begin{bmatrix} q_1^T q_1 & q_1^T q_2 & \cdots & q_1^T q_n \\ q_2^T q_1 & q_2^T q_2 & \cdots & q_2^T q_n \\ \vdots & \vdots & \ddots & \vdots \\ q_n^T q_1 & q_n^T q_2 & \cdots & q_n^T q_n \end{bmatrix}.$$

Отсюда следует, что

$$q_i^T q_j = \delta_{ij}, \ i, j = \overline{1, n},$$

где δ_{ij} — символ Кроннекера. Поскольку

$$(Q^T)^T Q^T = I,$$

то столбцы матрицы Q^T, т.е. строки матрицы Q, тоже ортонормированы. □

Свойство 1.3.2. *Умножение ортогональной матрицы на вектор не меняет его евклидовой нормы.*

Доказательство. Действительно

$$||Qx||_2^2 = (Qx)^T Qx = x^T Q^T Qx = ||x||_2^2.$$

□

Свойство 1.3.3. *Умножение слева матрицы A на ортогональную матрицу не меняет норму Фробениуса и 2-норму матрицы A.*

Доказательство. Используя свойство 1.3.2, получаем

$$||QA||_F^2 = |||[Qa_1, \ldots, Qa_n]|||_F^2 = ||Qa_1||_2^2 + \cdots + ||Qa_n||_2^2 =$$

$$= ||a_1||_2^2 + \cdots + ||a_n||_2^2 = ||A||_F^2,$$

и

$$||QA||_2^2 = \max_{||x||_2 \neq 0} \frac{||QAx||_2^2}{||x||_2^2} = \max_{||x||_2 \neq 0} \frac{(QAx)^T (QAx)}{||x||_2^2} =$$

$$= \max_{||x||_2 \neq 0} \frac{x^T A^T Q^T QAx}{||x||_2^2} = \max_{||x||_2 \neq 0} \frac{x^T A^T Ax}{||x||_2^2} == \max_{||x||_2 \neq 0} \frac{||Ax||_2^2}{||x||_2^2} = ||A||_2^2.$$

□

В качестве примеров ортогональных матриц рассмотрим матрицы вращения, перестановок и отражения.

В двумерном случае матрица вращения определяется следующим образом. Рассмотри в R^2 вектор $x = [x_1, x_2]^T$, который с осью Ox_1 образует угол α (см. рис. 1.1). Поворачивая этот вектор против часовой стрелки на угол φ, получим вектор $y = [y_1, y_2]^T$. Поскольку

$$y_1 = |y| \cos(\alpha + \varphi) = |x|(\cos\alpha \cos\varphi - \sin\alpha \sin\varphi) = x_1 \cos\varphi - x_2 \sin\varphi,$$
$$y_2 = |y| \sin(\alpha + \varphi) = |x|(\sin\alpha \cos\varphi - \cos\alpha \sin\varphi) = x_1 \sin\varphi + x_2 \cos\varphi,$$

то

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Матрица

$$Q_{12}(\varphi) = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix}$$

называется *матрицей вращения*. Таким образом, умножение матрицы вращения $Q_{12}(\varphi)$ на вектор поворачивает этот вектор против часовой стрелки на угол φ.

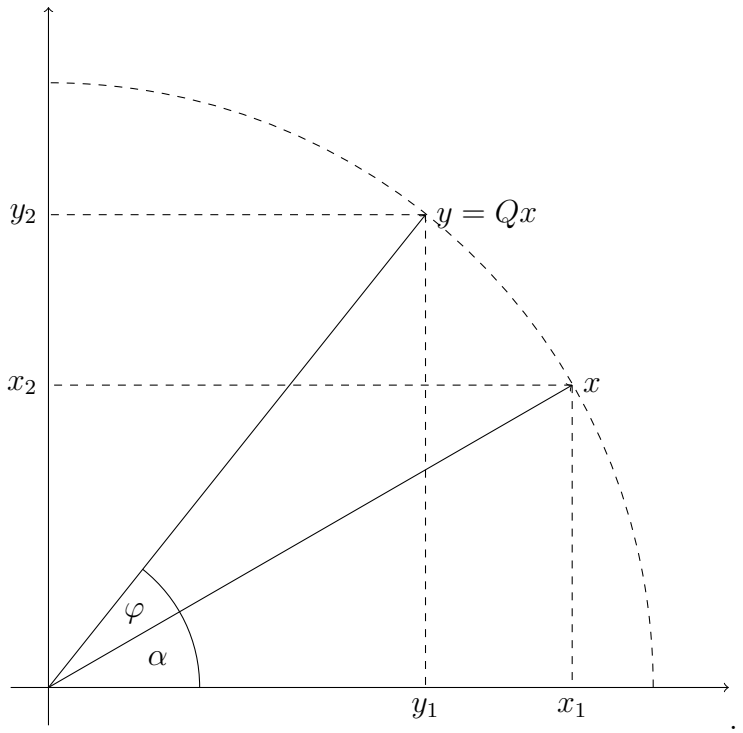

Рис. 1.1. Умножение матрицы вращения на вектор

В n-мерном случае матрица вращения

$$Q_{kl}(\varphi) = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & \cos\varphi & \dots & -\sin\varphi & & \\ & & \vdots & \ddots & \vdots & & \\ & & \sin\varphi & \dots & \cos\varphi & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \begin{matrix} \\ \\ k \\ \\ l \\ \\ \\ \end{matrix} \tag{1.10}$$

поворачивает n-мерный вектор в плоскости $x_k 0 x_l$ на угол φ.

Элементарной матрицей перестановок называется матрица P^{i,γ_i}, которая получается из единичной матрицы перестановкой i-й и γ_i-й строк.

Умножение матрицы P^{i,γ_i} на вектор x меняет местами в векторе x i-ю и γ_i-ю компоненты.

Умножение матрицы P^{i,γ_i} на матрицу A слева (справа) меняет местами в матрице A i-ю и γ_i-ю строки (столбцы).

Пусть $\gamma = (\gamma_1, \dots, \gamma_n)$ — некоторая перестановка натуральных чисел $1, \dots, n$. *Матрицей перестановок* P^γ называется матрица у которой i-я

строка — это γ_i-я строка единичной матрицы. Любая матрица перестановок может быть представлена в виде произведения конечного числа элементарных матриц перестановок. Очевидно, что произведение любого числа матриц перестановок есть матрица перестановок.

Если матрицу P^γ умножить на вектор x,то в векторе x i-й компонентой станет γ_i-я компонента.

Если матрицу A умножить слева на P^γ (справа на $(P^\gamma)^T$), то γ_i-я строка станет i-й (γ_i-й столбец станет i-м).

Пример 1.3.1. Пусть $\gamma = (2, 3, 4, 1)$, тогда

$$P^\gamma = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

и

$$P^\gamma A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \\ a_{11} & a_{12} & a_{13} & a_{14} \end{bmatrix}.$$

Матрица отражения (матрица Хаусхолдера) определяется следующим образом. Произвольный вектор $u \in R^n$ однозначно определяет $(n-1)$–мерную плоскость $P = \mathrm{span}\{u\}^\perp$ в R^n. Матрица отражения $H(u)$ строится так, что вектор $H(u)v$ — отражение относительно плоскости P произвольного вектора $v \in R^n$ (см. рис. 1.2). Обозначим через v_1 проек-

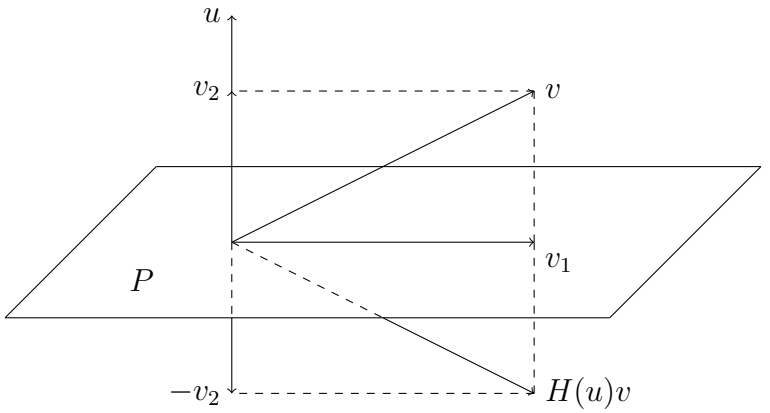

Рис. 1.2. Умножение матрицы отражения $H(u)$ на вектор

цию вектора v на плоскость P, а через v_2 проекцию вектора v на $\mathrm{span}\{u\}$, тогда

$$v = v_1 + v_2.$$

Нормируя вектор u,

$$\widetilde{u} = \frac{u}{\sqrt{(u^T u)}} = \frac{u}{||u||_2},$$

получаем

$$v_2 = \widetilde{u}^T v \widetilde{u} = \widetilde{u}\widetilde{u}^T v.$$

Тогда

$$H(u)v = H(u)(v_1 + v_2) = H(u)v_1 + H(u)v_2 = v_1 - v_2 =$$

$$= v - 2v_2 = v - 2\widetilde{u}\widetilde{u}^T v = (I - 2\widetilde{u}\widetilde{u}^T)v.$$

Отсюда

$$H(u) = (I - \gamma u u^T), \tag{1.11}$$

где

$$\gamma = \frac{2}{u^T u}.$$

Очевидно, что матрица отражения $H(u)$ симметрична. Легко доказать ее ортогональность:

$$H(u)H(u)^T = H(u)H(u) = (I - \gamma u u^T)(I - \gamma u u^T) =$$

$$= I^2 - 2\gamma u u^T + \gamma^2 u u^T u u^T = I - 2\gamma u u^T + 2\gamma u u^T = I.$$

Для любого вектора v из $(n-1)$–мерной плоскости P

$$H(u)v = v.$$

Значит, $\lambda = 1$ есть собственное значение матрицы $H(u)$ и этому собственному значению соответствует $n-1$ собственный вектор. Если же $v \in \mathrm{span}(u)$, то

$$H(u)v = -v.$$

Поэтому $\lambda = -1$ — собственное значение матрицы $H(u)$ кратности 1.

Пример 1.3.2. Вычислим матрицу отражения $H(u)$ для вектора $u = [1,1,2]^T$.

$$H(u) = I - \frac{2}{u^T u} u u^T = I - \frac{2}{6} \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} [1,1,2] =$$

$$= I - \frac{1}{3} \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 2 \\ 2 & 2 & 4 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 2 & -1 & -2 \\ -1 & 2 & -2 \\ -2 & -2 & -1 \end{bmatrix}.$$

1.4. Операторы проектирования

Оператор $P : R^n \to R^n$ называется проектором, если

$$P^2 = P. \tag{1.12}$$

Пространство R^n можно разложить в прямую сумму

$$R^n = \mathrm{Null}(P) \oplus \mathrm{Image}(P).$$

Поэтому, если M и S такая пара подпространств, что

$$R^n = S \oplus M,$$

то можно единственным образом определить проектор P, так, что

$$\mathrm{Null}(P) = S \quad \text{и} \quad \mathrm{Image}(P) = M,$$

т. е.

$$Px \in M, \tag{1.13}$$

$$x - Px \in S, \quad \forall x \in R^n. \tag{1.14}$$

В этом случае говорят, что P проектирует вектор x на M параллельно S.

Подпространство S можно определить через его ортогональное дополнение $L = S^{\perp}$. Тогда условия (1.13)–(1.14) определения оператора проектирования P примут вид

$$Px \in M, \tag{1.15}$$

$$x - Px \perp L, \quad \forall x \in R^n. \tag{1.16}$$

Условия (1.15)–(1.16) определяют оператор проектирования на подпространство M ортогонально пространству L.

Пусть столбцы матрицы $V = [v_1, \ldots, v_m]$ образуют базис пространства M, а столбцы матрицы $W = [w_1, \ldots, w_m]$ — базис пространства S. Два базиса называются биортогональными, если

$$W^T V = I.$$

Поскольку вектор $Px \in M$, то его можно разложить по базису подпространства M,

$$Px = y_1 v_1 + \cdots + y_m v_m = Vy. \tag{1.17}$$

Здесь $y = (y_1, \ldots, y_m)^T$). Тогда условие (1.16) приобретает вид,

$$w_j^T (x - Vy) = 0, \quad j = \overline{1, m},$$

или

$$W^T(x - Vy) = 0.$$

Отсюда

$$y = (W^T V)^{-1} W^T x.$$

Тогда, учитывая (1.17),

$$Px = V(W^T V)^{-1} W^T x,$$

и

$$P = V(W^T V)^{-1} W^T.$$

Таким образом, мы получили матричное представление оператора проектирования. Если же базисы подпространств M и L биортогональны, то

$$P = VW^T.$$

Легко проверить, что в обоих случаях условие (1.12) выполняется.

Рассмотренный нами проектор называется наклонным проектором.

Если $M = L$, т.е., если

$$\mathrm{Null}(P) = \mathrm{Image}(P)^{\perp},$$

проектор P есть ортогональный проектор на подпространство M.

Таким образом, ортогональный проектор определяется условиями:

$$Px \in M, \tag{1.18}$$

$$x - Px \perp M, \quad \forall x \in R^n. \tag{1.19}$$

Пусть столбцы матрицы $V = [v_1, \ldots, v_m]$ образуют базис пространства M. Поскольку вектор $Px \in M$, то его можно разложить по базису подпространства M,

$$Px = Vy. \tag{1.20}$$

Тогда условие (1.19) приобретает вид,

$$V^T(x - Vy) = 0.$$

Отсюда

$$y = (V^T V)^{-1} V^T x.$$

Тогда, учитывая (1.20),

$$Px = V(V^T V)^{-1} V^T x,$$

и

$$P = V(V^T V)^{-1} V^T.$$

Таким образом, мы получили матричное представление ортогонального проектора на подпространство M. Если же базис подпространства M ортонормирован, то

$$P = VV^T.$$

Легко проверить, что в обоих случаях условие (1.12) выполняется, кроме того, матрица P — симметричная.

Если P — ортогональный проектор, то векторы Px и $(I-P)x$ в разложении

$$x = Px + (I-P)x$$

ортогональны, а значит, по теореме Пифагора

$$||x||_2^2 = ||Px||_2^2 + ||(I-P)x||_2^2.$$

Отсюда

$$||Px||_2 \leq ||x||_2.$$

Поэтому

$$||P||_2 = \max_{||x||_2=1} ||Px||_2 \leq 1.$$

Теорема 1.4.1. *Пусть P — ортогональный проектор на подпространство M. Тогда для любого $x \in R^n$*

$$\min_{y \in M} ||x-y||_2 = ||x-Px||_2.$$

Доказательство. Пусть вектор $y \in M$. Тогда вектор $(Px-y)$ тоже принадлежит подпространству M, и по теореме Пифагора

$$||x-y||_2^2 = ||(x-Px)+(Px-y)||_2^2 = ||x-Px||_2^2 + ||Px-y||_2^2.$$

Очевидно, что минимум $||x-y||_2^2$ достигается при $y = Px$. □

Пример 1.4.1. Построим проектор $P : R^3 \to R^3$ на подпространство $M = \text{span}\{v_1, v_2\}$, где $v_1 = [1,2,1]^T$, $v_2 = [0,-1,1]^T$, параллельно подпространству

$$S = \{x \in R^3 \,:\, x = \alpha[1,1,1]^T,\ \alpha \in R\}.$$

Вначале построим базис $\{w_1, w_2\}$ подпространства $L \perp S$. Поскольку должны выполняться равенства:

$$w_1^T[1,1,1]^T = 0,$$
$$w_2^T[1,1,1]^T = 0,$$

то можно положить:

$$w_1 = [1,0,-1]^T, \quad w_2 = [0,1,-1]^T.$$

Значит,

$$P = V(W^TV)^{-1}W^T =$$

$$= \begin{bmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 1 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -2 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} = \begin{bmatrix} -2 & 1 & 1 \\ -3 & 2 & 1 \\ -3 & 1 & 2 \end{bmatrix}.$$

Легко проверить, что $P^2 = P$, т.е., P — проектор. Заметим, что $\forall x \in R^3$

$$Px = \begin{bmatrix} -2 & 1 & 1 \\ -3 & 2 & 1 \\ -3 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2x_1 + x_2 + x_3 \\ -3x_1 + 2x_2 + x_3 \\ -3x_1 + x_2 + 2x_3 \end{bmatrix} =$$

$$= (-2x_1 + x_2 + x_3) \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + (-x_1 + x_3) \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix},$$

т. е., $Px \in M$, и

$$x - Px = \begin{bmatrix} 3x_1 - x_2 - x_3 \\ 3x_1 - x_2 - x_3 \\ 3x_1 - x_2 - x_3 \end{bmatrix} = (3x_1 - x_2 - x_3) \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

т. е., $x - Px \in S$.

1.5. Собственные значения матриц

Собственные значения матрицы $A \in C^{n\times n}$ — это n корней ее *характеристического многочлена*

$$p_n(\lambda) = \det(A - \lambda I). \tag{1.21}$$

Множество всех корней называют *спектром* и обозначают $\lambda(A)$.

Если $\lambda \in \lambda(A)$, то ненулевой вектор $x \in C^n$, удовлетворяющий уравнению

$$Ax = \lambda x,$$

называют *собственным вектором.*

Алгебраическая кратность собственного значения λ — это кратность корня λ характеристического многочлена (1.21)

Геометрической кратностью собственного значения λ называется число линейно независимых собственных векторов, ему соответствующих.

Лемма 1.5.1. (Лемма Гершгорина). *Все собственные значения комплексной матрицы $A \in C^{n\times n}$ находятся в объединении кругов с центрами в точках a_{ii} и радиусами*

$$r_{ii} = \sum_{j=1\,(j\neq i)}^{n} |a_{ij}|, \quad i = \overline{1,n}.$$

Доказательство. Пусть $\{\lambda, x\}$ — собственная пара матрицы A и

$$|x_i| = \max_j |x_j|.$$

Тогда

$$a_{i1}x_1 + \cdots + a_{in}x_n = \lambda x_i,$$

или

$$(\lambda - a_{ii})x_i = a_{i1}x_1 + \cdots + a_{i,i-1}x_{i-1} + a_{i,i+1}x_{i+1} + \cdots + a_{in}x_n.$$

Отсюда

$$|\lambda - a_{ii}| \leq \sum_{j=1\,(j\neq i)}^{n} |a_{ij}|, \quad i = \overline{1,n}.$$

□

Лемма Гершгорина дает, как правило, завышенные оценки для границ спектра.

Собственный вектор определяет одномерное подпространство, инвариантное по отношению к умножению слева на матрицу A. В общем случае подпространство $S \subseteq C^n$, обладающее свойством

$$x \in S \Rightarrow Ax \in S,$$

называют инвариантным подпространством матрицы A.

Лемма 1.5.2. *Пусть $X \in C^{n\times k}$, $B \in C^{k\times k}$ $(k \leq n)$. Если*

$$AX = XB,$$

то подпространство $S = \mathrm{range}(X)$ является инвариантным. Если к тому же матрица X имеет полный столбцовый ранг, то $\lambda(B) \subseteq \lambda(A)$.

Доказательство. Если $x \in S = \mathrm{range}(X)$, то существует такой вектор $y \in C^k$, что $x = Xy$. Тогда

$$Ax = AXy = XBy = Xz \in S = \mathrm{range}(X).$$

Пусть $\lambda \in \lambda(B)$. Тогда существует такой вектор $y \neq 0$, что $By = \lambda y$. Тогда

$$AXy = XBy = \lambda Xy$$

и поскольку матрица X полного столбцового ранга, вектор Xy отличен от нуля. Значит $\{\lambda, Xy\}$ — собственная пара матрицы A. □

Заметим, что, если квадратная матрица X из предыдущей леммы является невырожденной, то $\lambda(B) = \lambda(A)$ и говорят, что матрицы A и $B = X^{-1}AX$ подобны. В этом случае X называют преобразованием подобия. Таким образом, преобразование подобия сохраняет собственные значения матрицы.

Многие алгоритмы вычисления собственных значений используют разбиение исходной задачи на ряд задач меньшего размера. Следующий результат является основой таких преобразований.

Лемма 1.5.3. *Если матрица $T \in C^{n\times n}$ представима в блочном виде*

$$T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}$$

то $\lambda(T) = \lambda(T_{11}) \cup \lambda(T_{22})$.

Доказательство. Пусть

$$Tx = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

где $x_1 \in C^p$ и $x_2 \in C^q$. Если $x_2 \neq 0$, то $T_{22}x_2 = \lambda x_2$ и $\lambda \in \lambda(T_{22})$. Если $x_2 = 0$, то $T_{11}x_1 = \lambda x_1$ и $\lambda \in \lambda(T_{11})$. Следовательно, $\lambda(T) \in \lambda(T_{11}) \cup \lambda(T_{22})$. Но, поскольку множества $\lambda(T)$ и $\lambda(T_{11}) \cup \lambda(T_{22})$ имеют одинаковое число элементов, эти два множества равны. □

Лемма 1.5.4. *Если матрицы $A \in C^{n\times n}$, $X \in C^{n\times p}$, $B \in C^{p\times p}$ $(p < n)$ удовлетворяют условиям*

$$AX = XB, \quad \operatorname{rank}(X) = p, \tag{1.22}$$

то существует унитарная матрица $Q \in C^{n\times n}$ такая, что

$$Q^T AQ = T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix},$$

где $\lambda(T_{11}) = \lambda(A) \cap \lambda(B)$.

Доказательство. Пусть

$$X = Q \begin{bmatrix} R \\ 0 \end{bmatrix}, \quad Q \in C^{n\times n},\ R \in C^{p\times p}$$

есть QR-разложение матрицы X. Подставляя его в (1.22) и преобразуя полученное выражение, имеем

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix} B,$$

где

$$Q^T A Q = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}$$

Из невырожденности матрицы R и уравнений $T_{21}R = 0$, $T_{11}R = RB$ следует, что $T_{21} = 0$ и $\lambda(T_{11}) = \lambda(B)$. Окончательный результат получаем, воспользовавшись леммой 1.5.3, согласно которой $\lambda(A) = \lambda(T) = \lambda(T_{11}) \cup \lambda(T_{22})$. □

Лемма 1.5.4 утверждает, что произвольная квадратная матрица может быть приведена к блочно-треугольной форме, если известно одно из ее инвариантных подпространств.

Теорема 1.5.1. (Разложение Шура). *Для любой матрицы $A \in C^{n\times n}$ существует такая унитарная матрица $Q \in C^{n\times n}$, что*

$$Q^H A Q = T = D + N,$$

где $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ и $N \in C^{n\times n}$ есть строго верхняя треугольная матрица. Более того, матрицу Q можно выбрать так, что собственные значения λ_i будут расположены в любом заданном порядке.

Доказательство. Теорему докажем по индукции. Теорема верна при $n = 1$. Предположим, что теорема верна для всех матриц порядка $n - 1$ или меньше. Если $Ax = \lambda x$ $(x \neq 0)$ то, согласно лемме 1.5.4 (при $B = (\lambda)$), существует такая унитарная матрица U, что

$$U^H A U = \begin{bmatrix} \lambda & w^H \\ 0 & C \end{bmatrix}.$$

По предположению индукции, найдется такая унитарная матрица $\widetilde{U} \in C^{(n-1)\times(n-1)}$, что матрица $\widetilde{U}^H C \widetilde{U}$ — верхняя треугольная. Следовательно, если $Q = U\mathrm{diag}(1, \widetilde{U})$, то матрица $Q^H A Q$ — верхняя треугольная. □

Теорема 1.5.2. (Разложение Жордана). *Если $A \in C^{n\times n}$, то существует такая невырожденная матрица $X \in C^{n\times n}$, что*

$$X^{-1}AX = \mathrm{diag}(J_1, \ldots, J_k),$$

где

$$J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix} \in C^{m_i\times m_i}$$

и $m_1 + \cdots + m_k = n$.

Доказательство. См., например, [20]. □

Блоки J_i называют жордановыми. Число и размеры жордановых блоков, связанных с каждым собственным значением, определяются однозначно. Их расположение вдоль диагонали неоднозначно.

Вещественное разложение Шура представляет собой блочную верхнюю треугольную матрицу с диагональными блоками размера 1×1 и 2×2.

Теорема 1.5.3. (Вещественное разложение Шура). *Для любой вещественной матрицы $A \in R^{n\times n}$ существует такая ортогональная матрица $Q \in R^{n\times n}$, что*

$$Q^T AQ = \begin{bmatrix} R_{11} & R_{11} & \ldots & R_{1m} \\ 0 & R_{22} & \ldots & R_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & R_{mm} \end{bmatrix}, \tag{1.23}$$

где каждый блок R_{ii} — либо 1×1–матрица, либо 2×2–матрица, имеющая комплексно сопряженные собственные значения.

Доказательство. Комплексные собственные значения матрицы A должны входить в $\lambda(A)$ комплексно сопряженными парами, ибо характеристический многочлен $\det(A - \lambda I)$ имеет вещественные коэффициенты. Пусть k — число комплексно сопряженных пар в $\lambda(A)$. Если $k = 0$, то доказательство теоремы можно построить так же, как доказывалась теорема 1.5.1 с использованием леммы 1.5.4, учитывая, что коэффициенты и собственные значения матрицы A вещественны.

Предположим теперь, что $k \geq 1$. Если $\lambda = \gamma + i\mu \in \lambda(A)$ и $\mu \neq 0$, то в R^n существуют такие векторы y и z ($z \neq 0$), что

$$A(y + iz) = (\gamma + i\mu)(y + iz).$$

Приравнивая вещественные и мнимые части, получаем

$$A[y,z] = [y,z] \begin{bmatrix} \gamma & \mu \\ -\mu & \gamma \end{bmatrix}.$$

Значит, векторы y и z образуют двумерное вещественное инвариантное подпространство матрицы A. Из леммы 1.5.4 следует, что существует такая ортогональная матрица $U \in R^{n\times n}$, что

$$U^T A U = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix},$$

где $\lambda(T_{11}) = \{\lambda, \overline{\lambda}\}$. Согласно предположению индукции, существует такая ортогональная матрица $\widetilde{U}$, что матрица $\widetilde{U}^T T_{22} \widetilde{U}$ имеет требуемую структуру. Полагая $Q = U\mathrm{diag}(I_2, \widetilde{U})$, получаем искомое разложение. □

Теорема 1.5.3 показывает, что любая вещественная матрица ортогонально подобна некоторой верхней почти треугольной матрице. Вещественные и мнимые части комплексных собственных значений могут быть легко найдены из диагональных блоков размера 2×2.

Теорема 1.5.4. (Теорема о спектральном разложении). *Для любой вещественной симметричной матрицы $A \in R^{n\times n}$ существует такая ортогональная матрица $Q \in R^{n\times n}$, что*

$$Q^T A Q = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n). \tag{1.24}$$

Доказательство. Доказательство теоремы следует из теоремы 1.5.3. Действительно, поскольку матрица A симметричная, то в разложении (1.23) теоремы 1.5.3 матрица $Q^T A Q$ тоже симметрична, а значит, и диагональна. □

Из теоремы 1.5.4 следует, что вещественная симметричная матрица размера $n \times n$ имеет n вещественных собственных значений и n собственных векторов, из которых можно построить ортонормированный базис пространства R^n.

Умножая (1.24) слева на Q, а справа на Q^T, получаем

$$A = Q\Lambda Q^T = \lambda_1 q_1 q_1^T + \lambda_2 q_2 q_2^T + \cdots + \lambda_n q_n q_n^T = \sum_{i=1}^{n} \lambda_i H_i, \tag{1.25}$$

где $H_i = q_i q_i^T$ — ортогональный проектор на одномерное собственное подпространство собственного вектора q_i. Если собственному значению λ соответствует несколько собственных векторов (т.е. собственное значение кратное), то эти векторы определяются неоднозначно. В то же время,

подпространство, натянутое на эти собственные векторы, определяется однозначно. Поэтому, для кратного собственного значения λ определим ортогональный проектор на его собственное подпространство

$$H_\lambda = \sum_{\lambda_j=\lambda} H_j.$$

Теперь из (1.25) получаем разложение

$$A = \sum_{j=1}^{m} \lambda_j H_{\lambda_j}, \tag{1.26}$$

где все λ_j ($j = \overline{1,m}$) различны. Разложение (1.26) называют спектральным разложением симметричной вещественной матрицы A. Заметим также, что для симметричной матрицы алгебраическая и геометрическая кратности собственного значения совпадают.

Поскольку, согласно предыдущей теореме, все собственные значения вещественной симметричной матрицы вещественны, то из леммы Гершгорина можно вывести такое следствие.

Следствие 1.5.1. *Все собственные значения вещественной симметричной матрицы $A \in R^{n\times n}$ лежат в объединении промежутков*

$$[a_{ii} - r_{ii}, a_{ii} + r_{ii}],\ \textit{где}\ r_{ii} = \sum_{j=1\,(j\neq i)}^{n} |a_{ij}|, \quad i = \overline{1,n}.$$

Пример 1.5.1. Матрица

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

имеет собственные значения $\lambda_1 = 0$, $\lambda_2 = \lambda_3 = 2$, $\lambda_4 = 4$. Собственному значению λ_1 соответствуют собственный вектор $q_1 = 1/2[1, -1, -1, 1]^T$ и ортогональный проектор

$$H_{\lambda_1} = q_1 q_1^T = \frac{1}{4} \begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

Собственному значению λ_4 соответствуют собственный вектор

$$q_4 = 1/2[1, 1, 1, 1]^T$$

и ортогональный проектор

$$H_{\lambda_4} = q_4 q_4^T = \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Собственному значению $\lambda = 2$ (кратности 2) соответствуют два собственных вектора. Их можно определить по–разному. Например,

$$q_2 = 1/2[1, -1, 1, -1]^T \quad \text{и} \quad q_3 = 1/2[1, 1, -1, -1]^T$$

или

$$q_2 = 1/\sqrt{2}[1, 0, 0, -1]^T \quad \text{и} \quad q_3 = 1/\sqrt{2}[0, 1, -1, 0]^T.$$

Легко проверить, что в обоих случаях ортогональный проектор на собственное подпространство собственного значения $\lambda = \lambda_2 = \lambda_3 = 2$ определяется однозначно

$$H_\lambda = q_2 q_2^T + q_3 q_3^T = \frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}.$$

Докажем следующую лемму.

Лемма 1.5.5. *Для симметричной матрицы $A \in R^{n\times n}$ и произвольного вектора $x \neq 0$*

$$\min_\sigma ||Ax - \sigma x||_2 = ||Ax - \rho(x)x||_2,$$

где $\rho(x) = \frac{x^T Ax}{x^T x}$ называется отношением Релея матрицы A.

Доказательство. Действительно, функция

$$f(\sigma) = ||Ax - \sigma x||_2^2 = x^T A^2 x - 2\sigma x^T Ax + \sigma^2 x^T x$$

достигает своего минимума в точке $\sigma = \frac{x^T Ax}{x^T x}$ □

На основании леммы можно сделать следующий вывод. Если $x \neq 0$ является приближением к некоторому собственному вектору симметричной матрицы A, то наилучшее приближение к соответствующему собственному значению, которое можно вычислить зная вектор x, дает отношение Релея $\rho(x)$.

Пусть $\{q_i\}_{i=1}^n$ — ортонормированный базис R^n составленный из собственных векторов матрицы A. Тогда любой вектор $x \in R^n$ можно представить в виде

$$x = \alpha_1 q_1 + \alpha_2 q_2 + \cdots + \alpha_n q_n.$$

Значит,

$$\rho(x) = \frac{x^T Ax}{x^T x} = \frac{\sum\limits_{i=1}^{n} \lambda_i \alpha_i^2}{\sum\limits_{i=1}^{n} \alpha_i^2}.$$

Отсюда сразу следует, что

$$\lambda_{min} \leq \frac{x^T Ax}{x^T x} \leq \lambda_{max} \quad \forall x \in R^n \setminus 0.$$

1.6. Сингулярное разложение матрицы

Теорема 1.6.1. (Сингулярное разложение (SVD-разложение)). *Для любой вещественной матрицы $A \in R^{m\times n}$ ($m \geq n$) существуют такие матрица $U \equiv [u_1, \ldots, u_n] \in R^{m\times n}$ с ортонормированными столбцами и ортогональная матрица $V \equiv [v_1, \ldots, v_n] \in R^{v\times n}$, что*

$$A = U\Sigma V^T, \tag{1.27}$$

где

$$\Sigma \equiv \mathrm{diag}(\sigma_1, \ldots, \sigma_n) \text{ и } \sigma_1 \geq, \ldots, \geq \sigma_n \geq 0.$$

Доказательство. Будем считать, что $A \neq 0$. В противном случае можно положить, что $\Sigma = 0$, и взять в качестве U и V произвольные матрицы с ортонормированными столбцами.

Доказательство проведем с помощью математической индукции. При $n = 1$, поскольку $m \geq n$, можно положить $U = A/||A||_2$, $V = 1$ и $\Sigma = ||A||_2$.

Предположим, что SVD–разложение существует для матриц размера $(m-1)\times(n-1)$, и докажем его существование для матриц размера $m \times n$.

По определению матричной 2-нормы

$$||A||_2 = \max_{||x||_2=1} ||Ax||_2 = ||A\widetilde{x}||_2 = \sigma_1.$$

Положим

$$y = A\widetilde{x}, \quad \widetilde{y} = \frac{y}{||y||_2} = \frac{y}{\sigma_1}.$$

Тогда

$$A\widetilde{x} = \sigma_1 \widetilde{y}, \tag{1.28}$$

где $||\widetilde{x}||_2 = 1$, $||\widetilde{y}||_2 = 1$, $\sigma_1 = ||A||_2$. Поскольку любой ортонормированный набор векторов можно дополнить до ортонормированного базиса всего

пространства, то существуют такие матрицы $\widetilde{U}$ и $\widetilde{V}$, что столбцы матриц $U_0 = [\widetilde{y}, \widetilde{U}] \in R^{m\times n}$ и $V_0 = [\widetilde{x}, \widetilde{V}] \in R^{n\times n}$ ортонормированы. Тогда

$$U_0^T A V_0 = \begin{bmatrix} \widetilde{y}^T \\ \widetilde{U}^T \end{bmatrix} A[\widetilde{x}, \widetilde{V}] = \begin{bmatrix} \widetilde{y}^T A\widetilde{x} & \widetilde{y}^T A\widetilde{V} \\ \widetilde{U}^T A\widetilde{x} & \widetilde{U}^T A\widetilde{V} \end{bmatrix} = \begin{bmatrix} \sigma_1 & w^T \\ 0 & \widetilde{U}^T A\widetilde{V} \end{bmatrix} \equiv A_1,$$

ибо

$$\widetilde{y}^T A\widetilde{x} = \sigma_1 \widetilde{y}^T \widetilde{y} = \sigma_1,$$

и

$$\widetilde{U}^T A\widetilde{x} = \sigma_1 \widetilde{U}^T \widetilde{y} = 0.$$

Так как

$$\left\| A_1 \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|_2^2 \geq (\sigma_1^2 + w^T w)^2,$$

то

$$||A_1||_2^2 = \max_{x\neq 0} \frac{||A_1 x||_2^2}{||x||_2^2} \geq \frac{\left\| A_1 \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|_2^2}{\left\| \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|_2^2} \geq \frac{(\sigma_1^2 + w^T w)^2}{\sigma_1^2 + w^T w} = \sigma_1^2 + w^T w.$$

Отсюда, учитывая тот факт, что умножение на ортогональную матрицу не меняет матричную 2-норму, получаем

$$\sigma_1^2 + w^T w \leq ||A_1||_2^2 = ||A||_2^2 = \sigma_1^2.$$

Значит $w = 0$ и

$$U_0^T A V_0 = \begin{bmatrix} \sigma_1 & 0^T \\ 0 & \widetilde{U}^T A\widetilde{V} \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0^T \\ 0 & \widetilde{A} \end{bmatrix}.$$

По предположению индукции для матрицы $\widetilde{A}$ существует SVD–разложение. Пусть $\widetilde{A} = U_1 \Sigma_1 V_1^T$, где U_1, V_1 и Σ_1 — матрицы размера $(m-1)\times(n-1)$, $(n-1)\times(n-1)$ и $(n-1)\times(n-1)$ соответственно. Тогда

$$U_0^T A V_0 = \begin{bmatrix} \sigma_1 & 0^T \\ 0 & U_1 \Sigma_1 V_1^T \end{bmatrix} = \begin{bmatrix} 1 & 0^T \\ 0 & U_1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0^T \\ 0 & \Sigma_1 \end{bmatrix} \begin{bmatrix} 1 & 0^T \\ 0 & V_1^T \end{bmatrix},$$

и

$$A = \left(U_0 \begin{bmatrix} 1 & 0^T \\ 0 & U_1 \end{bmatrix} \right) \begin{bmatrix} \sigma_1 & 0^T \\ 0 & \Sigma_1 \end{bmatrix} \left(V_0 \begin{bmatrix} 1 & 0^T \\ 0 & V_1 \end{bmatrix} \right)^T.$$

□

Числа σ_i называются сингулярными числами матрицы A, а векторы u_i и v_i — сингулярными векторами, соответственно, левыми и правыми.

Если известно сингулярное разложение матрицы A и

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0,$$

то ранг, ядро и образ матрицы A определяются следующим образом:

$$\begin{aligned} &\mathrm{rank}(A) = r, \\ &\mathrm{Null}(A) = \mathrm{span}\{v_{r+1}, \ldots, v_n\}, \\ &\mathrm{Range}(A) = \mathrm{span}\{u_1, \ldots, u_r\}. \end{aligned}$$

Кроме того, матрица A представима в виде разложения

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T,$$

и, поскольку ортогональные преобразования не меняют норму Фробениуса и матричную 2-норму, то

$$\begin{aligned} &||A||_F^2 = \sigma_1^2 + \cdots + \sigma_r^2, \\ &||A||_2 = \sigma_1. \end{aligned}$$

Число обусловленности прямоугольной матрицы $A \in R^{m \times n}$ $(m > n)$ в смысле матричной 2-нормы определяется как

$$\mathrm{cond}_2(A) \equiv \frac{\sigma_{max}}{\sigma_{min}}.$$

Наконец, заметим следующее. Пусть известно SVD-разложение матрицы $A = U\Sigma V^T$. Тогда

$$A^T A = (U\Sigma V^T)^T U\Sigma V^T = V\Sigma^2 V^T,$$

т. е.,

$$A^T AV = V\Sigma^2.$$

А это значит, что v_i — собственные векторы матрицы $A^T A$, соответствующие собственным значениям σ_i^2 $(i = \overline{1, n})$. Аналогично,

$$AA^T = U\Sigma V^T (U\Sigma V^T)^T = U\Sigma^2 U^T,$$

т. е.,

$$AA^T U = U\Sigma^2.$$

А это значит, что u_i — собственные векторы матрицы AA^T, соответствующие собственным значениям σ_i^2 $(i = \overline{1, n})$.

