

разложениях (методы отражений, вращений и ортогонализации), метод прогонки для случая системы с трехдиагональной матрицей; дается строгое теоретическое обоснование этих методов. Рассматриваются итерационные методы решения систем линейных уравнений, проводится подробный анализ вопросов, связанных с получением необходимых и достаточных условий сходимости различных методов (в том числе классических методов простой итерации и Зейделя), исследуется их сходимость с точки зрения канонической формы одншаговых итерационных процессов, затрагивается проблема оптимизации скорости сходимости; дается понятие об итерационных методах, основанных на вариационных принципах. Аналогичные вопросы изучаются при рассмотрении методов решения другой задачи линейной алгебры — алгебраической проблемы собственных значений, когда требуется вычислить собственные значения матрицы и соответствующие им собственные векторы. Ряд положений в книге иллюстрируются примерами. В конце каждой главы приводятся задачи для самостоятельной работы.

Автор благодарен рецензентам профессору А. К. Синицыну и профессору Л. А. Яновичу за замечания и предложения, способствовавшие улучшению книги, доценту кафедры вычислительной математики В. И. Репникову за по постоянное обсуждение данной проблематики, сотрудникам кафедры вычислительной математики доценту П. А. Вакульчику, старшему преподавателю И. С. Пукиной, а также сотрудникам Управления редакционно-издательской работы БГУ за помощь в подготовке рукописи. Автор особо признателен профессору А. Д. Егорову за моральную поддержку и ценные методические рекомендации.

ГЛАВА I

ОБУСЛОВЛЕННОСТЬ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

§ 1. Устойчивость систем линейных алгебраических уравнений

Пусть дана система линейных алгебраических уравнений (СЛАУ)

$$Ax = f, \quad (1)$$

где $A = (n \times n)$ -матрица с элементами a_{ij} , $i, j = 1, 2, \dots, n$, $x = (x_1, x_2, \dots, x_n)^T$ — искомый вектор-столбец с n -компонентами, $f = (f_1, f_2, \dots, f_n)^T$ — заданный вектор-столбец с n -компонентами.

Прежде чем использовать какой-либо метод для решения той или иной математической задачи, в вычислительной математике принято рассматривать вопрос о ее корректности.

Определение 1. Будем говорить, что задача поставлена корректно, если:

- 1) решение задачи существует и единствено;
- 2) решение задачи непрерывно зависит от входных данных (устойчиво относительно входных данных).

Как известно, для задачи (1) первое требование в определении 1 будет выполнено, если $\det A \neq 0$. В этом случае можно определить матрицу A^{-1} , обратную матрице A , и записать решение в виде

$$x = A^{-1}f.$$

Второе же требование в определении корректности применительно к (1) нуждается в некоторой детализации. Входными данными в задаче (1), очевидно, являются компоненты вектора f и элементы a_{ij} ,

§ 1. УСТОЙЧИВОСТЬ СЛАУ

9

Сравнивая два полученных соотношения, приходим к (3).

Если теперь рассматривать норму вектора как функцию переменной x , то из неравенства (3) сразу следует равномерная непрерывность нормы в пространстве R^n .

Определение 3. Говорят, что последовательность векторов x^k сходится к вектору x (сходится по норме), если $\lim_{k \rightarrow \infty} \|x^k - x\| = 0$.

Можно показать, что в конечномерном пространстве определения сходимости по норме и сходимости по компонентам* являются эквивалентны. При этом из условия $x^k \rightarrow x$ следует, что $\|x^k\| \rightarrow \|x\|$. Действительно, в силу (3) получим

$$\|x^k\| - \|x\| \leq \|x^k - x\| \rightarrow 0.$$

Существует ряд способов введения нормы вектора. Важный класс векторных норм представляют p -нормы, определяемые равенством

$$\|x\|_p = (\|x_1\|^p + \|x_2\|^p + \dots + \|x_n\|^p)^{1/p}, \quad p \geq 1.$$

Наиболее употребительными из p -норм являются C -норма, 1-норма (октаэдрическая) и 2-норма (сферическая или евклидова):

$$\|x\|_C = \max_{1 \leq i \leq n} |x_i|, \quad (4)$$

$$\|x\|_1 = |x_1| + |x_2| + \dots + |x_n|, \quad (5)$$

$$\|x\|_2 = \sqrt{\langle x, x \rangle} = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}. \quad (6)$$

Норма (4), которую иногда называют *кубической*, представляет собой предельный случай p -нормы, т. е. $\|x\|_C = \lim_{p \rightarrow \infty} \|x\|_p$.

Отметим, что в конечномерном пространстве все нормы эквивалентны, т. е. для двух норм $\|\cdot\|_\alpha$ и $\|\cdot\|_\beta$ существуют положительные постоянные c_1 и c_2 такие, что

$$c_1\|x\|_\alpha \leq \|x\|_\beta \leq c_2\|x\|_\alpha \quad \forall x \in R^n. \quad (7)$$

*Наше внимание в этой книге сконцентрировано в основном на работе с вещественными векторами и матрицами, во-первых, для упрощения изложения, а во-вторых, виду того, что большинство реальных задач связано с использованием вещественных данных. Однако в некоторых разделах, где это необходимо, рассматриваются вычисления с комплексными данными.

В частности, при $x \in \mathbb{R}^n$ для p -норм имеют место легко проверяемые соотношения

$$\|x\|_C \leq \|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \|x\|_2 \leq n \|x\|_C. \quad (8)$$

Неравенства (7) показывают, что из сходимости последовательности векторов в α -норме следует сходимость в β -норме и наоборот.

2. Нормы матриц. Подчиненность норм. Рассмотрим понятие нормы матрицы. Обозначим через $\mathbb{R}^{n \times n}$ линейное пространство вещественных числовых $(n \times n)$ -матриц.

Определение 4. *Нормой (мультитипликативной) квадратной матрицы A называется поставленное в соответствие этой матрице неотрицательное число $\|A\|$, удовлетворяющее аксиомам:*

- 1) $\|A\| > 0 \quad \forall A \in \mathbb{R}^{n \times n}, \quad A \neq 0; \quad \|0\| = 0;$
- 2) $\|\alpha A\| = |\alpha| \|A\| \quad \forall \alpha \in \mathbb{R}, \quad \forall A \in \mathbb{R}^{n \times n};$
- 3) $\|A + B\| \leq \|A\| + \|B\| \quad \forall A, B \in \mathbb{R}^{n \times n};$
- 4) $\|AB\| \leq \|A\| \|B\| \quad \forall A, B \in \mathbb{R}^{n \times n}.$

Как и в случае векторов, условие $\|A_k - A\| \rightarrow 0$ является необходимым и достаточным условием сходимости по элементам $A_k \rightarrow A$, а из неравенства

$$\|A\| - \|B\| \leq \|A - B\|,$$

аналогичного (3), следует, что если $A_k \rightarrow A$, то $\|A_k\| \rightarrow \|A\|$.

Норма матрицы также может быть определена различными способами. Например, рассматривая матрицу A как n^2 -мерный вектор с вещественными компонентами, получим очевидные обобщения C и 2-норм:

$$\|A\|_M = n \max_{1 \leq i, j \leq n} |a_{ij}|, \quad (9)$$

$$\|A\|_E = \sqrt{\sum_{j=1}^n a_{ij}^2}. \quad (10)$$

Нормы (9) и (10) называются *максимальной* и *сферической* (евклидовой) соответственно.

Определение 5. Если для любой квадратной матрицы A и любого вектора x , размерность которого равна порядку матрицы, выполняется неравенство

$$\|Ax\| \leq \|A\| \|x\|, \quad (11)$$

то будем говорить, что норма матрицы *согласована* с данной нормой вектора.

Если в неравенстве (11) определить

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}, \quad (12)$$

то введенная таким образом норма матрицы A будет наименьшей из всех норм, согласованных с нормой вектора. Эту норму называют *подчиненной* данной норме вектора. Очевидно, что (12) есть наибольшая из норм векторов, полученных действием матрицы A на векторы единичной длины:

$$\|A\| = \sup_{x \neq 0} \|A \frac{x}{\|x\|}\| = \max_{\|x\|=1} \|Ax\|.$$

Поскольку множество $\{x \in \mathbb{R}^n : \|x\| = 1\}$ ограничено и замкнуто, то в силу непрерывности нормы максимальное значение $\|Ax\|$ достигается, т. е.

$$\|A\| = \max_{\|x\|=1} \|Ax\| = \|Ax^*\|$$

для некоторого $x^* \in \mathbb{R}^n$ с единичной нормой.

Нетрудно показать, что норма (10) согласована с нормой $\|x\|_2$, а норма (9) согласована со всеми рассмотренными выше нормами вектора. Однако ни максимальная, ни сферическая нормы матрицы не подчинены ни одной из норм (4)–(6). Это следует из того, что в силу (12) для любой подчиненной нормы $\|E\| = 1$, где E – единичная матрица, тогда как $\|E\|_M = n$, $\|E\|_E = \sqrt{n}$.

Подчиненными нормам вектора (4)–(6) в пространстве $\mathbb{R}^{n \times n}$ являются соответственно нормы

$$\|A\|_C = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|, \quad (13)$$

$$\|A\|_1 = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|, \quad (14)$$

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}, \quad (15)$$

где $\lambda_{\max}(A^T A)$ – наибольшее собственное значение матрицы $A^T A$.