Пример 1.6.1. Построим SVD–разложение матрицы

$$A = \begin{bmatrix} 2 & 0 \\ 1 & 2 \\ 0 & -1 \end{bmatrix}.$$

Вычислим

$$A^T A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & -1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 1 & 2 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 5 & 2 \\ 2 & 5 \end{bmatrix}.$$

Вычисляя собственные значения и собственные векторы матрицы $A^T A$ находим, что

$$\begin{aligned} \sigma_1^2 = 7, \quad & v_1 = \tfrac{1}{\sqrt{2}}[1,1]^T, \\ \sigma_2^2 = 3, \quad & v_2 = \tfrac{1}{\sqrt{2}}[1,-1]^T. \end{aligned}$$

Далее,

$$AA^T = \begin{bmatrix} 2 & 0 \\ 1 & 2 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & -1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 5 & -2 \\ 0 & -2 & 1 \end{bmatrix}.$$

Вычисляя собственные векторы матрицы AA^T, соответствующие собственным значениям $\sigma_1^2 = 7$ и $\sigma_2^2 = 3$, находим

$$u_1 = \tfrac{1}{\sqrt{14}}[2,3,-1]^T, \quad u_2 = \tfrac{1}{\sqrt{6}}[2,-1,1]^T.$$

Таким образом, SVD–разложение матрицы A имеет вид,

$$A = \begin{bmatrix} \frac{2}{\sqrt{14}} & \frac{2}{\sqrt{6}} \\ \frac{3}{\sqrt{14}} & \frac{-1}{\sqrt{6}} \\ \frac{-1}{\sqrt{14}} & \frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} \sqrt{7} & 0 \\ 0 & \sqrt{3} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}^T.$$

А также

$$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T = \sqrt{7} \begin{bmatrix} \frac{2}{\sqrt{14}} \\ \frac{3}{\sqrt{14}} \\ \frac{-1}{\sqrt{14}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}^T + \sqrt{3} \begin{bmatrix} \frac{2}{\sqrt{6}} \\ \frac{-1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}^T.$$

Приведем геометрическое истолкование сингулярного разложения матрицы. Пусть известно SVD-разложение (1.28) матрицы A и

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0.$$

Рассмотрим отображение

$$y = Ax, \quad x \in R^n, \ y \in R^m. \tag{1.29}$$

В пространствах R^m и R^n введем преобразования координат:

$$y' = [U, \widetilde{U}]^T y, \quad x' = V^T x,$$

где $\widetilde{U}$ — такая матрица с ортонормированными столбцами, что столбцы матрицы $[U, \widetilde{U}]$ образуют ортонормированный базис пространства R^m. Тогда

$$y = [U, \widetilde{U}]y', \quad x = Vx',$$

и из (1.29) получаем

$$[U, \widetilde{U}]y' = AVx',$$

или

$$y' = \begin{bmatrix} U^T \\ \widetilde{U}^T \end{bmatrix} AVx' = \begin{bmatrix} U^T \\ \widetilde{U}^T \end{bmatrix} U\Sigma V^T V x' = \begin{bmatrix} \Sigma x' \\ 0 \end{bmatrix}.$$

Значит, отображение (1.29) в новых системах координат определяется так,

$$\begin{aligned} &y'_1 = \sigma_1 x'_1, \\ &\vdots \\ &y'_r = \sigma_r x'_r, \\ &y'_{r+1} = \cdots = y'_n = 0. \end{aligned} \tag{1.30}$$

Таким образом, согласно формулам (1.30), отображение (1.29) единичный шар, который в системе координат $x'_1, \ldots, x'_n$ пространства R^n, определяется неравенством

$$(x'_1)^2 + \cdots + (x'_n)^2 \leq 1,$$

преобразует в область, ограниченную r-мерным эллипсоидом с полуосями $\sigma_1, \ldots, \sigma_r$ в координатной системе $y'_1, \ldots y'_m$ пространства R^m. Действительно

$$\left(\frac{y'_1}{\sigma_1}\right)^2 + \cdots + \left(\frac{y'_r}{\sigma_r}\right)^2 \leq (x'_1)^2 + \cdots + (x'_n)^2 \leq 1.$$

Пример 1.6.2. Пусть $m = n = r = 2$. Тогда круг

$$(x'_1)^2 + (x'_2)^2 \leq 1,$$

преобразуется в область, ограниченную в эллипсом

$$\left(\frac{y'_1}{\sigma_1}\right)^2 + \left(\frac{y'_2}{\sigma_2}\right)^2 = (x'_1)^2 + (x'_2)^2 \leq 1.$$

В этом случае число обусловленности матрицы A

$$\mathrm{cond}_2(A) = \frac{\sigma_1}{\sigma_2}.$$

Его геометрический смысл — мера вытянутости эллипса.

Пусть $m = n = 2$, $r = 1$. В этом случае матрица A вырождена, $\sigma_2 = 0$ и единичный круг отображается в отрезок:

$$\left(\frac{y_1'}{\sigma_1}\right)^2 \leq (x_1')^2 + (x_2')^2 \leq 1.$$

При этом

$$\mathrm{cond}_2(A) = \frac{\sigma_1}{\sigma_2} = \infty.$$

1.7. Положительно определенные матрицы

Матрица $A \in R^{n\times n}$ называется *положительно определенной*, если

$$x^T Ax > 0 \quad \forall x \in R^{n\times n},\ x \neq 0.$$

Матрица $A \in R^{n\times n}$ называется *положительно полуопределенной*, если

$$x^T Ax \geq 0 \quad \forall x \in R^{n\times n}.$$

Теорема 1.7.1. *Для того, чтобы матрица $A \in R^{n\times n}$ была положительно определенной, необходимо и достаточно, чтобы положительно определенной была ее симметрическая часть.*

Доказательство. Матрицу A можно представить в виде,

$$A = \frac{A + A^T}{2} + \frac{A - A^T}{2} = A_c + A_k.$$

Покажем, что

$$x^T A_k x = 0 \quad \forall x \in R^n.$$

Действительно

$$x^T A_k x = x^T \left(\frac{A - A^T}{2}\right) x = \frac{1}{2} x^T Ax - \frac{1}{2} x^T A^T x =$$

$$= \frac{1}{2} x^T Ax - \frac{1}{2} (Ax)^T x = \frac{1}{2} x^T Ax - \frac{1}{2} x^T Ax = 0.$$

Следовательно

$$x^T Ax = x^T (A_c + A_k) x = x^T A_c x.$$

Что и требовалось доказать. □

Пример 1.7.1. Покажем, что матрица

$$A = \begin{bmatrix} 2 & -1 & 2 & \dots & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & 0 & \dots & -1 & 2 \end{bmatrix}$$

положительно определена. Действительно

$$x^T Ax = 2x_1^2 - x_1x_2 - x_1x_2 + 2x_2^2 - x_3x_2 + \cdots - 2x_nx_{n-1} + 2x_n^2 =$$

$$= x_1^2 + (x_1 - x_2)^2 + \cdots + (x_{n-1} - x_n)^2 + x_n^2 > 0 \quad \forall x \in R^n.$$

Кроме того, из $x^T Ax = 0$ следует $x = 0$.

Перечислим некоторые свойства положительно определенных матриц.

Свойство 1.7.1. *Положительно определенная матрица является невырожденной.*

Доказательство. Предположим, что положительно определенная матрица A вырождена. Тогда однородная система

$$Ax = 0$$

имеет нетривиальное решение $\widetilde{x} \neq 0$ и

$$\widetilde{x}^T A\widetilde{x} = 0.$$

Получили противоречие. □

Свойство 1.7.2. *Любая главная подматрица положительно определенной матрицы положительно определена.*

Свойство 1.7.3. *Все диагональные элементы положительно определенной матрицы положительны.*

Доказательство. Действительно, из определения положительно определенной матрицы следует, что

$$e_i^T Ae_i = a_{ii} > 0 \quad i = \overline{1, n}.$$

□

Свойство 1.7.4. *Максимальный по модулю элемент симметричной положительно определенной матрицы находится на главной диагонали.*

Доказательство. Предположим противное. Пусть максимальный по модулю элемент находится в позиции (i, j) $(i \neq j)$, т.е.

$$|a_{ij}| = \max_{1 \leq k,l \leq n} |a_{kl}|.$$

Предположим, что $a_{ij} < 0$. Положим

$$x = [0, \ldots, 0, \overset{i}{1}, 0, \ldots, 0, \overset{j}{1}, 0, \ldots, 0]^T,$$

тогда

$$x^T A x = a_{ii} + a_{jj} + 2a_{ij} < 0.$$

Получили противоречие.

Если же $a_{ij} > 0$, то положив

$$x = [0, \ldots, 0, \overset{i}{1}, 0, \ldots, 0, \overset{j}{-1}, 0, \ldots, 0]^T,$$

снова приходим к противоречию, ибо

$$x^T A x = a_{ii} + a_{jj} - 2a_{ij} < 0.$$

□

Свойство 1.7.5. *Матрица A положительно определена тогда и только тогда, когда положительно определена матрица BAB^T, где* $\det B \neq 0$.

1.8. Вопросы и задания

1. Показать, что для любой матрицы $A \in R^{m \times n}$ верно $\dim(\text{null}(A) + \text{rank}(A) = n$.
2. Показать, что $\text{rank}(A) = \text{rank}(A^T)$.
3. Доказать, что матричная 1–норма согласована с векторной 1–нормой.
4. Доказать, что матричная 1–норма порождается векторной 1–нормой.
5. Показать, что функционал $F : R^n \to R$, определяемый следующим образом
$$F(x) = ||Ax|| \quad \forall A \in R^{n \times n},$$
непрерывен.

6. Для заданной матрицы A вычислить $||A||_\infty$, $||A||_1$, $||A||_F$, $||A||_C$.

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

7. Проверить выполнение неравенства $||AB|| \leq ||A||\,||B||$ для норм $||\cdot||_\infty$, $||\cdot||_1$, $||\cdot||_F$, $||\cdot||_C$ и заданных матриц A и B.

$$A = \begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix}$$

8. Показать, что модуль собственного значения матрицы не больше любой ее нормы.

9. Показать, что умножение на ортогональную матрицу не меняет 2–норму вектора, спектральную норму и норму Фробениуса матрицы.

10. Построить SVD–разложение матрицы

$$A^T = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}^T.$$

11. Найти SVD–разложение заданных матриц: .

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}.$$

Построить геометрические интерпретации найденных разложений.

12. Доказать, что матрица $A \in \mathbb{R}^{n\times n}$ положительно определена тогда и только тогда, когда для любой невырожденной матрицы $X \in \mathbb{R}^{n\times n}$ матрица X^TAX положительно определена.

13. Доказать свойство 1.7.2.

14. Доказать, что симметричная матрица $A \in \mathbb{R}^{n\times n}$ положительно определена тогда и только тогда, когда все ее собственные значения положительны.

15. Проверить, является ли матрица A положительно определенной

$$A = \begin{bmatrix} 2 & -1 & -1 \\ 2 & 3 & 0 \\ 0 & -1 & 2 \end{bmatrix}.$$

2. Основы вычислений в арифметике с плавающей точкой

В курсе математического анализа было введено определение вещественного числа. Все утверждения предыдущей главы из линейной алгебры и теории матриц использовали вещественные числа. В дальнейшем мы рассмотрим методы решения некоторых основных задач линейной алгебры с помощью компьютеров. Поскольку в общем случае вещественные числа представимы бесконечными десятичными дробями, то для их представления на компьютере необходимо ввести некоторую аппроксимацию.

2.1. Система чисел с плавающей точкой

В современных компьютерах для представления вещественных чисел используется *система чисел с плавающей точкой*. Множество F чисел с плавающей точкой характеризуется четырьмя целыми числами: β, t, L и U. Число x множества F чисел с плавающей точкой представимо следующим образом

$$x = \pm \left(\frac{\alpha_1}{\beta} + \frac{\alpha_2}{\beta^2} + \cdots + \frac{\alpha_t}{\beta^t} \right) \beta^l = \pm 0.\underbrace{\alpha_1 \alpha_2 \ldots \alpha_t}_{\text{мантисса}} \times \beta^l, \quad L \le l \le U,$$

где целые числа α_i удовлетворяют неравенствам

$$0 \le \alpha_i \le \beta - 1, \quad i = \overline{1,t}.$$

Здесь β — основание системы счисления, t — количество разрядов (точность), числа L и U определяют интервал в котором изменяется целочисленный показатель l. Если для каждого отличного от нуля $x \in F$ $\alpha_1 \neq 0$, то система чисел с плавающей точкой называется *нормализованной*. Легко показать, что нормализованная система F содержит

$$2(\beta - 1)\beta^{t-1}(U - L + 1) + 1$$

число. Эти числа расположены неравномерно. Однако между β^{l-1} и β^l для любого $l \in [L, U]$ числа расположены равномерно с интервалом β^{l-t}.

Число x нормализованной системы F заключено в следующих границах

$$m \le |x| \le M,$$

где

$$m = \left(\frac{1}{\beta} + \frac{0}{\beta^2} + \cdots + \frac{0}{\beta^t}\right)\beta^L = \beta^{L-1},$$

$$M = (\beta - 1)\left(\frac{1}{\beta} + \frac{1}{\beta^2} + \cdots + \frac{1}{\beta^t}\right)\beta^U = (1 - \beta^{-t})\beta^U.$$

Пример 2.1.1. Нормализованная система F чисел с плавающей точкой, определяемая параметрами $\beta = 2$, $t = 3$, $L = -1$, $U = 2$, содержит 33 числа. В следующей таблице представлены все числа системы F кроме нуля.

Мантисса	$l = -1$	$l = 0$	$l = 1$	$l = 2$
$\pm 0.100 = \pm\frac{1}{2}$	$\pm\frac{1}{4}$	$\pm\frac{1}{2}$	± 1	± 2
$\pm 0.101 = \pm\frac{5}{8}$	$\pm\frac{5}{16}$	$\pm\frac{5}{8}$	$\pm\frac{5}{4}$	$\pm\frac{5}{2}$
$\pm 0.110 = \pm\frac{3}{4}$	$\pm\frac{3}{8}$	$\pm\frac{3}{4}$	$\pm\frac{3}{2}$	± 3
$\pm 0.111 = \pm\frac{7}{8}$	$\pm\frac{7}{16}$	$\pm\frac{7}{8}$	$\pm\frac{7}{4}$	$\pm\frac{7}{2}$

При этом

$$m = \beta^{L-1} = 2^{-2} = \frac{1}{4},$$

$$M = (1 - \beta^{-t})\beta^U = (1 - 2^{-3})2^2 = \frac{7}{2}.$$

2.2. Приближение вещественных чисел

При представлении вещественного числа x числом $\widetilde{x}$ из системы F чисел с плавающей точкой могут возникнуть следующие ситуации.

1. Если $|x| > M$, то вещественное число x не может быть представлено числом из системы F. Говорят, что в этом случае произошло *переполнение порядка* (overflow).

2. Если $|x| < m$, то говорят, что произошло *исчезновение порядка* (underflow). В этом случае число x обычно заменяется нулем.

3. Если $x \in R$ совпадает с одним из чисел системы F, то в этом случае вещественное число x представимо в F точно.

4. Если же $m \leq |x| \leq M$ и x не совпадает ни с одним числом системы F, то в этом случае число x приближается некоторым числом системы F.

Последний случай рассмотрим более подробно. В дальнейшем будем обозначать через $fl(x)$ представление вещественного числа x в системе F чисел с плавающей точкой. В современных компьютерах при представлении вещественных чисел наиболее часто используется округление. А именно, число $x \in R$ представляется ближайшим к нему числом из множества F. В случае неопределенности из двух ближайших чисел выбирается, например, то, у которого больше абсолютная величина.

Абсолютной погрешностью представления вещественного числа x называется величина

$$\Delta x = fl(x) - x.$$

Относительной погрешностью представления вещественного числа x называют величину

$$\delta x = \frac{|\Delta x|}{|x|}.$$

Если $x \in R$ находится между двумя числами f_1 и f_2 системы F и

$$\beta^{l-1} \leq |f_1| < |f_2| \leq \beta^l,$$

для некоторого $l \in [L, U]$, то

$$|\Delta x| \leq \frac{1}{2}\beta^{l-t},$$

поскольку числа системы F на промежутке $[\beta^{l-1}, \beta^l]$ находятся на одинаковом расстоянии β^{l-t} друг от друга. В этом случае для относительной погрешности представления вещественного числа получаем оценку

$$\delta x = \frac{|\Delta x|}{|x|} \leq \frac{\frac{1}{2}\beta^{l-t}}{\beta^{l-1}} = \frac{1}{2}\beta^{1-t}. \tag{2.1}$$

Теорема 2.2.1. *Если вещественное число x удовлетворяет неравенствам*

$$m \leq |x| \leq M, \tag{2.2}$$

то для его представления $fl(x)$ в системе F чисел с плавающей точкой имеет место равенство

$$fl(x) = x(1 + \delta), \tag{2.3}$$

где

$$|\delta| \leq \frac{1}{2}\beta^{1-t}.$$

Доказательство. Пусть вещественное число x удовлетворяет неравенствам (2.2). Обозначим

$$\delta = \frac{fl(x) - x}{x}.$$

Тогда (2.3) следует из (2.1). □

Из теоремы 2.2.1 следует, что относительная погрешность представления вещественного числа не зависит от самого числа, а зависит только от основания β и количества разрядов t системы F. Исключение составляют случаи, когда число x не принадлежит области чисел представимых в системе F.

Величина $\frac{1}{2}\beta^{1-t}$ называется *относительной погрешностью арифметики с плавающей точкой*. С этой величиной совпадает *машинный эпсилон*, который определяется как наименьшее число $\varepsilon \in F$, для которого выполняется неравенство

$$fl(1 + \varepsilon) > 1.$$

Действительно, поскольку

$$1 = \left(\frac{1}{\beta} + \frac{0}{\beta^2} + \cdots + \frac{0}{\beta^t}\right)\beta,$$

то

$$\varepsilon = \frac{1}{2}\frac{1}{\beta^t}\beta.$$

Пример 2.2.1. Представим вещественное число $x = 3.3$ в системе F из примера 2.1.1. Поскольку ближайшее к x в этой системе число $\frac{7}{2}$, то

$$fl(3.3) = 0.111 \times 2^2.$$

Абсолютная погрешность представления числа

$$\Delta x = fl(x) - x = 3.5 - 3.3 = 0.2.$$

Вычисленная с точностью до четырех значащих цифр относительная погрешность представления числа

$$\delta x = \frac{|\Delta x|}{|x|} = \frac{0.2}{3.3} = 0.06061.$$

Относительная погрешность арифметики с плавающей точкой равна

$$\frac{1}{2}\beta^{1-t} = \frac{1}{2}2^{1-2} = 0.25.$$

2.3. Арифметические операции в системе чисел с плавающей точкой

Через $*$ будем обозначать одну из арифметических операций $+$, $-$, $/$, $\times$. Точный результат выполнения арифметической операции $*$ над числами x и y системы F будем обозначать $x * y$. Если

$$m \leq |x * y| \leq M,$$

то результат арифметической операции $*$ для чисел x и y в системе F чисел с плавающей определяется так

$$x \odot y = fl(x * y).$$

Из теоремы 2.2.1 следует, что для арифметических операций в системе F чисел с плавающей точкой выполняется равенство

$$x \odot y = (x * y)(1 + \delta), \tag{2.4}$$

где $\delta \leq \frac{1}{2}\beta^{1-t}$.

Систему F чисел с плавающей точкой с определенными в ней арифметическими операциями называют *машинной арифметикой*.

В машинной арифметике могут не выполняются ассоциативные законы умножения и сложения, а также дистрибутивный закон умножения относительно сложения. Коммутативные законы сложения и умножения выполняются.

Пример 2.3.1. Пусть машинная арифметика определяется параметрами $\beta = 10$, $t = 3$, $L = -10$, $U = 10$ и $x = 10^{-3}$, $y = 1$, $z = -1$. Тогда в точной арифметике

$$x + y + z = 10^{-3}.$$

В машинной арифметике

$$(x \oplus y) \oplus z = fl(fl(x + y) + z) = fl(fl(1.001) - 1.00) = 0.00,$$

с другой стороны

$$x \oplus (y \oplus z) = fl(fl(x + y) + z) = fl(0.001 + fl(1.00 - 1.00)) = 0.001.$$

Таким образом, ассоциативный закон для сложения нарушается.

Нарушение законов арифметики приводит к тому, что алгоритм численного решения математической задачи, полученный в точной арифметике, в машинной арифметике можно реализовать по-разному, изменяя

порядок арифметических операций. Более того, может оказаться, что машинные реализации алгоритма не дают ожидаемого результата.

Отметим еще одну особенность машинной арифметики, которая состоит в том, что при вычитании близких чисел происходит *катастрофическая потеря значащих цифр.*

Рассмотрим следующий пример. Пусть машинная арифметика определяется параметрами $\beta = 10$, $t = 5$, $L = -10$, $U = 10$. Рассмотрим фрагмент программы, реализующей некоторый численный алгоритм, на компьютере с этой арифметикой:

```
. . .
x = 0.1234623456;    y = 0.1234512345;
z = x − y;
. . .
```

Вычисления по программе приведут к таким результатам:

$x = fl(0.1234623456) = 0.12346,$
$y = fl(0.1234512345) = 0.12345,$
$z = fl(0.12346 - 0.12345) = 0.10000 \cdot 10^{-4}.$

Таким образом, z будет вычислено с одной значащей цифрой (точный результат $z = 0.111111 \cdot 10^{-4}$) и при вычитании близких чисел потеряно четыре значащих цифры. Поэтому, в дальнейшем вычисления в программе будут производится с одной значащей цифрой. Отметим, что если разрядность t машинной арифметики увеличить до 10, то при вычитании удалось бы сохранить шесть значащих цифр.

2.4. IEEE-стандарт арифметики с плавающей точкой

В процессе развития вычислительной техники различные компьютеры использовали различные системы чисел с плавающей точкой. В настоящее время почти общепринятым является IEEE-стандарт двоичной арифметики. Стандарт реализован рабочими станциями Sun, DEC, HP, IBM и всеми персональными компьютерами с процессорами Intel. IEEE-арифметика предусматривает два типа чисел с плавающей точкой: числа обычной точности (в нотации языка C – float) и числа двойной точности (в нотации языка C – double).

Для представления числа с одинарной точностью используется 32 бита, для представления числа с двойной точностью — 64 бита. Числа обоих типов — это числа с плавающей точкой с основанием системы счисления $\beta = 2$.

Число с одинарной точностью представляется в виде

$$(-1)^s(1+f)2^{l-127},$$

где s — 1-битовый знак числа, l — 8-битовый показатель и $f < 1$ — 23-битовая мантисса. Относительная погрешность арифметики чисел с одинарной точностью равна $2^{-24} \approx 6 \cdot 10^{-8}$. Область положительных нормализованных чисел простирается от $m = 2^{-126} \approx 10^{-38}$ до $M = 2^{128} \cdot (2 - 2^{-23}) \approx 10^{38}$.

Число с двойной точностью имеет представление

$$(-1)^s(1+f)2^{l-1023},$$

где s — 1-битовый знак числа, l — 11-битовый показатель и $f < 1$ — 52-битовая мантисса. Область положительных нормализованных чисел простирается от $m = 2^{-1022} \approx 10^{-308}$ до $M = 2^{1022} \cdot (2 - 2^{-52}) \approx 10^{308}$.

IEEE-арифметика включает в себя *субнормальные числа*, т.е. ненормализованные числа с плавающей точкой, имеющие наименьший из возможных показателей.

Пример 2.4.1. Субнормальные числа для нормализованной системы F чисел с плавающей точкой, определяемой параметрами $\beta = 2$, $t = 3$, $L = -1$, $U = 2$, имеют вид:

$$\pm 0.001 \cdot 2^{-1},\ \pm 0.010 \cdot 2^{-1},\ \pm 0.011 \cdot 2^{-1}.$$

Наличие субнормальных чисел в машинной арифметике обеспечивает, что $fl(x - y) = 0$ тогда и только тогда, когда $x = y$. Если бы это свойство отсутствовало, то формулу (2.4) пришлось бы изменить так, чтобы она учитывала появление машинных нулей при вычитании.

IEEE-арифметика обеспечивает, что

$$fl(\sqrt{x}) = \sqrt{x}(1+\delta),$$

где $|\delta| \leq \varepsilon = \frac{1}{2}\beta^{1-t}$ и ε — относительная погрешность машинной арифметики (машинный эпсилон).

IEEE-арифметика предусматривает также символы $\pm\infty$ и NaN (Not a Number — не число). Символы $\pm\infty$ генерируются при переполнении и в дальнейшем ведут себя по правилам:

- $x/\pm\infty = 0$ для всякого числа x с плавающей точкой;
- $x/0 = \pm\infty$ для всякого числа x с плавающей точкой;
- $+\infty + \infty = +\infty$, и т.д.

Любая операция, результат которой, конечный или бесконечный не определен корректно, генерирует символ NaN. Например, $\infty - \infty$, $\frac{\infty}{\infty}$, $\frac{0}{0}$, $\sqrt{-1}$, $NaN \odot x$, и т.п.

В каждом из следующих случаев:

- арифметическая операция некорректна и порождает NaN;
- произошло переполнение;
- встретилось деление на ноль, в следствие чего генерируется $\pm\infty$;
- получен машинный нуль,

генерируется исключение (exseption), которое может быть обработано программой пользователя (по умолчанию обработчик исключения завершает выполнение программы).

Эти особенности позволяют разрабатывать более надежные программы численных алгоритмов.

2.5. Обусловленность задач

Многие математические задачи состоят в том, что по исходным данным u ищется решение z. При этом считается, что u и z связаны функциональной зависимостью $z = f(u)$. Задача называется *корректной*, если выполнены следующие условия (условия корректности):

1. задача имеет решение при любых допустимых исходных данных (условие существования решения);
2. каждым исходным данным u соответствует только одно решение (единственность решения задачи);
3. решение устойчиво.

Смысл первого условия заключается в том, что среди исходных данных нет противоречащих друг другу условий, что исключало бы возможность решения задачи. Второе условие означает, что исходных данных достаточно для однозначной разрешимости задачи. Третье условие заключается в следующем. Если $u1$ и $u2$ — два различных набора исходных данных, мера уклонения которых друг от друга достаточно мала, то мера уклонения решений $z1 = f(u1)$ и $z2 = f(u2)$ меньше любой наперёд заданной меры точности. Задачи, не удовлетворяющие хотя бы одному условию корректности, называются *некорректными*.

Математические задачи корректные в точной арифметике в машинной арифметике могут стать некорректными или близкими к некорректным (*плохо обусловленными*).

Числом обусловленности задачи называют величину, которая является мерой близости задачи к некорректной. Для различных задач число обусловленности определяется по–разному.

Пример 2.5.1. [9] Пусть $f(x)$ — веществennозначная дифференцируемая функция вещественной переменной x. Требуется вычислить $f(x)$, при этом значение x точно не известно. Предположим, что вместо x заданы значение $x+\delta x$ и граница для величины δx. В этом случае, можно сделать следующее. Вначале вычисляем $f(x+\delta x)$, затем пытаемся оценить погрешность $|f(x)-f(x+\delta x)|$. Воспользовавшись формулой Тейлора, имеем

$$|f(x)-f(x+\delta x)| \approx |f'(x)||\delta x|,$$

или

$$\frac{|f(x)-f(x+\delta x)|}{|f(x)|} \approx \frac{|f'(x)|\,|x|}{|f(x)|}\frac{|\delta x|}{|x|}. \tag{2.5}$$

Числом обусловленности этой задачи является величина $\frac{|f'(x)|\,|x|}{|f(x)|}$. Если эта величина очень большая, то для данной задачи нарушается, например, условие 3) корректности задачи, ибо небольшие изменения в исходных данных приводят к большим изменениям в решениях задачи, т.е. решение задачи неустойчиво.

Следующий пример плохо обусловленной задачи принадлежит Уилкинсону [23].

Пример 2.5.2. Пусть

$$p(x) = (x-1)(x-2)\dots(x-19)(x-20) = x^{20} - 210x^{19} + \dots . \tag{2.6}$$

Корни многочлена (2.6) $x_i = i$ $(i = \overline{1,20})$ хорошо отделены. Уилкинсон вычислял корни этого многочлена следующим образом. В программе коэффициенты многочлена округлялись до 30-значного двоичного числа. Коэффициент при x^{19} вводился с ошибкой в последнем разряде, а именно, вместо -210 было введено число $-210+2^{-23}$. Чтобы определить, какое воздействие имеет это незначительное изменение коэффициента на корни многочлена, вычисления производились очень точно с использованием машинной арифметики с параметрами $\beta = 2$, $t = 90$. Были получены следующие результаты, правильно округленные до указанно-

го числа цифр:

$$
\begin{aligned}
&x_1 = 1.000000000, \quad x_{10} = 10.095266145 \pm 0.643500904i,\\
&x_2 = 2.000000000, \quad x_{12} = 11.793633881 \pm 1.652329728i,\\
&x_3 = 3.000000000, \quad x_{14} = 13.992358137 \pm 2.518830070i,\\
&x_4 = 4.000000000, \quad x_{17} = 16.730737466 \pm 2.812624894i,\\
&x_5 = 4.999999928, \quad x_{19} = 19.502439400 \pm 1.940330347i,\\
&x_6 = 6.000006944, \quad x_{20} = 20.846908101.\\
&x_7 = 6.999697234,\\
&x_8 = 8.007267603,\\
&x_9 = 8.917250249,
\end{aligned}
$$

Заметим, что малое изменение в коэффициенте при x^{19} имело следствием то, что девять корней стали комплексными и некоторые из них отодвинулись от действительной оси более чем на 2.51.

Полученные результаты можно объяснить следующим образом. Полином (2.6) запишем в виде

$$p(x) = x^{20} - \alpha x^{19} + \dots$$

и найдем частную производную по α для каждого корня полинома (2.6). Для этого продифференцируем уравнение

$$p(x, \alpha) = 0$$

по α:

$$\frac{\partial p(x,\alpha)}{\partial x}\frac{\partial x}{\partial \alpha} + \frac{\partial p(x,\alpha)}{\partial \alpha} = 0.$$

Отсюда

$$\frac{\partial x}{\partial \alpha} = -\frac{\partial p/\partial \alpha}{\partial p/\partial x} = \frac{x^{19}}{\sum\limits_{i=1}^{20}\prod\limits_{j=1,\, j\neq i}^{20}(x-j)}.$$

Вычисляя это выражение для каждого корня, получаем

$$\left.\frac{\partial x}{\partial \alpha}\right|_{x=i} = \frac{i^{19}}{\sum\limits_{i=1}^{20}\prod\limits_{j=1,\, j\neq i}^{20}(i-j)}, \quad i = \overline{1,20}.$$

Таким образом, чувствительность корней к изменению коэффициента α растет с увеличением номера корня.

2.6. Устойчивость алгоритмов

Обозначим через $\mathrm{alg}(x)$ алгоритм вычисления в машинной арифметике значения функции $f(x)$.

Алгоритм называется *обратно устойчивым* (*устойчивым*), если для всякого x найдется такое "малое" δx, что

$$\mathrm{alg}(x) = f(x + \delta x),$$

т.е., если алгоритм дает точное решение $f(x+\delta x)$ для слабо возмущенной задачи $(x + \delta x)$. Величина δx называется *обратной ошибкой*. Используя (2.5), для погрешности алгоритма получаем формулу:

$$\frac{|f(x) - \mathrm{alg}(x)|}{|f(x)|} = \frac{|f(x) - f(x + \delta x)|}{|f(x)|} \approx \frac{|f'(x)|\,|x|}{|f(x)|}\frac{|\delta x|}{|x|}.$$

Отсюда следует, что если алгоритм устойчив и число обусловленности задачи мало, то небольшие возмущения исходных данных приведут к небольшим возмущениям в решении построенном алгоритмом.

Алгоритм может быть неустойчивым по следующим причинам. Из–за потери значащих цифр при вычитании близких чисел. В следствие умножения на большое число погрешностей, которые имеют два источника. Погрешности могут содержать исходные данные задачи, а также погрешности появляются в следствие округлений при выполнении арифметических операций в машинной арифметике.

Пример 2.6.1. Рассмотрим пример алгоритма неустойчивого по причине умножения погрешности на большое число.

Предположим, надо вычислить значения интегралов

$$E_n = \int_0^1 x^n e^{x-1}\,dx, \quad n = 1, 2, \dots\ .$$

Интегрируя по частям, получим

$$\int_0^1 x^n e^{x-1}\,dx = x^n e^{x-1}|_0^1 - n\int_0^1 x^{n-1} e^{x-1}\,dx.$$

Отсюда получаем рекуррентную формулу

$$E_n = 1 - nE_{n-1}, \quad n = 2, 3, \dots\ , \tag{2.7}$$

где $E_1 = 1/e$. Если для вычисления первых девяти интегралов использовать формулу (2.7), то в машинной арифметике с параметрами $\beta = 10$ и $t = 6$ получим:

$$\begin{array}{lll} E_1 \approx 0.367879, & E_4 \approx 0.170904, & E_7 \approx 0.110160, \\ E_2 \approx 0.264242, & E_5 \approx 0.145480, & E_8 \approx 0.118720, \\ E_3 \approx 0.207274, & E_6 \approx 0.127120, & E_9 \approx -0.0684800. \end{array}$$

Хотя подинтегральное выражение $x^n e^{1-x}$ положительно на всем интервале $(0, 1)$, вычисленное значение E_9 отрицательно. Такой отрицательный результат был получен по следующей причине. При вычислении E_1 в следствие округления $1/e$ до шести значащих цифр возникает погрешность приблизительно равная $4.412 \cdot 10^{-7}$. При вычислении E_9 по рекуррентным формулам (2.7) эта погрешность умножается на $2 \cdot 3 \cdot \cdots \cdot 9 = 9!$. Поэтому ошибка вычисления E_9 приблизительно составляет $4.412 \cdot 10^{-7} \cdot 9! \approx 0.1601$, что превышает значение $E_9 \approx 0.0916123$ точное до шести значащих цифр.

Построим другой алгоритм вычисления интегралов E_n. Из (2.7) получаем следующие рекуррентные соотношения

$$E_{n-1} = \frac{1 - E_n}{n}, \quad n = \ldots, 3, 2. \tag{2.8}$$

Заметим, что

$$E_n = \int\limits_0^1 x^n e^{x-1}\, dx \le \int\limits_0^1 x^n\, dx = \left.\frac{x^{n+1}}{n+1}\right|_0^1 = \frac{1}{n+1}.$$

Значит

$$\lim_{n\to\infty} E_n = 0.$$

Положим $E_{20} = 0$ и будем вычислять E_9 по рекуррентным формулам (2.8) в той же машинной арифметике с параметрами $\beta = 10$ и $t = 6$:

$$\begin{array}{lll} E_{19} \approx 0.0500000, & E_{15} \approx 0.0590176, & E_{11} \approx 0.0773523, \\ E_{18} \approx 0.0500000, & E_{14} \approx 0.0627322, & E_{10} \approx 0.0838771, \\ E_{17} \approx 0.0527778, & E_{13} \approx 0.069477, & E_9 \approx 0.0916123. \\ E_{16} \approx 0.0557190, & E_{12} \approx 0.0717733, & \end{array}$$

По этому алгоритму E_9 вычислено точно до шести значащих цифр. Первоначальная погрешность, возникшая при замене E_{20} нулем, не превосходит $1/21$. При вычислении по формулам формулам (2.8) эта погрешность делится на $20 \cdot 19 \cdot \cdots \cdot 10$ и уменьшается до величины $4 \cdot 10^{-8}$, что меньше единственной ошибки округления.

2.7. Вопросы и задания

1. Нормализованная система F чисел с плавающей точкой определяется параметрами $\beta = 2$, $t = 3$, $L = -1$, $U = 2$.

   а. Сколько чисел в этой системе?

   б. Найти все числа и представить их на числовой прямой.

   в. Вычислить наименьшие и наибольшие по модулю числа системы F.

   г. Определить все возможные расстояния между числами системы F.

   д. Определить относительную погрешность ε системы F.

   е. Представить в системе F числа: 0.375; 3.14; 0.5; -1.6.

   ж. Вычислить относительную погрешность представления числа 2.17 в системе F.

   з. Вычислить $fl(fl(1.7) + fl(1.5))$ и найти такое δ, что

   $$fl(fl(1.7) + fl(1.5)) = (fl(1.7) + fl(1.5))(1 + \delta).$$

   Проверить, что $|\delta| < \varepsilon$.

   и. Привести пример нарушения законов арифметики для чисел из системы F.

   к. Найти все субнормальные числа системы F.

   л. Найти такие $x, y \in F$ $(x \neq y)$, что $fl(x - y) = 0$. Показать, что если использовать и субнормальные числа системы F, то $fl(x - y) \neq 0$.

2. Нормализованная система F чисел с плавающей точкой определяется параметрами $\beta = 10$, $t = 2$, $L = -1$, $U = 1$. Определить границы m и M для чисел системы F, количество чисел в системе F и все возможные расстояния между числами.

3. В нормализованной системе F чисел с плавающей точкой ($\beta = 2$, $t = 3$, $L = -1$, $U = 2$) для заданного числа x определить относительную ошибку δx представления числа. ($x = 2.11, x = 0.7, x = -0.3$).

4. В нормализованной системе F чисел с плавающей точкой ($\beta = 10$, $t = 3$, $L = -2$, $U = 2$) для заданного числа x определить относительную ошибку δx представления числа. ($x = 0.0011567, x = -2410.734$).

5. Нормализованная система F чисел с плавающей точкой определяется параметрами $\beta = 10$, $t = 3$, $L = -2$, $U = 2$. Для заданных x и y проверить выполнение равенства:

$$fl(x * y) = (x * y)(1 + \delta),$$

где $|\delta| \leq \frac{1}{2}\beta^{1-t}$. ($x = 12.1$, $y = 2.4$).

6. Нормализованная система чисел с плавающей точкой определяется параметрами $\beta = 10$, $t = 5$, $L = -10$, $U = 10$. Сколько значащих цифр будет утеряно при вычитании близких чисел x и y. Какова относительная погрешность вычислений при использовании этой разности в дальнейшем? ($x = 231.4671, y = 231.5631$).

7. Нормализованная система чисел с плавающей точкой определяется параметрами $\beta = 10$, t, $L = -10$, $U = 10$. При каком значении t следующая задача определения решения СЛАУ становится некорректной:

$$x_1 + 2.74661x_2 = 3.11345,$$

$$x_1 + 2.74637x_2 = 3.12345.$$

8. Известно, что

$$s = \sum_{k=1}^{\infty} k^{-2} = \frac{\pi^2}{6} = 1.64493406\ldots.$$

Составить программу на языке C, использующую тип данных float для приближенного вычисления суммы s. Выполнить следующее:

   а. Вычисляя s, суммировать по возрастанию k до тех пор, пока сумма s изменяется. Вывести значения s и k.

   б. Зная k, вычислить значение s, суммируя по убыванию k.

   в. Чему равно k? Обосновать полученные результаты.

3. Прямые методы решения систем линейных алгебраических уравнений

В этом разделе рассматриваются прямые методы решения систем линейных алгебраических уравнений $Ax = b$. Они называются прямыми потому, что в отсутствие ошибок округления дают точное решение системы за конечное число арифметических действий.

3.1. LU-разложение

Пусть $x \in R^n$ и $x_k \neq 0$. Если

$$\tau = [\underbrace{0, \ldots, 0}_{k}, \tau_{k+1}, \ldots, \tau_n]^T, \quad \tau_i = -\frac{x_i}{x_k},\ i = \overline{k+1, n}$$

и обозначить

$$N_k = I + \tau e_k^T, \tag{3.1}$$

то

$$N_k x = \begin{bmatrix} 1 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & \ldots & \tau_{k+1} & 1 & \ldots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \tau_n & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Матрица N_k называется *матрицей исключения* или матрицей *преобразования Гаусса.*

Главной ведущей подматрицей $A^{(k)}$ матрицы $A \in R^{n \times n}$ называется матрица, которая получается из матрицы A вычеркиванием всех одно-

именных строк и столбцов начиная с $k+1$. Так

$$A^{(1)} = [a_{11}],$$

$$A^{(2)} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

$$A^{(3)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

$$\dots$$

$$A^{(n)} = A.$$

Теорема 3.1.1. *Пусть все главные ведущие подматрицы матрицы $A \in R^{n\times n}$ невырождены, т. е.,*

$$\det A^{(k)} \neq 0, \quad k = \overline{1,n}, \tag{3.2}$$

тогда матрица A единственным образом представима в виде

$$A = LU, \tag{3.3}$$

где L — нижняя треугольная матрица с единицами на главной диагонали, U — верхняя треугольная матрица диагональные элементы которой вычисляются по формуле

$$u_{kk} = \frac{\det A^{(k)}}{\det A^{(k-1)}}, \quad k = \overline{1,n}, \tag{3.4}$$

где $\det A^{(0)} = 1$.