Докажем, что норма (13) подчинена норме $\|x\|_C$. Для любого вектора $x \in \mathbb{R}^n$ справедливо неравенство

$$\|Ax\|_C = \max_{1 \leq i \leq n} \left| \sum_{j=1}^n a_{ij} x_j \right| \leq \max_{1 \leq j \leq n} |x_j| \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|.$$

Следовательно,

$$\|Ax\|_C \leq \left(\max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}| \right) \|x\|_C. \quad (16)$$

Оценка (16) устанавливает согласованность норм $\|A\|_C$ и $\|x\|_C$. Для завершения доказательства достаточно построить такую вектор $x^* = (x_1^*, x_2^*, \dots, x_n^*)^T$, для которого выполняется равенство

$$\|Ax^*\|_C = \left(\max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}| \right) \|x^*\|_C. \quad (17)$$

Пусть функция $\varphi_i = \sum_{j=1}^n |a_{ij}|$, $i = 1, 2, \dots, n$, достигает максимума при $i = k$:

$$\max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}| = \sum_{j=1}^n |a_{kj}|.$$

Рассмотрим вектор x^* с компонентами

$$x_j^* = \begin{cases} 1, & \text{если } a_{kj} \geq 0, \\ -1, & \text{если } a_{kj} < 0. \end{cases}$$

Очевидно, что $\|x^*\|_C = 1$. Далее, исходя из определения вектора x^* , получим

$$\|Ax^*\|_C = \max_{1 \leq i \leq n} \left| \sum_{j=1}^n a_{ij} x_j^* \right| \geq \left| \sum_{j=1}^n a_{kj} x_j^* \right| = \sum_{j=1}^n a_{kj} x_j^* = \sum_{j=1}^n |a_{kj}|$$

и, следовательно,

$$\|Ax^*\|_C \geq \sum_{j=1}^n |a_{kj}| = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|.$$

Таким образом, найден вектор x^* , для которого

$$\|Ax^*\|_C \geq \left(\max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}| \right) \|x^*\|_C.$$

Сравнивая последнее неравенство с неравенством (16), справедливым для любого вектора x , заключаем, что для x^* выполняется равенство (17).

Аналогично можно доказать подчиненность нормы (14) октаэдрической норме вектора. Действительно,

$$\begin{aligned} \|Ax\|_1 &= \sum_{i=1}^n \left| \sum_{j=1}^n a_{ij} x_j \right| \leq \sum_{i=1}^n \sum_{j=1}^n |a_{ij}| |x_j| = \sum_{j=1}^n \left(\sum_{i=1}^n |a_{ij}| \right) |x_j| \leq \\ &\leq \left(\max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}| \right) \sum_{j=1}^n |x_j| = \left(\max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}| \right) \|x\|_1. \end{aligned}$$

Пусть функция $\varphi_j = \sum_{i=1}^n |a_{ij}|$, $j = 1, 2, \dots, n$, достигает максимума при $j = k$. Тогда для вектора

$$x^* = (0, 0, \dots, 0, 1, 0, 0, \dots, 0)^T, \quad \|x^*\|_1 = 1,$$

имеем

$$\|Ax^*\|_1 = \sum_{i=1}^n \left| \sum_{j=1}^n a_{ij} x_j^* \right| = \sum_{i=1}^n |a_{ik}| = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|.$$

Таким образом, приходим к равенству

$$\|Ax^*\|_1 = \left(\max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}| \right) \|x^*\|_1,$$

которое и доказывает сделанное утверждение.

Наконец, рассмотрим норму матрицы $\|A\|_2$, или *спектральную* норму. Отметим, что практическое вычисление 2-нормы значительно сложнее, чем вычисление C -нормы или 1-нормы, так как требует предварительного знания $\lambda_{\max}(A^T A)$. Способом нахождения собственных значений матрицы будет посвящена глава IV, а пока напомним, что число λ называется *собственным значением матрицы* A , если существует ненулевой вектор $x \in \mathbb{R}^n$, для которого

$$Ax = \lambda x.$$

Вектор x называется в этом случае *собственным вектором матрицы* A , соответствующим данному собственному значению λ .

Для доказательства подчиненности спектральной нормы сферической норме вектора предварительно покажем, что матрица $A^T A$ является симметрической и все ее собственные значения неотрицательны. В самом деле

$$(A^T A)^T = A^T (A^T)^T = A^T A.$$

Пусть λ — собственное значение матрицы $A^T A$, $x \neq 0$ — соответствующий ему собственный вектор. Умножим обе части равенства $A^T Ax = \lambda x$ скалярно на x :

$$(A^T Ax, x) = (\lambda x, x). \quad (18)$$

Учитывая, что $(A^T Ax, x) = (Ax, Ax) = \|Ax\|_2^2$ и $(\lambda x, x) = \lambda \|x\|_2^2$, равенство (18) можно записать в виде

$$\|Ax\|_2^2 = \lambda \|x\|_2^2,$$

откуда следует, что $\lambda \geq 0$.

Пусть, далее, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ — собственные значения матрицы $A^T A$; x_1, x_2, \dots, x_n — ортонормированная система собственных векторов этой матрицы. Разлагая произвольный вектор $x \in \mathbb{R}^n$ по базису $\{x_k\}_{k=1}^n$, получим

$$x = \sum_{k=1}^n c_k x_k, \quad \|x\|_2^2 = \sum_{k=1}^n c_k^2$$

и соответственно

$$A^T Ax = \sum_{k=1}^n c_k \lambda_k x_k, \quad \|Ax\|_2^2 = (A^T Ax, x) = \sum_{k=1}^n c_k^2 \lambda_k.$$

Отсюда имеем

$$\|Ax\|_2^2 \leq \lambda_1 \sum_{k=1}^n c_k^2 = \lambda_1 \|x\|_2^2,$$

или

$$\|Ax\|_2 \leq \sqrt{\lambda_1}.$$

С другой стороны, если в качестве вектора x взять собственный вектор x_1 , то

$$\|Ax_1\|_2^2 = (A^T Ax_1, x_1) = \lambda_1 \|x_1\|_2^2.$$

Поэтому

$$\|A\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \geq \sqrt{\lambda_1}.$$

Сравнение полученных оценок позволяет сделать вывод о справедливости равенства (18).

Отметим важный частный случай. Если A — симметричная матрица, то $\lambda_{\max}(A^T A) = \lambda_{\max}(A^2) = \lambda_{\max}^2(A)$, где $\lambda_{\max}(A)$ — наибольшее по модулю собственное значение матрицы A . Поэтому для $A = A^T$ справедливо равенство

$$\|A\|_2 = |\lambda_{\max}(A)|.$$

Число $\rho(A) = |\lambda_{\max}(A)|$ будем называть *спектральным радиусом матрицы* A .

Для подчиненных норм матриц имеют место оценки, аналогичные неравенствам (8) для норм векторов:

$$\frac{1}{n} \|A\|_C \leq \frac{1}{\sqrt{n}} \|A\|_2 \leq \|A\|_1 \leq \sqrt{n} \|A\|_2 \leq n \|A\|_C. \quad (19)$$

В случае максимальной и сферической норм матриц справедливы следующие соотношения:

$$\frac{1}{n} \|A\|_M \leq \|A\|_p \leq \|A\|_M, \quad p = C, 1, 2, E; \quad (20)$$

$$\frac{1}{\sqrt{n}} \|A\|_E \leq \|A\|_p \leq \sqrt{n} \|A\|_E, \quad p = C, 1; \quad (21)$$

$$\frac{1}{\sqrt{n}} \|A\|_E \leq \|A\|_2 \leq \|A\|_E. \quad (22)$$

3. Оценки относительной погрешности решения. Перейдем к получению количественных оценок, связывающих погрешности δf , δA с погрешностью δx , и тем самым проанализируем второе требование в определении корректировки. Сначала будем считать, что в матрице A возмущений не вносится, т. е. $\delta A = 0$. Тогда из (1), (2) следует уравнение для погрешности

$$A(\delta x) = \delta f,$$

откуда имеем

$$\delta x = A^{-1}(\delta f)$$

и, следовательно,

$$\|\delta x\| \leq \|A^{-1}\| \|\delta f\|. \quad (23)$$

Неравенство (23) устанавливает непрерывную зависимость решения от правой части, т. е. показывает, что условие $\|\delta f\| \rightarrow 0$ влечет за собой и $\|\delta x\| \rightarrow 0$.

В оценку (23) входят нормы абсолютных погрешностей решения и правой части. В практических приложениях более естественной является связь между нормами относительных погрешностей

$$\frac{\|\delta x\|}{\|x\|}, \quad \frac{\|\delta f\|}{\|f\|}.$$

Поскольку из (1) следует, что

$$\|f\| \leq \|A\| \|x\|, \quad (24)$$

то, перемножив (23) и (24), приходим к оценке устойчивости для относительных погрешностей

$$\frac{\|\delta x\|}{\|x\|} \leq \|A^{-1}\| \|A\| \frac{\|\delta f\|}{\|f\|}. \quad (25)$$

Для получения оценки коэффициентной устойчивости системы (1) нам понадобится следующая теорема.