Доказательство. Разложение (3.3) построим следующим образом. Положим

$$A_0 = A.$$

На первом шаге разложения построим матрицу исключения N_1 так, чтобы все поддиагональные элементы первого столбца матрицы

$$A_1 = N_1 A_0$$

равнялись нулю. Для этого элементы матрицы исключения N_1 выбираем следующим образом

$$n_{i,1} = -\frac{a^{(0)}_{i,1}}{a^{(0)}_{1,1}} \quad i = \overline{2,n}. \tag{3.5}$$

Здесь $a_{1,1}^{(0)}$ — ведущий элемент первого шага разложения. Таким образом,

$$A_1 = \begin{bmatrix} a_{11}^{(0)} & a_{12}^{(0)} & \dots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{bmatrix}.$$

Матрица

$$\begin{bmatrix} a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \vdots & \ddots & \vdots \\ a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{bmatrix}$$

называется активной подматрицей первого шага разложения. Ее элементы пересчитываются по формуле

$$a_{ij}^{(1)} = a_{ij}^{(0)} + n_{i1} a_{1j}^{(0)}, \quad i, j = \overline{2, n}. \tag{3.6}$$

Запишем в матричном виде второй шаг разложения.

$$A_2 = N_2 A_1.$$

Здесь матрица исключения N_2 строится так, чтобы все поддиагональные элементы во втором столбце матрицы A_2 равнялись нулю:

$$A_2 = \begin{bmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \dots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ 0 & 0 & a_{32}^{(2)} & \dots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{bmatrix}.$$

Элементы активной подматрицы второго шага пересчитываются по формуле

$$a_{ij}^{(2)} = a_{ij}^{(1)} + n_{i2} a_{1j}^{(2)}, \quad i, j = \overline{3, n}, \tag{3.7}$$

где

$$n_{i,2} = -\frac{a_{i,2}^{(1)}}{a_{2,2}^{(1)}} \quad i = \overline{3, n}, \tag{3.8}$$

и $a_{2,2}^{(1)}$ — ведущий элемент второго шага разложения.

Запишем k-й шаг разложения.

$$A_k = N_k A_{k-1},$$

$$A_k = \begin{bmatrix} a_{11}^{(0)} & \dots & a_{1k}^{(0)} & a_{1,k+1}^{(0)} & \dots & a_{1n}^{(0)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & a_{kk}^{(k-1)} & a_{k,k+1}^{(k-1)} & \dots & a_{kn}^{(k-1)} \\ 0 & \dots & 0 & a_{k+1,k+1}^{(k)} & \dots & a_{k+1,n}^{(k)} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & a_{n,k+1}^{(k)} & \dots & a_{nn}^{(k)} \end{bmatrix}.$$

Элементы активной подматрицы k–шага пересчитываются по формуле

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} + n_{ik} a_{kj}^{(k-1)}, \quad i,j = \overline{k+1,n}, \tag{3.9}$$

где

$$n_{i,k} = -\frac{a_{i,k}^{(k-1)}}{a_{k,k}^{(k-1)}} \quad i = \overline{k+1,n}, \tag{3.10}$$

и $a_{k,k}^{(k-1)}$ — ведущий элемент k-го шага разложения.

Выполнив $(n-1)$ шаг, получаем верхнюю треугольную матрицу

$$A_{n-1} = \begin{bmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \dots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \dots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{nn}^{(n-1)} \end{bmatrix} \equiv U.$$

Заметим, что здесь $u_{kk} = a_{kk}^{(k-1)}$ — ведущий элемент k-го шага разложения ($k = \overline{1,n-1}$).

Таким образом,

$$U = A_{n-1} = N_{n-1}A_{n-2} = N_{n-1}N_{n-2}A_{n-3} = \dots = N_{n-1}N_{n-2}\dots N_1 A_0$$

или

$$U = N_{n-1}N_{n-2}\dots N_1 A. \tag{3.11}$$

Умножая матричное равенство (3.11) слева на $(N_{n-1}N_{n-2}\dots N_1)^{-1}$, имеем

$$A = (N_{n-1}N_{n-2}\dots N_1)^{-1}U = N_1^{-1}\dots N_{n-2}^{-1}N_{n-1}^{-1}U = LU,$$

где

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ -n_{21} & 1 & 0 & 0 & \dots & 0 \\ -n_{31} & -n_{32} & 1 & 0 & \dots & 0 \\ -n_{41} & -n_{42} & -n_{43} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -n_{n1} & -n_{n2} & -n_{n3} & -n_{n4} & \dots & 1 \end{bmatrix}. \tag{3.12}$$

Здесь мы воспользовались легко проверяемым свойством матрицы исключения:

$$N_k^{-1} = \begin{bmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & n_{k+1,k} & 1 & \dots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & n_{n,k} & 0 & \dots & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & -n_{k+1,k} & 1 & \dots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & -n_{n,k} & 0 & \dots & 1 \end{bmatrix}.$$

Однако для построения матрицы исключения N_k надо доказать, что ведущий элемент $a_{kk}^{(k-1)}$ на k-м шаге разложения отличен от нуля.

Для этого нам понадобится следующая лемма.

Лемма 3.1.1. *Определители главных ведущих подматриц не меняются при умножении на матрицы исключения.*

Доказательство. Очевидно достаточно доказать лемму для первого шага разложения. Пусть $A_1^{(k)}$ — главная ведущая подматрица матрицы

$$A_1 = N_1 A_0.$$

Легко проверить, что

$$A_1^{(k)} = N_1^{(k)} A_0^{(k)}, \quad k = \overline{1,n}.$$

Значит

$$\det A_1^{(k)} = \det(N_1^{(k)} A_0^{(k)}) = \det N_1^{(k)} \det A_0^{(k)} = \det A_0^{(k)}.$$

□

Продолжим доказательство теоремы. Поскольку

$$N_{n-1} N_{n-2} \dots N_1 A = U,$$

то из условия (3.2) теоремы и доказанной леммы следует, что

$$\det A^{(k)} = \det(N_{n-1} N_{n-2} \dots N_1 A)^{(k)} = \det U^{(k)} = u_{11} \dots u_{kk} \neq 0,$$

$$\det A^{(k-1)} = \det(N_{n-1} N_{n-2} \dots N_1 A)^{(k-1)} = \det U^{(k-1)} =$$

$$= u_{11} \dots u_{k-1,k-1} \neq 0, \quad k = \overline{2,n}.$$

Отсюда

$$u_{kk} = \frac{\det A^{(k)}}{\det A^{(k-1)}}, \quad k = \overline{2,n}.$$

При $k = 1$

$$\det A^{(1)} = \det((N_{n-1}N_{n-2}\dots N_1 A)^{(1)}) = \det U^{(1)} = u_{11} \neq 0.$$

Таким образом, мы доказали утверждение (3.4) теоремы и показали, что на k-м шаге разложения ведущий элемент $a_{kk}^{(k-1)} = u_{kk}$ отличен от нуля, что дает возможность построить матрицу исключения N_k, а значит описанный выше алгоритм построения LU-разложения законен.

Докажем теперь единственность разложения (3.3). Предположим, что существует два различных разложения, т.е.

$$A = L_1U_1 = L_2U_2.$$

Умножая равенство

$$L_1U_1 = L_2U_2$$

слева на L_1^{-1}, а справа на U_1^{-1}, получаем

$$I = L_1^{-1}L_2U_2U_1^{-1}.$$

Поскольку $L_1^{-1}L_2$ — нижняя треугольная матрица, $U_2U_1^{-1}$ — верхняя треугольная матрица, а их произведение есть единичная матрица, то $L_1^{-1}L_2$ и $U_2U_1^{-1}$ обязаны быть единичными матрицами. Таким образом,

$$L_1^{-1}L_2 = I \;\Rightarrow\; L_2 = L_1,$$

и

$$U_2U_1^{-1} = I \;\Rightarrow\; U_2 = U_1.$$

Теорема полностью доказана. □

Подсчитаем количество арифметических действий которое необходимо выполнить для построения LU-разложения. На первом шаге разложения для построения матрицы исключения (формула (3.5)) необходимо $n-1$ деление, а для вычисления элементов активной подматрицы (формула (3.6)) — $(n-1)^2$ сложений и умножений, т.е. первый шаг требует

$$(n-1)D + (n-1)^2(A+M)$$

арифметических операций. Здесь для обозначения арифметических действий приняты следующие обозначения: D — деление (divide), A — сложение (add), M — умножение (multiply). Второй шаг разложения (формулы (3.7)–(3.8)) требует

$$(n-2)D + (n-2)^2(A+M)$$

арифметических действий. Общее количество арифметических действий LU-разложения:

$$(n-1)D+(n-1)^2(A+M)+(n-2)D+(n-2)^2(A+M)+\cdots+D+(A+M)=$$

$$=\frac{n(n-1)}{2}D+\frac{(n-1)n(2n-1)}{6}(A+M)=\frac{2n^3}{3}+O\left(n^2\right) \text{ а. д.}$$

Здесь мы воспользовались формулой для суммы арифметической прогрессии и равенством

$$\sum_{k=1}^{n}k^2=\frac{n(n+1)(2n+1)}{6},$$

которое легко доказывается по индукции.

Приведем теперь алгоритм вычисления LU-разложения. В этом алгоритме множители L и U сохраняются на месте исходной матрицы A.

Алгоритм 3.1.1. ***LU-разложение***

1. $For\ k=1\ To\ n-1$
2. $\quad If\ a_{kk}=0\ Then$
3. $\quad\quad print(\text{"Деление на 0."});\ Break$
4. $\quad End\ If$
5. $\quad For\ i=k+1\ To\ n$
6. $\quad\quad N=-\frac{a_{ik}}{a_{kk}}$
7. $\quad\quad a_{ik}=-N$
8. $\quad\quad For\ j=k+1\ To\ n$
9. $\quad\quad\quad a_{ij}=a_{ij}+Na_{kj}$
10. $\quad\quad End\ For$
11. $\quad End\ For$
12. $End\ For$

Пусть надо решить систему линейных алгебраических уравнений

$$Ax=b. \tag{3.13}$$

Если для матрицы системы (3.13) известно LU-разложение, то решение системы сводится к поиску решения двух систем с треугольными матрицами:

$$Ux=y, \tag{3.14}$$

$$Ly=b. \tag{3.15}$$

Вначале решаем систему (3.15) с нижней треугольной матрицей L с единицами на главной диагонали. Приведем алгоритм решения.

Алгоритм 3.1.2. ***Решение системы*** $Ly = b$**.**

1. $y_1 = b_1$
2. $For\ i = 2\ To\ n$
3. $\quad y_i = b_i - \sum_{j=1}^{i-1} l_{ij} y_j$
4. $End\ For$

Вычислив решение y системы (3.15), находим решение системы (3.14) с верхней треугольной матрицей U. Алгоритм решения следующий.

Алгоритм 3.1.3. ***Решение системы*** $Ux = y$**.**

1. $x_n = y_n$
2. $For\ i = n-1\ To\ 1\ Step\ -1$
3. $\quad x_i = (y_i - \sum_{j=i+1}^{n} u_{ij} x_j)/u_{ii}$
4. $End\ For$

3.2. Метод Гаусса

Рассмотрим теперь метод Гаусса решения систем линейных алгебраических уравнений. Пусть матрица $A \in R^{n\times n}$ системы

$$Ax = b \tag{3.16}$$

удовлетворяет условиям теоремы 3.1.1 об LU-разложении. Тогда метод Гаусса состоит в следующем. Вначале систему (3.16) сводим к системе с верхней треугольной матрицей. В матричном виде этот процесс можно записать так:

$$N_{n-1} \dots N_2 N_1 A x = N_{n-1} \dots N_2 N_1 b,$$

т. е., получаем

$$Ux = f, \tag{3.17}$$

где $U = N_{n-1} \dots N_2 N_1 A$ — верхняя треугольная матрица,

$$f = N_{n-1} \dots N_2 N_1 b,$$

N_k $(k = \overline{1, n-1})$ — матрицы исключения. Этот шаг решения называется *прямым ходом* метода Гаусса. *Обратный ход* метода состоит в решении системы (3.17) с верхней треугольной матрицей по алгоритму 3.1.3.

3.3. Теория возмущений для СЛАУ

Рассмотрим систему линейных алгебраических уравнений

$$Ax = b. \tag{3.18}$$

Если возмутить исходные данные системы 3.18, то решение тоже изменится и вместо 3.18 будем иметь

$$(A + \delta A)(x + \delta x) = b + \delta b. \tag{3.19}$$

Наша цель — оценить норму вектора δx. Обозначим $\widehat{x} = x + \delta x$. Вычитая 3.18 из 3.19 получаем

$$A\delta x + \delta A\widehat{x} = \delta b.$$

Отсюда

$$\delta x = A^{-1}(-\delta A\widehat{x} + \delta b)$$

и

$$||\delta x|| \leq ||A^{-1}||(||\delta A||\,||\widehat{x}|| + ||\delta b||).$$

Разделив обе части последнего неравенства на $||\widehat{x}||$ получаем

$$\frac{||\delta x||}{||\widehat{x}||} \leq \mathrm{cond}(A)\left(\frac{||\delta A||}{||A||} + \frac{||\delta b||}{||A||\,||\widehat{x}||}\right), \tag{3.20}$$

где $\mathrm{cond}(A) = ||A||\,||A^{-1}||$ — *число обусловленности* матрицы A.

Если $\mathrm{cond}(A)$ мало, то из (3.20) следует, что малым изменениям исходных данных системы (3.18) будут соответствовать малые изменения решения, т.е. задача (3.18) хорошо обусловлена.

Перечислим наиболее важные свойства числа обусловленности, некоторые из которых очевидны.

Свойство 3.3.1. *Величина числа обусловленности зависит от выбора матричной нормы.*

Свойство 3.3.2. $\mathrm{cond}(AB) \leq \mathrm{cond}(A)\mathrm{cond}(B)$.

Свойство 3.3.3. $\mathrm{cond}(A) \geq 1$.

Доказательство. Действительно,

$$1 \leq ||I|| = ||AA^{-1}|| \leq ||A||\,||A^{-1}|| = \mathrm{cond}(A).$$

□

Свойство 3.3.4. $\mathrm{cond}(\alpha A) = |\alpha|\mathrm{cond}(A), \quad \forall \alpha \in R \setminus 0$.

Свойство 3.3.5. *Для симметричной матрицы*

$$\mathrm{cond}_C(A) = ||A||_C||A^{-1}||_C = \frac{\max|\lambda(A)|}{\min|\lambda(A)|}.$$

Свойство 3.3.6. *Для ортогональной матрицы* Q $\mathrm{cond}_C(Q) = 1$.

Доказательство. Действительно,

$$||Q||_C^2 = \lambda_{\max}(Q^T Q) = \lambda_{\max}(I) = 1,$$

$$||Q^{-1}||_C^2 = ||Q^T||_C^2 = \lambda_{\max}(QQ^T) = \lambda_{\max}(I) = 1.$$

□

В дальнейшем нам понадобится следующая лемма.

Лемма 3.3.1. *Если* $||X|| < 1$*, то матрица* $I - X$ *обратима,*

$$(I - X)^{-1} = \sum_{i=0}^{\infty} X^i \tag{3.21}$$

и

$$||(I - X)^{-1}|| \leq \frac{1}{1 - ||X||}. \tag{3.22}$$

Доказательство. По определению матричный ряд

$$\sum_{i=0}^{\infty} X^i \tag{3.23}$$

сходится тогда и только тогда, когда сходится каждый числовой ряд из элементов матриц,

$$\sum_{i=0}^{\infty} (X^i)_{kl}, \qquad \forall k, l = \overline{1, n}. \tag{3.24}$$

Используя тот факт, что для произвольной матричной нормы существует такая константа $C > 0$, что

$$|a_{kl}| \leq C||A||,$$

получаем

$$|(X^i)_{kl}| \leq C||X^i|| \leq C||X||^i.$$

Отсюда следует, что ряды (3.24) мажорируются геометрической прогрессией

$$C \sum_{i=0}^{\infty} ||X||^i = \frac{C}{1 - ||X||},$$

а значит сходятся, и сходится матричный ряд 3.23. Поэтому последовательность частичных сумм

$$S_n = \sum_{i=0}^{n} X^i$$

при $n \to \infty$ стремится к некоторой матрице S, которая является суммой матричного ряда (3.23). Рассматривая предел при $n \to \infty$ очевидного равенства

$$(I - X)S_n = (I - X)(I + X + X^2 + \cdots + X^n) = I - X^{n+1},$$

получаем

$$(I - X)S = I$$

Равенство (3.23) доказано. Наконец,

$$||(I - X)^{-1}|| = \left|\left|\sum_{i=0}^{\infty} X^i\right|\right| \leq \sum_{i=0}^{\infty} ||X||^i = \frac{1}{1 - ||X||}.$$

□

Следующая теорема дает геометрическую интерпретацию числа обусловленности матрицы.

Теорема 3.3.1. *Пусть матрица $A \in R^{n\times n}$ невырождена. Тогда*

$$\min_{\delta A,\, \det(A+\delta A)=0} \frac{||\delta A||_C}{||A||_C} = \frac{1}{\mathrm{cond}_C(A)}. \tag{3.25}$$

Доказательство. Достаточно показать, что

$$\min_{\delta A,\, \det(A+\delta A)=0} ||\delta A||_C = \frac{1}{||A^{-1}||_C}. \tag{3.26}$$

Убедимся, что указанный минимум не меньше $\frac{1}{||A^{-1}||_C}$. Действительно, если $||\delta A||_C < \frac{1}{||A^{-1}||_C}$, то

$$1 > ||\delta A||_C ||A^{-1}||_C \geq ||A^{-1}\delta A||_C$$

и по предыдущей лемме матрица $I + A^{-1}\delta A$ обратима, а значит обратима и матрица $A + \delta A$, что противоречит условию теоремы.

Покажем, что минимум в (3.26) равен $\frac{1}{||A^{-1}||_C}$. Для этого построим такое δA с нормой $\frac{1}{||A^{-1}||_C}$, чтобы $\det(A + \delta A) = 0$. Поскольку

$$||A^{-1}||_C = \max_{||x||_2=1} ||A^{-1}x||_2,$$

то найдется такой вектор x с единичной евклидовой нормой, что

$$||A^{-1}||_C = ||A^{-1}x||_2 > 0.$$

Положим,

$$y = \frac{A^{-1}x}{||A^{-1}x||_2} = \frac{A^{-1}x}{||A^{-1}||_C}.$$

Заметим, что $||y||_2 = 1$. Положим,

$$\delta A = -\frac{xy^T}{||A^{-1}||_C}.$$

Тогда,

$$||\delta A||_C = \max_{||z||_2=1} \frac{||xy^Tz||_2}{||A^{-1}||_C} = \max_{||z||_2=1} \frac{|y^Tz|\,||x||_2}{||A^{-1}||_C} = \frac{1}{||A^{-1}||_C},$$

ибо по неравенству Коши–Буняковского $|y^Tz| \le ||y||_2||z||_2 = 1$ и максимум достигается при $z = \pm y$. Матрица $A + \delta A$ вырождена, ибо

$$(A + \delta A)y = Ay - \frac{xy^Ty}{||A^{-1}||_C} = \frac{x}{||A^{-1}||_C} - \frac{x}{||A^{-1}||_C} = 0.$$

□

Рассмотрим следующий пример. Вектор $x = [1, 1]^T$ является точным решением системы

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.0001 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2.0001 \end{bmatrix}. \tag{3.27}$$

Точным решением возмущенной системы

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.0001 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2.0002 \end{bmatrix}.$$

является вектор $x = [0, 2]^T$. Чтобы найти ответ на вопрос, почему небольшие изменения в исходных данных задачи привели к существенному изменению результата, оценим число обусловленности матрицы задачи (3.27), используя предыдущую теорему:

$$\frac{1}{\mathrm{cond}_C(A)} = \min_{\delta A,\, \det(A+\delta A)=0} \frac{||\delta A||_C}{||A||_C} \le \frac{\left|\left|\left|\begin{bmatrix} 0 & 0 \\ 0 & -0.0001 \end{bmatrix}\right|\right|\right|_C}{1} = 10^{-4}.$$

Значит, $\mathrm{cond}_C(A) \ge 10^4$, т. е., система (3.27) плохо обусловлена.

3.4. LU-разложение и метод Гаусса в арифметике с плавающей точкой

Начнем с примера. Предположим в машинной арифметике с параметрами $\beta = 10$, $t = 3$, относительная погрешность которой $\varepsilon = \frac{1}{2}10^{-2} = 0.005$, требуется решить систему

$$\begin{bmatrix} 0.0001 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.0001 \\ 2 \end{bmatrix}, \tag{3.28}$$

построив LU-разложение матрицы A системы. Точное решение системы $x = [1, 1]^T$. Поскольку $\mathrm{cond}_\infty(A) = ||A||_\infty ||A^{-1}||_\infty \approx 4$, то система (3.28) хорошо обусловлена, а значит система должна решаться с хорошей точностью.

Построим LU-разложение матрицы системы.

$$L = \begin{bmatrix} 1 & 0 \\ fl(\frac{1}{10^{-4}}) & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 10^4 & 1 \end{bmatrix},$$

$$U = \begin{bmatrix} 0.0001 & 1 \\ 0 & fl(1 - 10^4) \end{bmatrix} = \begin{bmatrix} 10^{-4} & 1 \\ 0 & -10^4 \end{bmatrix}.$$

Тогда

$$LU = \begin{bmatrix} 1 & 0 \\ 10^4 & 1 \end{bmatrix} \begin{bmatrix} 10^{-4} & 1 \\ 0 & -10^4 \end{bmatrix} = \begin{bmatrix} 10^{-4} & 1 \\ 1 & 0 \end{bmatrix}.$$

Перемножив LU мы не получили матрицу, близкую к матрице A. Действительно, $||A - LU||_\infty \approx 1$, что сравнимо с $||A||_\infty \approx 2$, а требуется, чтобы $||A - LU||_\infty \approx \varepsilon ||A||_\infty \approx 0.005 * 2 = 0.01$.

Заметим, что элемент a_{22} был полностью утрачен для дальнейших вычислений, когда из него вычитали число 10^4. Мы бы получили те же самые множители L и U, если бы элемент a_{22} равнялся бы любому такому числу, что $fl(a_{22} - 10^4) = -10^4$. Таким образом, решая систему для различных матриц, мы получили бы один и тот же результат. Следовательно, в данном случае невозможно гарантировать хорошую точность результата. Значит, алгоритм LU-разложения, а также метод Гаусса, являются *численно неустойчивыми* в машинной арифметике.

Попробуем решить систему (3.28), используя полученное LU-разложение. Из

$$\begin{bmatrix} 1 & 0 \\ 10^4 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1.0001 \\ 2 \end{bmatrix},$$

находим

$$\begin{aligned} y_1 &= fl(1.0001) = 1, \\ y_2 &= fl(2 - 10^4 * 1) = -10^4. \end{aligned}$$

Из

$$\begin{bmatrix} 10^4 & 1 \\ 0 & -10^4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -10^4 \end{bmatrix},$$

находим

$$\begin{aligned} x_2 &= fl(1.0001) = 1, \\ x_1 &= fl((1-1)/-10^4) = 0. \end{aligned}$$

Полученное решение $x = [0,1]^T$ существенно отличается от точного.

Еще один признак численной неустойчивости алгоритма LU-разложения можно получить, сравнивая число обусловленности матрицы A и числа обусловленности матриц L, U, которые много больше первого. Действительно

$$\mathrm{cond}_\infty(L) = \left|\left|\begin{bmatrix} 1 & 0 \\ 10^4 & 1 \end{bmatrix}\right|\right|_\infty \left|\left|\begin{bmatrix} 1 & 0 \\ -10^4 & 1 \end{bmatrix}\right|\right|_\infty \approx 10^8,$$

$$\mathrm{cond}_\infty(U) = \left|\left|\begin{bmatrix} 10^{-4} & 1 \\ 0 & -10^4 \end{bmatrix}\right|\right|_\infty \left|\left|\begin{bmatrix} 10^4 & 1 \\ 0 & -10^{-4} \end{bmatrix}\right|\right|_\infty \approx 10^8.$$

Таким образом, можно сделать следующий вывод. Если промежуточные величины, возникающие при вычислении LU-разложения, велики по сравнению с $||A||$, то информация, содержащаяся в элементах матрицы A, будет утеряна, когда эти элементы будут складываться с этими большими величинами. Заметим, что в нашем случае большая величина получилась из-за деления на очень малый диагональный элемент.

Поменяем местами строки системы (3.28).

$$\begin{bmatrix} 1 & 1 \\ 0.0001 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1.0001 \end{bmatrix}, \tag{3.29}$$

Построим LU-разложение матрицы системы (3.29).

$$L = \begin{bmatrix} 1 & 0 \\ 10^{-4} & 1 \end{bmatrix},$$

$$U = \begin{bmatrix} 1 & 1 \\ 0 & fl(1 - 1 * 10^{-4}) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Тогда

$$LU = \begin{bmatrix} 1 & 0 \\ 10^{-4} & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 10^{-4} & 1 + 10^{-4} \end{bmatrix}$$

и

$$||A - LU||_\infty = \left|\left|\begin{bmatrix} 0 & 0 \\ 0 & 10^{-4} \end{bmatrix}\right|\right|_\infty = 10^{-4} < \varepsilon ||A||_\infty = 0.01,$$

т. е., произведение LU дает матрицу очень близкую к матрице A. Кроме того, легко проверить, что матрицы L и U хорошо обусловлены. Решим теперь систему (3.29), используя полученное LU-разложение. Из

$$\begin{bmatrix} 1 & 0 \\ 10^{-4} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1.0001 \end{bmatrix},$$

находим

$$\begin{aligned} y_1 &= 2, \\ y_2 &= fl(1.0001 - 10^{-4} * 2) = 1. \end{aligned}$$

Из

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix},$$

находим

$$\begin{aligned} x_2 &= 1, \\ x_1 &= 1. \end{aligned}$$

Получили точное решение $x = [1, 1]^T$.

3.5. PLU-разложение невырожденной матрицы

Теорема 3.5.1. *Любая невырожденная матрица $A \in R^{n\times n}$ может быть представлена в виде*

$$A = PLU, \tag{3.30}$$

где P — матрица перестановок, L — нижняя треугольная матрица с единицами на главной диагонали, U — невырожденная верхняя треугольная матрица.

Доказательство. Доказательство конструктивно. Положим

$$A_0 = A.$$

Найдем наибольший по модулю элемент в первом столбце матрицы A_0:

$$|a^{(0)}_{\gamma_1 1}| = \max_{1\le i\le n} |a^{(0)}_{i1}|.$$

Если таких элементов несколько, выбираем любой из них. В матрице A_0 меняем местами первую строку и строку с номером γ_1, которая содержит найденный наибольший по модулю элемент. Для этого умножаем

слева матрицу A_0 на элементарную матрицу перестановок $P^{1\gamma_1}$. В первом столбце полученной матрицы обнуляем все элементы кроме первого:

$$A_1 = N_1 P^{1\gamma_1} A_0.$$

Здесь N_1 — матрица исключения.

На k-м шаге находим наибольший по модулю элемент среди $n-k+1$ элементов k-го столбца матрицы A_{k-1}:

$$|a^{(k-1)}_{\gamma_k k}| = \max_{k \le i \le n} |a^{(k-1)}_{i1}|.$$

Если таких элементов несколько, выбираем любой из них. В матрице A_{k-1} меняем местами k-ю строку и строку с номером γ_k, которая содержит найденный наибольший по модулю элемент. Для этого умножаем слева матрицу A_{k-1} на элементарную матрицу перестановок $P^{k\gamma_k}$. В k-м столбце полученной матрицы обнуляем все элементы начиная с $k+1$-го:

$$A_k = N_k P^{k\gamma_k} A_{k-1}.$$

Здесь N_k — матрица исключения.

После $n-1$ шага получаем верхнюю треугольную матрицу

$$U = N_{n-1} P^{n-1,\gamma_{n-1}} \cdots N_2 P^{2\gamma_2} N_1 P^{1\gamma_1} A. \tag{3.31}$$

Заметим, что в точной арифметике этот алгоритм прерваться не может. Действительно, предположим, что на k-м шаге элементы k-го столбца, начиная с k-й строки матрицы A_{k-1}, равны нулю (в этом случае невозможно определить матрицу исключения N_k). Тогда

$$0 = \det(A_{k-1}) = \det(N_{k-1} P^{k-1,\gamma_{k-1}} \cdots N_2 P^{2\gamma_2} N_1 P^{1\gamma_1} A) = \pm \det(A) \neq 0.$$

Получили противоречие.

Заметим, что

$$(N_{n-1} P^{n-1,\gamma_{n-1}} \cdots N_2 P^{2\gamma_2} N_1 P^{1\gamma_1})^{-1} = P^{1\gamma_1} N_1^{-1} P^{2\gamma_2} N_2^{-1} \cdots P^{n-1,\gamma_{n-1}} N_{n-1}^{-1} =$$

$$= (P^{1\gamma_1} P^{2\gamma_2} \cdots P^{n-1,\gamma_{n-1}})(P^{n-1,\gamma_{n-1}} \cdots P^{2\gamma_2} N_1^{-1} P^{2\gamma_2} \cdots P^{n-1,\gamma_{n-1}}) \times$$

$$\times (P^{n-1,\gamma_{n-1}} \cdots P^{3\gamma_3} N_2^{-1} P^{3\gamma_3} \cdots P^{n-1,\gamma_{n-1}}) \cdots N_{n-1}^{-1} = PL,$$

где $P = P^{1\gamma_1} P^{2\gamma_2} \cdots P^{n-1,\gamma_{n-1}}$ — матрица перестановок, L — нижняя треугольная матрица с единицами на главной диагонали, ибо

$$L = \widetilde{N}_1 \widetilde{N}_2 \cdots \widetilde{N}_{n-2} N_{n-1}^{-1},$$

где

$$\widetilde{N}_k = P^{n-1,\gamma_{n-1}} \dots P^{k+1\gamma_{k+1}} N_k^{-1} P^{k+1\gamma_{k+1}} \dots P^{n-1,\gamma_{n-1}}, \quad k = \overline{1, n-2},$$

— некоторые матрицы исключения. Наконец, умножая слева матричное равенство (3.31) на

$$(N_{n-1} P^{n-1,\gamma_{n-1}} \dots N_2 P^{2\gamma_2} N_1 P^{1\gamma_1})^{-1},$$

получаем разложение (3.30). □

Если получено разложение матрицы $A = PLU$, то систему $Ax = b$ можно заменить двумя системами

$$\begin{cases} Ly = P^T b, \\ Ux = y, \end{cases}$$

которые легко решить, поскольку их матрицы треугольные. Поскольку $P = P^{1\gamma_1} P^{2\gamma_2} \dots P^{n-1,\gamma_{n-1}}$, то $P^T = P^{n-1,\gamma_{n-1}} \dots P^{2\gamma_2} P^{1\gamma_1}$. Перестановку γ, которой определяется матрица P^T можно определить следующим образом:

$$g = [1, 2, \dots n]^T, \qquad g = P^{k\gamma_k} g, \ k = \overline{1, n-1}, \qquad \gamma = g.$$

Таким образом, чтобы определить перестановку γ надо на k-м шаге разложения поменять местами k-ю и γ_k-ю компоненты вектора g. Заметим, что $(P^T b)(i) = b(g(i)), \ i = \overline{1, n}$.

3.6. Метод Гаусса с частичным выбором главного элемента

Используя приведенные выше обозначения, метод Гаусса с частичным выбором главного элемента (выбором главного элемента по столбцу) решения системы $Ax = b$ можно в матричном виде описать следующим образом. Прямой ход метода:

$$N_{n-1} P^{n-1,\gamma_{n-1}} \dots N_2 P^{2\gamma_2} N_1 P^{1\gamma_1} Ax = N_{n-1} P^{n-1,\gamma_{n-1}} \dots N_2 P^{2\gamma_2} N_1 P^{1\gamma_1} b.$$

Обратный ход метода состоит в нахождении решения полученной системы, матрица которой является верхней треугольной.

Приведем некоторые основные положения анализа ошибок метода Гаусса с частичным выбором главного элемента в арифметике с плавающей точкой[9].

Метод Гаусса с частичным выбором главного элемента в арифметике с плавающей точкой находит приближенное решение $\widehat{x}$ системы $Ax = b$. На первом этапе анализа, анализируют ошибки округлений с тем, чтобы показать, что приближенное решение $\widehat{x}$ является точным решением возмущенной системы

$$(A + \delta A)\widehat{x} = b + \delta b. \tag{3.32}$$

с малыми возмущениями δA и δb. Такие исследования называются *обратным анализом ошибок*, а δA и δb называются *обратными ошибками*. Мы не будем приводить обратный анализ ошибок для метода Гаусса с частичным выбором главного элемента, ограничимся некоторыми результатами.

Теорема 3.6.1. *Метод Гаусса с частичным выбором главного элемента гарантирует, что коэффициент роста*

$$g = \frac{\max\limits_{i,j} |u_{ij}|}{\max\limits_{i,j} |a_{ij}|} \leq 2^{n-1}.$$

Эта оценка достижима. Обратная ошибка δA приближенного решения удовлетворяет оценке

$$||\delta A||_\infty \leq 3gn^3\varepsilon||A||_\infty, \tag{3.33}$$

где ε — относительная погрешность арифметики с плавающей точкой.

Оценка (3.33) сильно завышена. Так, например, даже при коэффициенте роста $g = 1$ для матрицы умеренного размера $n = 150$ при $\varepsilon = 10^{-7}$ (относительная погрешность арифметики с плавающей точкой одинарной точности стандарта IEEE) имеем $3gn^3\varepsilon > 1$, т.е. $||\delta A||_\infty \leq ||A||_\infty$. Последнее неравенство означает, что все верные знаки в вычисленном решении могут быть потеряны. Т. о., метод Гаусса с частичным выбором главного элемента является относительно устойчивым алгоритмом. Метод Гаусса с полным выбором главного элемента более устойчив. Однако метод Гаусса с полным выбором главного элемента требует $O(n^3)$ сравнений для выбора главного элемента на каждом шаге, а метод Гаусса с частичным выбором — всего $O(n^2)$. Поэтому метод Гаусса с полным выбором главного элемента используют редко. Заметим, что в большинстве случаев для прикладных задач ошибка приближенного решения полученного методом Гаусса с частичным выбором главного элемента много меньше, чем в оценке (3.33).

После того, как показано, что приближенное решение $\widehat{x}$ является решением возмущенной системы (3.32), для оценки ошибки приближенного решения $\widehat{x}$ на втором этапе анализа применяют теорию возмущений для

СЛАУ. Можно воспользоваться оценкой (3.20). Однако, чаще всего ошибку приближенного решения оценивают следующим образом. Вектор

$$r = b - A\widehat{x}$$

называют невязкой (residual) приближенного решения $\widehat{x}$. Поскольку

$$\delta x = x - \widehat{x} = A^{-1}b - \widehat{x} = A^{-1}b - (A^{-1}b - A^{-1}r) = A^{-1}r,$$

то $||\delta x|| \leq ||A^{-1}||\,||r||$ и

$$\frac{||\delta x||}{||\widehat{x}||} \leq ||A^{-1}||\frac{||r||}{||\widehat{x}||} \tag{3.34}$$

3.7. Разложение Холесского

Докажем следующую теорему.

Теорема 3.7.1. *Любую симметричную положительно определенную матрицу $A \in R^{n\times n}$ можно единственным образом представить в виде*

$$A = LL^T, \tag{3.35}$$

где L — нижняя треугольная матрица с положительными диагональными элементами.

Доказательство. Поскольку матрица A положительно определена, то все ее главные ведущие подматрицы тоже положительно определены и, следовательно, невырождены. Поэтому для матрицы A существует единственное LU-разложение:

$$A = LU = LD\widetilde{U}, \tag{3.36}$$

где $\overline{u}_{ij} = \frac{u_{ij}}{u_{ii}}$, $i = \overline{1,n}$, $j = \overline{i,n}$, $D = \mathrm{diag}(u_{11}, u_{22}, \ldots u_{nn})$. Поскольку матрица A симметрична,

$$A^T = (LD\widetilde{U})^T = \widetilde{U}^T D L^T = A = LD\widetilde{U},$$

т. е.,

$$\widetilde{U}^T D L^T = LD\widetilde{U}.$$

Домножая последнее матричное равенство слева на $(\widetilde{U}^T)^{-1}$ и справа на $(L^T)^{-1}$, получим

$$D = (\widetilde{U}^T)^{-1} L D \widetilde{U} (L^T)^{-1}. \tag{3.37}$$

Матричное равенство (3.37) возможно тогда и только тогда, когда

$$(\widetilde{U}^T)^{-1}L = I \ \text{ и } \ \widetilde{U}(L^T)^{-1} = I,$$

т. е., если $\widetilde{U} = L^T$. Значит разложение (3.36) можно записать в виде

$$A = LDL^T. \tag{3.38}$$

Поскольку матрица A симметрична и положительно определена, матрица L невырождена, то из (3.38) следует, что матрица D тоже положительно определена и ее диагональные элементы положительны. Значит разложение (3.38) можно записать так

$$A = LD^{\frac{1}{2}}(LD^{\frac{1}{2}})^T,$$

где $D^{\frac{1}{2}} = \mathrm{diag}(\sqrt{d_{11}}, \sqrt{d_{22}}, \ldots \sqrt{d_{nn}})$. □

Разложение (3.35) называется разложением Холесского симметричной положительно определенной матрицы, а матрица L — множителем Холесского.

Рассмотрим теперь алгоритм построения разложения Холесского симметричной положительно определенной матрицы A. Первый шаг алгоритма можно получить из следующих очевидных равенств:

$$A = \begin{bmatrix} a_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} \sqrt{a_{11}} & 0^T \\ \frac{A_{21}}{\sqrt{a_{11}}} & I \end{bmatrix} \begin{bmatrix} 1 & 0^T \\ 0 & \widetilde{A}_{22} \end{bmatrix} \begin{bmatrix} \sqrt{a_{11}} & \frac{A_{21}^T}{\sqrt{a_{11}}} \\ 0 & I \end{bmatrix},$$

где

$$\widetilde{A}_{22} = A_{22} - \frac{A_{21}A_{21}^T}{a_{11}}.$$

Заметим, что матрица $\widetilde{A}_{22}$ (активная подматрица первого шага) — симметрична и положительно определена. Первым столбцом множителя Холесского является вектор $[\sqrt{a_{11}}, \frac{A_{21}^T}{\sqrt{a_{11}}}]^T$. Записывая аналогичное разложение для матрицы $\widetilde{A}_{22}$, получим второй столбец множителя Холесского. Повторяя n раз описанный выше процесс, найдем все n столбцов множителя Холесского. Таким образом, алгоритм разложения Холесского можно записать следующим образом.

Алгоритм 3.7.1. ***Разложение Холесского***

1. *For* $i = 1$ *To* n
2. *For* $j = 1$ *To* i
3. $l_{ij} = a_{ij}$
4. *End For*

5. *End For*
6. *For* $k = 1$ *To* n
7. $\quad l_{kk} = \sqrt{l_{kk}}$
8. $\quad$ *For* $i = k + 1$ *To* n
9. $\quad\quad l_{ik} = l_{ik}/l_{kk}$
10. $\quad\quad$ *For* $j = k + 1$ *To* i
11. $\quad\quad\quad l_{ij} = l_{ij} - l_{ki}l_{kj}$
12. $\quad\quad$ *End For*
13. $\quad$ *End For*
14. *End For*

Теорема 3.7.2. *Алгоритм разложения Холесского устойчив в арифметике с плавающей точкой.*

Доказательство. Из 11-й строки алгоритма 3.7.1 следует, что

$$l_{ii} = l_{ii} - l_{ki}^2,$$

т. е., диагональные элементы активных подматриц $\widetilde{A}_{kk}$ не возрастают. Поскольку матрицы $\widetilde{A}_{kk}$ положительно определены, то их диагональные элементы l_{ii} остаются положительными, а модули внедиагональных элементов не возрастают, поскольку максимальный по модулю элемент симметричной положительно определенной матрицы находится на диагонали. Таким образом, в процессе построения разложения Холесского получаемые значения не растут по абсолютной величине. Значит, алгоритм устойчив в арифметике с плавающей точкой. □

3.8. Итерационное уточнение

Из-за ошибок округленияй прямой метод находит приближенное решение x_0 системы

$$Ax = b. \tag{3.39}$$

Положим

$$x = x_0 + \Delta x,$$

где x — точное решение системы (3.39). Тогда для определения Δx получаем систему

$$A\Delta x = r_0, \tag{3.40}$$

где $r_0 = b - Ax_0$ — невязка приближенного решения x_0. Решая уравнение (3.40) тем же прямым методом, находим приближенное значение Δx, которое обозначим через Δx_0. Положим, $x_1 = x_0 + \Delta x_0$. Продолжая этот процесс, получаем алгоритм, который называется итерационным уточнением.

Алгоритм 3.8.1. ***k*-я итерация итерационного уточнения**

1. $r_k = b - Ax_k$
2. $A\Delta x_k = r_k$
3. $x_{k+1} = x_k + \Delta x_k$

Известны следующие результаты [9].

Теорема 3.8.1. *Предположим, что невязка r_k вычисляется в арифметике двойной точности и*

$$\mathrm{cond}(A)\varepsilon < c \equiv \frac{1}{3n^3 g + 1} < 1, \tag{3.41}$$

где n — порядок матрицы, g — коэффициент роста. Тогда, повторяя шаги итерационного уточнения, получим такой вектор x_k, что

$$\frac{||x_k - A^{-1}b||_\infty}{||A^{-1}b||_\infty} = O(\varepsilon). \tag{3.42}$$

Заметим, что в оценке ошибки (3.42) не присутствует число обусловленности матрицы A. Это означает, что решение системы можно вычислять с высокой точностью независимо от числа обусловленности матрицы, если выполняется условие (3.41). На практике, константа c из условия (3.41) является слишком завышенной верхней границей и алгоритм часто работает успешно, даже если $\mathrm{cond}(A)\varepsilon > c$.

Если нет возможности использовать арифметику удвоенной точности, например, вычисления уже ведутся в арифметике двойной точности, итерационное уточнение тоже можно использовать.

Теорема 3.8.2. *Пусть невязка вычисляется с обычной точностью и*

$$||A||_\infty ||A^{-1}||_\infty \frac{\max\limits_i(|A|\,|x_0|)}{\min\limits_i(|A|\,|x_0|)} \cdot \varepsilon < 1. \tag{3.43}$$

Тогда один шаг итерационного уточнения дает такое приближение $x_1 = x_0 + \Delta x_0$, что

$$(A + \delta A)x_1 = b + \delta b,$$

где

$$\begin{aligned} |\delta a_{ij}| &= O(\varepsilon)|a_{ij}|, \\ |\delta b_i| &= O(\varepsilon)|b_i|, \quad 1 \le i, j \le n. \end{aligned}$$

Иными словами, теорема утверждает, что покомпонентная относительная обратная ошибка для x_1 имеет максимально возможный порядок малости.

3.9. Уравновешивание

Повысить точность решения системы линейных алгебраических уравнений можно с помощью одной процедуры, которая называется уравновешиванием или масштабированием.

Матрица A называется равновесной по строкам (по столбцам), если $\beta^{-1} \leq ||(a_{i,*})^T|| \leq 1,\ i = \overline{1,n}$ ($\beta^{-1} \leq ||a_{*,j}|| \leq 1,\ j = \overline{1,n}$), где β — основание используемой системы счисления.

Матрица называется равновесной, если она равновесна и по строкам и по столбцам.

Процедура масштабирования состоит в следующем. Систему

$$Ax = b$$

заменяют системой

$$D_1 Ax = D_1 b$$

или

$$D_1 A D_2 y = D_1 b,$$

где $y = D_2^{-1}x$. Матрицы D_1 и D_2 подбирают так, чтобы матрица решаемой системы была равновесной и число обусловленности равновесной матрицы было меньше числа обусловленности исходной матрицы.

Отметим, что для заданной матрицы не существует единственной равновесной.

Пример 3.9.1. Рассмотрим матрицу

$$A = \begin{bmatrix} 1 & 1 & 2 \cdot 10^9 \\ 2 & -1 & 10^9 \\ 1 & 2 & 0 \end{bmatrix}.$$

Поскольку $||A||_1 = 3 \cdot 10^9$, то $\mathrm{cond}_1(A) \geq 3 \cdot 10^9$, т.е. матрица A плохо обусловлена.

Масштабируя матрицу A сначала по столбцам, а затем по строкам (при $\beta = 10$), получим равновесную матрицу

$$\begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.1 \end{bmatrix} \left(\begin{bmatrix} 1 & 1 & 2 \cdot 10^9 \\ 2 & -1 & 10^9 \\ 1 & 2 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^{-9} \end{bmatrix} \right) =$$

$$= \begin{bmatrix} 0.1 & 0.1 & 0.2 \\ 0.2 & -0.1 & 0.1 \\ 0.1 & 0.2 & 0 \end{bmatrix} = D_1 A D_2.$$

Число обусловленности матрицы $\mathrm{cond}_1(D_1AD_2) = 4$. Значит, равновесная матрица хорошо обусловлена. Если же матрицу A масштабировать сначала по строкам, то сразу получим равновесную матрицу

$$\begin{bmatrix} 10^{-10} & 0 & 0 \\ 0 & 10^{-10} & 0 \\ 0 & 0 & 0.1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 2\cdot 10^9 \\ 2 & -1 & 10^9 \\ 1 & 2 & 0 \end{bmatrix} =$$

$$= \begin{bmatrix} 10^{-10} & 10^{-10} & 0.2 \\ 2\cdot 10^{-10} & -10^{-10} & 0.1 \\ 0.1 & 0.2 & 0 \end{bmatrix} = D_1A.$$

Однако, число обусловленности матрицы $\mathrm{cond}_1(D_1A) = 2\cdot 10^9$, т. е., равновесная матрица тоже плохо обусловлена.

3.10. Вопросы и задания

1. Для заданных матриц проверить выполнение условий теоремы об LU-разложении:

$$A_1 = \begin{bmatrix} 2 & 1 & 1 \\ 2 & 1 & -1 \\ 0 & 1 & 2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 1 & 2 \\ 2 & 1 & -1 \end{bmatrix}.$$

2. Используя матрицы исключения, записать в матричном виде алгоритм LU-разложения матрицы A. Найти первый столбец матрицы L LU-разложения матрицы

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 2 & 4 & 0 \\ 1 & 2 & -3 \end{bmatrix}.$$

3. Построить LU-разложение заданной матрицы A. Используя полученное разложение, решить систему $Ax = b$ для заданной правой части. Вычисления производить с тремя значащими цифрами.

$$A = \begin{bmatrix} 2 & 1 & -1 \\ 0 & 2 & 1 \\ 1 & 1 & 4 \end{bmatrix}, \quad b = \begin{bmatrix} 1 & 1 & 5 \end{bmatrix}^T.$$

4. Методом Гаусса решить систему $Ax = b$. Вычисления производить с тремя значащими цифрами.

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 1 & 2 \\ 2 & 1 & -1 \end{bmatrix}, \quad b = \begin{bmatrix} 4 & 3 & 2 \end{bmatrix}^T.$$

5. Определить количество арифметических действий обратного хода метода Гаусса.

6. Определить количество арифметических действий, необходимых для построения LU-разложения ленточной матрицы $A \in R^{n\times n}$ с шириной ленты $k + l + 1 < n$.

7. Пусть
$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$
где $k \times k$–подматрица A_{11} невырождена. Тогда матрица $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ называется дополнением Шура подматрицы A_{11} в A или просто дополнением Шура. Показать, что после k шагов гауссова исключения без выбора главных элементов на месте матрицы A_{22} находится матрица S.

8. Найти PLU-разложение матрицы
$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 3 & 0 & 1 \end{bmatrix}$$
и решить систему $Ax = f$, где $f = [1; 4; 1]^T$. Вычисления производить с тремя значащими цифрами.

9. Методом Гаусса с частичным выбором главного элемента найти решение системы $Ax = f$, где $f = [1; 4; 1]^T$,
$$A = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 1 & 1 \\ 2 & 0 & 1 \end{bmatrix}$$
Вычисления производить с тремя значащими цифрами.

10. Для матрицы
$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ -1 & 1 & 0 & \cdots & 0 & 1 \\ -1 & -1 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & -1 & -1 & \cdots & 1 & 1 \\ -1 & -1 & -1 & \cdots & -1 & 1 \end{bmatrix}$$
построить PLU-разложение. Вычислить коэффициент роста
$$g_{pp} = ||U||_{\max}/||A||_{\max}, \quad \text{где} \quad ||A||_{\max} = \max_{ij} |a_{ij}|.$$
Вычислить $\mathrm{cond}_\infty(A)$, $\mathrm{cond}_\infty(U)$, $\mathrm{cond}_\infty(L)$.

11. Построить разложение Холесского матрицы

$$A = \begin{bmatrix} 1 & -1 & 1 & 1 \\ -1 & 5 & -3 & -3 \\ 1 & -3 & 6 & -2 \\ 1 & -3 & -2 & 7 \end{bmatrix}.$$

12. Построить алгоритм разложения Холесского в форме скалярных произведений (приравнять j-е столбцы в матричном уравнении $A = LL^T$).

4. Линейная задача наименьших квадратов

Рассмотрим систему линейных алгебраических уравнений

$$Ax = b, \tag{4.1}$$

где $A \in R^{m \times n}$ — матрица полного столбцового ранга, $b \in R^m$, $m > n$. Поскольку система (4.1) содержит больше уравнений, чем неизвестных, то она переопределена. Точного решения системы (4.1) может не существовать. Однако, среди приближенных решений системы (4.1) имеет смысл найти такое, 2-норма невязки которого наименьшая. Приходим к линейной задаче наименьших квадратов:

$$\min_{x \in R^n} ||Ax - b||_2. \tag{4.2}$$

Решение задачи минимизации (4.2) называют *псевдорешением* переопределенной системы (4.1).

Поскольку функционал

$$F(x) = ||Ax - b||_2^2, \tag{4.3}$$

квадратичный, и его градиент

$$\nabla F(x) = 2A^T(Ax - b),$$

то необходимое и достаточное условие минимума функционала $F(x)$ имеет вид

$$\nabla F(x) = 2A^T(Ax - b) = 0,$$

т. е.,

$$A^T Ax = A^T b. \tag{4.4}$$

Уравнения системы (4.4) называют *нормальными* уравнениями.

Поскольку матрица A системы (4.1) полного столбцового ранга, матрица $A^T A$ симметрична и положительно определена. Значит, система (4.4) имеет единственное решение. Это решение является единственным решением линейной задачи наименьших квадратов (4.2).

Систему (4.4) можно решить, используя, например, разложение Холесского матрицы $A^T A$.

Общая стоимость вычисления $A^T A$, $A^T b$ и последующего разложения Холесского матрицы $A^T A$ составляет $n^2 m + \frac{1}{3}n^3 + O(n^2)$ арифметических

действий. Так как $m > n$, то в общей стоимости преобладает $n^2 m$ — цена формирования матрицы $A^T A$.

Существуют и другие методы решения решения линейной задачи наименьших квадратов. Метод, использующий нормальные уравнения, наиболее быстр, но и наименее точен. Метод, использующий QR-разложение матрицы A, является наиболее употребительным, но его стоимость может превышать стоимость первого в два раза. Метод, использующий SVD-разложение матрицы A, наиболее полезен для плохо обусловленных задач. Его стоимость может в несколько раз превышать стоимость упомянутых методов [9].

4.1. QR-разложение

Теорема 4.1.1. (QR-разложение). *Для любой вещественной матрицы $A \in R^{m\times n}$ $(m \geq n)$ полного столбцового ранга существуют такие матрица $Q \in R^{m\times n}$ с ортонормированными столбцами и верхняя треугольная матрица $R \in R^{n\times n}$ с положительными диагональными элементами, что*

$$A = QR. \tag{4.5}$$

Доказательство. Процесс Грама-Шмидта ортогонализации столбцов матрицы $A = [a_1, \ldots, a_n]$ в порядке возрастания их номеров вычисляет ортонормированные векторы $q_1, \ldots, q_n$. Ниже приведены два алгоритма процесса Грама–Шмидта.

Алгоритм 4.1.1. ***Классический (CGS) и модифицированный (MGS) алгоритмы Грама–Шмидта вычисления разложения*** $A = QR$

1. $For\ i = 1\ To\ n$
2. $\quad q_i = a_i$
3. $\quad For\ j = 1\ To\ i - 1$
4. $\quad\quad r_{ji} = q_j^T a_i\ (CGS)\ \{r_{ji} = q_j^T q_i\ (MGS)\}$
5. $\quad\quad q_i = q_i - r_{ji} q_j$
6. $\quad End\ For$
7. $\quad r_{ii} = ||q_i||_2$
8. $\quad If\ r_{ii} = 0\ Then\ Stop$
9. $\quad q_i = q_i / r_{ii}$
10. $End\ For$

Поскольку матрица A имеет полный столбцовый ранг, то алгоритм в точной арифметике прерваться не может. Из строк 5 и 9 алгоритма 4.1.1

получаем формулы

$$r_{ii}q_i = a_i - \sum_{j=1}^{i-1} r_{ji}q_j, \quad i = \overline{1,n},$$

или

$$a_i = \sum_{j=1}^{i} r_{ji}q_j, \quad i = \overline{1,n}. \tag{4.6}$$

Формулы (4.6) — это разложение (4.5) матрицы A, записанное по столбцам. □

К сожалению, в арифметике с плавающей точкой алгоритм CGS численно неустойчив. В большинстве случаев происходит потеря ортогональности вычисленных векторов q_i. Модифицированный метод Грама-Шмидта (MGS) более устойчив. Алгоритм MGS требует $2mn^2$ арифметических операций.

Рассмотрим еще один алгоритм построения QR-разложение матрицы A.

Решим вначале следующую задачу. Для заданного вектора v найдем такой вектор u и число μ, что

$$H(u)v = \mu e_1, \tag{4.7}$$

где $H(u)$ — матрица отражения Хаусхолдера.

Поскольку

$$H(u)v = (I - \gamma uu^T)v = v - \gamma u(u^T v) = \mu e_1, \quad \text{где } \gamma = \frac{2}{u^T u},$$

то

$$v - (\gamma u^T v)u = \mu e_1. \tag{4.8}$$

Потребуем чтобы

$$\gamma u^T v = 1. \tag{4.9}$$

Тогда из (4.8) следует, что

$$u = v - \mu e_1. \tag{4.10}$$

Из (4.9) получаем,

$$2u^T v = u^T u.$$

Подставляя в это равенство выражение для u из (4.10), получаем

$$2(v - \mu e_1)^T v = (v - \mu e_1)^T (v - \mu e_1).$$

Упрощая полученное равенство, находим

$$\mu^2 = ||v||_2^2,$$

т. е.,

$$\mu = \pm||v||_2. \tag{4.11}$$

Формулы (4.10) и (4.11) дают решение задачи (4.7).

Заметим, что при вычислении первой компоненты вектора u по формуле (4.10) возникает опасность потери значащих цифр из-за вычитания близких чисел. Чтобы ее избежать, решение задачи (4.7) находят по следующим формулам:

$$\mu = -\text{sign}(v_1)||v||_2, \quad u = v + \text{sign}(v_1)||v||_2 e_1. \tag{4.12}$$

Построим теперь QR-разложение матрицы $A \in R^{m\times n}$ $(m > n)$, используя матрицы отражения Хаусхолдера. Как и ранее, будем предполагать, что матрица A полного столбцового ранга.

Первый шаг алгоритма. Положим

$$A_1 = A.$$

Найдем такой вектор u_1, что

$$H(u_1)a^{(1)}_{*,1} = \mu_1 e_1,$$

где $a^{(1)}_{*,1}$ — первый столбец матрицы A_1. Положим

$$A_1 = H(u_1)A_1.$$

Первый столбец обновленной матрицы A_1 уже вычислен. Пересчитаем остальные столбцы матрицы A_1:

$$a^{(1)}_{*,j} = H(u_1)a^{(1)}_{*,j} = (I - \gamma u_1 u_1^T)a^{(1)}_{*,j} = a^{(1)}_{*,j} - \gamma(u_1^T a^{(1)}_{*,j})u_1,\ j = \overline{2,n},$$

где $\gamma = \frac{2}{u_1^T u_1}$.

Второй шаг алгоритма. Положим

$$A_2 = A_1(2:m, 2:n).$$

Найдем такой вектор u_2, что

$$H(u_2)a^{(2)}_{*,1} = \mu_2 e_1,$$

где $a^{(2)}_{*,1}$ — первый столбец матрицы A_2. Положим

$$A_2 = H(u_2)A_2.$$

Первый столбец обновленной матрицы A_2 уже вычислен. Пересчитаем остальные столбцы матрицы A_2:

$$a^{(2)}_{*,j} = H(u_2)a^{(2)}_{*,j} = (I - \gamma u_2 u_2^T)a^{(2)}_{*,j} = a^{(2)}_{*,j} - \gamma(u_2^T a^{(2)}_{*,j})u_2,\ j = \overline{2, n-1},$$

где $\gamma = \frac{2}{u_2^T u_2}$.

Запишем k-й шаг алгоритма. Положим

$$A_k = A_{k-1}(2 : (m-k+2), 2 : (n-k+2)).$$

Найдем такой вектор u_k, что

$$H(u_k)a^{(k)}_{*,1} = \mu_k e_1,$$

где $a^{(k)}_{*,1}$ — первый столбец матрицы A_k. Положим

$$A_k = H(u_k)A_k.$$

Первый столбец обновленной матрицы A_k уже вычислен. Пересчитаем остальные столбцы матрицы A_k:

$$a^{(k)}_{*,j} = H(u_k)a^{(k)}_{*,j} = (I - \gamma u_k u_k^T)a^{(k)}_{*,j} = a^{(k)}_{*,j} - \gamma(u_k^T a^{(k)}_{*,j})u_k,\ j = \overline{2, n-k+1},$$

где $\gamma = \frac{2}{u_k^T u_k}$.

В матричном виде n шагов этого алгоритма можно записать так:

$$HA = \widetilde{R}, \tag{4.13}$$

где

$$H = H_n \cdots H_2 H_1$$

— произведение симметричных ортогональных матриц

$$H_k = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & H(u_k) & \end{bmatrix},\ k = \overline{1, n}, \quad \widetilde{R} = \begin{bmatrix} \mu_1 & a^1_{12} & a^1_{13} & \dots & a^1_{1n} \\ 0 & \mu_2 & a^2_{23} & \dots & a^2_{2n} \\ 0 & 0 & \mu_3 & \dots & a^3_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \vdots & \mu_n \\ 0 & 0 & 0 & \vdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \vdots & 0 \end{bmatrix}.$$

Умножая матричное равенство(4.13) слева на матрицу H и учитывая, что H — симметричная ортогональная матрица, получаем

$$A = H\widetilde{R}.$$

Из этого разложения получаем QR-разложение матрицы A, составив матрицу Q из первых n столбцов матрицы H, а матрицу R — из первых n строк матрицы $\widetilde{R}$. Построение QR-разложения с использованием матриц Хаусхолдера требует $4mn^2 - \frac{4n^3}{3}$ арифметических операций. Значит метод MGS построения QR-разложения в два раза эффективнее, чем использование матриц Хаусхолдера. Однако, вычисленная методом MGS матрица Q удовлетворяет условиям

$$Q^TQ = I + E_{MGS}, \quad ||E_{MGS}||_2 \approx \varepsilon \operatorname{cond}_2(A),$$

в то время как, используя матрицы Хаусхолдера, получаем такую матрицу Q, что

$$Q^TQ = I + E_H, \quad ||E_H||_2 \approx \varepsilon.$$

Поэтому, метод MGS применяют лишь в тех случаях, когда столбцы матрицы A "достаточно"линейно независимы[7].

Используя введенные выше обозначения, легко описать метод отражений решения системы линейных алгебраических уравнений. Сначала исходная система преобразуется в систему с верхней треугольной матрицей:

$$H_{n-1} \cdots H_2 H_1 Ax = H_{n-1} \cdots H_2 H_1 b. \tag{4.14}$$

На вычисление верхней треугольной матрицы этой системы необходимо $\frac{4n^3}{3} + O(n^2)$ арифметических операций. Решение треугольной системы (4.14) требует $O(n^2)$ арифметических операций. Значит, метод отражений требует в два раза больше арифметических операций, чем метод Гаусса с частичным выбором главного элемента. Однако метод отражений устойчив в машинной арифметике, поскольку ортогональные преобразования устойчивы к округлениям.

4.2. Решение ЛЗНК с помощью QR-разложения

Пусть

$$A = QR$$

— QR-разложение матрицы A. Положим

$$Ax - b = QRx - b = QRx - (QQ^T + I - QQ^T)b = Q(Rx - Q^Tb) + (QQ^T - I)b.$$

Поскольку векторы $Q(Rx - Q^Tb)$ и $(QQ^T - I)b$ ортогональны, ибо

$$(Q(Rx - Q^Tb))^T(QQ^T - I)b = (Rx - Q^Tb)^TQ^T(QQ^T - I)b = 0,$$

то по теореме Пифагора

$$||Ax-b||_2^2 = ||Q(Rx-Q^Tb)||_2^2+||(QQ^T-I)b||_2^2 = ||Rx-Q^Tb||_2^2+||(QQ^T-I)b||_2^2.$$

Последняя сумма будет минимальной, если ее первое слагаемое будет равно нулю, т. е., если

$$Rx = Q^Tb.$$

Значит, решение линейной задачи наименьших квадратов (4.2) можно записать в следующем виде:

$$x = R^{-1}Q^Tb.$$

Стоимость данного решения определяется стоимостью QR-разложения.

4.3. Решение ЛЗНК с помощью SVD-разложения

Пусть $A = U\Sigma V^T$ — сингулярное разложение матрицы A и $[U, \widetilde{U}] \in R^{m\times m}$ — ортогональная матрица. Тогда

$$||Ax-b||_2^2 = ||U\Sigma V^Tx-b||_2^2 = \left|\left|\left[\begin{array}{c} U^T \\ \widetilde{U}^T \end{array}\right](U\Sigma V^Tx-b)\right|\right|_2^2 =$$

$$= \left|\left|\left[\begin{array}{c} \Sigma V^Tx-U^Tb \\ -\widetilde{U}^Tb \end{array}\right]\right|\right|_2^2 = ||\Sigma V^Tx-U^Tb||_2^2 + ||\widetilde{U}^Tb||_2^2.$$

Последняя сумма будет минимальной, если первое слагаемое равно нулю, т. е., если

$$\Sigma V^Tx = U^Tb.$$

Отсюда находим решение линейной задачи наименьших квадратов

$$x = V\Sigma^{-1}U^Tb.$$

Для $m \gg n$ стоимость этого решения такая же, как и решения построенного с помощью QR-разложения, найденного методом MGS. Для меньших m стоимость равна $4n^2m - \frac{4}{3}n^3 + O(n^2)$.

Если матрица $A \in R^{m\times n}$ $(m \geq n)$ полного столбцового ранга и

$$A = QR = U\Sigma V^T,$$

то матрица

$$A^+ = (A^TA)^{-1}A^T = R^{-1}Q^T = V\Sigma^{-1}U^T$$

называется *псевдобратной* (Мура-Пенроуза) для матрицы A.

Псевдообратная матрица позволяет записать решение линейной задачи наименьших квадратов (4.2) с матрицей A полного столбцового ранга в виде

$$x = A^+b.$$

4.4. Обусловленность прямоугольных матриц

Обусловленность прямоугольной матрицы полного столбцового ранга определяется величиной

$$\mathrm{cond}_2(A) = \frac{\sigma_{max}}{\sigma_{min}}. \tag{4.15}$$

Следующий пример демонстрирует, что решение линейной задачи наименьших квадратов (4.2) с плохо обусловленной матрицей чувствительно к возмущениям.

Пример 4.4.1.

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 10^{-6} \\ 0 & 0 \end{bmatrix}, \quad \delta A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 10^{-6} \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad \delta b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Число обусловленности матрицы A равно $\mathrm{cond}_2(A) = 2 \cdot 10^6$.

Решение линейной задачи наименьших квадратов

$$\min_{x \in R^n} ||Ax - b||_2$$

равно

$$x = [1, 0]^T,$$

при этом невязка решения

$$r = b - Ax = [0, 0, 1]^T.$$

Решение возмущенной линейной задачи наименьших квадратов

$$\min_{x \in R^n} ||(A + \delta A)x - (b + \delta b)||_2$$

равно

$$\widetilde{x} = [1, 0.9999 \cdot 10^4]^T,$$

при этом невязка решения

$$\widetilde{r} = (b + \delta b) - (A + \delta A)\widetilde{x} = [0, -0.9999 \cdot 10^{-2}, 0.9999]^T.$$

4.5. Вопросы и задания

1. Доказать, что матрица $A^T A$ — симметрична и положительно определена, если матрица $A \in R^{m \times n}$ $(m \geq n)$ полного столбцового ранга.

2. Доказать, что классический (CGS) и модифицированный (MGS) алгоритмы Грама-Шмидта вычисления разложения $A = QR$ математически эквивалентны.

3. Используя классический метод Грама-Шмидта, построить QR-разложение матрицы
$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}^T.$$
Вычисления производить с тремя значащими цифрами.

4. Используя модифицированный метод Грама-Шмидта, построить QR-разложение матрицы
$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}^T.$$
Вычисления производить с тремя значащими цифрами.

5. Используя матрицы отражения, построить QR-разложение матрицы
$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}^T.$$
Вычисления производить с тремя значащими цифрами.

6. Решить линейную задачу наименьших квадратов
$$\min_x ||Ax - b||_2,$$
используя:

   а. нормальные уравнения;

   б. QR–разложение матрицы A,

   где
$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}^T, \quad b = [1, 1, 1]^T.$$

7. Решить линейную задачу наименьших квадратов
$$\min_x ||Ax - b||_2,$$
используя полученное SVD-разложение матрицы A.
$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}^T, \quad b = [1, 1, 1]^T.$$

8. Пусть для матрицы $A \in R^{m \times n}$ $(m \geq n)$ известно сингулярное разложение $A = U\Sigma V^T$. Выразить через U, V и Σ сингулярные разложения следующих матриц:

$$(A^T A)^{-1};$$
$$(A^T A)^{-1} A^T;$$
$$A(A^T A)^{-1};$$
$$A(A^T A)^{-1} A^T.$$

9. Пусть матрица $A \in R^{m \times n}$ $(m \geq n)$. Показать, что выбор $X = A^+$ (A^+ — псевдообратная матрица Мура-Пенроуза) минимизирует функцию $||AX - I||_F$ на множестве всех $n \times m$-матриц. Каково значение этого минимума?

10. Показать, что матрица A^+ — псевдообратная матрица Мура-Пенроуза для матрицы A — удовлетворяет следующим соотношениям:

$$AA^+A = A;$$
$$A^+AA^+ = A^+;$$
$$A^+A = (A^+A)^T;$$
$$AA^+ = (AA^+)^T.$$

5. Итерационные методы решения систем линейных алгебраических уравнений

Итерационный метод решения СЛАУ

$$Ax = b,\ A \in R^{n\times n},\ \det(A) \neq 0, \tag{5.1}$$

строит последовательность

$$x_0, x_1, \ldots, x_k, \ldots$$

приближенных решений системы (5.1). Если

$$\lim_{k\to\infty} ||x_k - x_*|| = 0,$$

где x_* — решение системы (5.1), итерационный метод называется сходящимся.

Для завершения итерационного процесса обычно используют условие

$$||x_{k+1} - x_k|| \leq \varepsilon ||x_k|| \tag{5.2}$$

или

$$||r_k|| \leq \varepsilon ||r_0||, \tag{5.3}$$

где ε — точность искомого решения, $r_k = b - Ax_k$ — невязка k–го приближения.

Сначала мы рассмотрим классические итерационные методы, а затем методы подпространства Крылова, которые в настоящее время широко используются для решения больших разреженных систем.

5.1. Классические итерационные методы

5.1.1. Метод простой итерации

Сначала систему,

$$Ax = b, \tag{5.4}$$

приведем к эквивалентной системе вида,

$$x = Gx + f. \tag{5.5}$$

Это можно сделать многими способами. Например, если положить

$$x = x - \alpha(Ax - b),\ \alpha \neq 0,$$

то $G = I - \alpha A,\ f = \alpha b$, где α — вещественный параметр, если же

$$x = x - B^{-1}(Ax - b),$$

то $G = I - B^{-1}A,\ f = B^{-1}b$. Как выбрать параметр α или невырожденную матрицу B, называемую предобуславливателем, будет объяснено позже.

Для нахождения решения системы (5.5) можно построить итерационный процесс,

$$x_{k+1} = Gx_k + f, \quad k = 0, 1, 2, \ldots, \tag{5.6}$$

здесь x_0 — начальное приближение, которое обычно выбирают произвольно. Итерационный процесс (5.6) называют методом простой итерации решения системы (5.5). Исследуем сходимость метода простой итерации.

Лемма 5.1.1. *Для того, чтобы последовательность матриц*

$$G^k = \underbrace{G \cdot \ldots \cdot G}_{k}$$

сходилась к нулевой матрице необходимо и достаточно, чтобы радиус спектра матрицы G был меньше единицы:

$$\rho(G) = \max_i |\lambda_i(G)| < 1.$$

Доказательство. Преобразованием подобия матрицу G можно привести к нормальной жордановой форме J, т. е., существует такая невырожденная матрица X, что

$$G = XJX^{-1}.$$

Тогда

$$G^k = \underbrace{XJX^{-1}XJX^{-1}\ldots XJX^{-1}}_{k} = XJ^kX^{-1}.$$

Заметим, что J — блочно-диагональная матрица, на диагонали которой расположены ящики Жордана

$$J_{\lambda_i} = \begin{bmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \lambda_i & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda_i \end{bmatrix}.$$

При возведении в степень матрицы J возводится в степень каждый ящик Жордана J_{λ_i}. Например,

$$\begin{bmatrix} \lambda & 1 & \\ & \lambda & 1 \\ & & \lambda \end{bmatrix}^2 = \begin{bmatrix} \lambda & 1 & \\ & \lambda & 1 \\ & & \lambda \end{bmatrix} \cdot \begin{bmatrix} \lambda & 1 & \\ & \lambda & 1 \\ & & \lambda \end{bmatrix} = \begin{bmatrix} \lambda^2 & 2\lambda & 1 \\ & \lambda^2 & 2\lambda \\ & & \lambda^2 \end{bmatrix},$$

$$\begin{bmatrix} \lambda & 1 & \\ & \lambda & 1 \\ & & \lambda \end{bmatrix}^3 = \begin{bmatrix} \lambda & 1 & \\ & \lambda & 1 \\ & & \lambda \end{bmatrix} \cdot \begin{bmatrix} \lambda^2 & 2\lambda & 1 \\ & \lambda^2 & 2\lambda \\ & & \lambda^2 \end{bmatrix} = \begin{bmatrix} \lambda^3 & 3\lambda^2 & 3\lambda \\ & \lambda^3 & 3\lambda^2 \\ & & \lambda^3 \end{bmatrix},$$

$$\begin{bmatrix} \lambda & 1 & \\ & \lambda & 1 \\ & & \lambda \end{bmatrix}^4 = \begin{bmatrix} \lambda & 1 & \\ & \lambda & 1 \\ & & \lambda \end{bmatrix} \cdot \begin{bmatrix} \lambda^3 & 3\lambda^2 & 3\lambda \\ & \lambda^3 & 3\lambda^2 \\ & & \lambda^3 \end{bmatrix} = \begin{bmatrix} \lambda^4 & 4\lambda^3 & 6\lambda^2 \\ & \lambda^4 & 4\lambda^3 \\ & & \lambda^4 \end{bmatrix}.$$

Значит, $J_{\lambda_i}^k \to 0$ при $k \to \infty$ тогда и только тогда, когда $|\lambda_i| < 1$. Следовательно, $G^k = XJ^kX^{-1} \to 0$ при $k \to \infty$ тогда и только тогда, когда

$$\max_i |\lambda_i(G)| < 1.$$

□

Теорема 5.1.1. (Необходимое и достаточное условие сходимости простой итерации). *Для того, чтобы итерационный процесс (5.6) сходился к решению уравнения (5.5) для любых начального приближения x_0 и правой части f необходимо и достаточно, чтобы радиус спектра матрицы перехода G был меньше единицы:*

$$\rho(G) = \max_i |\lambda_i(G)| < 1. \tag{5.7}$$

Доказательство. Достаточность. Пусть x_* — решение задачи (5.5), т. е.,

$$x_* = Gx_* + f.$$

Вычитая из (5.6) последнее равенство, получаем

$$x_k - x_* = G(x_{k-1} - x_*) = G^2(x_{k-2} - x_*) = \cdots = G^k(x_0 - x_*).$$

Значит

$$||x_k - x_*|| \le ||G^k||\,||x_0 - x_*||.$$

Поскольку $\rho(G) < 1$, то по предыдущей лемме из этого неравенства получаем

$$\lim_{k\to\infty} ||x_k - x_*|| = 0.$$

Значит, простая итерация является сходящимся итерационным процессом.

Необходимость. Предположим, что простая итерация сходится и тем не менее, существует такая собственная пара $\{\lambda, u\}$ матрицы G, что

$$|\lambda| \geq 1, \quad ||u|| = 1.$$

Положим, $x_0 = x_* + u$, где x_* — решение задачи (5.5). Тогда

$$x_k - x_* = G^k(x_0 - x_*) = G^k u = \lambda^k u.$$

Значит,

$$||x_k - x_*|| = |\lambda^k|.$$

Рассматривая предел этого равенства при $k \to \infty$, получаем противоречие, ибо левая часть сходится к нулю, а правая — к нулю не сходится. □

На практике необходимое и достаточное условие сходимости простой итерации проверить сложно, поскольку надо определять границы спектра матрицы G, что само по себе является сложной задачей.

Рассмотрим достаточные условия сходимости простой итерации, легко проверяемые в практических вычислениях.

Теорема 5.1.2. *Если $||G|| < 1$, то метод простой итерации (5.6) сходится.*

Доказательство. Пусть $\{\lambda, u\}$ ($||u|| = 1$) — собственная пара матрицы перехода G, т. е., $Gu = \lambda u$. Тогда

$$|\lambda|\,||u|| = ||Gu|| \leq ||G||\,||u||.$$

Значит,

$$|\lambda| \leq ||G||.$$

Поэтому, если $||G|| < 1$, то все собственные значения матрицы G тоже меньше 1 и по теореме 5.1.1 итерационный процесс (5.6) сходится. □

Получим еще одно достаточное условие сходимости простой итерации. Матрицу A системы

$$Ax = b \tag{5.8}$$

представим в виде

$$A = L + D + U,$$

где $D = \mathrm{diag}\{a_{11}, a_{22}, \ldots, a_{nn}\}$, L — нижняя треугольная матрица с нулевой диагональю, U — верхняя треугольная матрица с нулевой диагональю. Систему (5.8) можно записать так

$$x = -D^{-1}(L + U)x + D^{-1}b.$$

По теореме 5.1.2 итерационный процесс

$$x_{k+1} = -D^{-1}(L+U)x_k + D^{-1}b, \quad k = 0, 1, 2, \ldots \tag{5.9}$$

сходится, если

$$||D^{-1}(L+U)||_\infty < 1,$$

т. е., если

$$\max_i \frac{1}{|a_{ii}|} \sum_{j=1, j\neq i}^{n} |a_{ij}| < 1.$$

Это условие можно записать так:

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| < |a_{ii}| \quad i = \overline{1, n}. \tag{5.10}$$

Если вместо матричной ∞-нормы использовать матричную 1-норму, то аналогично можно получить еще одно условие сходимости итерационного процесса (5.9)

$$\sum_{i=1, i\neq j}^{n} |a_{ij}| < |a_{jj}| \quad j = \overline{1, n}. \tag{5.11}$$

Матрица A, удовлетворяющая условию (5.10) или (5.11) называется матрицей со строгим диагональным преобладанием.

Таким образом, мы получили еще одно достаточное условие сходимости простой итерации.

Теорема 5.1.3. *Если матрица A является матрицей со строгим диагональным преобладанием, то итерационный процесс (5.9) сходится.*

5.1.2. Показатель сходимости итерационного процесса

Вначале докажем теорему.

Теорема 5.1.4. *Для любой матрицы A*

$$\rho(A) = \lim_{k\to\infty} ||A^k||^{\frac{1}{k}}. \tag{5.12}$$

Доказательство. Поскольку

$$\rho(A)^k = \rho(A^k) \leq ||A^k||,$$

то

$$\rho(A) \leq ||A^k||^{\frac{1}{k}}. \tag{5.13}$$

Возьмем произвольное малое $\varepsilon > 0$. Спектральный радиус матрицы

$$\widetilde{A} = (\rho(A) + \varepsilon)^{-1} A$$

меньше единицы. Значит,

$$\lim_{k\to\infty} ||\widetilde{A}^k|| = 0.$$

Поэтому существует такое $N = N(A, \varepsilon)$, что

$$||\widetilde{A}^k|| < 1 \quad \forall k \geq N.$$

Значит,

$$||A^k|| < (\rho(A) + \varepsilon)^k \quad \forall k \geq N,$$

или

$$||A^k||^{\frac{1}{k}} < (\rho(A) + \varepsilon) \quad \forall k \geq N. \tag{5.14}$$

Тогда из (5.13) и (5.14) (в силу произвольности ε) следует (5.12)

□

Ранее для погрешности приближения x_k решения x_* системы

$$x = Gx + f,$$

найденного из итерационного процесса

$$x_{k+1} = Gx_k + f, \quad k = 0, 1, 2, \cdots, \tag{5.15}$$

была получена оценка

$$||x_k - x_*|| \leq ||G^k||\,||x_0 - x_*||. \tag{5.16}$$

Поскольку по теореме 5.1.4

$$||G^k|| \approx \rho(G)^k,$$

то из (5.16) следует, что количество итераций k итерационного процесса (5.15), необходимое для уменьшения начальной ошибки в e раз, можно найти из уравнения

$$\rho(G)^k = \frac{1}{e}.$$

Отсюда

$$k \ln \rho(G) = -1,$$

или

$$k = -\frac{1}{\ln \rho(G)}. \tag{5.17}$$

Величину $\nu = -\ln \rho(G)$ называют показателем сходимости итерационного процесса (5.15). Из (5.17) следует, что величина, обратная к показателю сходимости итерационного процесса, равна числу итераций, которые необходимо уменьшат начальную ошибку $||x_0 - x_*||$ в e раз. Таким образом, чем больше показатель сходимости итерационного процесса, тем быстрее итерационный процесс сходится.