Теорема 1. Пусть C — квадратная матрица, удовлетворяющая условию $\|C\| < 1$, E — единичная матрица. Тогда существует обратная матрица $(E + C)^{-1}$, причем

$$\|(E + C)^{-1}\| \leq \frac{1}{1 - \|C\|}. \quad (26)$$

Доказательство. Для любого вектора x имеем

$$\begin{aligned} \|(E + C)x\| &= \|x + Cx\| \geq \|x\| - \|Cx\| \geq \|x\| - \|C\| \|x\| = \\ &= (1 - \|C\|) \|x\| = q \|x\|, \end{aligned}$$

где $q = 1 - \|C\| > 0$. Поэтому если $(E + C)x = 0$, то $q \|x\| \leq 0$ и $x = 0$, т. е. однородное уравнение $(E + C)x = 0$ имеет только тривиальное решение. Значит, соответствующее неоднородное уравнение имеет единственное решение и таким образом, существует матрица $(E + C)^{-1}$. Поэтому в неравенстве $\|(E + C)x\| \geq q \|x\|$ можно

обозначить $(E + C)x = y$, $x = (E + C)^{-1}y$. Тогда

$$\|y\| \geq q \|(E + C)^{-1}y\|,$$

откуда получим

$$\|(E + C)^{-1}y\| \leq \frac{1}{q} \|y\| = \frac{1}{1 - \|C\|} \|y\|,$$

следовательно, неравенство (26) выполнено. Теорема доказана.

Пусть $\delta f = 0$, $\delta A \neq 0$. Будем также дополнительно предполагать, что δA удовлетворяет неравенству

$$\|\delta A\| < \|A^{-1}\|^{-1}. \quad (27)$$

Это условие обеспечивает невырожденность матрицы $A + \delta A$ и единственность решения возмущенной системы. Из уравнения (2) следуют соотношения

$$Ax + A\delta x + \delta A x + \delta A \delta x = f,$$

или с учетом равенства $Ax = f$

$$(A + \delta A)x + \delta A x = f.$$

БИБЛИОТЕКА

БГУ

1824386

Отсюда имеем

$$\delta x = -(A + \delta A)^{-1} \delta A x = -(E + A^{-1} \delta A)^{-1} A^{-1} \delta A x.$$

Переходя в последнем равенстве к нормам, получим

$$\|\delta x\| \leq \| (E + A^{-1} \delta A)^{-1} \| \| A^{-1} \| \| \delta A \| \| x \| . \quad (28)$$

Так как в силу аксиомы 4) нормы матрицы и неравенства (27)

$$\| A^{-1} \delta A \| \leq \| A^{-1} \| \| \delta A \| < 1,$$

то, согласно теореме 1, матрица $E + A^{-1} \delta A$ обратима и выполняется соотношение

$$\| (E + A^{-1} \delta A)^{-1} \| \leq \frac{1}{1 - \| A^{-1} \| \| \delta A \|}.$$

Тогда из (28) следует, что

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\| A^{-1} \| \| A \| \| \delta A \|}{1 - \| A^{-1} \| \| A \| \| \delta A \| \| A \|}. \quad (29)$$

Оценка (29) означает коэффициентную устойчивость системы (1), т. е. $\|\delta x\| \rightarrow 0$ при $\|\delta A\| \rightarrow 0$. Таким образом, условие $\det A \neq 0$ и неравенство (27) обеспечивают корректную постановку исходной задачи.

§ 2. Обусловленность матриц. Регуляризация

1. Число обусловленности и его свойства. В § 1 были получены оценки устойчивости по правой части и коэффициентной устойчивости для СЛАУ

$$Ax = f. \quad (1)$$

Входящее в эти оценки число

$$\kappa(A) = \| A^{-1} \| \| A \| \quad (2)$$

называется числом обусловленности матрицы A . Оно характеризует степень влияния относительной погрешности входных данных на относительную погрешность искомого решения.

Отметим следующие свойства числа обусловленности (2):

$$1. \kappa(A) \geq \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|} \geq 1, \quad (3)$$

где $\lambda_{\max}(A)$ и $\lambda_{\min}(A)$ — наибольшее и наименьшее по модулю собственные значения матрицы A соответственно;

$$2. \kappa(AB) \leq \kappa(A)\kappa(B).$$

Докажем свойство 1. Рассмотрим собственный вектор x , соответствующий наибольшему по модулю собственному значению. Имеет место равенство

$$Ax = \lambda_{\max}(A)x,$$

из которого получим

$$\|A\| \|x\| \geq |\lambda_{\max}(A)| \|x\|.$$

Поэтому

$$\|A\| \geq \rho(A), \quad (4)$$

т. е. любая норма матрицы A , согласованная с данной нормой вектора, не меньше ее спектрального радиуса. Поскольку собственные значения матриц A и A^{-1} взаимно обратны, то

$$\|A^{-1}\| \geq \max_{\lambda} \frac{1}{|\lambda(A)|} = \frac{1}{|\lambda_{\min}(A)|}.$$

Отсюда из неравенства (4) следует свойство 1.

Для симметричной матрицы и спектральной нормы свойство 1 выполняется со знаком равенства. Действительно, как показано выше, 2-норма симметричной матрицы совпадает с ее спектральным радиусом, т. е. $\|A\|_2 = \rho(A)$. С другой стороны,

$$\|A^{-1}\|_2 = \max_{\lambda} \frac{1}{|\lambda(A)|} = \frac{1}{|\lambda_{\min}(A)|}.$$

Таким образом, в случае нормы $\|\cdot\|_2$ и $A = A^T$

$$\kappa(A) = \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|}. \quad (5)$$

Формулы (3), (5) могут быть использованы для количественных оценок числа обусловленности $\kappa(A)$.

Что касается свойства 2, то оно вытекает из аксиомы 4) в определении нормы матрицы.

Матрицы с большим числом обусловленности $\kappa(A)$ называются плохими обусловленными матрицами. Заметим, что это свойство зависит от выбора нормы. Так, например, в случае спектральной нормы

$$\kappa_2(A) = \sqrt{\lambda_{\max}(A^T A)} \sqrt{\lambda_{\min}((A^{-1})^T A^{-1})} = \sqrt{\frac{\mu_1}{\mu_n}},$$

где μ_1 и μ_n — наибольшее и наименьшее собственные значения матрицы $A^T A$. Однако в пространстве $\mathbb{R}^{n \times n}$ любые два числа обусловленности эквивалентны в том смысле, что найдутся положительные константы c_1 и c_2 такие, что

$$c_1 \kappa_2(A) \leq \kappa_2(A) \leq c_2 \kappa_2(A) \quad \forall A \in \mathbb{R}^{n \times n}.$$

В частности, справедлива цепочка неравенств

$$\frac{1}{n^2} \kappa_C(A) \leq \frac{1}{n} \kappa_2(A) \leq \kappa_1(A) \leq n \kappa_2(A) \leq n^2 \kappa_C(A). \quad (6)$$

Поэтому если матрица плохо обусловлена в α -норме, то она будет плохо обусловленной и в β -норме с поправкой на соответствующие константы. Системы вида (1) с плохо обусловленной матрицей также называются плохо обусловленными. Для плохо обусловленных систем линейных алгебраических уравнений недопустимо велико правая часть в неравенствах (25), (29) (см. § 1), поэтому лишь очень малые погрешности входных данных гарантируют приемлемую относительную погрешность решения.

Пример 1. Пусть дана матрица

$$A = \begin{pmatrix} 3,0000 & -7,0001 \\ 3,0000 & -7,0000 \end{pmatrix}. \quad (7)$$

Воспользуемся для оценки числа обусловленности этой матрицы неравенством (3). Собственные значения матрицы A являются корнями характеристического уравнения

$$\det(A - \lambda E) = 0.$$

В данном случае имеем квадратное уравнение

$$\lambda^2 + 4\lambda + 0,0003 = 0,$$

решая которое, находим корни

$$\lambda_1 = -3,9999, \quad \lambda_2 = -0,0001.$$

Таким образом, оценка для числа обусловленности имеет вид

$$\kappa(A) \geq 39,999 = 4 \cdot 10^4.$$

Рассмотрим теперь исходную систему (1) с матрицей (7). Нетрудно проверить, что если в качестве правой части взять вектор

$$f = \begin{pmatrix} 0,9998 \\ 1,0000 \end{pmatrix}, \quad \text{то } x = \begin{pmatrix} 5,0000 \\ 2,0000 \end{pmatrix},$$

а если взять вектор

$$\tilde{f} = \begin{pmatrix} 1,0000 \\ 1,0000 \end{pmatrix}, \quad \text{то } \tilde{x} = \begin{pmatrix} 0,3333 \\ 0,0000 \end{pmatrix}$$

т. е. малая погрешность в задании лишь одной компоненты вектора-столбца правой части влечет за собой недопустимое искажение искомого решения.

Этот результат достаточно точно предсказывает оценкой (25) предыдущего параграфа. Действительно, если число обусловленности матрицы A является равным $\kappa(A) = 4 \cdot 10^4$, то, несмотря на малость погрешности

$$\frac{\|\delta f\|}{\|f\|} = 1,4143 \cdot 10^{-4}$$

(норма вектора сферической), относительная погрешность в решении велика в соответствии с оценкой

$$1,3569 = \frac{\|\delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\delta f\|}{\|f\|} = 5,6772.$$

2. Обусловленность матриц и близость к вырожденности. В последнем примере может создаться впечатление, что причиной плохой обусловленности задачи является малость определителя $\det A$. В каком-то смысле это так, поскольку известно, что

$$\det A = \lambda_1 \lambda_2 \dots \lambda_n.$$

Будем считать, что

$$|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|.$$

Если теперь зафиксировать λ_1 и рассмотреть последовательность таких матриц A_k , для которых $|\lambda_n^{(k)}| \rightarrow 0$, то

$$\det A_k \rightarrow 0, \quad \kappa(A_k) \rightarrow \infty.$$

Однако в действительности величина определителя $\det A$ и обусловленность системы линейных алгебраических уравнений $Ax = f$ слабо связаны между собой. Проиллюстрируем этот факт следующими примерами.