5.1.3. Простая итерация с оптимальным параметром

Рассмотрим систему

$$Ax = b \tag{5.18}$$

с симметричной положительно определенной матрицей $A \in R^{n\times n}$. Представив систему (5.18) в виде

$$x = x - \alpha(Ax - b),$$

построим итерационный процесс

$$x_{k+1} = (I - \alpha A)x_k + \alpha b, \quad k = 0, 1, 2, \cdots, \tag{5.19}$$

где α — параметр, подлежащий определению. По теореме 5.1.1 итерационный процесс (5.19) сходится тогда и только тогда, когда радиус спектра матрицы перехода $G = (I - \alpha A)$ меньше единицы, т. е., если

$$\max_i |\lambda_i(I - \alpha A)| < 1. \tag{5.20}$$

Это условие будет выполняться, если

$$\begin{cases} 1 - \alpha\lambda_i(A) < 1, \\ -(1 - \alpha\lambda_i(A)) < 1, \quad i = \overline{1, n}, \end{cases}$$

или

$$\begin{cases} 0 < \alpha\lambda_i(A), \\ \alpha < \frac{2}{\lambda_i(A)}, \quad i = \overline{1, n}. \end{cases} \tag{5.21}$$

Поскольку матрица A системы (5.18) симметрична и положительно определена, то

$$0 < m \le \lambda_i(A) \le M, \quad i = \overline{1, n}. \tag{5.22}$$

Из (5.21) и (5.22) следует, что необходимое и достаточное условие сходимости итерационного процесса (5.19) выполняется, если

$$0 < \alpha < \frac{2}{M}. \tag{5.23}$$

Параметр α, удовлетворяющий условию (5.23), выберем так, чтобы показатель сходимости итерационного процесса (5.19)

$$\nu = -\ln \rho(G) = -\ln \max_i |1 - \alpha\lambda_i(A)|$$

был максимальным. Таким образом, для выбора параметра α получаем условие

$$\min_{0<\alpha<\frac{2}{M}} \max_i |1 - \alpha\lambda_i(A)|. \tag{5.24}$$

Поскольку собственные значения $\lambda_i(A)$ матрицы A неизвестны (задача определения собственных значений матрицы A много сложнее задачи нахождения решения системы $Ax = b$), то вместо задачи (5.24) рассмотрим более общую задачу:

$$\min_{0<\alpha<\frac{2}{M}} \max_{m\le\lambda\le M} |1-\alpha\lambda|. \tag{5.25}$$

Чтобы решить эту задачу, рассмотрим графики функции

$$\varphi_\alpha(\lambda) = 1 - \alpha\lambda$$

при различных параметрах $\alpha \in (0, M)$ (см. рис.5.1). Очевидно, что наименьший из максимумов

$$\max_{m\le\lambda\le M} |1-\alpha\lambda|$$

достигается при таком значении параметра α, что

$$\varphi_\alpha\left(\frac{m+M}{2}\right) = 1 - \alpha\frac{m+M}{2} = 0.$$

Значит, оптимальное значение параметра α равно

$$\alpha_{opt} = \frac{2}{m+M}. \tag{5.26}$$

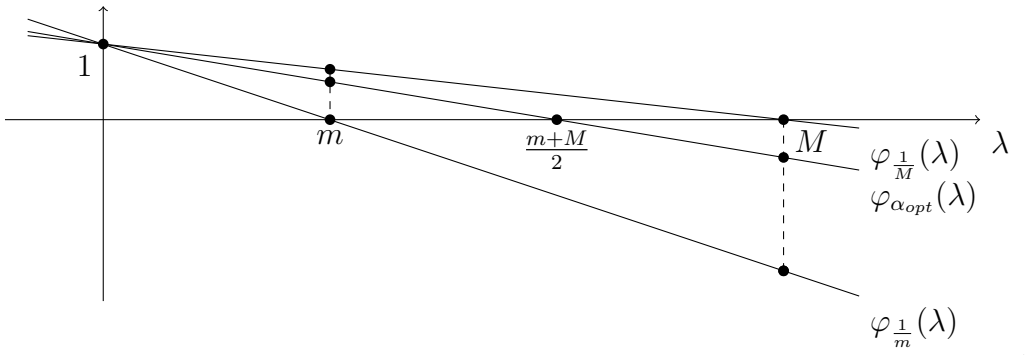

Рис. 5.1. Графики функции $\varphi_\alpha(\lambda) = 1 - \alpha\lambda$ при различных значениях параметра α

Определим теперь показатель сходимости простой итерации с оптимальным параметром. Поскольку

$$\rho(G) = \rho(I - \alpha_{opt}A) = \varphi_{\alpha_{opt}}(m) = -\varphi_{\alpha_{opt}}(M) =$$

$$= 1 - \frac{2}{m+M}m = \frac{M-m}{M+m} = \frac{1-\frac{m}{M}}{1+\frac{m}{M}} = \frac{1-\frac{1}{\mathrm{cond}_c(A)}}{1+\frac{1}{\mathrm{cond}_c(A)}},$$

то показатель сходимости

$$\nu = -\ln \rho(G) = -\ln \frac{1 - \frac{1}{\mathrm{cond}_c(A)}}{1 + \frac{1}{\mathrm{cond}_c(A)}} =$$

$$= \ln\left(1 + \frac{1}{\mathrm{cond}_c(A)}\right) - \ln\left(1 - \frac{1}{\mathrm{cond}_c(A)}\right) =$$

$$= \frac{2}{\mathrm{cond}_c(A)} + O\left(\frac{2}{\mathrm{cond}_c^2(A)}\right).$$

Таким образом, простая итерация с оптимальным выбором параметра имеет показатель сходимости

$$\nu \approx \frac{2}{\mathrm{cond}_c(A)}. \tag{5.27}$$

5.1.4. Метод Зейделя

Приведем систему

$$Ax = b \tag{5.28}$$

к эквивалентному виду

$$x = Gx + f. \tag{5.29}$$

Матрицу G представим в виде суммы двух матриц

$$G = G_1 + G_2,$$

где G_1 — нижняя треугольная матрица с нулями на главной диагонали, G_2 — верхняя треугольная матрица. Методом Зейделя (Гаусса-Зейделя) решения системы (5.28) (или эквивалентной системы (5.29)) называется итерационный процесс

$$x_{k+1} = G_1 x_{k+1} + G_2 x_k + f, \quad k = 0, 1, 2, \cdots . \tag{5.30}$$

Заметим, что переставив уравнения в системе (5.28) мы получим другой итерационный процесс (5.30). Существует $n!$ различных вариантов метода Зейделя решения системы (5.28) (ибо столько же существует перестановок n переменных). Вопрос о том, какая из перестановок переменных приводит к методу Зейделя с наивысшей скоростью сходимости, требует дополнительных исследований с учетом свойств матрицы A.

Теорема 5.1.5. *Метод Зейделя (5.29) сходится тогда и только тогда, когда радиус спектра матрицы $(I - G_1)^{-1}G_2$ меньше единицы:*

$$\rho((I - G_1)^{-1}G_2) = \max_i |\lambda_i((I - G_1)^{-1}G_2)| < 1. \tag{5.31}$$

Доказательство. Метод Зейделя (5.30) можно записать следующим образом:

$$x_{k+1} = (I - G_1)^{-1} G_2 x_k + (I - G_1)^{-1} f, \quad k = 0, 1, 2, \cdots .$$

А это метод простой итерации с матрицей перехода $(I - G_1)^{-1} G_2$. Поэтому, утверждение теоремы следует из теоремы 5.1.1 о необходимом и достаточном условии сходимости простой итерации. □

Рассмотрим систему

$$Ax = b \tag{5.32}$$

с симметричной положительно определенной матрицей $A \in R^{n \times n}$. Представив матрицу A в виде суммы

$$A = L + D + L^T,$$

где $D = \mathrm{diag}\{a_{11}, a_{22}, \ldots, a_{nn}\}$, L — нижняя треугольная матрица с нулями на диагонали, систему (5.32) запишем в эквивалентном виде

$$Dx = -Lx - L^T x + b.$$

Для этого уравнения построим итерационный процесс

$$x_{k+1} = -D^{-1} L x_{k+1} - D^{-1} L^T x_k + D^{-1} b, \quad k = 0, 1, 2, \ldots . \tag{5.33}$$

Теорема 5.1.6. *Если матрица $A \in R^{n \times n}$ симметрична и положительно определена, то метод Зейделя (5.33) сходится при любом x_0.*

Доказательство. Итерационный процесс (5.33) можно записать в виде простой итерации

$$x_{k+1} = -(D + L)^{-1} L^T x_k + (D + L)^{-1} b, \quad k = 0, 1, 2, \ldots . \tag{5.34}$$

Поэтому, учитывая теорему 5.1.1 о необходимом и достаточном условии сходимости простой итерации, достаточно показать, что все собственные значения матрицы $G = -(D + L)^{-1} L^T$ по модулю меньше единицы. Поскольку матрица G и матрица

$$G_1 = -D^{\frac{1}{2}} (D + L)^{-1} L^T D^{-\frac{1}{2}} =$$

$$= -D^{\frac{1}{2}} (D^{\frac{1}{2}} (I + D^{-\frac{1}{2}} L D^{-\frac{1}{2}}) D^{\frac{1}{2}})^{-1} L^T D^{-\frac{1}{2}} =$$

$$= -D^{\frac{1}{2}} D^{-\frac{1}{2}} (I + D^{-\frac{1}{2}} L D^{-\frac{1}{2}})^{-1} D^{-\frac{1}{2}} L^T D^{-\frac{1}{2}} =$$

$$= -(I + L_1)^{-1} L_1^T,$$

где $L_1 = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$, имеют одинаковые собственные значения, то надо убедиться, что

$$\rho(G_1) < 1.$$

Если $\{\lambda, x\}$ ($\lambda \in C$, $x \in C^n$, $x^H x = 1$) — произвольная собственная пара матрицы G, т. е.,

$$G_1 x = \lambda x,$$

то

$$-L_1^T x = \lambda(I + L_1)x.$$

Следовательно,

$$-x^H L_1^T x = \lambda(I + x^H L_1 x).$$

Положив

$$x^H L_1^T x = a + ib,$$

получаем

$$|\lambda|^2 = \left| \frac{a+ib}{1+a+ib} \right|^2 = \frac{a^2+b^2}{1+2a+a^2+b^2}.$$

Однако, в силу положительной определенности матрицы

$$D^{-\frac{1}{2}} A D^{-\frac{1}{2}} = I + L_1 + L_1^T$$

нетрудно показать, что

$$0 < I + x^H L_1 x + x^H L_1^T x = 1 + 2a.$$

Откуда следует, что $|\lambda| < 1$. □

5.1.5. Геометрическая интерпретация метода Зейделя

Рассмотрим пример. Применим метод Зейделя для нахождения решения системы

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}. \tag{5.35}$$

Точное решение системы $x_* = [2; 1]^T$. Поскольку матрица системы симметрична и положительно определена (легко проверить, что матрица системы имеет положительные собственные значения: $\lambda_1 = 1$; $\lambda_2 = 3$), то метод Зейделя, в данном случае, можно записать в виде сходящегося итерационного процесса (5.33):

$$x_{k+1} = -\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix} x_{k+1} - \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} x_k + \\ + \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 3 \\ 0 \end{bmatrix}, \quad k = 0, 1, 2, \ldots,$$

т. е.,

$$\begin{cases} x_1^{(k+1)} = 0.5x_2^{(k)} + 1.5, \\ x_2^{(k+1)} = 0.5x_1^{(k+1)}, \quad k = 0, 1, 2, \ldots. \end{cases} \tag{5.36}$$

Пусть начальное приближение $x_0 = [4; 3]^T$. Выполним два шага итерационного процесса (5.36):

$$\begin{aligned} x_1^{(1)} &= 0.5 * 3 + 1.5 = 3, \\ x_2^{(1)} &= 0.5 * 3 = 1.5, \\ x_1^{(2)} &= 0.5 * 1.5 + 1.5 = 2.25, \\ x_2^{(2)} &= 0.5 * 2.25 = 1.125. \end{aligned}$$

Очевидно, что итерационный процесс сходится к решению x_*. Рисунок 5.2 иллюстрирует сходимость метода Зейделя для данного примера.

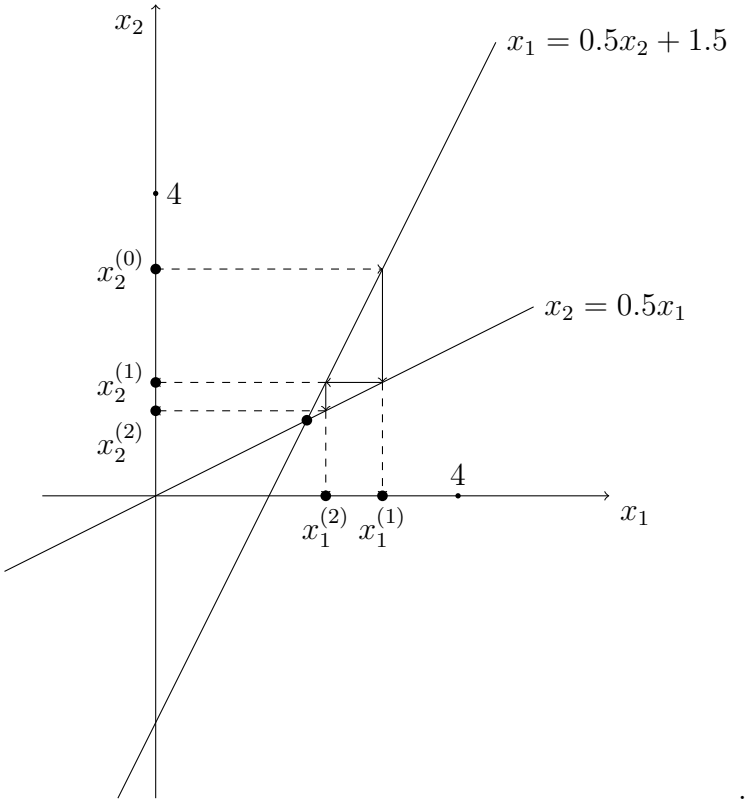

Рис. 5.2. Сходимость метода Зейделя (5.36)

Поменяв местами уравнения в системе (5.35), получим систему

$$\begin{bmatrix} -1 & 2 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \end{bmatrix}.$$

Метод Зейделя для этой системы можно записать так:

$$\begin{cases} x_1^{(k+1)} = 2x_2^{(k)}, \\ x_2^{(k+1)} = 2x_1^{(k+1)} - 3, \quad k = 0, 1, 2, \ldots. \end{cases} \tag{5.37}$$

Здесь

$$G_1 = \begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix}, \quad G_2 = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}.$$

Поскольку радиус спектра матрицы $(I - G_1)^{-1}G_2$ равен 4, то из теоремы 5.1.5 о необходимом и достаточном условии сходимости метода Зейделя следует, что итерационный процесс (5.37) расходится. Действительно, пусть $x_0 = [2.25; 1.25]$. Выполним два шага итерационного процесса (5.37):

$$\begin{aligned} x_1^{(1)} &= 2 * 1.25 = 2.5, \\ x_2^{(1)} &= 2 * 2.5 - 3 = 2, \\ x_1^{(2)} &= 2 * 4 = 4, \\ x_2^{(2)} &= 2 * 4 - 3 = 5. \end{aligned}$$

Рисунок 5.3 иллюстрирует расходимость метода Зейделя для данного примера.

В предыдущих примерах сходимость и расходимость метода Зейделя были односторонними. Рисунок 5.4 иллюстрирует двустороннюю сходимость, а рисунок 5.5 — двустороннюю расходимость метода Зейделя для одной и той же системы.

На примере метода Зейделя легко показать, что близость двух соседних итераций не всегда означает их близость к решению (см. рис 5.6). Заметим, что на этом рисунке прямые, изображающие уравнения системы почти параллельны, т. е. уравнения системы почти совпадают, а значит, система плохо обусловлена.

5.2. Многочлены Чебышева

Многочлен Чебышева степени $n \geq 0$ на отрезке $[-1, 1]$ определяется формулой

$$T_n(x) = \cos(n \arccos x). \tag{5.38}$$

В частности

$$\begin{aligned} T_0(x) &= \cos(0 \cdot \arccos x) = 1, \\ T_1(x) &= \cos(1 \cdot \arccos x) = x. \end{aligned} \tag{5.39}$$

Используя формулу произведения косинусов

$$\cos\alpha \cos\beta = \frac{1}{2}(cos(\alpha - \beta) + cos(\alpha + \beta)),$$

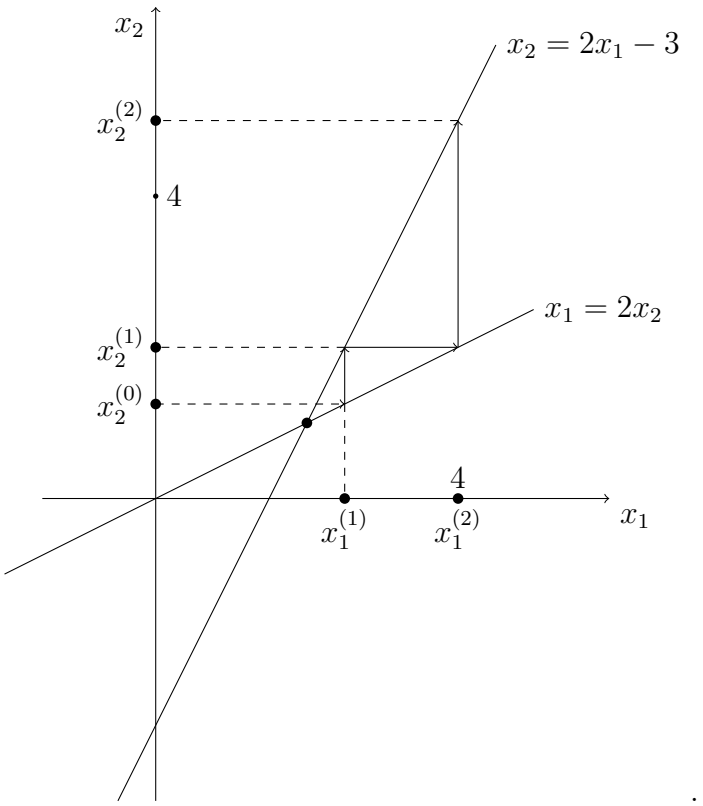

Рис. 5.3. Расходимость метода Зейделя (5.37)

находим

$$\cos(n+1)\varphi = 2\cos\varphi \cdot \cos n\varphi - \cos(n-1)\varphi.$$

Полагая $\varphi = \arccos x$, в соответствии с (5.38), имеем

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n = 0, 1, 2, \ldots. \tag{5.40}$$

Учитывая (5.39), из этой рекуррентной формулы получаем

$$\begin{aligned}
&T_0(x) = 1,\\
&T_1(x) = x,\\
&T_2(x) = 2x^2 - 1,\\
&T_3(x) = 4x^3 - 3x,\\
&T_4(x) = 8x^4 - 8x^2 + 1,\\
&\cdots .
\end{aligned}$$

Таким образом, $T_n(x)$ действительно является многочленом степени n. Ниже приведены основные свойства многочленов Чебышева, некоторые из которых очевидны.

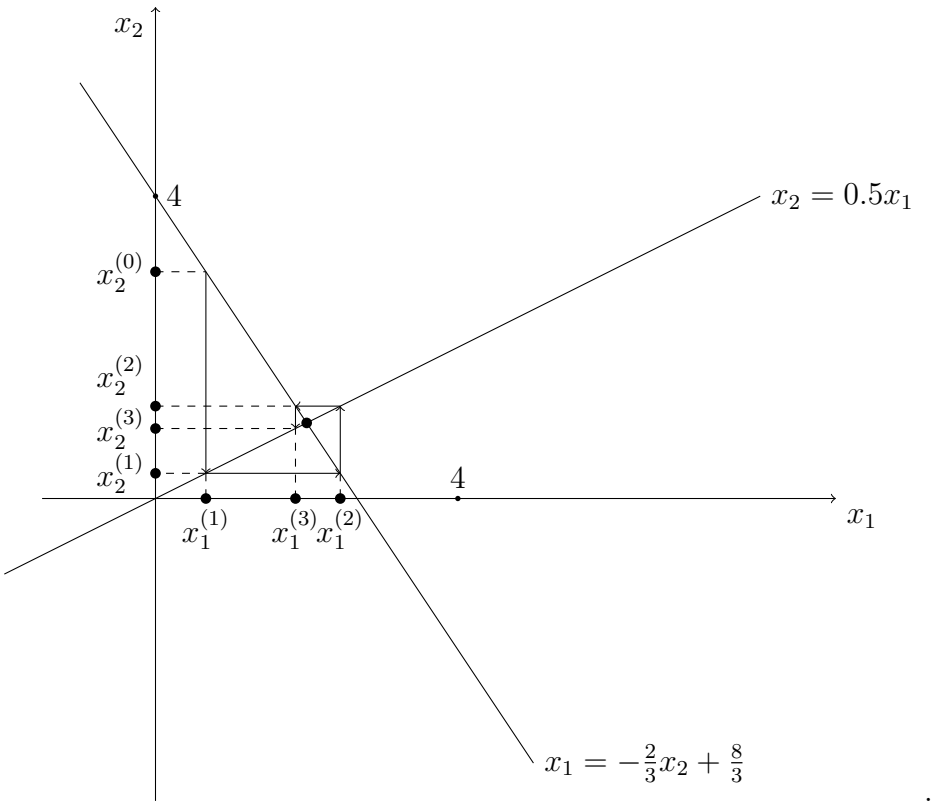

Рис. 5.4. Двусторонняя сходимость метода Зейделя

Свойство 5.2.1. *При четном (нечетном) n многочлен Чебышева $T_n(x)$ является четной (нечетной) функцией.*

Свойство 5.2.2. *Старший коэффициент многочлена Чебышева $T_n(x)$ равен 2^{n-1}.*

Свойство 5.2.3. *Многочлен Чебышева $T_n(x)$ имеет n корней в интервале $(-1, 1)$, выражаемых формулой*

$$x_i = \cos\frac{(2i+1)\pi}{2n}, \quad i = \overline{0, n-1}. \tag{5.41}$$

Доказательство. Действительно,

$$T_n(x_i) = \cos(n \arccos x_i) = \cos\frac{(2i+1)\pi}{2} = 0, \quad i = \overline{0, n-1}.$$

Поскольку у многочлена степени n ровно n корней, то других корней нет. □

Свойство 5.2.4.

$$\max_{x\in[-1,1]} |T_n(x)| = 1, \tag{5.42}$$

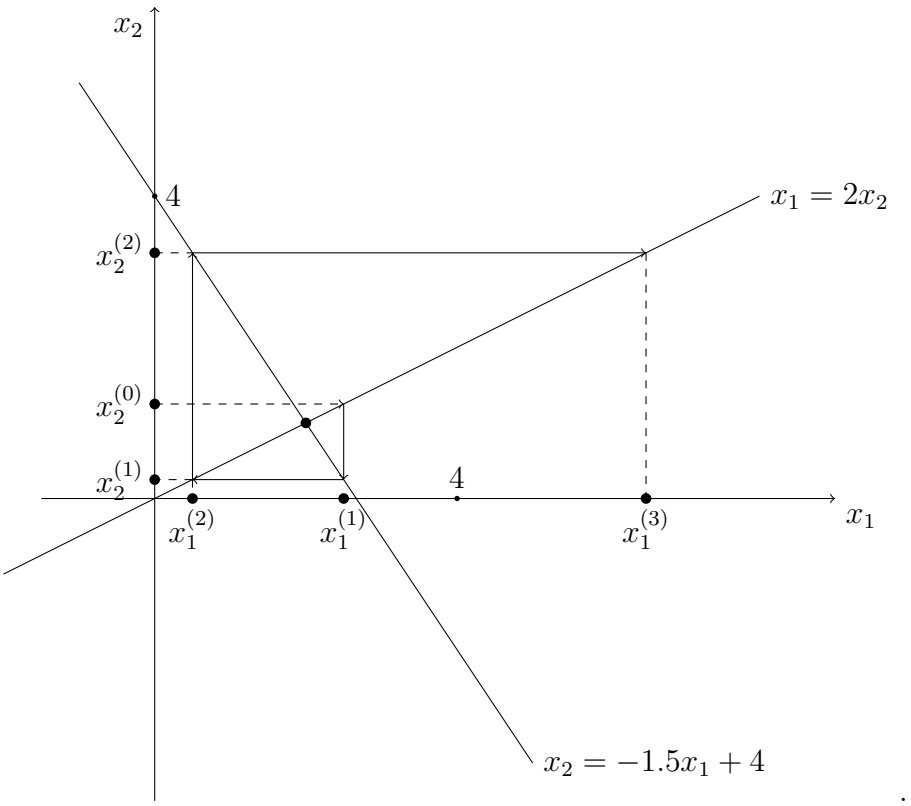

Рис. 5.5. Двусторонняя расходимость метода Зейделя

причем $T_n(x_m) = (-1)^m$, *где*

$$x_m = \cos\left(\frac{m\pi}{n}\right), \quad m = \overline{0,n}. \tag{5.43}$$

Доказательство. Действительно,

$$T_n(x_m) = \cos(m\pi) = (-1)^m.$$

С другой стороны, согласно (5.38) $|T_n(x)| \le 1$ при $x \in [-1,1]$. □

Свойство 5.2.5. *Пусть* $P_n(x)$ *— произвольный многочлен степени* n *со старшим коэффициентом равным 1, тогда*

$$\max_{x\in[-1,1]} |P_n(x)| \ge \max_{x\in[-1,1]} |\overline{T}_n(x)| = 2^{1-n}, \tag{5.44}$$

где $\overline{T}_n(x) = 2^{1-n}T_n(x)$.

Доказательство. Предположим противное. Пусть существует такой многочлен $P_n(x) = x^n + \alpha_{n-1}x^{n-1} + \cdots + \alpha_1 x + \alpha_0$, что

$$\max_{x\in[-1,1]} |P_n(x)| < \max_{x\in[-1,1]} |\overline{T}_n(x)| = 2^{1-n}.$$

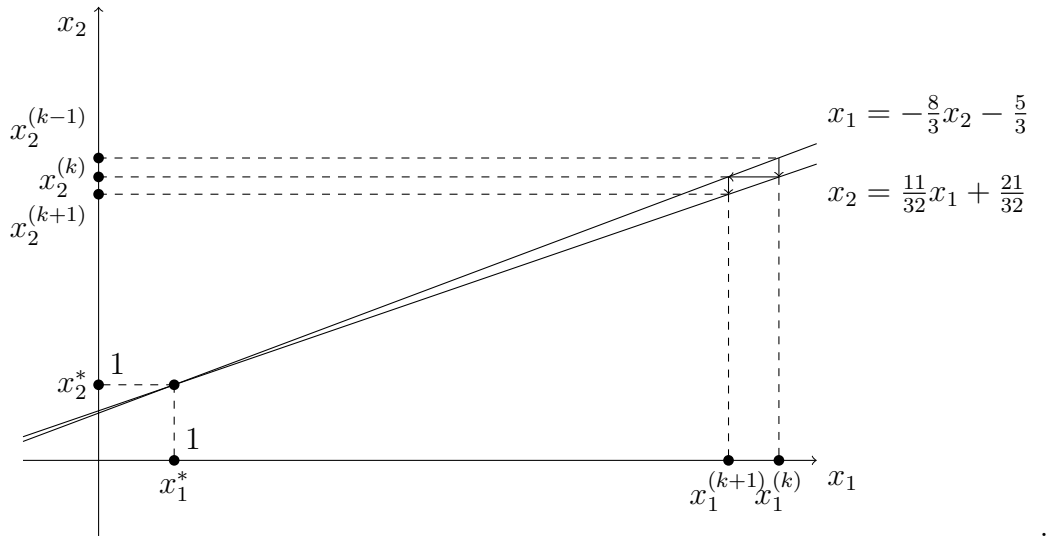

Рис. 5.6. Сходимость метода Зейделя для плохо обусловленной системы

Тогда многочлен $\overline{T}_n(x) - P_n(x)$ имеет степень $n-1$ и отличен от нуля. В то же время

$$\operatorname{sign}(\overline{T}_n(x_m) - P_n(x_m)) = \operatorname{sign}((-1)^m 2^{1-n} - P_n(x_m)) = (-1)^m,\ m = \overline{0,n},$$

где x_m — точки экстремума (5.43) многочлена $T_n(x)$. Таким образом, многочлен $\overline{T}_n(x) - P_n(x)$ между точками x_m и x_{m+1} $(m = \overline{0,n-1})$ меняет знак. Значит, этот многочлен имеет n корней. Получили противоречие, ибо отличный от нуля многочлен степени $n-1$ не может иметь n корней. □

Многочлен $\overline{T}_n(x) = 2^{1-n}T_n(x)$ называется многочленом наименьшего отклонения от нуля.

Свойство 5.2.6. *Многочлен Чебышева $T_n(x)$ можно представить следующим образом*

$$T_n(x) = \frac{(x - \sqrt{x^2-1})^n}{2} + \frac{(x + \sqrt{x^2-1})^n}{2}, \tag{5.45}$$

причем $T_n(-1) = (-1)^n$, $T_n(1) = 1$.

Доказательство будет приведено позже.

5.3. Метод Ричардсона

Метод Ричардсона решения системы

$$Ax = b \tag{5.46}$$

с симметричной положительно определенной матрицей $A \in R^{n\times n}$ определяется следующим итерационным процессом:

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k(Ax_k - b), \quad k = \overline{0, p-1}, \\ \text{после каждых } & \quad p \quad \text{шагов полагают } x_0 = x_p. \end{aligned} \tag{5.47}$$

Пусть x_* — решение задачи (5.46). Тогда

$$x_* = x_* - \alpha_k(Ax_* - b), \quad k = \overline{0, p-1}. \tag{5.48}$$

Вычитая (5.48) из (5.47), получим

$$x_{k+1} - x_* = x_k - x_* - \alpha_k A(x_k - x_*) = (I - \alpha_k A)(x_k - x_*) =$$

$$= (I - \alpha_k A)(I - \alpha_{k-1}A)(x_{k-1} - x_*) = \cdots =$$
$$= (I - \alpha_k A)(I - \alpha_{k-1}A)\dots(I - \alpha_0 A)(x_0 - x_*).$$

Отсюда, положив $k = p - 1$, имеем

$$x_p - x_* = G_p(x_0 - x_*),$$

где

$$G_p = (I - \alpha_{p-1}A)(I - \alpha_{p-2}A)\dots(I - \alpha_0 A).$$

Значит, если цикл (5.47) повторить m раз, то получим приближение x_{mp}. Причем

$$x_{mp} - x_* = G_p(x_{m(p-1)} - x_*) = G_p^2(x_{m(p-2)} - x_*) = \cdots = G_p^m(x_0 - x_*).$$

Значит,

$$||x_{mp} - x_*|| \le ||G_p^m||\,||x_0 - x_*||. \tag{5.49}$$

На основании леммы 5.1.1 отсюда следует, что метод Ричардсона (5.47) сходится при фиксированном p и $m \to \infty$, если

$$\rho(G_p) < 1. \tag{5.50}$$

По теореме 5.1.4 о радиусе спектра матрицы

$$||G_p^m|| \approx \rho^m(G_p)$$

при больших m. Тогда из (5.49) следует, что количество итераций $k = mp$ достаточных для уменьшения начальной ошибки $||x_0 - x_|$ в e раз можно найти из уравнения

$$\rho^m(G_p) = \frac{1}{e}.$$

Отсюда

$$mp\frac{\ln \rho(G_p)}{p} = -1,$$

$$k = mp = \frac{1}{-\ln \rho(G_p)/p},$$

т. е.,

$$\nu = -\frac{\ln \rho(G_p)}{p} \tag{5.51}$$

— показатель сходимости метода Ричардсона.

Параметры $\alpha_0, \ldots, \alpha_{p-1}$ метода Ричардсона выбирают так, чтобы показатель сходимости был наибольшим, т. е., из условия

$$\min_{\alpha_0,\ldots,\alpha_{p-1}} \max_i |\lambda_i(G_p)|. \tag{5.52}$$

Заметим, что

$$\lambda(G_p) = (1 - \alpha_{p-1}\lambda(A))(1 - \alpha_{p-2}\lambda(A)) \cdots (1 - \alpha_0\lambda(A)) \equiv P_p(\lambda(A)).$$

Поэтому задачу (5.52) можно записать в таком виде

$$\min_{\alpha_0,\ldots,\alpha_{p-1}} \max_i |P_p(\lambda_i(A))|. \tag{5.53}$$

Чтобы решить эту задачу, необходимо знать собственные значения матрицы A. Поскольку матрица A симметрична и положительно определена, то

$$0 < m \le \lambda_i(A) \le M, \quad i = \overline{1, n}.$$

Предполагая, что константы m и M известны, задачу (5.53) заменяют более общей задачей:

$$\min_{\alpha_0,\ldots,\alpha_{p-1}} \max_{m \le \lambda \le M} |P_p(\lambda)|. \tag{5.54}$$

Решим эту задачу. Заменой переменных

$$x = \frac{2}{M-m}\lambda - \frac{M+m}{M-m} \quad \left(\lambda = \frac{M-m}{2}x + \frac{M+m}{2}\right)$$

промежуток $[m, M]$ отображается в промежуток $[-1, 1]$. Решением задачи (5.54) будет многочлен, равный многочлену наименее уклоняющемуся от нуля:

$$(1-\alpha_{p-1}\lambda)(1-\alpha_{p-2}\lambda)\cdots(1-\alpha_0\lambda) = \frac{T_p(\frac{2}{M-m}\lambda - \frac{M+m}{M-m})}{T_p(-\frac{M+m}{M-m})}. \qquad (5.55)$$

Корни многочлена в левой части равенства (5.55):

$$\lambda_i = \frac{1}{\alpha_i}, \quad i = \overline{0, p-1}.$$

Корни многочлена в правой части равенства(5.55):

$$\lambda_i = \frac{M-m}{2}\cos\frac{(2i+1)\pi}{2p} + \frac{M+m}{2}, \quad i = \overline{0, p-1}.$$

Из условия совпадения этих корней получаем

$$\alpha_i = \left(\frac{M-m}{2}\cos\frac{(2i+1)\pi}{2p} + \frac{M+m}{2}\right)^{-1}, \quad i = \overline{0, p-1}. \qquad (5.56)$$

Этот набор параметров итерационного процесса (5.48) называется чебышевским. При таком выборе параметров итерационного процесса (5.48)

$$\rho(G_p) = \max_{m\le\lambda\le M}|(1-\alpha_{p-1}\lambda)(1-\alpha_{p-2}\lambda)\cdots(1-\alpha_0\lambda)| = \frac{1}{|T_p(-\frac{M+m}{M-m})|}. \qquad (5.57)$$

Поскольку

$$\frac{M+m}{M-m} = 1 + \frac{2m}{M-m} > 1,$$

то

$$\left|T_p\left(-\frac{M+m}{M-m}\right)\right| = T_p\left(\frac{M+m}{M-m}\right) > 1.$$

Поэтому условие (5.50) сходимости итерационного процесса выполняется.

Вычислим теперь показатель сходимости метода Ричардсона (5.51). Для вычисления

$$T_p\left(\frac{M+m}{M-m}\right)$$

воспользуемся следующим представлением многочлена Чебышева:

$$T_p(x) = \frac{(x-\sqrt{x^2-1})^p}{2} + \frac{(x+\sqrt{x^2-1})^p}{2}.$$

Если положить

$$x = \frac{M+m}{M-m} = \frac{1+\frac{1}{\mathrm{cond}_c A}}{1-\frac{1}{\mathrm{cond}_c A}} = \frac{1+\xi}{1-\xi},$$

где $\xi = \frac{1}{\mathrm{cond}_c A}$, то

$$x + \sqrt{x^2-1} = \frac{1+\xi}{1-\xi} + \sqrt{\frac{(1+\xi)^2}{(1-\xi)^2} - 1} = \frac{1+\xi}{1-\xi} + \frac{2\sqrt{\xi}}{1-\xi} = \frac{1+\sqrt{\xi}}{1-\sqrt{\xi}},$$

$$x - \sqrt{x^2-1} = \frac{1+\xi}{1-\xi} - \frac{2\sqrt{\xi}}{1-\xi} = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}.$$

Полагая

$$\eta = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}$$

имеем

$$T_p\left(\frac{M+m}{M-m}\right) = \frac{\eta^p}{2} + \frac{1}{2\eta^p} = \frac{\eta^{2p}+1}{2\eta^p}. \tag{5.58}$$

Значит,

$$\rho(G_p) = \frac{1}{T_p(\frac{M+m}{M-m})} = \frac{2\eta^p}{\eta^{2p}+1}.$$

Тогда коэффициент сходимости метода Ричардсона

$$\nu = -\frac{\ln \rho(G_p)}{p} = -\frac{1}{p}\ln\frac{2\eta^p}{\eta^{2p}+1} = -\frac{1}{p}\ln\eta^p + \frac{1}{p}\ln\frac{\eta^{2p}+1}{2} = -\ln\eta + \frac{1}{p}\ln\frac{\eta^{2p}+1}{2}.$$

Поэтому при больших p можно положить

$$\nu \approx -\ln\eta = -\ln\frac{1-\sqrt{\xi}}{1+\sqrt{\xi}} = 2\sqrt{\xi} + O(\xi).$$

Значит коэффициент сходимости метода Ричардсона

$$\nu \approx 2\sqrt{\xi} = \frac{2}{\sqrt{\mathrm{cond}_c A}}. \tag{5.59}$$

Наконец, сравнивая коэффициенты сходимости (5.27) и (5.59), можно сделать вывод, что для системы с положительно определенной симметричной матрицей метод Ричардсона сходится быстрее простой итерации с оптимальным параметром.

5.4. Итерационные методы подпространств Крылова

Для матрицы $A \in R^{n\times n}$ и произвольного вектора $v \in R^n$ m-е подпространство Крылова определяется следующим образом:

$$\mathcal{K}_m(A, v) = \text{span}\,\{v, Av, \ldots, A^{m-1}v\}. \tag{5.60}$$

Хорошо известно, что подпространства Крылова образуют вложенную последовательность подпространств, наибольшая размерность которых

$$d \equiv \dim \mathcal{K}_n(A, v) \leq n,$$

т. е.,

$$\mathcal{K}_1(A, v) \subset \cdots \subset \mathcal{K}_d(A, v) = \cdots = \mathcal{K}_n(A, v).$$

В частности, для любого $m \leq d$ размерность подпространства Крылова $\mathcal{K}_m(A, v)$ равна m.

Алгоритм Арнольди строит ортонормированный базис подпространства Крылова $\mathcal{K}_m(A, v)$. Фактически Алгоритм Арнольди — это процесс Грама-Шмидта построения ортонормированного базиса $\{v_1, v_2, \ldots, v_m\}$ для множества векторов $\{v, Av, \ldots, A^{m-1}v\}$.

Алгоритм 5.4.1. ***Алгоритм Арнольди***

1. $v_1 = v/||v||_2$
2. *For* $j = 1, 2, \ldots, m$ *Do*
3. *For* $i = 1, 2, \ldots, j$ *Do*
4. $h_{ij} = (Av_j, v_i)$
5. *End Do*
6. $w_j = Av_j - \sum_{i=1}^{j} h_{ij}v_i$
7. $h_{j+1,j} = ||w_j||_2$
8. *If* $h_{j+1,j} = 0$ *Then Stop*
9. $v_{j+1} = w_j/h_{j+1,j}$
10. *End Do*

Теорема 5.4.1. *Имеют место следующие равенства:*

$$AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^T = \tag{5.61}$$

$$= V_{m+1}\overline{H}_m, \tag{5.62}$$

$$V_m^T A V_m = H_m. \tag{5.63}$$

Здесь $V_m = [v_1, v_2, \ldots, v_m]$, H_m *— верхняя Хессенбергова матрица размера* $m \times m$ *с элементами* h_{ij}, *матрица* $\overline{H}_m$ *получается из матрицы* H_m *добавлением строки* $[0, \ldots, 0, h_{m+1,m}]$.

Доказательство. Из строк с номерами 6 и 9 алгоритма Арнольди следуют равенства

$$Av_j = \sum_{i=1}^{j+1} h_{ij} v_i \qquad j = \overline{1, m}. \tag{5.64}$$

Записывая равенства (5.64) в матричном виде, получаем (5.61) и (5.62). Равенство (5.63) следует из (5.61), поскольку вектор v_{m+1} ортогонален столбцам матрицы V_m. □

5.4.1. Метод Арнольди

Метод Арнольди является проекционным методом решения системы линейных алгебраических уравнений $Ax = b$. Другое его название — FOM (Full Orthogonalization Method). Метод строится следующим образом. Пусть x_0 — произвольное начальное приближение, $r_0 = b - Ax_0$ — невязка начального приближения, $\mathcal{K}_m(A, r_0)$ — подпространство Крылова и $\dim \mathcal{K}_m(A, r_0) = m$. Найдем такой вектор

$$x_m \in x_0 + \mathcal{K}_m(A, r_0), \tag{5.65}$$

что

$$r_m \equiv b - Ax_m \perp \mathcal{L} = \mathcal{K}_m(A, r_0). \tag{5.66}$$

Используя алгоритм Арнольди, построим ортонормированный базис

$$V_m = [v_1, v_2, \ldots, v_m]$$

подпространства Крылова $\mathcal{K}_m(A, r_0)$, где $v_1 = r_0/\beta$, $\beta = ||r_0||_2$. Тогда

$$V_m^T r_0 = V_m^T \beta v_1 = \beta e_1.$$

Вектор x_m будем искать в виде

$$x_m = x_0 + V_m y_m. \tag{5.67}$$

Используя условие (5.66) для определения вектора y_m, получаем цепочку равенств

$$V_m^T r_m = V_m^T (b - Ax_m) = V_m^T (b - A(x_0 + V_m y_m)) = V_m^T r_0 - V_m^T A V_m y_m =$$

$$= V_m^T \beta v_1 - H_m y_m = \beta e_1 - H_m y_m = 0.$$

Из последнего равенства в этой цепочке получаем уравнение для определения вектора y_m

$$H_m y_m = \beta e_1. \tag{5.68}$$

Отсюда

$$y_m = H_m^{-1}\beta e_1.$$

Зная y_m, по формуле (5.67) находим приближение x_m к решению системы $Ax = b$, построенное за m шагов метода Арнольди.