Пример 2. Рассмотрим последовательность матриц $A_n = A_n^T$, $n = 2m$, $m = 1, 2, \dots$, таких, что собственными значениями матрицы A_n являются числа λ_1 и λ_2 , каждое из которых имеет кратность m . Тогда

$$\det A_n = (\lambda_1 \lambda_2)^m, \quad \kappa_2(A_n) = \frac{|\lambda_1|}{|\lambda_2|}.$$

Если $\lambda_1 \lambda_2 < 1$, то величина определителя матрицы A_n может быть сделана меньше любого наперед заданного числа. Однако число обусловленности такой матрицы постоянно.

Пример 3. Рассмотрим верхнюю треугольную матрицу

$$A_n = \begin{pmatrix} 1 & -1 & -1 & \cdots & -1 \\ & 1 & -1 & \cdots & -1 \\ 0 & & 1 & \cdots & -1 \\ & & & \ddots & \vdots \\ & & & & 1 \end{pmatrix}.$$

Это плохо обусловленная матрица, так как $\kappa_C(A_n) = n^{2m-1}$. В то же время ее определитель равен $\det A_n = 1$.

Пример 4. Диагональная матрица $D_n = \text{diag}[\varepsilon, \varepsilon, \dots, \varepsilon]$, где $\varepsilon > 0$ — малая константа, имеет малый определитель $\det D_n = \varepsilon^n$, а ее число обусловленности $\kappa_2(D_n) = 1$.

Знание числа обусловленности позволяет проанализировать ситуацию, связанную с погрешностью округления чисел. Наиболее распространенной формой записи действительных чисел в современных компьютерах является их представление в виде чисел с плавающей точкой

$$x = \pm q^p \sum_{k=1}^t b_k q^{-k} = \pm q^p(\beta_1, \beta_2, \dots, \beta_t),$$

где $0 \leq \beta_k < q$, $k = 1, 2, \dots, t$. Тогда каждая компонента вектора f в правой части (1) или каждый элемент матрицы A округляется с относительной погрешностью $O(q^{-t})$. Следовательно,

$$\frac{\|\delta x\|}{\|x\|} \leq \kappa(A) O(q^{-t}).$$

Таким образом, на реальной вычислительной машине решение (1) не может быть найдено с точностью, большей чем $\kappa(A) O(q^{-t})$.

3. Геометрическая интерпретация понятия обусловленности. Метод регуляризации. Пусть $A = A^T$, x_1, x_2, \dots, x_n — собственные векторы матрицы A , образующие базис в пространстве \mathbb{R}^n , $\lambda_1, \lambda_2, \dots, \lambda_n$ — соответствующие этим векторам собственные значения. Рассмотрим задачу (1) и разложим x и f в ряд по собственным функциям:

$$x = \sum_{k=1}^n c_k x_k, \quad f = \sum_{k=1}^n b_k x_k,$$

где коэффициенты c_k определяются равенством

$$c_k = \frac{b_k}{\lambda_k} = \frac{(f, x_k)}{\lambda_k}. \quad (8)$$

Если матрица A плохо обусловлена, то среди λ_k имеются малые числа такие, что малым изменениям b_k будут соответствовать недопустимо большие изменения c_k . Другими словами, решение задачи (1) недопустимо искажается в направлении векторов x_k , соответствующих малым λ_k . Заметим, что указанного искажения может и не быть, если вектор f ортогонален (или почти ортогонален) "плохим" векторам x_k , т. е.

$$(f, x_k) = 0 \quad \text{или} \quad (f, x_k) \approx 0.$$

Бороться с искажениями решения в направлении векторов x_k , соответствующих малым λ_k , можно, например, следующим образом.

Предположим, что все собственные значения $\lambda_k(A) > 0$, $k = 1, 2, \dots, n$. Зададим некоторое $\varepsilon > 0$ и вместо (1) рассмотрим систему

$$(A + \varepsilon E)x^{(\varepsilon)} = f.$$

Тогда вместо равенства (8) получим

$$c_k^{(\varepsilon)} = \frac{b_k}{\lambda_k + \varepsilon}.$$

Очевидно, что если для всех k величина $\lambda_k + \varepsilon \neq 0$, то $\lim_{\varepsilon \rightarrow 0} x^{(\varepsilon)} = x$. Поскольку

$$c_k - c_k^{(\varepsilon)} = \frac{\varepsilon b_k}{\lambda_k(\lambda_k + \varepsilon)},$$

то для больших λ_k введение параметра ε не оказывает существенного влияния на c_k . Если же $\lambda_k \ll \varepsilon$, то

$$|c_k^{(\varepsilon)}| = \left| \frac{b_k}{\lambda_k + \varepsilon} \right| \ll \left| \frac{b_k}{\lambda_k} \right| = |c_k|$$

и вклад слагаемых, соответствующих малым числам λ_k , весьма незначителен.

Приведенный подход получил название *метода регуляризации*. Основная проблема здесь заключается в выборе оптимального значения ε . Как правило, это осуществляется экспериментально путем сравнения результатов расчетов при различных ε .

Если матрица A не является симметричной, то для регуляризации плохо обусловленных систем можно использовать вариационный принцип (более детально он будет рассмотрен в главе III применительно к итерационным методам решения СЛАУ). Перепишем исходную задачу (1) в виде

$$(Ax - f, Ax - f) = 0. \quad (9)$$

Если компоненты правой части f или элементы матрицы A заданы не точно, то вместо уравнения (9) мы фактически имеем приближенное уравнение $(Ax - f, Ax - f) \approx 0$. Потребуем, чтобы решение этой задачи как можно меньше отклонялось от заданного вектора x^* , т. е. чтобы скалярное произведение $(x - x^*, x - x^*)$ было минимальным. Тогда регуляризованная задача формулируется следующим образом:

$$(Ax - f, Ax - f) + \varepsilon(x - x^*, x - x^*) \rightarrow \min,$$

или в эквивалентной форме

$$(x, A^T Ax) - 2(x, A^T f) + (f, f) + \varepsilon[(x, x) - 2(x, x^*) + (x^*, x^*)] \rightarrow \min, \quad (10)$$

где $\varepsilon > 0$ — малый параметр. Варьируя x в (10), получим систему уравнений

$$(A^T A + \varepsilon E)x = A^T f + \varepsilon x^*,$$

из которой находим решение $x^{(\varepsilon)}$, зависящее от параметра ε . Оптимальное значение ε , как и в предыдущем случае, определяется экспериментально.

Задачи к главе I

1. Доказать, что при $x \in \mathbb{R}^n$ справедливо равенство $\|x\|_C = \lim_{p \rightarrow \infty} \|x\|_p$.

2. Доказать неравенство Коши — Буняковского

$$|(x, y)| \leq \|x\|_2 \|y\|_2$$

(рассмотреть неравенство $(ax + by, ax + by) \geq 0$ для подходящих чисел a и b).

3. Проверить, что $\|\cdot\|_\infty$, $\|\cdot\|_1$ и $\|\cdot\|_2$ являются векторными нормами, и доказать неравенства (8) § 1. Найти векторы, на которых в этих неравенствах достигается равенство.

4. Доказать, что $\max_{1 \leq i \leq n} \left(\sum_{j=1}^n x_{ij} \right)$ является нормой вектора x . Найти норму матрицы, подчиненную данной норме вектора.

5. Пусть $\|\cdot\|$ — норма вектора в \mathbb{R}^n . Доказать, что равенство

$$\|x\|_* = \sup_{y \neq 0} \frac{(x, y)}{\|y\|}$$

также задает норму в \mathbb{R}^n , называемую *действительной* к $\|\cdot\|$.

6. Пусть $\|\cdot\|$ — некоторая норма в пространстве \mathbb{R}^n , $A \in \mathbb{R}^{n \times n}$ — невырожденная матрица. Показать, что $\|x\|_* = \|Ax\|$ также является нормой в \mathbb{R}^n .

7. Пусть $x, y \in \mathbb{R}^n$. Определить функцию $\psi: \mathbb{R} \rightarrow \mathbb{R}$ равенством $\psi(t) = \|x - ty\|$. Показать, что минимум этой функции достигается при значении $t = (x, y)/\|x\|^2$.

8. Проверить, что $\|x\|_p = ((|x_1|^p + |x_2|^p + \dots + |x_n|^p)^{1/p})$, $p \geq 1$, является нормой вектора в комплексном векторном пространстве \mathbb{C}^n . Доказать, что при $x \in \mathbb{C}^n$ имеет место неравенство

$$\|x\|_p \leq c(\|Re x\|_p + \|Im x\|_p), \quad c = \text{const} > 0.$$

Найти такую постоянную $c_0 > 0$, что $c_0(\|Re x\|_2 + \|Im x\|_2) \leq \|x\|_2 \quad \forall x \in \mathbb{C}^n$.