Имеет место следующая теорема.

Теорема 5.4.2. *Если x_m — приближенное решение системы $Ax = b$ построенное за m шагов метода Арнольди, то*

$$r_m \equiv b - Ax_m = -h_{m+1,m}e_m^T y_m v_{m+1}, \tag{5.69}$$

поэтому

$$||b - Ax_m||_2 = h_{m+1,m}|e_m^T y_m|. \tag{5.70}$$

Доказательство. Из (5.67), (5.61), (5.63) и (5.68)

$$b - Ax_m = b - Ax_0 - AV_m y_m = r_0 - AV_m y_m =$$

$$= \beta v_1 - V_m H_m y_m - h_{m+1,m}e_m^T y_m v_{m+1} =$$

$$= \beta v_1 - V_m \beta e_1 - h_{m+1,m}e_m^T y_m v_{m+1} = -h_{m+1,m}e_m^T y_m v_{m+1}.$$

Для доказательства (5.70) осталось заметить, что вектор v_{m+1} единичной длины. □

Из (5.70) следует, что для вычисления нормы невязки r_m нет необходимости вычислять x_m и саму невязку $r_m = b - Ax_m$. Итерации метода Арнольди повторяют, пока не выполняется неравенство

$$||r_m||_2 = h_{m+1,m}|e_m^T y_m| \le \varepsilon ||r_0||_2$$

для заданного ε. После завершения итерационного процесса вычисляем $x_m = V_m y_m$.

Приведем алгоритм метода Арнольди.

Алгоритм 5.4.2. ***Метод Арнольди (FOM)***

1. $r_0 = b - Ax_0; \beta = ||r_0||_2; v_1 = r_0/\beta$
2. $j = 0$
3. *Repeat*
4. $\quad j = j + 1$
5. $\quad$ *For* $i = 1, 2, \ldots, j$ *Do*
6. $\quad\quad h_{ij} = (Av_j, v_i)$
7. $\quad$ *End Do*
8. $\quad w_j = Av_j - \sum_{i=1}^{j} h_{ij}v_i$
9. $\quad h_{j+1,j} = ||w_j||_2$

10. $If\ h_{j+1,j} = 0\ Then\ Stop$
11. $v_{j+1} = w_j / h_{j+1,j}$
12. $y_j = H_j^{-1} \beta e_1$
13. $Until\ h_{j+1,j} |e_j^T y_j| < \varepsilon ||r_0||_2$
14. $x_j = x_0 + V_j y_j$

5.4.2. Метод обобщенной минимизации невязки

В англоязычной литературе этот метод носит название GMRES (Generalized Minimum Residual Method).

Пусть x_0 — начальное приближение к решению системы $Ax = b$; $r_0 = b - Ax_0$ — невязка начального приближения x_0; столбцы матрицы V_m, построенной алгоритмом Арнольди, образуют ортонормированный базис подпространства Крылова $\mathcal{K}_m(A, r_0)$. Тогда любой вектор

$$x \in x_0 + \mathcal{K}_m(A, r_0)$$

можно представить в виде

$$x = x_0 + V_m y.$$

Определим функционал

$$J(y) = ||b - Ax||_2 = ||b - A(x_0 + V_m y)||_2.$$

Из (5.62) следует, что

$$b - Ax = b - A(x_0 + V_m y) = r_0 - AV_m y = \beta v_1 - V_{m+1} \overline{H}_m y =$$

$$= V_{m+1}(\beta e_1 - \overline{H}_m y).$$

Поскольку столбцы матрицы V_{m+1} ортонормированны, то матрица V_{m+1} сохраняет 2–норму вектора. Поэтому

$$J(y) = ||b - Ax||_2 = ||V_{m+1}(\beta e_1 - \overline{H}_m y)||_2 = ||\beta e_1 - \overline{H}_m y||_2. \qquad (5.71)$$

Рассмотрим линейную задачу наименьших квадратов

$$\min_{y \in \mathcal{K}_m(A, r_0)} ||\beta e_1 - \overline{H}_m y||_2. \qquad (5.72)$$

Пусть $y_m = \overline{H}_m^+ \beta e_1$ — решение задачи (5.72), где $\overline{H}_m^+$ — псевдообратная матрица. Тогда

$$x_m = x_0 + V_m y_m$$

является приближенным решением системы $Ax = b$, построенным за m шагов алгоритма GMRES. Из (5.71),(5.72) следует, что метод GMRES минимизирует невязку на подпространстве Крылова $\mathcal{K}_m(A, r_0)$. Более того, из (5.71) получаем

$$||r_m||_2 = ||b - Ax_m||_2 = ||\beta e_1 - \overline{H}_m y_m||_2.$$

Приведем теперь алгоритм метода GMRES.

Алгоритм 5.4.3. ***GMRES***

1. $r_0 = b - Ax_0; \beta = ||r_0||_2; v_1 = r_0/\beta$
2. $j = 0$
3. *Repeat*
4. $\quad j = j + 1$
5. $\quad For\ i = 1, 2, \ldots, j\ Do$
6. $\quad\quad h_{ij} = (Av_j, v_i)$
7. $\quad End\ Do$
8. $\quad w_j = Av_j - \sum_{i=1}^{j} h_{ij} v_i$
9. $\quad h_{j+1,j} = ||w_j||_2$
10. $\quad If\ h_{j+1,j} = 0\ Then\ Stop$
11. $\quad v_{j+1} = w_j/h_{j+1,j}$
12. $\quad y_j = \overline{H}_j^{+} \beta e_1$
13. $Until\ ||\beta e_1 - \overline{H}_j y_j||_2 < \varepsilon ||r_0||_2$
14. $x_j = x_0 + V_j y_j$

5.4.3. Метод Ланцоша

В начале рассмотрим алгоритм Ланцоша построения ортонормированного базиса подпространства Крылова $\mathcal{K}_m(A, v)$ для симметричной матрицы A. Алгоритм Ланцоша — это алгоритм Арнольди для симметричной матрицы A.

Итак, пусть $A = A^T \in R^{n \times n}$. Напомним, что алгоритм Арнольди строит матрицу

$$V_m = [v_1, v_2, \ldots, v_m],$$

столбцы которой образуют ортонормированный базис $\mathcal{K}_m(A, v)$. При этом

$$V_m^T A V_m = H_m, \tag{5.73}$$

где H_m — верхняя Хессенбергова матрица. Поскольку матрица A симметрична, то матрица H_m тоже симметричная, а значит, трехдиагональная. Обозначим

$$\alpha_i = h_{ii}, \quad i = \overline{1, m},$$
$$\beta_i = h_{i-1,i}, \quad i = \overline{2, m},$$

где h_{ij} — элементы матрицы H_m. Определим матрицу

$$T_m = \begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \cdot & \cdot & \cdot & \\ & & \beta_{m-1} & \alpha_{m-1} & \beta_m \\ & & & \beta_m & \alpha_m \end{bmatrix}$$

Тогда (5.73) приобретет вид

$$V_m^T A V_m = T_m. \tag{5.74}$$

Равенство (5.61) в новых обозначениях выглядит так

$$AV_m = V_m T_m + \beta_{m+1} v_{m+1} e_m^T. \tag{5.75}$$

Записывая матричное равенство (5.75) по-векторно, получим следующий алгоритм Ланцоша.

Алгоритм 5.4.4. ***Алгоритм Ланцоша***
1. $v_1 = v/||v||_2; \beta_1 = 0; v_0 = 0$
2. $For\ j = 1, 2, \ldots, m\ Do$
3. $\quad w_j = Av_j - \beta_j v_{j-1}$
4. $\quad \alpha_j = w_j^T v_j$
5. $\quad w_j = w_j - \alpha_j v_j$
6. $\quad \beta_{j+1} = ||w_j||_2$
7. $\quad If\ \beta_{j+1} = 0\ Then\ Stop$
8. $\quad v_{j+1} = w_j/\beta_{j+1}$
9. $End\ Do$

Действительно, из (5.75) имеем

$$Av_j = \beta_j v_{j-1} + \alpha_j v_j + \beta_{j+1} v_{j+1}, \quad j = \overline{1, m}, \tag{5.76}$$

где $v_0 = 0$, $\beta_1 = 1$. Предположим, что

$$v_1, v_2, \ldots, v_j, \quad \alpha_1, \alpha_2, \ldots, \alpha_{j-1}, \quad \beta_2, \ldots, \beta_j$$

уже определены и на j–м шаге необходимо определить α_j, β_{j+1} и v_{j+1}. Вначале обозначим

$$w_j = \alpha_j v_j + \beta_{j+1} v_{j+1}. \tag{5.77}$$

Тогда из (5.76) можно определить

$$w_j = Av_j - \beta_j v_{j-1}.$$

Учитывая (5.77),

$$\alpha_j = w_j^T v_j.$$

Положим теперь

$$w_j = w_j - \alpha_j v_j.$$

Снова учитывая (5.77), находим

$$\beta_{j+1} = ||w_j||_2, \quad v_{j+1} = w_j/\beta_{j+1}.$$

На этом j–й шаг алгоритма заканчивается.

Рассмотрим теперь метод Ланцоша решения систем линейных алгебраических уравнений. Фактически — это метод Арнольди (FOM) для симметричной матрицы. Как и метод Арнольди, метод Ланцоша за m шагов находит приближенное решение

$$x_m = x_0 + V_m y_m,$$

где

$$y_m = H_m^{-1}(\beta e_1) = T_m^{-1}(\beta e_1).$$

При этом формулы (5.69),(5.70) приобретают вид

$$r_m \equiv b - Ax_m = -\beta_{m+1} e_m^T y_m v_{m+1}, \tag{5.78}$$

поэтому

$$||b - Ax_m||_2 = \beta_{m+1}|e_m^T y_m|. \tag{5.79}$$

Запишем теперь алгоритм метода Ланцоша.

Алгоритм 5.4.5. ***Метод Ланцоша решения СЛАУ***

1. $r_0 = b - Ax_0; \beta = ||r_0||_2; v_1 = r_0/\beta$
2. $j = 0; \beta_1 = 0; v_0 = 0$
3. *Repeat*
4. $\quad j = j + 1$
5. $\quad w_j = Av_j - \beta_j v_{j-1}$
6. $\quad \alpha_j = w_j^T v_j$
7. $\quad w_j = w_j - \alpha_j v_j$
8. $\quad \beta_{j+1} = ||w_j||_2$
9. $\quad If\ \beta_{j+1} = 0\ Then\ Stop$
10. $\quad v_{j+1} = w_j/\beta_{j+1}$
11. $\quad t_{jj} = \alpha_j; t_{j+1,j} = \beta_{j+1}$
12. $\quad y_j = T_j^{-1}(\beta e_1)$
13. $\quad t_{j,j+1} = \beta_{j+1}$
14. *Until* $\beta_{j+1}|e_j^T y_j| < \varepsilon||r_0||_2$
15. $x_j = x_0 + V_j y_j$

5.4.4. Метод сопряженных градиентов

Метод сопряженных градиентов (CG — Conjugate Gradient) получим, усовершенствуя метод Ланцоша. Если матрица системы $Ax = b$ симметрична и положительно определена, то симметричная трехдиагональная матрица T_m тоже положительно определена. Действительно, из (5.74) $\forall x \in R^m$, $x \neq 0$, получаем

$$x^T T_m x = x^T V_m^T A V_m x = (V_m x)^T A V_m x > 0.$$

Значит, для матрицы T_m существует LU-разложение

$$T_m = L_m U_m.$$

Тогда

$$x_m = x_0 + V_m T_m^{-1}(\beta e_1) = x_0 + V_m U_m^{-1} L_m^{-1}(\beta e_1).$$

Обозначим

$$P_m = V_m U_m^{-1}, \tag{5.80}$$
$$z_m = L_m^{-1}\beta e_1.$$

Тогда

$$x_m = x_0 + P_m z_m. \tag{5.81}$$

Из (5.80) следует, что последний столбец матрицы V_m можно выразить через предпоследний и последний столбцы матрицы P_m:

$$p_{m-1}u_{m-1,m} + p_m u_{mm} = v_m,$$

или

$$p_m = \frac{1}{u_{mm}}(v_m - u_{m-1,m}p_{m-1}). \tag{5.82}$$

Лемма 5.4.1. *Пусть x_m — приближенное решение системы $Ax = b$ ($A = A^T \in R^{n\times n}$), построенное за m шагов метода Ланцоша. Тогда*

$$r_j^T r_i = 0, \quad i \neq j, \ (i, j = \overline{1, m}), \tag{5.83}$$

векторы p_i A-сопряжены, т. е.,

$$p_i^T A p_j = 0, \quad i \neq j, \ (i, j = \overline{1, m}). \tag{5.84}$$

Доказательство. (5.83) сразу следует из (5.78), поскольку векторы v_i $(i = \overline{1, m+1})$ ортонормированы. Учитывая (5.80), получаем

$$P_m^T A P_m = (V_m U_m^{-1})^T A V_m U_m^{-1} = U_m^{-T} V_m^T A V_m U_m^{-1} =$$
$$= U_m^{-T} T_m U_m^{-1} = U_m^{-T} L_m U_m U_m^{-1} = U_m^{-T} L_m.$$

Матрица $U_m^{-T} L_m$ нижняя треугольная, как произведение нижних треугольных матриц, с другой стороны, — симметрична, значит, она диагональная. □

Приближенное решение x_m системы линейных алгебраических уравнений будем искать в виде (5.81). Примем обозначение

$$z_m = [\alpha_1, \alpha_2, \ldots, \alpha_m]^T.$$

Компоненты α_i вектора z_m подлежат определению. Из (5.81) следует

$$x_j = x_{j-1} + \alpha_j p_j \quad j = 1, 2, \ldots \tag{5.85}$$

Отсюда

$$r_j = r_{j-1} - \alpha_j A p_j \quad j = 1, 2, \ldots. \tag{5.86}$$

По предыдущей лемме векторы r_j ортогональны друг другу. Поэтому

$$r_j^T r_{j-1} = (r_{j-1} - \alpha_j A p_j)^T r_{j-1} = r_{j-1}^T r_{j-1} - \alpha_j (A p_j)^T r_{j-1} = 0.$$

Из последнего равенства

$$\alpha_j = \frac{r_{j-1}^T r_{j-1}}{p_j^T A r_{j-1}}. \tag{5.87}$$

Из (5.78) и (5.82) следует, что вектор p_m является линейной комбинацией векторов p_{m-1} и r_{m-1}:

$$p_m = \frac{1}{u_{mm}} \left(\frac{r_{m-1}}{-\beta_m e_{m-1}^T y_{m-1}} - u_{m-1,m} p_{m-1} \right).$$

Отсюда получаем рекуррентную формулу

$$\widetilde{p}_{j+1} = r_j + \beta_j \widetilde{p}_j, \quad j = 1, 2, \ldots,$$

где $\widetilde{p}_j$ — это вектор p_j, умноженный на некоторое число. Заметим, что векторы $\widetilde{p}_j$ тоже A-сопряжены. В дальнейшем вместо векторов p_j будем использовать векторы $\widetilde{p}_j$, сохранив предыдущие обозначения. Таким образом, имеем рекуррентную формулу

$$p_{j+1} = r_j + \beta_j p_j, \quad j = 1, 2, \ldots. \tag{5.88}$$

Учитывая A-сопряженность векторов p_j и симметричность матрицы A, получаем

$$(A p_j)^T r_j = (A p_j)^T (p_{j+1} - \beta_j p_j) = -\beta_j p_j^T A p_j.$$

Теперь можно определить

$$\beta_j = -\frac{p_j^T A r_j}{p_j^T A p_j}. \tag{5.89}$$

Из формул (5.85)–(5.89) получаем одну из реализаций метода сопряженных градиентов (CG).

Алгоритм 5.4.6. ***Метод сопряженных градиентов (CG)***

1. $r_0 = b - Ax_0; p_1 = r_0$
2. $j = 0$
3. *Repeat*
4. $\quad j = j + 1$
5. $\quad \alpha_j = \frac{r_{j-1}^T r_{j-1}}{p_j^T A r_{j-1}}$
6. $\quad x_j = x_{j-1} + \alpha_j p_j$
7. $\quad r_j = r_{j-1} - \alpha_j A p_j$
8. $\quad \beta_j = -\frac{p_j^T A r_j}{p_j^T A p_j}$
9. $\quad p_{j+1} = r_j + \beta_j p_j$
10. *Until* $||r_j||_2 < \varepsilon ||r_0||_2$

5.4.5. Сходимость метода сопряженных градиентов

Лемма 5.4.2. *Пусть матрица $A \in R^{n\times n}$ симметрична и положительно определена. Тогда приближенное решение x_m системы $Ax = b$, построенное за m шагов метода сопряженных градиентов минимизирует невязку $r_m = b - Ax_m$ в норме $||\cdot||_{A^{-1}}$ на подпространстве Крылова $\mathcal{K}_m(A, r_0)$.*

Доказательство. Поскольку матрица A симметрична и положительно определена, то и обратная матрица A^{-1} тоже симметрична и положительно определена. Поэтому в пространстве R^n можно определить векторную норму следующим образом:

$$||x||_{A^{-1}}^2 = x^T A^{-1} x.$$

Пусть $\widehat{x}_m = x_m + P_m \widehat{y}_m$. Тогда $\widehat{r}_m = b - A\widehat{x}_m = r_m - AP_m\widehat{y}_m$ и

$$||\widehat{r}_m||_{A^{-1}}^2 = ||r_m - AP_m\widehat{y}_m||_{A^{-1}}^2 =$$

$$= ||r_m||_{A^{-1}}^2 - 2r_m^T A^{-1} AP_m\widehat{y}_m + ||AP_m\widehat{y}_m||_{A^{-1}}^2.$$

Столбцы матрицы P_m образуют базис подпространства Крылова $\mathcal{K}_m(A, r_0)$, вектор r_m ортогонален этому подпространству в силу (5.78), поэтому $r_m^T \widehat{P}_m y_m = 0$. Значит,

$$||\widehat{r}_m||_{A^{-1}}^2 = ||r_m||_{A^{-1}}^2 + ||AP_m\widehat{y}_m||_{A^{-1}}^2. \tag{5.90}$$

Из (5.90) следует, что задача минимизации

$$\min_{\widehat{y}_m \in \mathcal{K}_m(A,r_0)} ||\widehat{r}_m||_{A^{-1}}^2$$

имеет решение, если $AP_m\widehat{y}_m = 0$, т. е., если $\widehat{y}_m = 0$. □

Пусть начальное приближение $x_0 = 0$, тогда невязка этого приближения $r_0 = b - Ax_0 = b$. По предыдущей лемме, x_m минимизирует невязку r_m в норме $||\cdot||_{A^{-1}}$ на подпространстве $\mathcal{K}_m(A, r_0)$, т. е., x_m минимизирует функцию

$$f(z) \equiv ||b - Az||^2_{A^{-1}} = (b - Az)^T A^{-1}(b - Az) = (Ax - Az)^T A^{-1}(Ax - Az) =$$

$$= (A(x - z))^T (x - z) = (x - z)^T A(x - z)$$

по всем векторам $z \in \mathcal{K}_m(A, r_0) = \mathcal{K}_m(A, b)$. Поскольку

$$\mathcal{K}_m(A, b) = \text{span}\,\{b, Ab, A^2b, \ldots, A^{m-1}b\},$$

вектор z представим в виде

$$z = \sum_{j=0}^{m-1} \alpha_j A^j b = p_{m-1}(A)b = p_{m-1}(A)Ax,$$

где $p_{m-1}(t)$ — многочлен степени $m - 1$. Поэтому

$$f(z) = ((I - p_{m-1}(A)A)x)^T A((I - p_{m-1}(A)A)x) = (q_m(A)x)^T A(q_m(A)x) =$$

$$= x^T q_m(A) A q_m(A) x,$$

ибо $(q_m(A))^T = q_m(A)$ в силу симметричности матрицы A. Здесь $q_m(t)$ — многочлен степени m, такой, что $q_m(0) = 1$. Обозначим через Q_m множество всех многочленов степени m, принимающих значение 1 в нуле. Тогда

$$f(x_m) = \min_{z \in \mathcal{K}_m(A,b)} f(z) = \min_{q_m \in Q_m} x^T q_m(A) A q_m(A) x. \qquad (5.91)$$

Чтобы упростить (5.91), используем спектральное разложение $A = U\Lambda U^T$. Положим $y = U^T x$. Тогда

$$f(x_m) = \min_{q_m \in Q_m} x^T q_m(A) A q_m(A) x =$$

$$= \min_{q_m \in Q_m} x^T q_m(U\Lambda U^T) U\Lambda U^T q_m(U\Lambda U^T) x =$$

$$= \min_{q_m \in Q_m} x^T U q_m(\Lambda)\Lambda q_m(\Lambda) U^T x =$$

$$= \min_{q_m \in Q_m} y^T q_m(\Lambda)\Lambda q_m(\Lambda) y =$$

$$= \min_{q_m \in Q_m} y^T \text{diag}\,(q_m(\lambda_i)\lambda_i q_m(\lambda_i)) y =$$

$$= \min_{q_m \in Q_m} \sum_{i=1}^{m} y_i^2 \lambda_i q_m^2(\lambda_i).$$

Таким образом,

$$f(x_m) = \min_{q_m \in Q_m} \sum_{i=1}^{m} y_i^2 \lambda_i q_m^2(\lambda_i). \tag{5.92}$$

Заметим, что

$$f(x_0) = x^T A x = x^T U \Lambda U^T x = y^T \Lambda y = \sum_{i=1}^{m} \lambda_i y_i^2.$$

Значит,

$$f(x_m) \le \left[\min_{q_m \in Q_m} \{ \max_{\lambda_i \in \lambda(A)} q_m^2(\lambda_i) \} \right] \sum_{i=1}^{m} \lambda_i y_i^2 =$$

$$= \min_{q_m \in Q_m} \{ \max_{\lambda_i \in \lambda(A)} q_m^2(\lambda_i) \} f(x_0).$$

Поэтому

$$\frac{||r_m||_{A^{-1}}^2}{||r_0||_{A^{-1}}^2} = \frac{f(x_m)}{f(x_0)} \le \min_{q_m \in Q_m} \max_{\lambda_i \in \lambda(A)} q_m^2(\lambda_i),$$

т. е.,

$$\frac{||r_m||_{A^{-1}}}{||r_0||_{A^{-1}}} \le \min_{q_m \in Q_m} \max_{\lambda_i \in \lambda(A)} |q_m(\lambda_i)|, \tag{5.93}$$

Матрица A симметричная и положительно определена, значит, спектр $\lambda(A)$ матрицы A положителен, т. е., $0 < \lambda_{min} \le \lambda(A) \le \lambda_{max}$. Тогда

$$\min_{q_m \in Q_m} \max_{\lambda_i \in \lambda(A)} |q_m(\lambda_i)| \le \min_{q_m \in Q_m} \max_{[\lambda_{min}, \lambda_{max}]} |q_m(\lambda)|.$$

Учитывая свойство 5.2.5 многочлена Чебышева о наименьшем отклонении от нуля и (5.58), получаем

$$\min_{q_m \in Q_m} \max_{[\lambda_{min}, \lambda_{max}]} |q_m(\lambda)| = T_m^{-1} \left(\frac{\lambda_{max} + \lambda_{min}}{\lambda_{max} - \lambda_{min}} \right) = \frac{2\eta^m}{\eta^{2m} + 1},$$

где

$$\eta = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{1}{\mathrm{cond}_c(A)}.$$

Значит,

$$||r_m||_{A^{-1}} \le \frac{2\eta^m}{\eta^{2m} + 1} ||r_0||_{A^{-1}}.$$

Отсюда следует, что количество итераций m алгоритма сопряженных градиентов, достаточных для уменьшения начальной невязки в e раз, можно найти из уравнения

$$\frac{2\eta^m}{\eta^{2m} + 1} = \frac{1}{e}.$$

Тогда

$$m \ln \eta - \ln \frac{\eta^{2m} + 1}{2} = -1$$

и

$$m \approx \frac{-1 + \ln 0.5}{\ln \eta}.$$

Поэтому показатель сходимости метода сопряженных градиентов

$$\nu = \frac{\ln \eta}{-1 + \ln 0.5} = \frac{2}{1 - \ln 0.5} \sqrt{\xi} + O(\xi),$$

т. е.,

$$\nu \approx \frac{2}{1 - \ln 0.5} \frac{1}{\sqrt{\mathrm{cond}_c(A)}}. \tag{5.94}$$

Значит, показатель сходимости метода сопряженных градиентов только коэффициентом отличается от показателя (5.59) сходимости метода Ричардсона.

Во многих случаях оценка (5.93) сильно занижает скорость сходимости метода сопряженных градиентов. Например, предположим, что матрица A имеет $k \ll n$ различных собственных значений. Тогда при отсутствии округлений метод сопряженных градиентов построит решение за k шагов. Действительно, если в (5.91) в качестве $q_m(t)$ взять аннулирующий многочлен матрицы A

$$q_k(t) = \frac{\prod_{i=1}^{k} (\lambda_i - t)}{\prod_{i=1}^{k} \lambda_i},$$

то $q_k(A) = 0 \Rightarrow ||r_k||_{A^{-1}} = 0 \Rightarrow r_k = 0 \Rightarrow x_k$ — точное решение системы $Ax = b$.

5.5. Предобуславливание

Ранее было показано, что скорость сходимости метода сопряженных градиентов зависит от числа обусловленности матрицы или, более общо, от распределения ее собственных значений. Этим же свойством обладают и другие итерационные методы. Чтобы увеличить скорость сходимости итерационных методов, используют предобуславливание. Предобуславливание состоит в следующем. Система

$$Ax = b$$

заменяется системой

$$M^{-1}Ax = M^{-1}b.$$

Матрица M, называемая предобуславливателем, должна быть близкой к матрице A и обладать следующими свойствами:

1. матрица $M^{-1}A$ должна быть хорошо обусловлена, или ее собственные значения должны быть сосредоточены на периферии спектра;

2. система $My = f$ должна легко решаться.

Продуманный, зависящий от конкретной задачи, выбор матрицы M может сделать число обусловленности матрицы $M^{-1}A$ много меньшим числа обусловленности матрицы A, и тем самым в огромной степени ускорить сходимость итерационного алгоритма. Часто без предобуславливателя итерационный метод вообще не сходится.

Приведем некоторые приемы построения предобуславливателей.

Если диагональные элементы матрицы A сильно различаются по величине, то можно использовать простой диагональный предобуславливатель, который называется предобуславливателем Якоби:

$$M = \mathrm{diag}(a_{11}, \ldots, a_{nn}).$$

Блочный предобуславливатель Якоби строится следующим образом. Матрица A записывается в блочном виде с квадратными диагональными блоками

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1k} \\ \vdots & \ddots & \vdots \\ A_{k1} & \cdots & A_{kk} \end{bmatrix}.$$

Затем, полагают

$$M = \begin{bmatrix} A_{11} & & 0 \\ & \ddots & \\ 0 & & A_{kk} \end{bmatrix}.$$

Рассмотрим еще один способ построения предобуславливателей. Матрицу A представим в виде

$$A = LU + R, \tag{5.95}$$

где матрицы в правой части удовлетворяют следующим свойствам:

1. матрицы L и U являются нижнетреугольной с 1 на главной диагонали и верхнетреугольной соответственно;

2. $P_L \subset P_A$ и $P_U \subset P_A$;

3. $\forall (i,j) \in P_A : [LU]_{ij} = [A]_{ij}$;

4. $P_A \cap P_R = \emptyset$.

Здесь множество индексов $P_A = \{(i,j) : a_{ij} \neq 0\}$ — портрет матрицы A. Аналогично определяются P_L, P_U и P_R. Тогда приближенное представление $A \approx LU$ называется неполной (incomplete) LUфакторизацией матрицы A или ее ILU-разложением. Для нахождения матриц L и U будем генерировать их построчно. Предположим, что первые $(k-1)$ строк уже найдены и необходимо найти k-ю. Запишем в блочном виде первые k строк разложения (5.95):

$$\begin{bmatrix} A_{11} & A_{12} \\ a_{21}^T & a_{22}^T \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ l_{21}^T & 1 \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & u_{22}^T \end{bmatrix} + \begin{bmatrix} R_{11} & R_{12} \\ r_{21}^T & r_{22}^T \end{bmatrix}, \tag{5.96}$$

где l_{21}, u_{22}, r_{21} и r_{22} — некоторые векторы. Выполняя действия над матрицами в правой части (5.96), получим

$$\begin{bmatrix} A_{11} & A_{12} \\ a_{21}^T & a_{22}^T \end{bmatrix} = \begin{bmatrix} L_{11}U_{11} + R_{11} & L_{11}U_{12} + R_{12} \\ l_{21}^T U_{11} + r_{21}^T & l_{21}^T U_{12} + u_{22}^T + r_{22}^T \end{bmatrix}.$$

Из равенства матриц следует, что искомые векторы l_{21} и u_{22} должны удовлетворять условиям:

$$l_{21}^T U_{11} + r_{21}^T = a_{21}^T; \tag{5.97}$$

$$u_{22}^T + r_{22}^T = a_{22}^T - l_{21}^T U_{12}. \tag{5.98}$$

Решив эти системы, можно найти коэффициенты k-х строк матриц разложения $l_{k1}, \ldots, l_{k,k-1}$ и $u_{kk}, \ldots, u_{kn}$. Определим l_{kj} из (5.97) в предположении, что $l_{k1}, \ldots, l_{k,j-1}$ уже найдены. Согласно ранее сформулированным условиям, если $a_{kj} = 0$, то сразу $l_{kj} = 0$. В противном же случае $r_{kj} = 0$ и (5.97) можно записать в виде:

$$\sum_{i=1}^{j} l_{ki} u_{ij} = \sum_{i=1}^{j-1} l_{ki} u_{ij} + l_{kj} u_{jj} = a_{kj}. \tag{5.99}$$

Это позволяет вычислить l_{kj} следующим образом:

$$l_{kj} = \frac{1}{u_{jj}} \left(a_{kj} - \sum_{i=1}^{j-1} l_{ki} u_{ij} \right). \tag{5.100}$$

Аналогичными рассуждениями, учитывая, что $l_{jj} = 1$, из (5.98) можно получить выражение для u_{kj}:

$$u_{kj} = a_{kj} - \sum_{i=1}^{k-1} l_{ki} u_{ij} \tag{5.101}$$

(для тех случаев, когда $a_{kj} \neq 0$, иначе сразу $u_{kj} = 0$). На основании формул (5.100) и (5.101) можно записать следующий алгоритм ILU-разложения.

Алгоритм 5.5.1. ***Алгоритм ILU-разложения***

1. $For\ k = 1 : n$
2. $\quad For\ j = 1 : k - 1$
3. $\quad\quad If\ (k, j) \in P_A$
4. $\quad\quad\quad l_{kj} = \frac{1}{u_{jj}} \left(a_{kj} - \sum_{i=1}^{j-1} l_{ki} u_{ij} \right)$
5. $\quad\quad else$
6. $\quad\quad\quad l_{kj} = 0$
7. $\quad\quad end$
8. $\quad end$
9. $\quad l_{kk} = 1$
10. $\quad For\ j = k : n$
11. $\quad\quad If\ (k, j) \in P_A$
12. $\quad\quad\quad u_{kj} = a_{kj} - \sum_{i=1}^{k-1} l_{ki} u_{ij}$
13. $\quad\quad else$
14. $\quad\quad\quad u_{kj} = 0$
15. $\quad\quad end$
16. $\quad end$
17. end

Очевидно, что матрица LU удовлетворяет всем трем требованиям, предъявляемым к матрице предобуславливателя. Действительно: она приближает матрицу A, так как на множестве индексов P_A точно воспроизводит ее; она легко вычисляется по приведенному выше несложному алгоритму; наконец, она легко обратима, так как является произведением двух треугольных матриц. Таким образом, выбор $M = LU$ является достаточно хорошим способом предобуславливания.

Если матрица A исходной системы симметрична, то желательно чтобы матрица предобусловленной системы тоже была симметрична. Поэтому для системы $Ax = b$ предобусловленную систему ищут в виде

$$M_1^{-1} A M_2^{-2} y = M_1^{-1} b,$$

где $y = M_2 x$ и предобуславливатели M_1 и M_2 выбирают так, чтобы матрица $M_1^{-1} A M_2^{-2}$ была симметричной. Например, если для симметричной положительно определенной матрицы A построено неполное разложение Холесского

$$A = LL^T + R,$$

то полагают $M_1 = L$ и $M_2 = L^T$.

5.6. Вопросы и задания

1. При каких α и β сходится метод простой итерации $x_{k+1} = Gx_k + f$, где
$$G = \begin{bmatrix} \alpha & \beta & 0 \\ \beta & \alpha & \beta \\ 0 & \beta & \alpha \end{bmatrix}.$$

2. Пусть все собственные значения матрицы A вещественны и положительны. Доказать сходимость метода
$$\frac{x_{k+1} - x_k}{\tau} + Ax_k = b$$
при $\tau = ||A||^{-1}$ для любой матричной нормы.

3. Пусть все собственные значения невырожденной матрицы A порядка n известны. Построить итерационный метод с переменным параметром τ_k, который не более чем за n шагов приводил бы к точному решению системы $Ax = b$.

4. Представим матрицу системы $Ax = b$ в виде $A = L + D + R$, где D — диагональная матрица, L и R — соответственно левая нижняя и правая верхняя треугольные матрицы с нулевыми диагоналями. Методы Якоби и Гаусса-Зейделя записываются в виде:
$$Dx_{k+1} + (L + R)x_k = b,$$
$$(D + L)x_{k+1} + Rx_k = b.$$

   а. Найти область сходимости методов Якоби и Гаусса-Зейделя для систем с матрицами вида
$$A = \begin{bmatrix} \alpha & \beta & 0 \\ \beta & \alpha & \beta \\ 0 & \beta & \alpha \end{bmatrix}.$$

   б. Доказать, что для систем линейных уравнений второго порядка ($n = 2$) методы Якоби и Гаусса-Зейделя сходятся и расходятся одновременно.

   в. Показать, что существует система линейных уравнений третьего порядка, для которой метод Якоби сходится, а метод Гаусса-Зейделя расходится.

   г. Показать, что существует система линейных уравнений третьего порядка, для которой метод Гаусса-Зейделя сходится, а метод Якоби расходится.

5. Проверить, что подпространство Крылова $\mathcal{K}_m(A, y)$ тогда и только тогда имеет размерность m, когда при вычислении вектора q_m не происходит досрочного выхода из алгоритма Арнольди (Ланцоша).

6. Пусть $A \in R^{n \times n}$ — симметричная положительно определенная матрица и пусть матрица $Q \in R^{n \times k}$ имеет полный столбцовый ранг. Проверить, что матрица $T = Q^T AQ$ также симметрична и положительно определена.

6. Симметричная проблема собственных значений

6.1. Степенной метод и обратная итерация

Предположим, что максимальное по модулю собственное значение симметричной вещественной матрицы $A \in R^{n\times n}$ отделено, т. е.,

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

В этом случае степенной метод позволяет найти максимальное по модулю собственное значение λ_1 и соответствующий собственный вектор q_1.

Построим следующий итерационный процесс. Пусть y_0 — начальное приближение к собственному вектору q_1. По теореме 1.5.4 о спектральном разложении симметричной матрицы собственные векторы матрицы A образуют ортонормированный базис пространства $R^{n\times n}$. Разложим y_0 по этому базису,

$$y_0 = \sum_{i=1}^{n} \alpha_i q_i, \tag{6.1}$$

где $\alpha_i = q_i^T y_0$. Предположим, что $\alpha_1 \neq 0$. Нормируем вектор y_0,

$$x_0 = \frac{y_0}{||y_0||_2}.$$

Умножим матрицу A на вектор x_0,

$$y_1 = Ax_0 = \frac{\sum\limits_{i=1}^{n} \alpha_i \lambda_i q_i}{\left|\left|\sum\limits_{i=1}^{n} \alpha_i q_i\right|\right|_2}.$$

Нормируем вектор y_1,

$$x_1 = \frac{y_1}{||y_1||_2} = \frac{\sum\limits_{i=1}^{n} \alpha_i \lambda_i q_i}{\left|\left|\sum\limits_{i=1}^{n} \alpha_i \lambda_i q_i\right|\right|_2}.$$

После k итераций получаем

$$x_k = \frac{\sum\limits_{i=1}^{n} \alpha_i \lambda_i^k q_i}{\left|\left|\sum\limits_{i=1}^{n} \alpha_i \lambda_i^k q_i\right|\right|_2}, \tag{6.2}$$

$$y_{k+1} = Ax_k = \frac{\sum\limits_{i=1}^{n} \alpha_i \lambda_i^{k+1} q_i}{\left|\left| \sum\limits_{i=1}^{n} \alpha_i \lambda_i^k q_i \right|\right|_2}. \tag{6.3}$$

Положим

$$\lambda^{(k+1)} = y_{k+1}^T x_k.$$

Покажем, что $\lambda^{(k+1)} \to \lambda_1$ при $k \to \infty$. Из (6.2) и (6.3) получаем

$$\lambda^{(k+1)} = y_{k+1}^T x_k = \frac{\left(\sum\limits_{i=1}^{n} \alpha_i \lambda_i^{k+1} q_i\right)^T \left(\sum\limits_{i=1}^{n} \alpha_i \lambda_i^k q_i\right)}{\left|\left| \sum\limits_{i=1}^{n} \alpha_i \lambda_i^k q_i \right|\right|_2^2} = \frac{\sum\limits_{i=1}^{n} \alpha_i^2 \lambda_i^{2k+1}}{\sum\limits_{i=1}^{n} \alpha_i^2 \lambda_i^{2k}} =$$

$$= \frac{\lambda_1 + \frac{\alpha_2^2 \lambda_2}{\alpha_1^2} \left(\frac{\lambda_2}{\lambda_1}\right)^{2k} + \cdots + \frac{\alpha_n^2 \lambda_n}{\alpha_1^2} \left(\frac{\lambda_n}{\lambda_1}\right)^{2k}}{1 + \frac{\alpha_2^2}{\alpha_1^2} \left(\frac{\lambda_2}{\lambda_1}\right)^{2k} + \cdots + \frac{\alpha_n^2}{\alpha_1^2} \left(\frac{\lambda_n}{\lambda_1}\right)^{2k}} = \frac{\lambda_1 + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2k}\right)}{1 + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2k}\right)}.$$

Отсюда

$$\lambda^{(k+1)} - \lambda_1 = O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2k}\right). \tag{6.4}$$

Покажем теперь, что $x_k \to q_1$ при $k \to \infty$. Действительно,

$$x_k = \frac{\sum\limits_{i=1}^{n} \alpha_i \lambda_i^k q_i}{\left|\left| \sum\limits_{i=1}^{n} \alpha_i \lambda_i^k q_i \right|\right|_2} = \frac{\alpha_1 \lambda_1^k q_1 + \alpha_2 \lambda_2^k q_2 + \cdots + \alpha_n \lambda_n^k q_n}{|\alpha_1| \, |\lambda_1|^k \sqrt{1 + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2k}\right)}} =$$

$$= \frac{\pm q_1 + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{k}\right)}{\sqrt{1 + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2k}\right)}}.$$

Отсюда

$$x_k - q_1 = O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{k}\right). \tag{6.5}$$

Заметим, что скорость сходимости собственного значения в 2 раза больше, чем собственного вектора.

Запишем алгоритм степенного метода.

Алгоритм 6.1.1. ***Степенной метод***

1. $x_0 = y_0/||y_0||_2$
2. *For* $k = 0, 1, \ldots$ *Until Convergence Do*
3. $\quad y_{k+1} = Ax_k$
4. $\quad \lambda^{(k+1)} = y_{k+1}^T x_k$
5. $\quad x_{k+1} = y_{k+1}/||y_{k+1}||_2$
6. *End Do*

Тестом на сходимость может быть одно из условий,

$$\left|\lambda^{(k+1)} - \lambda^{(k)}\right| \le \varepsilon \left|\lambda^{(k+1)}\right|, \tag{6.6}$$

или

$$\left| \max_{(x_k)_i \neq 0} \frac{(y_{k+1})_i}{(x_k)_i} - \min_{(x_k)_i \neq 0} \frac{(y_{k+1})_i}{(x_k)_i} \right| \le \varepsilon \left| \max_{(x_k)_i \neq 0} \frac{(y_{k+1})_i}{(x_k)_i} \right|. \tag{6.7}$$

Заметим, шаг 3. алгоритма 6.1.1 требует $O(n^2)$ арифметических действий, все остальные шаги — $O(n)$ арифметических действий.