9. Доказать, что

$$\|AB\|_p \leq \|A\|_p \|B\|_p \quad \forall A, B \in \mathbb{R}^{n \times n}, \quad 1 \leq p \leq \infty.$$

10. Пусть B — любая подматрица матрицы A , $1 \leq p \leq \infty$. Показать, что $\|B\|_p \leq \|A\|_p$.
11. Проверить неравенства (19)–(22) § 1 и указать матрицы, для которых эти неравенства становятся точными.

12. Доказать справедливость неравенства

$$\|A\|_2 \leq \sqrt{\|A\|_1 \|A\|_C}.$$

13. Доказать, что если $A = A^T$, то

$$\|A\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

14. Доказать неравенство

$$\|A\|_2 \leq \|A\|^{1/2} \|A^T\|^{1/2}$$

для любой нормы матрицы A , подчиненной какой-либо норме вектора.

15. Показать, что для любого собственного значения $\lambda(A)$ невырожденной матрицы A справедливо неравенство

$$|\lambda(A)| \leq \inf_k \|A^k\|^{1/k}, \quad k \in \mathbb{N}.$$

16. Вещественная матрица A называется *нормальной*, если она перестановочна со своей транспонированной матрицей A^T , т. е. $AA^T = A^TA$. Доказать, что для нормальной матрицы имеет место равенство $\|A\|_2 = \rho(A)$, где $\rho(A)$ — спектральный радиус матрицы A .

17. Доказать, что если $A \in \mathbb{R}^{n \times n}$, $x \in \mathbb{R}^n$, $x \neq 0$, то

$$\left\| A \left(E - \frac{xx^T}{(x, x)} \right) \right\|_F^2 = \|A\|_F^2 - \frac{\|Ax\|_2^2}{(x, x)}.$$

18. Пусть $x, y \in \mathbb{R}^n$. Показать, что если $A = xy^T$, то

$$\|A\|_F = \|A\|_2 = \|x\|_2 \|y\|_2, \quad \|A\|_C \leq \|x\|_C \|y\|_1.$$

19. Пусть $A \in \mathbb{R}^{n \times n}$, $x, y \in \mathbb{R}^n$, $x \neq 0$. Показать, что среди всех матриц, удовлетворяющих уравнению $(A + B)x = y$, матрица $B = (y - Ax)x^T/(x, x)$ имеет наименьшую спектральную норму.

20. Доказать неравенства (6) § 2 для чисел обусловленности матрицы A .

21. Привести пример несимметричной матрицы, для которой выполняется равенство $\kappa^2(A) = \kappa(A^2)$.

22. Пусть дан жорданов блок порядка n :

$$A = \begin{pmatrix} a & 1 & & 0 \\ & a & \ddots & \\ & & \ddots & 1 \\ 0 & & & a \end{pmatrix}.$$

Вычислить число обусловленности матрицы A и оценить возмущение в компоненте x_i решения системы $Ax = f$, если компонента f_u вектора f имеет возмущение ε .

23. Оценить число обусловленности $\kappa_2(A)$ ($n \times n$ -матрицы

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

24. Оценить снизу и сверху число обусловленности $\kappa_C(A)$, используя грамини спектра невырожденной матрицы $A \in \mathbb{R}^{n \times n}$: $\lambda(A^T A) \in [\delta, \Delta]$.

25. Пусть A — квадратная матрица порядка n с элементами

$$a_{ij} = \begin{cases} p, & \text{если } j = i, \\ q, & \text{если } j = i+1, \\ 0, & \text{если } j \neq i, i+1. \end{cases}$$

а) Вычислить матрицу A^{-1} и доказать, что при $|q| < |p|$ матрица A хорошо обусловлена, а при $|q| > |p|$ и больших значениях n плохо обусловлена.

б) Выписать явно решение СЛАУ $Ax = f$ через правую часть f .

в) Выписать явно через правую часть f вектор $x^{(e)}$, минимизирующий функционал

$$F(x) = (Ax - f, Ax - f) + \varepsilon(x - x^*, x - x^*),$$

где $x^* \in \mathbb{R}^n$ — заданный вектор.

ГЛАВА II

ПРЯМЫЕ МЕТОДЫ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

§ 1. Методы Гаусса последовательного исключения неизвестных

Пусть дана система линейных алгебраических уравнений

$$Ax = f. \quad (1)$$

Методы решения задачи (1) подразделяются на три основные группы: прямые, итерационные и вероятностные. В прямых (или точных) методах решение исходной задачи находится за конечное число арифметических действий. Итерационные методы состоят в том, что решение x системы (1) определяется как предел при $k \rightarrow \infty$ некоторой последовательности приближений x^k , где k — номер итерации. Особое место занимают вероятностные методы, или методы Монте-Карло, используемые обычно для решения систем с большим числом неизвестных.

Отметим, что вследствие погрешностей округления при решении на ЭВМ прямые методы реалью не приводят к точному решению системы (1) и называть их точными можно лишь абстрагируясь от погрешностей округления. Прямые методы различаются по числу (либо по асимптотике числа) арифметических действий $Q(A)$, необходимых для нахождения решений. Мерой различия может также служить число арифметических действий $q(A)$, необходимых для вычисления одного неизвестного, так что $Q(A) = nq(A)$, где n — число неизвестных.

Однако, что вследствие погрешностей округления при решении на ЭВМ прямые методы реалью не приводят к точному решению системы (1) и называть их точными можно лишь абстрагируясь от погрешностей округления. Прямые методы различаются по числу (либо по асимптотике числа) арифметических действий $Q(A)$, необходимых для нахождения решений. Мерой различия может также служить число арифметических действий $q(A)$, необходимых для вычисления одного неизвестного, так что $Q(A) = nq(A)$, где n — число неизвестных.

§ 1. МЕТОДЫ ГАУССА

Из курса алгебры известно, что исходную систему можно решить по крайней мере двумя способами: либо по правилу Крамера, либо с помощью метода Гаусса. Решение задачи (1) по правилу Крамера записывается в виде

$$x_j = \frac{\det A_j}{\det A}, \quad j = 1, 2, \dots, n, \quad (2)$$

где A_j — матрица, получаемая из A путем замены j -го столбца матрицы A на столбец свободных членов f . Если в (2) вычислять определители классическим способом с применением известной теоремы Лапласа, то количество арифметических действий будет равно $Q(A) = O(n!n)$. С ростом размерности системы это число возрастает очень быстро и уже при $n = 30$ оно достигает $Q(A) \approx 10^{30}$. Использование для решения такой задачи современных вычислительных машин с производительностью 10^{10} флотов^{*} потребует порядка 10^{17} лет непрерывной работы компьютера.

В методах Гаусса и их различных модификациях, к изложению которых мы сейчас переходим, $Q(A) = O(n^3)$. Можно показать, что для произвольной невырожденной матрицы существуют методы с $Q(A) = Mn^p$, $p = \log_2 7$. Однако их логическая сложность и большая величина константы M не дают этим методам практических преимуществ перед методами Гаусса. Поэтому считается, что для невырожденной матрицы общего вида прямые методы решения задачи (1) с оценкой $q(A) = O(n^2)$ являются оптимальными.

Определение 1. Метод решения задачи (1) называется экономичным, если $q(A)$ не зависит (или слабо зависит) от общего числа неизвестных, т. е. от порядка матрицы A .

К числу экономичных методов относятся некоторые методы решения систем с матрицей специального вида. В частности, для трехдиагональных, пятидиагональных и т. д. матриц существуют прямые методы решения (1) с $q(A) = \text{const}$. Для более общих матриц специального вида существуют методы с оценкой $q(A) = O(\ln n)$.

Классификация подобного рода можно ввести и для итерационных методов. В этом случае $Q(A)$ будет зависеть также от задаваемых параметров.

*Под флотом понимается одна арифметическая операция с плавающей точкой (floating point operation).

емой точности

$$\|x^k - x\| < \varepsilon,$$

с которой находится решение задачи (1).

1. Теоретические основы метода Гаусса. В этом параграфе мы изложим наиболее простые и естественные методы решения системы (1), базирующиеся на идеи последовательного исключения неизвестных. В основе рассматриваемых методов лежит теорема об LU -разложении.

Обозначим через L нижнюю треугольную матрицу, U — верхнюю треугольную матрицу:

$$L = \begin{pmatrix} l_{11} & & 0 \\ \vdots & \ddots & \\ l_{n1} & \dots & l_{nn} \end{pmatrix}, \quad U = \begin{pmatrix} u_{11} & \dots & u_{1n} \\ 0 & \ddots & \vdots \\ 0 & \dots & u_{nn} \end{pmatrix}.$$

Теорема 1 (LU -факторизация). Пусть все угловые миноры матрицы A отличны от нуля, т. е.

$$a_{11} \neq 0, \quad \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \neq 0, \quad \dots, \quad \det A \neq 0.$$

Тогда матрица A представима в виде

$$A = LU. \quad (3)$$

Доказательство. Доказательство теоремы проведем методом математической индукции. Для $n = 1$ утверждение очевидно, так как с произвольным выбором любого из сомножителей l_{11} или u_{11}

$$a_{11} = l_{11}u_{11}.$$

Пусть теорема верна для матрицы порядка $n - 1$, т. е.

$$A_{n-1} = L_{n-1}U_{n-1}, \quad (4)$$

где матрицы L_{n-1}, U_{n-1} обладают указанными в теореме свойствами. Покажем, что теорема справедлива и для матрицы порядка n .

Представим матрицу A следующим образом:

$$A_n = \begin{pmatrix} A_{n-1} & z \\ v^T & a_{nn} \end{pmatrix},$$

$$A_{n-1} = \begin{pmatrix} a_{11} & \dots & a_{1,n-1} \\ \vdots & \ddots & \\ a_{n-1,1} & \dots & a_{n-1,n-1} \end{pmatrix}, \quad z = \begin{pmatrix} a_{1n} \\ \vdots \\ a_{n-1,n} \end{pmatrix}, \quad v = \begin{pmatrix} a_{n1} \\ \vdots \\ a_{n,n-1} \end{pmatrix}.$$

Будем искать разложение матрицы A в виде

$$\begin{pmatrix} A_{n-1} & z \\ v^T & a_{nn} \end{pmatrix} = \begin{pmatrix} L_{n-1} & 0 \\ y^T & l_{nn} \end{pmatrix} \begin{pmatrix} U_{n-1} & w \\ 0 & u_{nn} \end{pmatrix}, \quad (5)$$

где $y = (l_{1n}, l_{2n}, \dots, l_{n,n-1})^T$, $w = (u_{1n}, u_{2n}, \dots, u_{n-1,n})^T$ — неизвестные пока векторы. Перемножая матрицы в правой части (5), получим систему уравнений

$$\begin{aligned} L_{n-1}U_{n-1} &= A_{n-1}, & L_{n-1}w &= z, \\ y^T U_{n-1} &= v^T, & y^T w + l_{nn}u_{nn} &= a_{nn}. \end{aligned} \quad (6)$$

Первое из уравнений (6) выполняется вследствие предположения индукции (4). Поскольку $\det A_{n-1} \neq 0$, то матрицы L_{n-1}, U_{n-1} обратимы и, следовательно, из второго и третьего уравнений однозначно определяются векторы w и y :

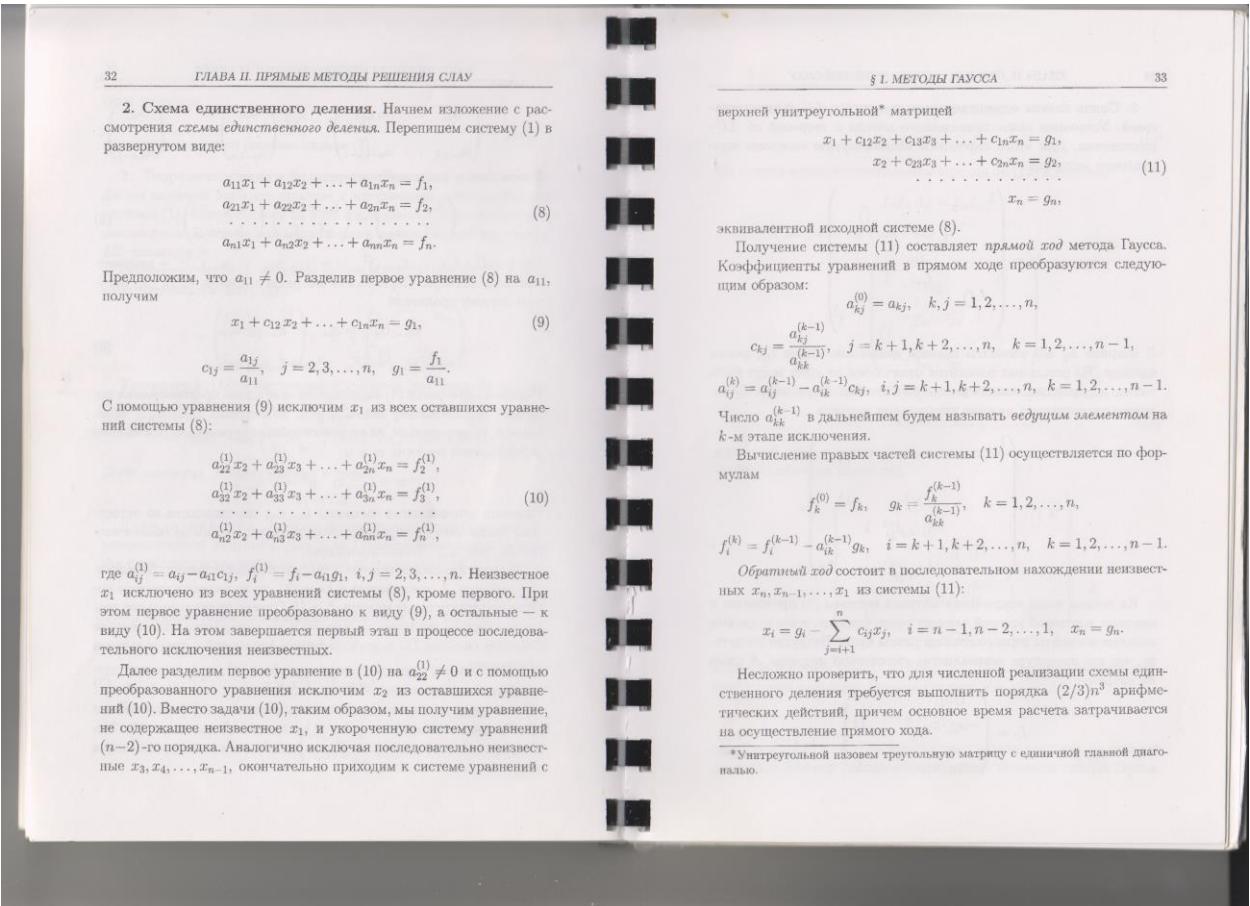
$$w = L_{n-1}^{-1}z, \quad y^T = v^T U_{n-1}^{-1}.$$

Наконец, диагональные элементы l_{nn}, u_{nn} определяются из четвертого уравнения (6) с произвольным выбором любого из сомножителей l_{nn} или u_{nn} . Теорема доказана.

Теорема 1 является типичной теоремой существования. Доказано, что матрица A представима в виде (3), но не указан алгоритм построения треугольных матриц L и U . Однако принципиальное значение теоремы заключается именно в доказательстве возможности сведения системы (1) к двум системам уравнений с треугольными матрицами

$$Ly = f, \quad Ux = y. \quad (7)$$

Практически не имеет смысла искать разложение (3). Достаточно преобразовать исходную задачу (1) к одной из задач вида (7). Это и осуществляется в *методах Гаусса последовательного исключения неизвестных*.



2. Схема единственного деления. Начнем изложение с рассмотрения *схемы единственного деления*. Перепишем систему (1) в развернутом виде:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= f_1, \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= f_2, \\ \dots & \dots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n &= f_n. \end{aligned} \quad (8)$$

Предположим, что $a_{11} \neq 0$. Разделив первое уравнение (8) на a_{11} , получим

$$x_1 + c_{12}x_2 + \dots + c_{1n}x_n = g_1, \quad (9)$$

$$c_{ij} = \frac{a_{ij}}{a_{11}}, \quad j = 2, 3, \dots, n, \quad g_1 = \frac{f_1}{a_{11}}.$$

С помощью уравнения (9) исключим x_1 из всех оставшихся уравнений системы (8):

$$\begin{aligned} a_{22}x_2 + a_{23}x_3 + \dots + a_{2n}x_n &= f_2^{(1)}, \\ a_{32}x_2 + a_{33}x_3 + \dots + a_{3n}x_n &= f_3^{(1)}, \\ \dots & \dots \\ a_{n2}x_2 + a_{n3}x_3 + \dots + a_{nn}x_n &= f_n^{(1)}, \end{aligned} \quad (10)$$

где $a_{ij}^{(1)} = a_{ij} - a_{11}c_{1j}$, $f_i^{(1)} = f_i - a_{11}g_1$, $i, j = 2, 3, \dots, n$. Неизвестное x_1 исключено из всех уравнений системы (8), кроме первого. При этом первое уравнение преобразовано к виду (9), а остальные — к виду (10). На этом завершается первый этап в процессе последовательного исключения неизвестных.

Далее разделим первое уравнение в (10) на $a_{22}^{(1)} \neq 0$ и с помощью преобразованного уравнения исключим x_2 из оставшихся уравнений (10). Вместо задачи (10), таким образом, мы получим уравнение, не содержащее неизвестное x_1 , и укороченную систему уравнений $(n-2)$ -го порядка. Аналогично исключая последовательно неизвестные x_3, x_4, \dots, x_{n-1} , окончательно приходим к системе уравнений с

верхней унитретугольной* матрицей

$$\begin{aligned} x_1 + c_{12}x_2 + c_{13}x_3 + \dots + c_{1n}x_n &= g_1, \\ x_2 + c_{23}x_3 + \dots + c_{2n}x_n &= g_2, \\ \dots & \dots \\ x_n &= g_n, \end{aligned} \quad (11)$$

эквивалентной исходной системе (8).

Получение системы (11) составляет *прямой ход* метода Гаусса. Коэффициенты уравнений в прямом ходе преобразуются следующим образом:

$$a_{kj}^{(0)} = a_{kj}, \quad k, j = 1, 2, \dots, n,$$

$$c_{kj} = \frac{a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}}, \quad j = k+1, k+2, \dots, n, \quad k = 1, 2, \dots, n-1,$$

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - a_{ik}^{(k-1)}c_{kj}, \quad i, j = k+1, k+2, \dots, n, \quad k = 1, 2, \dots, n-1.$$

Число $a_{kk}^{(k-1)}$ в дальнейшем будем называть *ведущим элементом* на k -м этапе исключения.