Рассмотрим теперь метод обратной итерации. Предположим, что наименьшее по модулю собственное значение вещественной симметричной матрицы $A \in R^{n\times n}$ отделено, т. е.,

$$|\lambda_1| \ge \cdots \ge |\lambda_{n-1}| > |\lambda_n|. \tag{6.8}$$

Заметим, что если $\{\lambda, x\}$ — собственная пара матрицы A, то $\{1/\lambda, x\}$ — собственная пара обратной матрицы A^{-1}. Очевидно, что $1/\lambda_n$ — максимальное по модулю собственное значение матрицы A^{-1}. Поэтому для нахождения $1/\lambda_n$ можно применить степенной метод для матрицы A^{-1}. В этом случае при $k \to \infty$

$$\lambda^{(k+1)} \to \frac{1}{\lambda_n}, \quad x_k \to q_n.$$

Таким образом, обратная итерация позволяет вычислить наименьшее по модулю собственное значение матрицы (и соответствующий собственный вектор), если оно отделено от остальных собственных значений.

Запишем алгоритм обратной итерации.

Алгоритм 6.1.2. ***Обратная итерация***

1. $x_0 = y_0/||y_0||_2$
2. *For* $k = 0, 1, \ldots$ *Until Convergence Do*
3. $\quad y_{k+1} = A^{-1}x_k$
4. $\quad \lambda^{(k+1)} = y_{k+1}^T x_k$
5. $\quad x_{k+1} = y_{k+1}/||y_{k+1}||_2$
6. *End Do*

Шаг 3. алгоритма обратной итерации сводится к решению системы $Ay_{k+1} = x_k$. Для решения этой системы требуется $O(n^3)$ арифметических действий. Однако, если известно некоторое разложение матрицы A (например, LU-разложение), то этот шаг будет требовать только $O(n^2)$ арифметических действий. Остальные шаги алгоритма требуют только $O(n)$ арифметических действий.

Замечание 6.1.1. Устанавливая сходимость степенного метода, мы предполагали (см. (6.1)), что коэффициент α_1 в разложении по собственным векторам начального приближения y_0 отличен от нуля. Даже если это не так, то, в следствие округлений при выполнении арифметических операций, на некоторой итерации k коэффициент α_1 в разложении y_k будет отличен от нуля.

6.2. Исчерпывание вычитанием

Пусть собственные значения вещественной симметричной матрицы $A \in R^{n\times n}$ удовлетворяют условиям

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \cdots \geq |\lambda_n|.$$

Предположим, что собственная пара $\{\lambda_1, q_1\}$ уже найдена. Необходимо найти $\{\lambda_2, q_2\}$.

Пусть $\widetilde{y}_0$ — начальное приближение и

$$\widetilde{y}_0 = \alpha_1 q_1 + \alpha_2 q_2 + \cdots + \alpha_n q_n,$$

причем $\alpha_2 \neq 0$. Положим

$$y_0 = \widetilde{y}_0 - \alpha_1 q_1 = \widetilde{y}_0 - \widetilde{y}_0^T q_1 q_1.$$

Повторяя доказательство сходимости степенного метода для выбранного таким образом начального приближения y_0, убедимся, что в точной арифметике при $k \to \infty$

$$\lambda^{(k+1)} \to \lambda_2, \quad x_k \to q_2.$$

Однако, из-за погрешностей округлений коэффициент при q_1 опять может появиться в разложении y_k. Поэтому этот коэффициент удаляют на каждой итерации, либо через несколько итераций. Таким образом, для нахождения собственной пары $\{\lambda_2, q_2\}$ можно применить степенной метод с исчерпыванием вычитанием.

Алгоритм 6.2.1. ***Степенной метод с исчерпыванием вычитанием***

1. $y_0 = \widetilde{y}_0 - \widetilde{y}_0^T q_1 q_1$
2. $x_0 = y_0/||y_0||_2$
3. *For* $k = 0, 1, \ldots$ *Until Convergence Do*
4. $\quad \widetilde{y}_{k+1} = Ax_k$
5. $\quad y_{k+1} = \widetilde{y}_{k+1} - \widetilde{y}_{k+1}^T q_1 q_1$
6. $\quad \lambda^{(k+1)} = y_{k+1}^T x_k$
7. $\quad x_{k+1} = y_{k+1}/||y_{k+1}||_2$
8. *End Do*

Так поочередно можно найти несколько максимальных по модулю собственных значений, если они отделены друг от друга и от остальной части спектра.

Очевидно, что для нахождения нескольких отделенных минимальных по модулю собственных значений можно использовать обратную итерацию.

6.3. Использование сдвигов

Пусть $A \in R^{n\times n}$ и $\sigma \in R$. Если $\{\lambda, x\}$ — собственная пара матрицы A, то $\{\lambda - \sigma, x\}$ является собственной парой матрицы $A - \sigma I$. Вещественное число σ называется сдвигом.

Вначале рассмотрим использование сдвига для увеличения скорости сходимости обратной итерации. Поскольку обратная итерация — это степенной метод для обратной матрицы A^{-1}, то

$$\lambda^{(k+1)} - \frac{1}{\lambda_n} = O\left(\left(\frac{\frac{1}{\lambda_{n-1}}}{\frac{1}{\lambda_n}}\right)^{2k}\right) = O\left(\left(\frac{\lambda_n}{\lambda_{n-1}}\right)^{2k}\right).$$

Отсюда следует, что скорость сходимости зависит от малости величины $|\frac{\lambda_n}{\lambda_{n-1}}|$. Последняя зависит от расстояния между $|\lambda_{n-1}|$ и $|\lambda_n|$, т. е., от отделимости наименьшего по модулю собственного значения, и от малости $|\lambda_n|$. Поэтому сдвиг σ выбирают так, чтобы наименьшее по модулю собственное значение $\widetilde{\lambda}_n = \lambda_n - \sigma$ матрицы $A - \sigma I$ было как можно ближе к нулю. Таким образом, для увеличения скорости сходимости обратной итерации сдвиг σ надо выбирать как можно ближе к искомому собственному значению λ_n. Если $\widetilde{x}$ — приближение к некоторому собственному вектору симметричной матрицы, то из леммы 1.5.5 следует, что наилучшим приближением к соответствующему собственному значению дает отношение Релея

$$\rho(\widetilde{x}) = \frac{\widetilde{x}^T A \widetilde{x}}{\widetilde{x}^T \widetilde{x}}.$$

Вышесказанное лежит в основе следующего алгоритма.

Алгоритм 6.3.1. ***Обратная итерация со сдвигом Релея***
1. $x_0 = y_0/||y_0||_2$
2. $For\ k = 0, 1, \ldots\ Until\ Convergence\ Do$
3. $\quad \sigma_{k+1} = x_k^T A x_k$
4. $\quad y_{k+1} = (A - \sigma_{k+1} I)^{-1} x_k$
5. $\quad x_{k+1} = y_{k+1}/||y_{k+1}||_2$
6. $End\ Do$

В качестве теста на сходимость можно использовать условие (6.7).

Заметим, что шаг 3. алгоритма требует $O(n^3)$ арифметических действий, ибо каждый раз надо решать систему линейных алгебраических уравнений с новой матрицей.

Рассмотрим теперь использование сдвига для вычисления собственных значений матрицы степенным методом (обратной итерацией) отличных от максимальных (минимальных) по модулю. Очевидно, что обратная итерация, вычисляя минимальное по модулю собственное значение матрицы $A - \sigma I$, фактически будет вычислять собственное значение матрицы A ближайшее к σ.

Определив границы спектра матрицы A (например, по лемме Гершгорина), можно так выбрать сдвиг σ, чтобы степенной метод или обратная итерация фактически вычисляли наименьшее (наибольшее) собственное значение матрицы A. Действительно (см. рис. 6.1) , если $\sigma \leq \lambda_{min}(A)$, то все собственные значения матрицы $A - \sigma I$ положительны, наименьшему собственному значению матрицы A соответствует минимальное (минимальное по модулю) собственное значение матрицы $A - \sigma I$, а максимальному собственному значению матрицы A — максимальное (максимальное по модулю) собственное значение матрицы $A - \sigma I$. Если же $\sigma \geq \lambda_{max}(A)$, то все собственные значения матрицы $A - \sigma I$ отрицательны, наименьшему собственному значению матрицы A соответствует минимальное (максимальное по модулю) собственное значение матрицы $A - \sigma I$, а максимальному собственному значению матрицы A — максимальное (минимальное по модулю) собственное значение матрицы $A - \sigma I$.

6.4. Метод Ланцоша

Пусть столбцы матрицы $V_m = [v_1, \ldots, v_n]$ — это построенный алгоритмом Ланцоша ортонормированный базис подпространства Крылова $\mathcal{K}_m(A, v)$ вещественной симметричной матрицы $A \in R^{n \times n}$ и произвольного вектора v. Тогда

$$AV_m = V_m T_m + \beta_{m+1} v_{m+1} e_m^T, \tag{6.9}$$

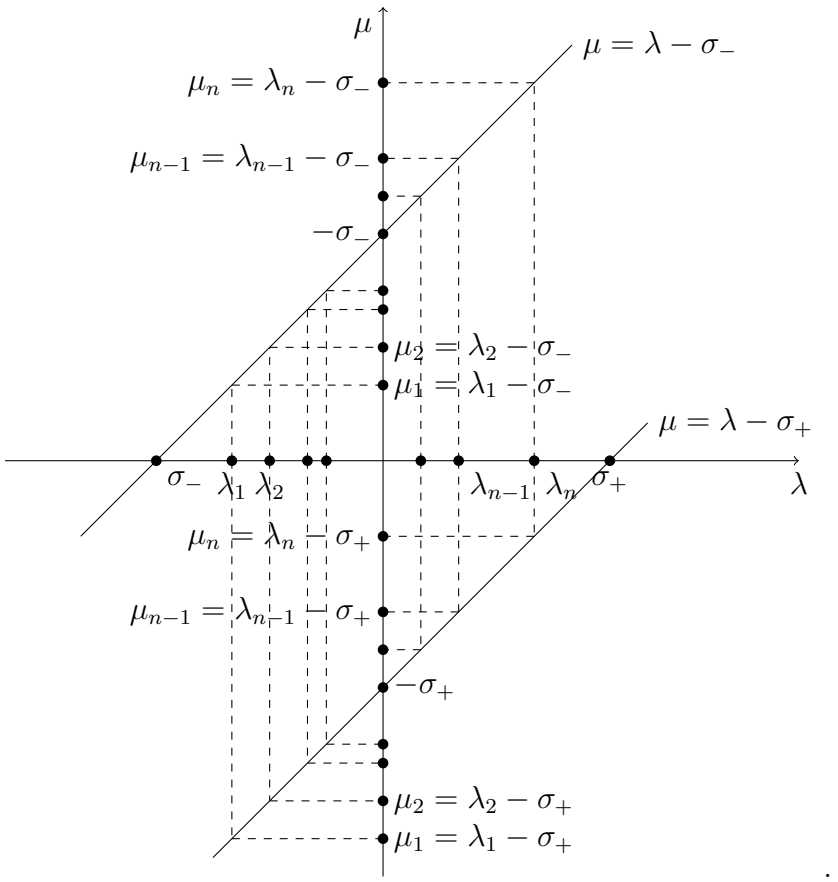

Рис. 6.1. Собственные значения μ матрицы $A - \sigma I$

$$V_m^T A V_m = T_m, \tag{6.10}$$

где

$$T_m = \begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \cdot & \cdot & \cdot & \\ & & \beta_{m-1} & \alpha_{m-1} & \beta_m \\ & & & \beta_m & \alpha_m \end{bmatrix} \tag{6.11}$$

Оказывается, что собственные значения матрицы T_m являются хорошими приближениями к собственным значениям матрицы A.

Теорема 6.4.1. *Пусть столбцы матрицы $V_m \equiv [v_1, \ldots, v_n]$ — это построенный алгоритмом Ланцоша ортонормированный базис подпространства Крылова $\mathcal{K}_m(A, v)$ вещественной симметричной матрицы $A \in R^{n \times n}$ и произвольного вектора v и*

$$S_m^T T_m S_m = \Theta_m = \mathrm{diag}(\vartheta_1, \ldots, \vartheta_m), \tag{6.12}$$

где $S_m \in R^{m\times m}$ — ортогональная матрица. Положим

$$Y_m \equiv [y_1, \ldots, y_m] = V_m S_m \in R^{n\times m},$$

тогда

$$||Ay_i - \vartheta_i y_i||_2 = \beta_{i+1}|s_{mi}|, \quad i = \overline{1,m}. \tag{6.13}$$

Доказательство. Из (6.12)

$$T_m = S_m \Theta_m S_m^T,$$

поэтому из (6.9) получаем

$$AV_m = V_m S_m \Theta_m S_m^T + \beta_{m+1} v_{m+1} e_m^T.$$

Умножая это равенства справа на S_m, получаем

$$AV_m S_m = V_m S_m \Theta_m + \beta_{m+1} v_{m+1} e_m^T S_m.$$

Учитывая, что $Y_m = V_m S_m$,

$$AY_m = Y_m \Theta_m + \beta_{m+1} v_{m+1} e_m^T S_m.$$

Приравнивая i-е столбцы в последнем матричном равенстве, получаем

$$Ay_i = \vartheta_i y_i + \beta_{m+1} v_{m+1} e_m^T S_m e_i.$$

Наконец, учитывая, что $||v_{m+1}||_2 = 1$ получаем (6.13). □

Пары $\{\vartheta_i, y_i\}$ называются парами Ритца, ϑ_i — число Ритца, y_i — вектор Ритца.

Если в конце m-го шага алгоритма Ланцоша $\beta_{m+1} = 0$, то из (6.9)

$$AV_m = V_m T_m.$$

А это значит, что столбцы матрицы V_m образуют инвариантное подпространство матрицы A, поэтому собственные значения матрицы T_m являются собственными значениями матрицы A. Если же $\beta_{m+1} \neq 0$, то собственные значения матрицы T_m являются лишь приближениями к собственным значениям матрицы A.

Таким образом, алгоритм Ланцоша сводит частичную проблему собственных значений симметричной вещественной матрицы A к полной проблеме собственных значений для симметричной трехдиагональной матрицы T_m меньшей размерности.

6.5. QL-алгоритм

QL-алгоритм решения полной проблемы собственных значений вещественной симметричной матрицы $A \in R^{n\times n}$ строит последовательность ортогонально подобных матриц $\{A_k\}_{k=1}^{\infty}$ так,что $A_k \to \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ при $k \to \infty$. Положим $A_1 = A$. Переход от k-й к $(k+1)$-й итерации состоит в следующем. Используя вращения Якоби, строим QL-разложение матрицы A_k:

$$A_k = Q_k L_k. \tag{6.14}$$

Вычисляем

$$A_{k+1} = L_k Q_k. \tag{6.15}$$

Здесь Q_k — ортогональная матрица, L_k — нижняя треугольная матрица.

Вначале заметим, что

$$A_{k+1} = L_k Q_k = Q_k^T A_k Q_k = \cdots = Q_k^T Q_{k-1}^T \ldots Q_1^T A_1 Q_1 \ldots Q_{k-1} Q_k.$$

Таким образом,

$$A_{k+1} = P_k^T A P_k, \tag{6.16}$$

где $P_k = Q_1 \ldots Q_{k-1} Q_k$ — ортогональная матрица. Значит собственные значения матриц A_{k+1} и A совпадают, поскольку они ортогонально подобны.

Установим связь между QL-алгоритмом, степенным методом и обратной итерацией.

QL-алгоритм порождает последовательность матриц $\{A_k\}_{k=1}^{\infty}$,

$$A_{k+1} = Q_k^T A_k Q_k = P_k^T A P_k, \tag{6.17}$$

где $A_k = Q_k L_k$, $A_1 = A$, $P_k = Q_1 \ldots Q_k$.

Степенной метод порождает последовательность векторов $\{v_k\}_{k=1}^{\infty}$,

$$v_{k+1} = \frac{A v_k}{\mu_k}, \tag{6.18}$$

где v_1 ($||v_1||_2 = 1$) — начальное приближение, μ_k — нормирующий множитель, который выбирают так, чтобы $||v_k||_2 = 1$ для любого k.

Обратная итерация порождает последовательность векторов $\{u_k\}_{k=1}^{\infty}$,

$$A u_{k+1} = \tau_k u_k, \tag{6.19}$$

где u_1 ($||u_1||_2 = 1$) — начальное приближение, τ_k — нормирующий множитель, который выбирают так, чтобы $||u_k||_2 = 1$ для любого k.

Теорема 6.5.1. *Пусть $u_1 = e_1$, $v_1 = e_n$, $P_0 = I$. Тогда последовательности векторов, порождаемые степенным методом и обратной итерацией, связаны с последовательностью матриц QL-алгоритма следующим образом. Для любого $k \geq 1$*

$$u_k = P_{k-1}e_1, \quad v_k = P_{k-1}e_n. \tag{6.20}$$

Доказательство. Используя (6.17), получаем

$$P_kL_k = P_{k-1}Q_kL_k = P_{k-1}A_k = P_{k-1}P_{k-1}^TAP_{k-1} = AP_{k-1},$$

т. е.,

$$P_kL_k = AP_{k-1}. \tag{6.21}$$

Приравнивая последние столбцы в матричном равенстве (6.21), получаем

$$P_ke_ne_n^TL_ke_n = AP_{k-1}e_n.$$

Сравнивая эту рекуррентную формулу с (6.18), находим, что $v_k = P_{k-1}e_n$, $\mu_k = e_n^TL_ke_n$.

Из (6.21)

$$L_k^TP_k^T = P_{k-1}^TA.$$

Домножая последнее равенство слева на P_{k-1} и справа на P_k, получаем

$$P_{k-1}L_k^T = AP_k.$$

Приравнивая первые столбцы в этом матричном равенстве, получаем

$$P_{k-1}e_1e_1^TL_ke_1 = AP_ke_1.$$

Сравнивая эту рекуррентную формулу с (6.20), находим, что $u_k = P_{k-1}e_1$, $\tau_k = e_1^TL_ke_1$. Теорема доказана. □

Докажем теперь сходимость QL-алгоритма, используя его связь со степенным методом и обратной итерацией.

Теорема 6.5.2. *Пусть собственные пары $\{\lambda_i, q_i\}$ $(i = \overline{1,n})$ симметричной вещественной матрицы $A \in R^{n\times n}$ удовлетворяют условиям*

$$|\lambda_1| < |\lambda_2| \leq \cdots \leq |\lambda_{n-1}| < |\lambda_n|, \tag{6.22}$$

$$e_1^Tq_1 \neq 0, \quad e_n^Tq_n \neq 0. \tag{6.23}$$

$\{A_k\}_{k=1}^{\infty}$ — последовательность QL-алгоритма. Тогда

$$A_ke_1 \to \lambda_1e_1, \ A_ke_n \to \lambda_ne_n \text{ при } k \to \infty.$$

Доказательство. По теореме 6.5.1 о связи QL-алгоритма с обратной итерацией, QL-алгоритм формирует последовательность векторов $\{u_k\}_{k=1}^{\infty}$ обратной итерации:

$$u_1 = e_1, \quad u_k = P_{k-1}e_1.$$

Из условий (6.22), (6.23) следует, что эта последовательность сходится и

$$u_k = q_1 + O\left(\left(\frac{\lambda_1}{\lambda_2}\right)^k\right).$$

Значит,

$$e_1 = P_{k-1}^T P_{k-1}e_1 = P_{k-1}^T u_k = P_{k-1}^T q_1 + O\left(\left(\frac{\lambda_1}{\lambda_2}\right)^k\right),$$

и

$$A_k e_1 = P_{k-1}^T A P_{k-1}e_1 = P_{k-1}^T A u_k = P_{k-1}^T(\lambda_1 q_1 + O\left(\left(\frac{\lambda_1}{\lambda_2}\right)^k\right) =$$

$$= \lambda_1 P_{k-1}^T q_1 + O\left(\left(\frac{\lambda_1}{\lambda_2}\right)^k\right) = \lambda_1 e_1 + O\left(\left(\frac{\lambda_1}{\lambda_2}\right)^k\right),$$

т. е.,

$$A_k e_1 = \lambda_1 e_1 + O\left(\left(\frac{\lambda_1}{\lambda_2}\right)^k\right). \tag{6.24}$$

Аналогично можно показать, что

$$A_k e_n = \lambda_n e_n + O\left(\left(\frac{\lambda_{n-1}}{\lambda_n}\right)^k\right).$$

□

QL-алгоритм можно применять для любой вещественной симметричной матрицы. Однако, наибольшая эффективность метода достигается для трехдиагональных симметричных матриц. Поэтому симметричную матрицу вначале приводят к трехдиагональному виду ортогональными преобразованиями, а затем применяют QL-алгоритм. QL-алгоритм для трехдиагональной симметричной матрицы A сходится за один шаг, если матрица вырождена.

Для увеличения скорости сходимости QL-алгоритма можно использовать сдвиги. Переход от k-й к $(k+1)$–й итерации QL-алгоритма со сдвигами состоит в следующем. Выбираем сдвиг σ_k. Используя вращения Якоби, строим QL-разложение матрицы $A_k - \sigma_k I$:

$$A_k - \sigma_k I = Q_k L_k.$$

Вычисляем

$$A_{k+1} = L_k Q_k + \sigma_k I.$$

Поскольку

$$A_{k+1} = L_k Q_k + \sigma_k I = Q_k^T (A_k - \sigma_k I) Q_k + \sigma_k I = Q_k^T A_k Q_k,$$

то

$$A_{k+1} = Q_k^T A_k Q_k = \cdots = Q_k^T Q_{k-1}^T \dots Q_1^T A_1 Q_1 \dots Q_{k-1} Q_k = P_k^T A P_k,$$

где $P_k = Q_1 \dots Q_{k-1} Q_k$ — ортогональная матрица. Значит матрицы A_k снова ортогонально подобны матрице A, т. е., их собственные значения совпадают с собственными значениями матрицы A. Из (6.24) следует, что чем ближе λ_1 к нулю, тем быстрее сходится QL-алгоритм. Поскольку $A_k e_1 \to \lambda_1 e_1$ при $k \to \infty$, то есть смысл в качестве сдвига σ_k брать $a_{11}^{(k)}$. Такой сдвиг называют сдвигом по Релею. Сдвиг по Уилкинсону — это ближайшее к $a_{11}^{(k)}$ собственное значение матрицы

$$\begin{bmatrix} a_{11}^{(k)} & a_{12}^{(k)} \\ a_{21}^{(k)} & a_{22}^{(k)} \end{bmatrix}.$$

Отметим, что QR-алгоритм отличается от QL-алгоритма лишь тем, что вместо QL-разложения матрицы A_k строится QR-разложение $A_k = Q_k R_k$, где Q_k — ортогональная матрица, R_k — верхняя треугольная матрица.

6.6. QL-алгоритм для трехдиагональной матрицы

Чтобы решить полную проблему собственных значений для симметричной матрицы A, последнюю посредством ортогональных преобразований приводят к трехдиагональной форме T. Для этого имеется ряд причин. Во–первых, собственные значения и собственные векторы симметричной трехдиагональной матрицы T можно найти со значительно меньшей затратой арифметических операций, чем заполненной матрицы A. Во-вторых, симметричную матрицу A можно привести к трехдиагональной форме T посредством конечного числа ортогональных преобразований, а чтобы диагонализовать матрицу A необходимо бесконечное число ортогональных преобразований. В-третьих, приводя посредством конечного

числа ортогональных преобразований симметричную матрицу A к трехдиагональной форме, получим блочно-диагональную матрицу

$$T = \begin{bmatrix} T_{11} & & \\ & \ddots & \\ & & T_{kk} \end{bmatrix},$$

где T_{ii} $(i = \overline{1,k})$ — неразложимые трехдиагональные матрицы. Поэтому, чтобы найти все собственные значения матрицы A, надо найти все собственные значения матриц T_{ii} $(i = \overline{1,k})$. Таким образом, полная проблема собственных значений для матрицы A сводится к задачам на собственные значения меньшей размерности. Отметим, что все собственные значения симметричной трехдиагональной неразложимой матрицы различны.

Симметричную матрицу A можно привести к трехдиагональной форме с помощью матриц отражения (матриц Хаусхолдера). Заметим, что для умножения матрицы A на матрицу отражения $H(u)$ слева (справа) нет необходимости вычислять матрицу отражения, ибо

$$H(u)A = (I - \gamma uu^T)A = A - \gamma u(Au)^T,$$

$$AH(u) = A(I - \gamma uu^T) = A - \gamma(Au)u^T.$$

Запишем теперь алгоритм приведения симметричной матрицы к трехдиагональной форме с использованием матрицы отражения.

Алгоритм 6.6.1. ***Приведение симметричной матрицы к трехдиагональной форме***

1. $For\ i = 1 : n - 2$
2. $\quad v = A(i + 1 : n, i)$
3. $\quad \beta = -sign(v(1)) * ||v||_2$
4. $\quad u = v - \beta e_1$
5. $\quad \gamma = 2/(u'u)$
6. $\quad A(i+1 : n, i+1 : n) = A(i+1 : n, i+1 : n) - \gamma u(A(i+1 : n, i+1 : n)u)'$
7. $\quad A(i+1 : n, i+1 : n) = A(i+1 : n, i+1 : n) - \gamma(A(i+1 : n, i+1 : n)u)u'$
8. $\quad A(i + 1, i) = \beta; A(i, i + 1) = \beta;$
9. $\quad For\ j = i + 2 : n$
10. $\quad\quad A(j, i) = 0; A(i, j) = 0;$
11. $\quad Next\ j$
12. $Next\ i$

Рассмотрим теперь QL-алгоритм для трехдиагональной матрицы. Если матрица A симметрична и трехдиагональна, то очевидно что A_{k+1} симметричная матрица. Более того, можно показать, что A_{k+1} тоже трехдиагональная. Последнее проверим на примере матрицы размера 4×4.

Приведем матрицу

$$A_k = \begin{bmatrix} * & * & 0 & 0 \\ * & * & * & 0 \\ 0 & * & * & * \\ 0 & 0 & * & * \end{bmatrix}$$

вращениями Якоби к нижнему треугольному виду

$$Q_{12}Q_{23}Q_{34}A_k = L_k = \begin{bmatrix} * & 0 & 0 & 0 \\ * & * & 0 & 0 \\ * & * & * & 0 \\ 0 & * & * & * \end{bmatrix}.$$

Далее

$$Q_{12}Q_{23} = \begin{bmatrix} * & * & 0 & 0 \\ * & * & 0 & 0 \\ 0 & 0 & * & 0 \\ 0 & 0 & 0 & * \end{bmatrix} \begin{bmatrix} * & 0 & 0 & 0 \\ 0 & * & * & 0 \\ 0 & * & * & 0 \\ 0 & 0 & 0 & * \end{bmatrix} = \begin{bmatrix} * & * & * & 0 \\ * & * & * & 0 \\ 0 & * & * & 0 \\ 0 & 0 & 0 & * \end{bmatrix},$$

$$(Q_{12}Q_{23})Q_{34} = \begin{bmatrix} * & * & * & 0 \\ * & * & * & 0 \\ 0 & * & * & 0 \\ 0 & 0 & 0 & * \end{bmatrix} \begin{bmatrix} * & 0 & 0 & 0 \\ 0 & * & 0 & 0 \\ 0 & 0 & * & * \\ 0 & 0 & * & * \end{bmatrix} = \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \end{bmatrix} = Q_k^T.$$

Значит

$$A_{k+1} = L_kQ_k = \begin{bmatrix} * & 0 & 0 & 0 \\ * & * & 0 & 0 \\ * & * & * & 0 \\ 0 & * & * & * \end{bmatrix} \begin{bmatrix} * & * & 0 & 0 \\ * & * & * & 0 \\ * & * & * & * \\ * & * & * & * \end{bmatrix} = \begin{bmatrix} * & * & 0 & 0 \\ * & * & * & 0 \\ 0 & * & * & * \\ 0 & 0 & * & * \end{bmatrix}.$$

В матрице A_{k+1} элементы $a_{31}^{(k+1)}$, $a_{41}^{(k+1)}$ и $a_{42}^{(k+1)}$ должны обратиться в нуль в силу симметрии матрицы.

Теперь $(k+1)$-й шаг QL-алгоритм для трехдиагональной симметричной матрицы T можно записать так

$$T_{k+1} = Q_{1,2}Q_{2,3}\cdots Q_{n-2,n-1}Q_{n-1,n}T_kQ_{n-1,n}^TQ_{n-2,n-1}^T\cdots Q_{2,3}^TQ_{1,2}^T,$$

где

$$T_k = \begin{bmatrix} \alpha_1^{(k)} & \beta_1^{(k)} & & & \\ \beta_1^{(k)} & \alpha_2^{(k)} & \beta_2^{(k)} & & \\ & \cdot & \cdot & \cdot & \\ & & \beta_{n-2}^{(k)} & \alpha_{n-1}^{(k)} & \beta_{n-1}^{(k)} \\ & & & \beta_{n-1}^{(k)} & \alpha_n^{(k)} \end{bmatrix}.$$

Вначале вычислим произведение $Q_{n-1,n}T_k$. Матрицу вращения $Q_{n-1,n}$ выбираем так, чтобы обнулить элемент $\beta_{n-1}^{(k)}$ в позиции $(n-1,n)$. При этом поменяются только две последних строки матрицы T_k:

$$\begin{bmatrix} c_{n-1} & -s_{n-1} \\ s_{n-1} & c_{n-1} \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & \beta_{n-2}^{(k)} & \alpha_{n-1}^{(k)} & \beta_{n-1}^{(k)} \\ 0 & \cdots & 0 & 0 & \beta_{n-1}^{(k)} & \alpha_n^{(k)} \end{bmatrix} =$$

$$\begin{bmatrix} 0 & \cdots & 0 & \gamma_{n-2} & \widetilde{\alpha}_{n-1}^{(k)} & 0 \\ 0 & \cdots & 0 & * & \widetilde{\beta}_{n-1}^{(k)} & \widetilde{\alpha}_n^{(k)} \end{bmatrix}.$$

Здесь c_{n-1} и s_{n-1} находим из уравнения

$$c_{n-1}\beta_{n-1}^{(k)} - s_{n-1}\alpha_n^{(k)} = 0.$$

Таким образом,

$$c_{n-1} = \alpha_n^{(k)}/\sqrt{D}, \quad s_{n-1} = \beta_{n-1}^{(k)}/\sqrt{D},$$

где

$$D = (\beta_{n-1}^{(k)})^2 + (\alpha_n^{(k)})^2.$$

$$\gamma_{n-2} = c_{n-1}\beta_{n-2}^{(k)},$$
$$\widetilde{\alpha}_{n-1}^{(k)} = c_{n-1}\alpha_{n-1}^{(k)} - s_{n-1}\beta_{n-1}^{(k)},$$
$$\widetilde{\beta}_{n-1}^{(k)} = s_{n-1}\alpha_{n-1}^{(k)} + c_{n-1}\beta_{n-1}^{(k)},$$
$$\widetilde{\alpha}_n^{(k)} = s_{n-1}\beta_{n-1}^{(k)} + c_{n-1}\alpha_n^{(k)}.$$

Теперь матрицу $Q_{n-1,n}T_k$ умножаем слева на матрицу вращения $Q_{n-2,n-1}$ чтобы обнулить элемент $\beta_{n-2}^{(k)}$ в позиции $(n-2,n-1)$. В результате умножения изменятся строки $n-2$ и $n-1$:

$$\begin{bmatrix} c_{n-2} & -s_{n-2} \\ s_{n-2} & c_{n-2} \end{bmatrix} \begin{bmatrix} 0 & \cdots & 0 & \beta_{n-3}^{(k)} & \alpha_{n-2}^{(k)} & \beta_{n-2}^{(k)} & 0 \\ 0 & \cdots & 0 & 0 & \gamma_{n-2} & \widetilde{\alpha}_{n-1}^{(k)} & 0 \end{bmatrix} =$$

$$\begin{bmatrix} 0 & \cdots & 0 & \gamma_{n-3} & \widetilde{\alpha}_{n-2}^{(k)} & 0 & 0 \\ 0 & \cdots & 0 & * & \widetilde{\beta}_{n-2}^{(k)} & \widetilde{\alpha}_{n-1}^{(k)} & 0 \end{bmatrix}.$$

Здесь c_{n-2} и s_{n-2} определяем из уравнения

$$c_{n-2}\beta_{n-2}^{(k)} - s_{n-2}\widetilde{\alpha}_{n-1}^{(k)} = 0.$$

Таким образом,

$$c_{n-2} = \widetilde{\alpha}_{n-1}^{(k)}/\sqrt{D}, \quad s_{n-2} = \beta_{n-2}^{(k)}/\sqrt{D},$$

где

$$D = (\beta_{n-2}^{(k)})^2 + (\widetilde{\alpha}_{n-1}^{(k)})^2.$$

$$\gamma_{n-3} = c_{n-2}\beta_{n-3}^{(k)},$$

$$\widetilde{\alpha}_{n-2}^{(k)} = c_{n-2}\alpha_{n-2}^{(k)} - s_{n-2}\gamma_{n-2},$$

$$\widetilde{\beta}_{n-2}^{(k)} = s_{n-2}\alpha_{n-2}^{(k)} + c_{n-2}\gamma_{n-2},$$

$$\widetilde{\widetilde{\alpha}}_{n-1}^{(k)} = s_{n-2}\beta_{n-2}^{(k)} + c_{n-2}\widetilde{\alpha}_{n-1}^{(k)}.$$

При дальнейшем умножении матрицы $Q_{n-2,n-1}Q_{n-1,n}T_k$ слева на матрицы вращения два последних столбца изменяться не будут. Поэтому эту матрицу можно домножить справа на $Q_{n-1,n}^T$. Умножение приведет к изменению двух последних столбцов:

$$\begin{bmatrix} \widetilde{\widetilde{\alpha}}_{n-1}^{(k)} & 0 \\ \widetilde{\beta}_{n-1}^{(k)} & \widetilde{\alpha}_n^{(k)} \end{bmatrix} \begin{bmatrix} c_{n-1} & s_{n-1} \\ -s_{n-1} & c_{n-1} \end{bmatrix} = \begin{bmatrix} \widetilde{\widetilde{\alpha}}_{n-1}^{(k)} & \beta_{n-1}^{(k+1)} \\ * & \alpha_n^{(k+1)} \end{bmatrix}.$$

Отметим, что последний столбец матрицы $Q_{n-2,n-1}Q_{n-1,n}T_kQ_{n-1,n}^T$ уже изменяться не будет. Таким образом мы вычислили последний столбец матрицы T_{k+1}, т. е., числа $\beta_{n-1}^{(k+1)}$, $\alpha_n^{(k+1)}$.

Повторяя вычисления описанные в последних двух абзацах, последовательно вычислим столбцы с номерами $n-1$, $n-2$, $\cdots$, 1. Заметим, что вычисление столбцов с номерами 2 и 1 требует некоторой модификации приведенных выше вычислений.

6.7. Метод вращений

Рассмотрим еще один метод решения полной проблемы собственных значений для симметричной вещественной матрицы $A \in R^{n\times n}$ называемый методом вращений или методом Якоби.

Метод строит последовательность ортогонально подобных матриц

$$\{A_k\}_{k=1}^{\infty} \quad (A_1 = A)$$

так, что

$$A_k \to \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \text{ при } k \to \infty.$$

Опишем переход от k-й к $(k+1)$-й итерации. В матрице A_k найдем максимальный по модулю внедиагональный элемент

$$|a_{pq}^{(k)}| = \max_{i\neq j} |a_{ij}^{(k)}|. \tag{6.25}$$

Построим матрицу вращения $Q_{pq}^{(k)}$, которая отличается от единичной матрицы элементами

$$q_{pp}^{(k)} = \cos\varphi_k, \quad q_{pq}^{(k)} = -\sin\varphi_k, \quad q_{qp}^{(k)} = \sin\varphi_k, \quad q_{qq}^{(k)} = \cos\varphi_k.$$

φ_k выберем так, чтобы в матрице

$$A_{k+1} = (Q_{pq}^{(k)})^T A_k Q_{qp}^{(k)}$$

в позициях (p, q) и (q, p) были нули. Отметим, что на последующих итерациях эти нули не сохраняются.

Обозначим

$$B = A_k Q_{pq}^{(k)},$$

тогда

$$A_{k+1} = (Q_{pq}^{(k)})^T B.$$

При переходе от A_k к B пересчитываются только p-й и q-й столбцы:

$$\begin{aligned} b_{ip} &= a_{ip}^{(k)} c + a_{iq}^{(k)} s, \\ b_{iq} &= -a_{ip}^{(k)} s + a_{iq}^{(k)} c, \quad i = \overline{1, n}, \end{aligned} \tag{6.26}$$

где $c = \cos\varphi_k$, $s = \sin\varphi_k$. При переходе от B к A_{k+1} пересчитываются только p-я и q-я строки:

$$\begin{aligned} a_{pj}^{(k+1)} &= b_{pj} c + b_{qj} s, \\ a_{qj}^{(k+1)} &= -b_{pj} s + b_{qj} c, \quad j = \overline{1, n}. \end{aligned} \tag{6.27}$$

Из (6.26), (6.27) получаем

$$a_{pq}^{(k+1)} = b_{pq} c + b_{qq} s = (-a_{pp}^{(k)} s + a_{pq}^{(k)} c) c + (-a_{qp}^{(k)} s + a_{qq}^{(k)} c) c =$$

$$= (a_{qq}^{(k)} - a_{pp}^{(k)}) sc + a_{pq}(c^2 - s^2).$$

Теперь условие,

$$a_{pq}^{(k+1)} = 0,$$

для определения матрицы вращения $Q_{pq}^{(k)}$ приобретает вид

$$(a_{qq}^{(k)} - a_{pp}^{(k)}) sc + a_{pq}(c^2 - s^2) = 0.$$

Разделив это уравнение на $-c^2$ и введя обозначения

$$t = \frac{s}{c}, \quad \tau = \frac{a_{pp}^{(k)} - a_{qq}^{(k)}}{2a_{pq}},$$

приходим к квадратному уравнению

$$t^2 + 2\tau t - 1 = 0.$$

В качестве t выберем меньший по модулю корень квадратного уравнения

$$t = \operatorname{sign}(\tau)(\sqrt{1+\tau^2} - |\tau|) = \frac{\operatorname{sign}(\tau)}{|\tau| + \sqrt{1+\tau^2}}. \tag{6.28}$$

Теперь $\sin\varphi_k$ и $\cos\varphi_k$ можно найти по формулам

$$\cos\varphi_k = \frac{1}{\sqrt{1+t^2}}, \quad \sin\varphi_k = t\cos\varphi_k. \tag{6.29}$$

Таким образом, переход от k-й к $(k+1)$-й итерации осуществляется так:

1. находим максимальный по модулю внедиагональный элемент матрицы A_k;
2. по формулам (6.28),(6.29) вычисляем $\cos\varphi_k$ и $\sin\varphi_k$;
3. элементы матрицы A_{k+1} вычисляем по формулам (6.26), (6.27).