Вычисление правых частей системы (11) осуществляется по формулам

$$f_k^{(0)} = f_k, \quad g_k = \frac{f_k^{(k-1)}}{a_{kk}^{(k-1)}}, \quad k = 1, 2, \dots, n,$$

$$f_i^{(k)} = f_i^{(k-1)} - a_{ik}^{(k-1)}g_k, \quad i = k+1, k+2, \dots, n, \quad k = 1, 2, \dots, n-1.$$

Обратный ход состоит в последовательном нахождении неизвестных x_n, x_{n-1}, \dots, x_1 из системы (11):

$$x_t = g_t - \sum_{j=t+1}^n c_{tj}x_j, \quad t = n-1, n-2, \dots, 1, \quad x_n = g_n.$$

Несложно проверить, что для численной реализации схемы единственного деления требуется выполнить порядка $(2/3)n^3$ арифметических действий, причем основное время расчета затрачивается на осуществление прямого хода.

*Унитретугольной назовем треугольную матрицу с единичной главной диагональю.

3. Связь схемы единственного деления с LU-факторизацией. Установим связь приведенного метода с теоремой об LU-разложении. Для этого определим элементарную нижнюю треугольную матрицу

$$L_j = \begin{pmatrix} 1 & & & & \\ \ddots & & & & 0 \\ & 1 & & & \\ & & l_{jj} & & \\ & & l_{j+1,j} & 1 & \\ 0 & & \vdots & & \ddots \\ & & l_{nj} & & 1 \end{pmatrix}.$$

В матрице L_j все элементы главной диагонали, кроме l_{jj} , равны единице. Из остальных элементов отличными от нуля могут быть только поддиагональные элементы j -го столбца. Обратной к L_j является элементарная нижняя треугольная матрица

$$L_j^{-1} = \begin{pmatrix} 1 & & & & \\ \ddots & & & & 0 \\ & 1 & & & \\ & & l_{jj}^{-1} & & \\ & & -l_{j+1,j}l_{jj}^{-1} & 1 & \\ 0 & & \vdots & & \ddots \\ & & -l_{nj}l_{jj}^{-1} & & 1 \end{pmatrix}.$$

На первом этапе исключения матрица системы (8) приводится к матрице, в которой угловой элемент равен единице, а все поддиагональные элементы первого столбца равны нулю. Нетрудно убедиться, что эта процедура эквивалентна умножению матрицы A слева на матрицу

$$L_1 = \begin{pmatrix} 1/a_{11} & & & & \\ -a_{21}/a_{11} & 1 & & & 0 \\ \vdots & & \ddots & & \\ -a_{n1}/a_{11} & & & & 1 \end{pmatrix}.$$

при этом обратная к L_k матрица имеет вид

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

Таким образом, связь схемы единственного деления с теоремой 1 установлена.

Замечание 1. Разложение $A = LU$, утверждаемое теоремой 1, неоднозначно, поскольку существует произвол в определении l_{ii}, u_{ii} . С другой стороны, LU-разложение, осуществляемое в схеме единственного деления, как показано выше, фиксирует диагональные элементы матриц L, U , а именно $l_{ii} = a_{ii}^{(0-1)}, u_{ii} = 1$. Поэтому если D — диагональная матрица с ведущими элементами метода исключения на главной диагонали, то вместо LU-разложения можно говорить об LDU-разложении матрицы A .

Теорема 2. Пусть все угловые миноры матрицы A отличны от нуля. Тогда матрица A единственным образом представима в виде произведения матриц

$$A = LDU, \quad (14)$$

где L — нижняя унитреугольная матрица; U — верхняя унитреугольная матрица; $D = \text{diag}[a_{11}, a_{22}^{(1)}, \dots, a_{nn}^{(n-1)}]$.

Доказательство. Существование разложения (14) следует непосредственно из теоремы 1, поскольку LD — нижняя треугольная матрица. Покажем, что такое разложение единственное. Предположим, что матрицу A можно разложить двумя способами:

$$A = L_1 D_1 U_1, \quad A = L_2 D_2 U_2.$$

Тогда имеем

$$U_1 U_2^{-1} = D_1^{-1} L_1^{-1} L_2 D_2. \quad (15)$$

Матрица в левой части (15) является верхней унитреугольной матрицей, а в правой части — нижней треугольной. Равенство таких

При этом система (8) преобразуется к виду

$$L_2 L_1 Ax = L_2 L_1 f. \quad (12)$$

На втором этапе осуществляется переход от (12) к системе

$$L_2 L_1 Ax = L_2 L_1 f,$$

в которой матрица L_2 имеет вид

$$L_2 = \begin{pmatrix} 1 & & & & \\ & 1/a_{22}^{(1)} & & & 0 \\ & -a_{32}^{(1)}/a_{22}^{(1)} & 1 & & \\ 0 & \vdots & & \ddots & \\ & -a_{n2}^{(1)}/a_{22}^{(1)} & & & 1 \end{pmatrix}.$$

Продолжая дальше этот процесс, приходим к системе

$$L_n L_{n-1} \dots L_1 Ax = L_n L_{n-1} \dots L_1 f, \quad (13)$$

в которой элементарная нижняя треугольная матрица L_k на k -м этапе исключения имеет вид

$$L_k = \begin{pmatrix} 1 & & & & & & 0 \\ & 1 & & & & & \\ & & 1/a_{kk}^{(k-1)} & & & & \\ & & -a_{k+1,k}^{(k-1)}/a_{kk}^{(k-1)} & 1 & & & \\ 0 & & \vdots & & \ddots & & \\ & & -a_{nk}^{(k-1)}/a_{kk}^{(k-1)} & & & & 1 \end{pmatrix}.$$

Матрица системы (13) $U = L_n L_{n-1} \dots L_1 A$ является верхней унитреугольной матрицей. Отсюда следует, что

$$A = LU,$$

где $L = L_1^{-1} L_2^{-1} \dots L_n^{-1}$ — нижняя треугольная матрица, на главной диагонали которой расположены ведущие элементы метода Гаусса,

матриц возможно лишь в случае, когда

$$U_1 U_2^{-1} = D_1^{-1} L_1^{-1} L_2 D_2 = E.$$

Отсюда $U_1 = U_2$, $L_1^{-1} L_2 D_2 = D_1$ и, следовательно, $L_1 = L_2$. Теорема доказана.

4. Метод Гаусса с выбором главного элемента. Матрицы перестановок. Возможность проведения процесса исключения в схеме единственного деления гарантируется условиями

$$a_{11} \neq 0, \quad \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \neq 0, \quad \dots, \quad \det A \neq 0.$$

Однако в расчетах заранее неизвестно, все ли угловые миноры матрицы A отличны от нуля. При этом может оказаться, что система (1) имеет единственное решение, несмотря на то что какой-либо из угловых миноров матрицы A равен нулю. Кроме того, фиксация ведущего элемента в случае его относительной малости может привести в процессе вычисления к сильному накоплению погрешностей. Избежать этих ситуаций позволяет метод Гаусса с выбором главного элемента. Основная идея метода заключается в том, что в качестве ведущего элемента на каждом этапе исключения выбирается *наибольший по модулю (главный) элемент*. На практике обычно используются следующие варианты метода Гаусса с выбором главного элемента.

a) **Метод Гаусса с выбором главного элемента по строке.** Эквивалентен схеме единственного деления, примененной к системе, в которой на каждом этапе исключения проводится перенумерация переменных.

b) **Метод Гаусса с выбором главного элемента по столбцу.** Эквивалентен применению схемы единственного деления к системе, в которой на каждом этапе исключения проводится перенумерация переменных, и перестановка уравнений.

Определение 2. Матрицей перестановок P называется квадратная матрица n -го порядка, каждая строка и каждый столбец которой содержит ровно одну единицу и $n - 1$ нулей.

Определение 3. Элементарной матрицей перестановок P_{km} называется матрица, полученная из единичной матрицы перестановки k -й и m -й строк.

Отметим свойства элементарных матриц перестановок, вытекающие непосредственно из определения.

1. Произведение любого числа элементарных матриц перестановок есть матрица перестановок (не обязательно элементарная).

2. Матрица $P_{km}A$ есть матрица A , строки которой с номерами k и m перестановлены.

3. Матрица AP_{km} есть матрица A , k -й и m -й столбцы которой перестановлены.

С учетом введенных определений нетрудно получить матричное представление рассматриваемых методов. Например, для метода Гаусса с выбором главного элемента по столбцу разложение (13) можно записать в виде

$$\begin{aligned} L_n L_{n-1} P_{n-1} \dots L_2 P_2 L_1 P_1 A x = \\ = L_n L_{n-1} P_{n-1} \dots L_2 P_2 L_1 P_1 f, \end{aligned} \quad (16)$$

где P_1, P_2, \dots, P_{n-1} — элементарные матрицы перестановок такие, что $P_k = P_{km}$, $k \leq m \leq n$; L_k , $k = 1, 2, \dots, n$, — элементарные нижние треугольные матрицы. Используем вытекающие из свойства $P_k^{-1} = P_k$ соотношения перестановки

$$P_k L_{k-1} = \tilde{L}_{k-1} P_k,$$

где $\tilde{L}_{k-1} = P_k L_{k-1} P_k$ — нижняя треугольная матрица, имеющая обратную. Тогда из (16) получим

$$L_n L_{n-1} \tilde{L}_{n-2} \dots \tilde{L}_2 \tilde{L}_1 P A x = L_n L_{n-1} \tilde{L}_{n-2} \dots \tilde{L}_2 \tilde{L}_1 P f, \quad (17)$$

т.е. метод Гаусса с выбором главного элемента по столбцу эквивалентен схеме единственного деления, примененной к системе

$$P A x = P f,$$

где P — некоторая результатирующая матрица перестановок.