Исследуем сходимость метода вращений. Обозначим через σ_k^2 сумму квадратов внедиагональных элементов матрицы A_k. Докажем, что

$$\sigma_k^2 \to 0 \text{ при } k \to \infty.$$

Поскольку ортогональные преобразования сохраняют норму Фробениуса, то

$$\sigma_{k+1}^2 = ||A_{k+1}||_F^2 - \sum_{i=1}^{n}(a_{ii}^{(k+1)})^2 = ||A_k||_F^2 - \sum_{i=1\, i\neq p,q}^{n}(a_{ii}^{(k)})^2 - (a_{pp}^{(k+1)})^2 - (a_{qq}^{(k+1)})^2.$$

Очевидно, что

$$\begin{bmatrix} a_{pp}^{(k+1)} & a_{pq}^{(k+1)} \\ a_{qp}^{(k+1)} & a_{qq}^{(k+1)} \end{bmatrix} = \begin{bmatrix} cos\varphi_k & -\sin\varphi_k \\ \sin\varphi_k & cos\varphi_k \end{bmatrix}^T \begin{bmatrix} a_{pp}^{(k)} & a_{pq}^{(k)} \\ a_{qp}^{(k)} & a_{qq}^{(k)} \end{bmatrix} \begin{bmatrix} cos\varphi_k & -\sin\varphi_k \\ \sin\varphi_k & cos\varphi_k \end{bmatrix}.$$

Поскольку $a_{pq}^{(k+1)} = a_{qp}^{(k+1)} = 0$, приравнивая нормы Фробениуса этих ортогонально подобных матриц, получаем

$$(a_{pp}^{(k+1)})^2 + (a_{qq}^{(k+1)})^2 = (a_{pp}^{(k)})^2 + 2(a_{pq}^{(k)})^2 + (a_{qq}^{(k)})^2.$$

Поэтому

$$\sigma_{k+1}^2 = ||A_k||_F^2 - \sum_{i=1\, i\neq p,q}^{n} (a_{ii}^{(k)})^2 - (a_{pp}^{(k+1)})^2 - (a_{qq}^{(k+1)})^2 =$$

$$= ||A_k||_F^2 - \sum_{i=1}^{n} (a_{ii}^{(k)})^2 - 2(a_{pq}^{(k)})^2 = \sigma_k^2 - 2(a_{pq}^{(k)})^2.$$

Очевидно, что

$$\sigma_k^2 \leq n(n-1)(a_{pq}^{(k)})^2.$$

Поэтому

$$\sigma_{k+1}^2 \leq \left(1 - \frac{2}{n(n-1)}\right) \sigma_k^2.$$

Применяя полученное неравенство последовательно, получаем

$$\sigma_{k+1}^2 \leq \left(1 - \frac{2}{n(n-1)}\right)^k \sigma_1^2.$$

Значит,

$$\sigma_{k+1}^2 \to 0 \text{ при } k \to \infty,$$

и сходимость метода линейная. Однако асимптотическая скорость сходимости значительно выше линейной. Так для больших k она квадратичная[9].

Метод Якоби исторически является старейшим для проблемы собственных значений. Обычно он много медленнее перечисленных выше методов, имея трудоемкость $O(n^3)$ арифметических операций. Однако интерес к методу сохраняется, потому что подчас он дает гораздо более точные результаты, чем другие методы [9].

Приведем алгоритм метода Якоби.

Алгоритм 6.7.1. ***Метод Якоби***

1. $Q = I$
2. $While\ True\ Do$
3. $\quad |a_{pq}| = \max\limits_{i\neq j} |a_{ij}|$
4. $\quad If\ |a_{pq}| \leq \varepsilon ||A||\ Then\ Break$
5. $\quad \tau = \frac{a_{pp} - a_{qq}}{2a_{pq}}$
6. $\quad t = \frac{sign(\tau)}{|\tau| + \sqrt{1+\tau^2}}$
7. $\quad c = 1/\sqrt{1+t^2}; s = tc$
8. $\quad For\ i = 1 : n$
9. $\quad\quad b_p(i) = a_{ip}c + a_{iq}s$
10. $\quad\quad b_q(i) = -a_{ip}s + a_{iq}c$

11. $Next\ i$
12. $a(:,p) = b_p;\ a(:,q) = b_q$
13. $For\ j = 1 : n$
14. $\quad b_p(j) = a_{pj}c + a_{qj}s$
15. $\quad b_q(i) = -a_{pj}s + a_{qj}c$
16. $Next\ j$
17. $a(p,:) = b_p;\ a(q,:) = b_q$
18. $For\ i = 1 : n$
19. $\quad b_p(i) = Q_{ip}c + Q_{iq}s$
20. $\quad b_q(i) = -Q_{ip}s + Q_{iq}c$
21. $Next\ i$
22. $Q(:,p) = b_p;\ Q(:,q) = b_q$
23. $End\ Do$

6.8. Вопросы и задания

1. Пусть $A \in R^{n\times n}$ — симметричная матрица, $\lambda \in R$, $x \in R^n$ — произвольное число и вектор, причем $||x|| = 1$. Доказать, что существует собственное число λ_k матрицы A, для которого $|\lambda_k - \lambda| \leq ||Ax - \lambda x||_2$.

2. Пусть $A = A^T > 0$. Доказать, что если $\lambda_{max}(A) = a_{kk}$ при некотором $1 \leq k \leq n$, то $a_{ik} = a_{kj}$ при всех $i \neq k$, $j \neq k$.

3. Показать, что любой собственный вектор матрицы A есть собственный вектор матрицы A^2. Используя разложение по собственным векторам или какой-либо иной способ, показать, что обратное верно лишь тогда, когда различные собственные значения матрицы A отображаются в одно и то же собственное значение матрицы A^2.

4. Пусть собственные значения симметричной матрицы A удовлетворяют неравенствам $0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1} < \lambda_n$. Оценить норму разности матрицы $A^k/||A^k||$ и близкой к ней матрицей ранга 1.

5. Пусть λ — собственное значение симметричной трехдиагональной матрицы T. Показать, что если λ имеет алгебраическую кратность k, то по крайней мере $k - 1$ поддиагональных элементов матрицы T равны нулю.

6. Показать, что собственные значения симметричной матрицы размера 2×2 должны быть вещественные.

7. Методы решения нелинейных уравнений

Рассмотрим задачу вычисления корней уравнения

$$f(x) = 0 \tag{7.1}$$

или, что то же самое, нулей функции $f(x)$, где $f : R \to R$ — алгебраическая или трансцендентная функция. Нелинейная функция $f(x)$ в своей области определения $D(f) \subseteq R$ может иметь конечное или бесконечное количество нулей или не иметь их вовсе. Большинство же численных методов нахождения нулей функции требует знания промежутков, где заведомо имеется единственный нуль функции. Задачи о существовании, единственности, нахождении границ и локализации корней нелинейной функции обычно решаются средствами математического анализа. Рассмотрим, например, графический метод локализации корней. Предположим, что уравнение (7.1) можно представить в виде

$$f_1(x) = f_2(x), \tag{7.2}$$

где функции $f_1(x)$ и $f_2(x)$ таковы, что можно построить графики $y = f_1(x)$ и $y = f_2(x)$. Тогда корни уравнения (7.2) — абсциссы точек пересечения этих графиков.

Пример 7.0.1. Представим уравнение

$$x^2 - \sin x - 1 = 0$$

в виде

$$x^2 - 1 = \sin x.$$

Построив графики функций $y = x^2 - 1$ и $y = \sin x$, устанавливаем, что рассматриваемое уравнение имеет два корня: $\xi_1 \in [-1, 0]$, $\xi_2 \in [1, \pi/2]$.

В дальнейшем будем предполагать, что нам известен промежуток $[a, b]$, содержащий единственный корень уравнения (7.1).

7.1. Методы дихотомии. Метод хорд

Пусть функция $f(x)$ определена и непрерывна на промежутке $[a, b]$ и $f(a)f(b) < 0$, тогда, согласно теоремы Больцано-Коши, на интервале

(a, b) она имеет хотя бы один корень. Возьмем произвольную точку $c \in (a, b)$. Вычислим $f(c)$. Возможны случаи: $f(c) = 0$ — корень найден; $f(a)f(c) < 0$ — корень находится на промежутке (a, c); $f(c)f(b) < 0$ — корень находится на промежутке (c, b). Если корень не найден, процесс повторяем для нового промежутка. Наиболее распространенным случаем метода дихотомии является метод половинного деления, согласно которому точка $c = (a + b)/2$. За один шаг метода половинного деления содержащий корень промежуток уменьшается вдвое. Через k шагов искомый корень ξ будет находится на промежутке $[a_k, b_k]$ длиной $(b-a)/2^k$ и

$$|\xi - x_k| < \frac{b-a}{2^k}, \tag{7.3}$$

где $x_k \in (a_k, b_k)$. Значит,

$$\lim_{k\to\infty} x_k = \xi.$$

С другой стороны, неравенство (7.3) позволяет определить количество шагов, достаточное для вычисления корня ξ с заданной точностью ε.

Можно предположить, что метод дихотомии будет сходится быстрее, если отрезок $[a, b]$ делить на части точкой c не пополам, а пропорционально величинам ординат $f(a)$ и $f(b)$ графика функции $f(x)$. Это означает, что точку c есть смысл находить как абсциссу точки пересечения оси $0x$ с прямой, проходящей через точки $(a, f(a))$, $(b, f(b))$:

$$\frac{y - f(a)}{f(b) - f(a)} = \frac{x - a}{b - a}.$$

Полагая $y = 0$, находим

$$c = a - \frac{f(a)(b-a)}{f(b) - f(a)}.$$

Метод дихотомии с таким выбором пробной точки c называют методом хорд.

Отметим, что в общем случае, если на функцию $f(x)$ не накладывать никаких дополнительных ограничений, может оказаться, что метод хорд сходится медленнее метода половинного деления. Чтобы в этом убедиться, достаточно проанализировать возможное поведение нескольких приближений по методу хорд на рис. 7.1.

7.2. Метод итерации

Уравнение

$$f(x) = 0$$

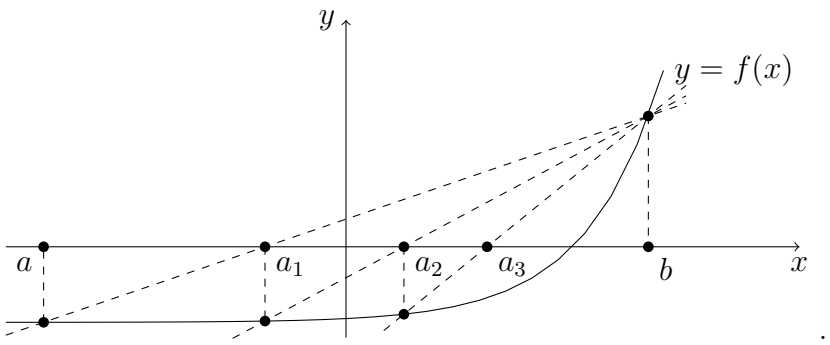

Рис. 7.1. Сходимость метода хорд

преобразуем к эквивалентному уравнению вида

$$x = \varphi(x). \tag{7.4}$$

Для нахождения корня уравнения (7.4) построим итерационный процесс, называемый методом простой итерации:

$$x_{k+1} = \varphi(x_k), \quad k = 0, 1, 2, \ldots, \tag{7.5}$$

где x_0 — некоторое начальное приближение.

Теорема 7.2.1. *Пусть функция $\varphi(x)$ определена и дифференцируема на отрезке $[a, b]$. Тогда, если выполняются условия:*

$$\varphi(x) \in [a, b], \quad \forall x \in [a, b], \tag{7.6}$$

$$\exists q : |\varphi'(x)| \le q < 1 \quad \forall x \in (a, b), \tag{7.7}$$

то для любого начального приближения $x_0 \in [a, b]$ последовательность $\{x_k\}$, определяемая методом простых итераций (7.5), сходится к единственному корню $\xi \in [a, b]$ уравнения (7.4). При этом справедливы следующие оценки:

$$|\xi - x_k| \le \frac{q}{1-q}|x_k - x_{k-1}|, \tag{7.8}$$

$$|\xi - x_k| \le \frac{q^k}{1-q}|x_1 - x_0|. \tag{7.9}$$

Доказательство. Пусть $x_0 \in [a, b]$. Из (7.6) следует, что $x_1 = \varphi(x_0) \in [a, b]$. По индукции получаем, что все члены последовательности $\{x_k\}$ принадлежат отрезку $[a, b]$. Из равенства (7.5) вычтем равенство $x_k = \varphi(x_{k-1})$ и к правой части полученного равенства

$$x_{k+1} - x_k = \varphi(x_k) - \varphi(x_{k-1})$$

применим формулу Лагранжа, согласно которой на интервале, определяемом точками x_{k-1} и x_k, найдется такая точка ϑ_k, что

$$\varphi(x_k) - \varphi(x_{k-1}) = \varphi'(\vartheta_k)(x_k - x_{k-1}).$$

Следовательно,

$$x_{k+1} - x_k = \varphi'(\vartheta_k)(x_k - x_{k-1}),$$

и, в силу условия (7.7), справедливо неравенство

$$|x_{k+1} - x_k| \le q|x_k - x_{k-1}|. \tag{7.10}$$

На основе (7.10) получаем неравенство

$$|x_{k+i} - x_{k+i-1}| \le q^i|x_k - x_{k-1}| \quad \forall i, k \in N,$$

в частности

$$|x_k - x_{k-1}| \le q^{k-1}|x_1 - x_0|. \tag{7.11}$$

Поэтому,

$$|x_{k+m} - x_k| \le |x_{k+m} - x_{k+m-1}| + |x_{k+m-1} - x_{k+m-2}| + \dots$$

$$\dots + |x_{k+2} - x_{k+1}| + |x_{k+1} - x_k| \le \left(q^m + q^{m-1} + \dots + q^2 + q\right)|x_k - x_{k-1}| =$$

$$= \frac{q - q^{m+1}}{1 - q}|x_k - x_{k-1}|,$$

т. е.,

$$|x_{k+m} - x_k| \le \frac{q - q^{m+1}}{1 - q}|x_k - x_{k-1}|. \tag{7.12}$$

Из (7.11) и (7.12) получаем

$$|x_{k+m} - x_k| \le \frac{q^k}{1 - q}(1 - q^m)|x_1 - x_0|. \tag{7.13}$$

Поскольку правая часть неравенства (7.13) при фиксированном m и $k \to \infty$ стремится к нулю, последовательность $\{x_k\}$ является фундаментальной, а значит, сходится к некоторому пределу ξ, причем $\xi \in [a, b]$, ибо все члены последовательности принадлежат замкнутому промежутку $[a, b]$. Так как дифференцируемая функция непрерывна, то, рассматривая предел равенства (7.5), находим, что ξ — решение уравнения (7.4).

Предположим, что существует еще один корень $\widetilde{\xi} \in [a, b]$ уравнения (7.4). Тогда

$$\xi - \widetilde{\xi} = \varphi(\xi) - \varphi(\widetilde{\xi}),$$

и, по формуле Лагранжа,

$$\xi - \widetilde{\xi} = \varphi'(\vartheta)(\xi - \widetilde{\xi}).$$

Последнее же равенство возможно лишь при $\xi = \widetilde{\xi}$, ибо по условию теоремы $\varphi'(\vartheta)$ не может равняться единице.

Переходя к пределу при $m \to \infty$ в неравенствах (7.12) и (7.13), получаем оценки (7.8) и (7.9), соответственно. □

В практических вычисления оценку (7.8) используют для завершения итерационного процесса (7.5). Если

$$|x_k - x_{k-1}| \le \frac{1-q}{q}\varepsilon, \tag{7.14}$$

то $\xi \approx x_k(\pm\varepsilon)$. Оценку же (7.9) используют для предварительного определения числа итераций, достаточных для вычисления корня с заданной точностью ε. Рисунок 7.2 иллюстрирует одностороннюю сходимость простой итерации, а рисунок 7.3 — двухстороннюю сходимость.

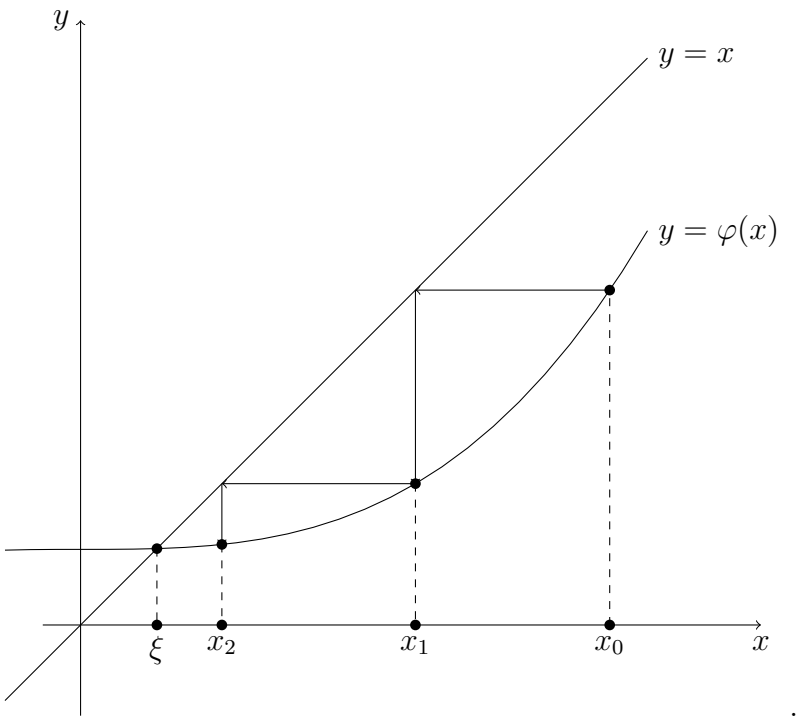

Рис. 7.2. Односторонняя сходимость простой итерации

В заключении отметим, что переход от уравнения

$$f(x) = 0$$

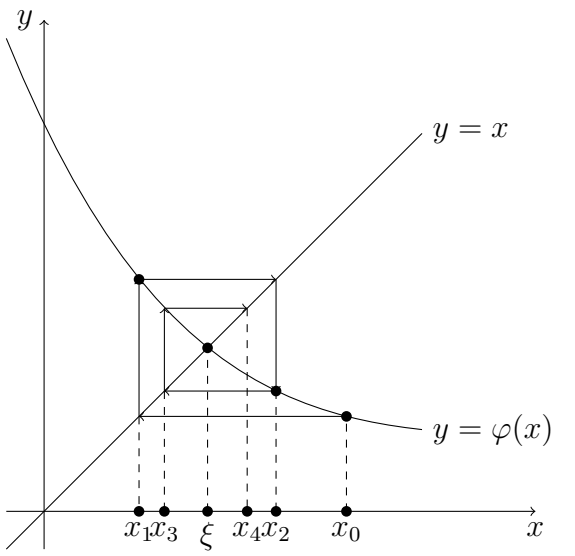

Рис. 7.3. Двухсторонняя сходимость простой итерации

к уравнению

$$x = \varphi(x)$$

можно осуществить так:

$$x = x - \lambda f(x),$$

где параметр λ выбираем таким образом, чтобы в нужной области производная от функции $\varphi(x)$

$$\varphi'(x) = 1 - \lambda f'(x)$$

была по модулю меньше 1, т. е., чтобы выполнялось условие (7.7) сходимости простой итерации.

Пример 7.2.1. Пусть

$$0 < \alpha \le f'(x) \le \beta < \infty \quad \forall x \in [a, b].$$

Тогда

$$1 - \lambda\beta \le \varphi'(x) \le 1 - \lambda\alpha$$

и

$$|\varphi'(x)| \le q(\lambda) = \max\{|1 - \lambda\alpha|, |1 - \lambda\beta|\}$$

для любого $x \in [a, b]$. Легко установить, что $q(\lambda) < 1$ при $\lambda \in (0, 2/\beta)$. Так при $\lambda = 1/\beta$

$$0 \le \varphi'(x) \le 1 - \frac{\alpha}{\beta}.$$

Оптимальное же значение параметра λ вычислим из уравнения

$$-(1-\lambda\beta) = 1-\lambda\alpha.$$

Тогда

$$\alpha_{opt} = \frac{2}{\alpha+\beta}$$

и

$$q(\alpha_{opt}) = \frac{\beta-\alpha}{\alpha+\beta}.$$

7.3. Метод Ньютона

Предположим, что в некоторой окрестности корня функция $f(x)$ строго возрастает и выпукла вниз. Выберем начальное приближение x_0 достаточно близко к корню ξ (см. рис. 7.4). Запишем уравнение касательной к графику функции $y = f(x)$ в точке x_0:

$$y = f(x_0) + f'(x_0)(x - x_0).$$

В качестве следующего приближения к корню возьмем точку пересечения этой касательной с осью $0x$:

$$0 = f(x_0) + f'(x_0)(x_1 - x_0).$$

Тогда

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

По приближению x_1 вычисляем приближение x_2, и т. д. Полученный итерационный процесс называется методом Ньютона:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, \ldots, \tag{7.15}$$

где x_0 — начальное приближение.

Теорема 7.3.1. *Пусть на отрезке $[a, b]$ функция $f(x)$ имеет первую и вторую производные постоянного знака и*

$$f(a)f(b) < 0. \tag{7.16}$$

Тогда, если начальное приближение x_0 выбрано на $[a, b]$ так, что

$$f(x_0)f''(x_0) > 0, \tag{7.17}$$

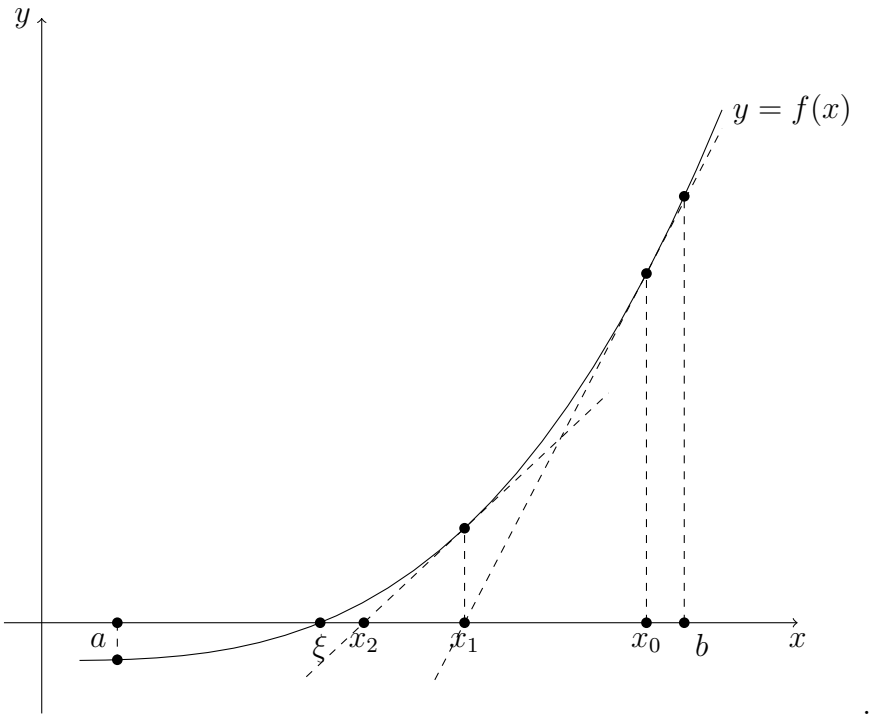

Рис. 7.4. Метод Ньютона

то последовательность $\{x_k\}$*, определяемая методом Ньютона (7.15), монотонно сходится к корню* $\xi \in (a, b)$ *уравнения* $f(x) = 0$*. Кроме того,*

$$|\xi - x_{k+1}| \le \frac{\beta}{2\alpha}|\xi - x_k|^2, \tag{7.18}$$

$$|\xi - x_{k+1}| \le \frac{\beta}{2\alpha}|x_{k+1} - x_k|^2, \tag{7.19}$$

где α *и* β *такие константы, что*

$$0 < \alpha \le |f'(x)|,\ |f''(x)| \le \beta < \infty\ \forall x \in [a, b].$$

Доказательство. Сначала заметим, что условия теоремы обеспечивают существование единственного корня $\xi \in (a, b)$. Положим для определенности, что $f'(x) > 0$, $f''(x) > 0$ для любого $x \in [a, d]$ и $f(a) < 0$, $f(b) > 0$. В этом случае в качестве x_0 можно взять любую точку из $(\xi, b]$. Очевидно, что $f(x_0) > 0$.

Запишем формулу Тейлора функции $f(x)$ в точке x_0:

$$f(\xi) = f(x_0) + f'(x_0)(\xi - x_0) + \frac{f''(\xi_0)}{2}(\xi - x_0)^2 = 0,$$

где $\xi_0 \in (a, b)$. Отсюда следует, что

$$f(x_0) + f'(x_0)(\xi - x_0) < 0.$$

Значит,

$$\xi < x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} < x_0.$$

Предположим теперь, что $\xi < x_k$ и $f(x_k) > 0$. Запишем формулу Тейлора функции $f(x)$ в точке x_k:

$$f(x_k) + f'(x_k)(\xi - x_k) + \frac{f''(\xi_k)}{2}(\xi - x_k)^2 = 0, \tag{7.20}$$

где $\xi_k \in (a, b)$. Отсюда следует, что

$$f(x_k) + f'(x_k)(\xi - x_k) < 0.$$

Значит,

$$\xi < x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} < x_k \quad \forall k \in N.$$

Таким образом, последовательность $\{x_k\}$, формируемая методом Ньютона (7.15), ограничена снизу и монотонно убывает. Значит,

$$\lim_{k\to\infty} x_k = \xi.$$

Из (7.15) получаем

$$f(x_k) + f'(x_k)(x_{k+1} - x_k) = 0. \tag{7.21}$$

Приравнивая левые части равенств (7.20) и (7.21), получаем

$$f(x_k) + f'(x_k)(\xi - x_k) + \frac{f''(\xi_k)}{2}(\xi - k_0)^2 = f(x_k) + f'(x_k)(x_{k+1} - x_k).$$

Значит,

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

Отсюда сразу следует оценка (7.18).

Запишем формулу Тейлора для функции $f(x)$ в точке x_k

$$f(x_{k+1}) = f(x_k) + f'(x_k)(x_{k+1} - x_k) + \frac{f''(\widetilde{\xi}_k)}{2}(x_{k+1} - x_k)^2,$$

где $\widetilde{\xi}_k \in (a, b)$. Отсюда, учитывая (7.21), получаем

$$f(x_{k+1}) = \frac{f''(\widetilde{\xi}_k)}{2}(x_{k+1} - x_k)^2.$$

Поэтому,

$$|f(x_{k+1})| \le \frac{\beta}{2}(x_{k+1} - x_k)^2. \tag{7.22}$$

По формуле Лагранжа

$$f(\xi) - f(x_{k+1}) = f'(\tau_{k+1})(\xi - x_{k+1}).$$

Значит,

$$|\xi - x_{k+1}| \le \frac{1}{\alpha}|f(x_{k+1})|. \tag{7.23}$$

Наконец, из (7.22) и (7.23) следует оценка (7.19). □

Чтобы более объективно судить о скорости сходимости итерационных процессов, вводят следующие понятия.

Говорят, что последовательность $\{x_k\}$ сходится к ξ по крайней мере с p-м порядком (итерационный процесс имеет по крайней мере p-й порядок скорости сходимости), если существуют такие константы $C > 0$, $p \ge 1$, $\nu \in (0, 1)$ и $k_0 \in N$, что

$$|\xi - x_{k+1}| \le C|\xi - x_k|^p \tag{7.24}$$

или

$$|x_{k+1} - x_k| \le C|x_k - x_{k-1}|^p \tag{7.25}$$

или

$$|\xi - x_k| \le C\nu^p \tag{7.26}$$

при всех $k \ge k_0$.

Если $p = 1$, то говорят, что итерационный процесс сходится линейно, или имеет первый порядок скорости сходимости. Из (7.10) следует, что простая итерация сходится линейно. К линейной сходимости применяют также термин "сходимость со скоростью геометрической прогрессии".

Если $p = 2$, то говорят, что итерационный процесс сходится квадратично, или имеет второй порядок скорости сходимости. Из (7.18) следует, что метод Ньютона имеет второй порядок скорости сходимости.

Сравнение различных определений скорости сходимости итерационных процессов можно найти в [17].

7.4. Методы решения систем нелинейных уравнений

Пусть требуется решить систему уравнений

$$\begin{cases} f_1(x_1, x_2, \ldots, x_n) = 0, \\ f_2(x_1, x_2, \ldots, x_n) = 0, \\ \qquad \ldots \\ f_n(x_1, x_2, \ldots, x_n) = 0, \end{cases} \tag{7.27}$$

где f_i $(i = \overline{1,n})$ — заданные нелинейные вещественнозначные функции вещественных переменных. Приняв обозначения

$$x = [x_1, x_2, \ldots, x_n]^T, F(x) = [f_1(x), f_2(x), \ldots, f_n(x)]^T,$$

систему (7.27) можно записать в виде:

$$F(x) = 0, \tag{7.28}$$

где $F : R^n \to R^n$ — векторная функция векторного аргумента.

Преобразуем систему (7.27) к эквивалентной системе вида

$$\begin{cases} x_1 = \varphi_1(x_1, x_2, \ldots, x_n) = 0, \\ x_2 = \varphi_2(x_1, x_2, \ldots, x_n) = 0, \\ \qquad \ldots \\ x_n = \varphi_n(x_1, x_2, \ldots, x_n) = 0, \end{cases} \tag{7.29}$$

или, в компактной форме:

$$x = \Phi(x), \tag{7.30}$$

где $\Phi(x) = [\varphi_1(x), \varphi_2(x), \ldots, \varphi_n(x)]^T$.

Метод простых итераций решения задачи (7.30) определяется рекуррентным равенством

$$x_{k+1} = \Phi(x_k), \quad k = 0, 1, 2, \ldots, \tag{7.31}$$

где x_0 — начальное приближение. Справедлива следующая теорема.

Теорема 7.4.1. *Пусть функция $\Phi(x)$ и замкнутое множество $M \subseteq D(\Phi) \subseteq R^n$ таковы, что:*

$$\Phi(x) \in M, \quad \forall x \in M; \tag{7.32}$$

$$\exists q < 1 : ||\Phi(x) - \Phi(\widetilde{x})|| \le q||x - \widetilde{x}|| \quad \forall x, \widetilde{x} \in M. \tag{7.33}$$

Тогда для любого начального приближения $x_0 \in M$ последовательность $\{x_k\}$, определяемая методом простых итераций (7.31), сходится к единственному корню $\xi \in M$ уравнения (7.30). При этом справедливы оценки:

$$||\xi - x_k|| \le \frac{q}{1-q}||x_k - x_{k-1}|| \le \frac{q^k}{1-q}||x_1 - x_0||. \tag{7.34}$$

Доказательство. Доказательство этой теоремы почти полностью повторяет доказательство теоремы 7.2.1 (см., например, [3]). □

Чтобы применить метод простых итераций (7.31) к системе (7.28), ее надо привести к виду (7.30), например, так:

$$x = x - AF(x), \tag{7.35}$$

где $A \in R^{n\times n}$ — невырожденная матрица, которую надо подобрать так, чтобы вектор-функция $\Phi(x) = x - AF(x)$ обладала нужными свойствами.

Пусть $\{A_k\}$ — некоторая последовательность невырожденных $n \times n$ -матриц. Очевидно, что задачи

$$x = x - A_k F(x), \quad k = 0, 1, 2, \ldots$$

имеют то же решение, что и задача (7.28). Для приближенного нахождения этого решения формально определим итерационный процесс:

$$x_k = x_k - A_k F(x_k), \quad k = 0, 1, 2, \ldots. \tag{7.36}$$

Заметим, что если $A_k \equiv A$, то итерационный процесс (7.36) — это метод простых итераций для уравнения (7.35). Если же A_k различны при разных k, то формула (7.36) определяет большое семейство итерационных методов.

Положим $A_k = (F'(x_k))^{-1}$, где

$$F'(x) = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} & \cdots & \frac{\partial f_2(x)}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n(x)}{\partial x_1} & \frac{\partial f_n(x)}{\partial x_2} & \cdots & \frac{\partial f_n(x)}{\partial x_n} \end{bmatrix}$$

— матрица Якоби для вектор-функции $F(x)$. Подставив эту матрицу A_k в (7.36), получим явную формулу метода Ньютона

$$x_{k+1} = x_k - (F'(x_k))^{-1} F(x_k), \tag{7.37}$$

обобщающего на многомерный случай скалярный метод Ньютона (7.15). Формулу (7.37) можно записать так

$$F'(x_k)(x_{k+1} - x_k) = -F(x_k). \tag{7.38}$$

Обозначив $p_k = x_{k+1} - x_k$, получаем неявную формулу метода Ньютона:

$$\begin{gathered} F'(x_k)p_k = -F(x_k), \\ x_{k+1} = x_k + p_k, \quad k = 0, 1, 2, \ldots. \end{gathered} \tag{7.39}$$

Сравнивая формулу (7.38) с формальным разложением функции $F(x)$ в ряд Тейлора

$$F(x) = F(x_k) + F'(x_k)(x - x_k) + \frac{1}{2!}F''(x_k)(x - x_k)^2 + \ldots,$$

видим, что последовательность $\{x_k\}$ в методе Ньютона получается в результате подмены на каждом шаге нелинейного уравнения $F(x) = 0$ линейным уравнением

$$F(x_k) + F'(x_k)(x - x_k) = 0,$$

т. е., пошаговой линеаризацией. Как следствие этого факта, можно рассчитывать, что при достаточной гладкости $F(x)$ и достаточно хорошем начальном приближении x_0, сходимость метода Ньютона будет квадратичной и в многомерном случае. Имеется ряд теорем, устанавливающих это при тех или иных предположениях (см., например, [1, 8, 21]).

7.5. Вопросы и задания

1. Методом простой итерации найти наименьший корень уравнения $e^x - 3x^2 + 5 = 0$ с точностью $\varepsilon = 10^{-6}$. Определить количество итераций, затраченных методом на вычисление корня с заданной точностью ε. Найти теоретически количество итераций, достаточное для вычисления корня с заданной точностью ε.

2. Методом Ньютона найти наименьший корень уравнения $e^x - 3x^2 + 5 = 0$ с точностью $\varepsilon = 10^{-6}$. Определить количество итераций, затраченных методом на вычисление корня с заданной точностью ε.

3. Доказать, что итерационный процесс $x_{k+1} = \cos x_k$ сходится для любого начального приближения $x_0 \in R$.

4. Исследовать сходимость метода простой итерации $x_{k+1} = x_k^2 - 2x_k + 2$ в зависимости от выбора начального приближения x_0.

5. Построить итерационный процесс Ньютона для вычисления $\sqrt[p]{a}$, $a > 0$, где p — вещественное число.

6. Пусть для решения уравнения $x^3 - x = 0$ применяется метод Ньютона. При каком начальном приближении он сходится и к какому корню?

7. Показать графически, что система уравнений

$$\begin{cases} x_1^2 + x_2^2 - 1 = 0, \\ x_1^2 - x_2 = 0 \end{cases}$$

имеет ровно два решения. В каких точках x матрица Якоби вектор-функции системы не вырождена. Методом Ньютона найти решения системы с точностью $\varepsilon = 10^{-6}$.

Список рекомендованной литературы

[1] Бахвалов Н. С., Жидков Н. П., Кобельков Г. М. Численные методы. — М.: Лаборатория базовых знаний, 2001. — 640 с.

[2] Бахвалов Н. С., Лапин А. В., Чижонков Е. В. Численные методы в задачах и упражнениях. — М.: Высшая школа, 2000. — 190 с.

[3] Вержбицкий В. М. Численные методы(линейная алгебра и нелинейные уравнения): Учеб. пособие. — М.: ОНИКС, 2005. — 432 с.

[4] Воеводин В. В. Линейная алгебра. — М.: Наука, 1974. — 336 с.

[5] Воеводин В.В. Вычислительные основы линейной алгебры. — М.: Наука, 1977. — 303 с.

[6] Воеводин В. В., Кузнецов Ю. А. Матрицы и вычисления. — М.: Наука, 1984. — 318 с.

[7] Голуб Дж., Ван Лоун Ч. Матричные вычисления. — М.: Мир, 1999. — 548 с.

[8] Демидович Б. П., Марон И. А. Основы вычислительной математики. — М.: Наука, 1966. — 664 с.

[9] Деммель Дж. Вычислительная линейная алгебра. Теория и приложения. — М.: Мир, 2002. — 429 с.

[10] Джордж А., Лю Дж. Численное решение больших разреженных систем. — М.: Мир. 1984. — 333 с.

[11] Ильин В. А., Позняк Э. Г. Линейная алгебра. — М.: Наука, 1974. — 296 с.

[12] Курош А. Г. Курс высшей алгебры. — М.: Наука, 1968. — 431 с.

[13] Лоусон Ч., Хенсон Р. Численное решение задач метода наименьших квадратов. — М. Наука, 1986. — 232 с.

[14] Масловская Л. В., Масловская О. М. Численные методы алгебры. Учебное пособие. — Одесса, Укрполиграф, 2006. — 146 с.

[15] Мэтьюз Д. Финк К. Численные методы. Использование MATLAB. — Вильямс, 2001. — 716 с.

[16] Ортега Дж. Введение в параллельные и векторные методы решения линейных систем. — М.: Мир, 1991. — 365 с.

[17] Ортега Дж., Рейнболдт В. Итерационные методы решения нелинейных систем уравнений со многими неизвестными. — М.: Мир, 1975. — 560 с.

[18] Парлетт Б. Симметричная проблема собственных значений. — М.: Мир, 1983. — 382 с.

[19] Уилкинсон Дж. Х. Алгебраическая проблема собственных значений. — М.: Наука, 1970. — 564 с.

[20] Хорн Р., Джонсон Ч. Матричный анализ. Пер. с англ. — М.: Мир, 1989. — 666 с.

[21] Цегелик Г. Г. Чисельні методи. — Львів: Світ, 2005. — 407 с.

[22] Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры. — М.: Государственное издательство физико-математическом литературы, 1963. — 655 с.

[23] Форсайт Дж., Малькольм М., Моулер К. Машинные методы математических вычислений. — М.: Мир, 1980. — 280 с.

[24] Форсайт Дж., Молер К. Численное решение систем линейных алгебраических уравнений. — М.: Мир, 1967. — 167 с.

[25] Чен К., Джиблин П., Ирвинг А. MATLAB в математических исследованиях. — М.: Мир, 2001. — 346 с.

[26] Bai Z., Demmel J., Dongarra J., Ruhe A., and H. van der Vorst. Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide. — SIAM, Philadelphia, PA, USA, 2000. — 410 p.

[27] Barrett R., Berry M. and others. Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods. — SIAM, Philadelphia, 1994. — 107 p.

[28] Dongarra J., Duff I., Sorensen D. and H. van der Vorst. Numerical Linear Algebra for High-Performance Computers. - SIAM, Philadelphia, PA, 1998. — 336 p.

[29] Higham Nicholas J., Functions of Matrices: Theory and Computation. - SIAM, 2008. — 425 p.

[30] Higham Nicholas J., Accuracy and Stability of Numerical Algorithms. — SIAM, Second edition, 2002. — 663 p.

[31] Higham Desmond J. and Higham Nicholas J., MATLAB Guide. — SIAM, Second edition, 2005. — 382 p.

[32] Meyer Carl D. Matrix analysis and applied linear algebra. — SIAM, 2000. — 718 p.

[33] Saad Y. Iterative Methods for Sparse Linear Systems, Second Edition. — SIAM, 2003. — 350 p.

[34] Saad Y. Numerical Methods for Large Eigenvalue Problems: Revised Edition. — SIAM, 2011. — 271 p.

[35] Vorst H. A. Computational Methods for large Eigenvalue Problems. — in P.G. Ciarlet and J.L. Lions (eds), Handbook of Numerical Analysis, Volume VIII, North-Holland (Elsevier), Amsterdam 2002, pp. 3-179.

[36] Vorst H. A. Iterative Krylov Methods for Large Linear systems. — Cambridge University Press, Cambridge, 2003. — 221 p.

Предметный указатель

Навчальне видання

Вербіцький Віктор Васильович
Реут Віктор Всеволодович

ВВЕДЕННЯ В ЧИСЛОВІ МЕТОДИ АЛГЕБРИ

Навчальний посібник
(*Російською мовою*)

За редакцією авторів

Підп. до друку 17.03.2015. Формат 70x108/16.
Умов.-друк. арк. — 14,52. Тираж 100 пр.
Зам. № 1088.

Видавець і виготовлювач
Одеський національний університет імені І. І. Мечникова
Свідоцтво суб'єкта видавничої справи ДК № 4215 від 22.11.2011 р.

Україна, 65082, м. Одеса, вул. Єлісаветинська, 12
Тел. (048) 723-28-39. E-mail: druk@onu.edu.ua