Теорема 3. Если $\det A \neq 0$, то существует матрица перестановок P такая, что матрица PA имеет отличные от нуля угловые миноры.

Доказательство. Докажем теорему 3 методом математической индукции. Для наглядности рассмотрим случай $n = 2$:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

Если $a_{11} \neq 0$, то теорема верна при $P = E$. Если $a_{11} = 0$, то $a_{21} \neq 0$, поскольку $\det A \neq 0$. Тогда у матрицы

$$P_{12} A = \begin{pmatrix} a_{21} & a_{22} \\ a_{11} & a_{12} \end{pmatrix}$$

все угловые миноры отличны от нуля.

Пусть утверждение теоремы выполняется для матрицы порядка $n - 1$. Покажем, что оно справедливо и для матрицы n -го порядка. Как и при доказательстве теоремы 1, представим матрицу A в блочном виде:

$$A_n = \begin{pmatrix} A_{n-1} & z \\ v^T & a_{nn} \end{pmatrix},$$

$$A_{n-1} = \begin{pmatrix} a_{11} & \dots & a_{1,n-1} \\ \dots & \dots & \dots \\ a_{n-1,1} & \dots & a_{n-1,n-1} \end{pmatrix}, \quad z = \begin{pmatrix} a_{1n} \\ \vdots \\ a_{n-1,n} \end{pmatrix}, \quad v = \begin{pmatrix} a_{n1} \\ \vdots \\ a_{n,n-1} \end{pmatrix}.$$

Рассмотрим два случая. Сначала предположим, что $\det A_{n-1} \neq 0$. По предположению индукции существует матрица перестановок P_{n-1} такая, что матрица $P_{n-1} A_{n-1}$ имеет отличные от нуля угловые миноры. Тогда для матрицы перестановок

$$P = \begin{pmatrix} P_{n-1} & 0 \\ 0 & 1 \end{pmatrix}$$

имеем

$$P A = \begin{pmatrix} P_{n-1} A_{n-1} & P_{n-1} z \\ v^T & a_{nn} \end{pmatrix},$$

и

$$P A x = P_{n-1} A_{n-1} x + P_{n-1} z.$$

или

$$\begin{pmatrix} 2 & 5 & 3 \\ 1 & 2 & 1 \\ 0 & 1 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 3 \end{pmatrix}.$$

Затем к системе (21) применим первый этап схемы единственного деления, т.е. умножим (21) на нижнюю треугольную матрицу $L_1 = \begin{pmatrix} 1/2 & 0 & 0 \\ -1/2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$:

$$L_1 P_{12} A x = L_1 P_{12} f,$$

или

$$\begin{pmatrix} 1 & 5/2 & 3/2 \\ 0 & -1/2 & -1/2 \\ 0 & 1 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 3 \end{pmatrix}.$$

Далее ищем наибольший по модулю элемент первого столбца матрицы укороченной системы

$$\begin{pmatrix} -1/2 & -1/2 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \end{pmatrix}.$$

Поскольку этот элемент находится во второй строке, меняем в (23) местами строки путем умножения (22) на матрицу перестановок $P_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$. Тем самым мы осуществляем переход от системы (22) к эквивалентной системе уравнений

$$P_{23} L_1 P_{12} A x = P_{23} L_1 P_{12} f,$$

(24)

при этом $\det(PA) = \pm \det A \neq 0$. Таким образом, все угловые миноры матрицы PA отличны от нуля.

Пусть теперь $\det A_{n-1} = 0$. Поскольку $\det A \neq 0$, то существует хотя бы один отличный от нуля минор порядка $n - 1$ матрицы A , полученный вычеркиванием последнего столбца и какой-либо строки, например:

$$\left| \begin{array}{ccc} a_{11} & \dots & a_{1,n-1} \\ \dots & \dots & \dots \\ a_{k-1,1} & \dots & a_{k-1,n-1} \\ a_{k+1,1} & \dots & a_{k+1,n-1} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{n,n-1} \end{array} \right| \neq 0, \quad k \neq n. \quad (18)$$

Переставляя в матрице A строки с номерами k и n , приходим к матрице $P_{kn}A$, у которой угловой минор $(n - 1)$ -го порядка имеет вид

$$\left| \begin{array}{ccc} a_{11} & \dots & a_{1,n-1} \\ \dots & \dots & \dots \\ a_{k-1,1} & \dots & a_{k-1,n-1} \\ a_{n1} & \dots & a_{n,n-1} \\ a_{k+1,1} & \dots & a_{k+1,n-1} \\ \dots & \dots & \dots \\ a_{n-1,1} & \dots & a_{n-1,n-1} \end{array} \right|$$

причем он отличается от (18) только перестановкой строк. Следовательно, этот минор не равен нулю. Теорема доказана.

Как следствие из теорем 2, 3 и представления (17) имеет место

Теорема 4. Пусть $\det A \neq 0$. Тогда существует матрица перестановок P такая, что справедливо единственное разложение

$$P A = L U, \quad (19)$$

где L — нижняя треугольная матрица, на главной диагонали которой стоят ведущие элементы метода Гаусса с выбором главного элемента по столбцу; U — верхняя унитреугольная матрица.

Замечание 2. Аналогичным образом можно сформулировать и доказать теоремы существования и единственности $L U$ ($L D U$)-разложений матрицы $A P$, реализуемых в методе Гаусса с выбором главного элемента по строке.

Пример 1. Рассмотрим применение элементарных нижних треугольных матриц и матриц перестановок для описания метода Гаусса с выбором главного элемента по столбцу. Пусть имеется СЛАУ третьего порядка

$$Ax = f, \quad A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 3 \\ 0 & 1 & 4 \end{pmatrix}, \quad f = \begin{pmatrix} 2 \\ 4 \\ 3 \end{pmatrix}. \quad (20)$$

Сначала ищем наибольший по модулю элемент первого столбца матрицы A (он находится во второй строке) и меняем местами первую и вторую строки исходной системы. Этой процедуре эквивалентна умножение (20) на матрицу перестановок P_{12} :

$$P_{12} A x = P_{12} f, \quad (21)$$

или

$$\begin{pmatrix} 2 & 5 & 3 \\ 1 & 2 & 1 \\ 0 & 1 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 3 \end{pmatrix}.$$

Затем к системе (21) применим первый этап схемы единственного деления, т.е. умножим (21) на нижнюю треугольную матрицу $L_1 = \begin{pmatrix} 1/2 & 0 & 0 \\ -1/2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$:

$$L_1 P_{12} A x = L_1 P_{12} f,$$

или

$$\begin{pmatrix} 1 & 5/2 & 3/2 \\ 0 & -1/2 & -1/2 \\ 0 & 1 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 3 \end{pmatrix}.$$

Далее ищем наибольший по модулю элемент первого столбца матрицы укороченной системы

$$\begin{pmatrix} -1/2 & -1/2 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \end{pmatrix}.$$

Поскольку этот элемент находится во второй строке, меняем в (23) местами строками путем умножения (22) на матрицу перестановок $P_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$. Тем самым мы осуществляем переход от системы (22) к эквивалентной системе уравнений

$$P_{23} L_1 P_{12} A x = P_{23} L_1 P_{12} f,$$

(24)

или

$$\begin{pmatrix} 1 & 5/2 & 3/2 \\ 0 & 1 & 4 \\ 0 & -1/2 & -1/2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 3/2 \end{pmatrix}.$$

Затем к системе (24) применим второй этап схемы единственного деления, т. е. умножим (24) на нижнюю треугольную матрицу $L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/2 & 1 \end{pmatrix}$:

$$L_2 P_{23} L_1 P_{12} A x = L_2 P_{23} L_1 P_{12} f, \quad (25)$$

или

$$\begin{pmatrix} 1 & 5/2 & 3/2 \\ 0 & 1 & 4 \\ 0 & 0 & 3/2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 3/2 \end{pmatrix}.$$

Наконец, к системе (25) применим третий этап схемы единственного деления, который состоит в умножении (25) на матрицу $L_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2/3 \end{pmatrix}$:

$$L_3 L_2 P_{23} L_1 P_{12} A x = L_3 L_2 P_{23} L_1 P_{12} f, \quad (26)$$

или

$$U x = \begin{pmatrix} 1 & 5/2 & 3/2 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}.$$

Таким образом, окончательно приходим к СЛАУ (26), из которой обратным ходом определим искомое решение: $x = (3, -1, 1)^T$.

Получим теперь факторизованное представление (19), утверждаемое теоремой 4. Для этого построим матрицу

$$\tilde{L}_1 = P_{23} L_1 P_{23} \quad (27)$$

путем перестановки сначала второй и третьей строк матрицы L_1 (находим матрицу $P_{23} L_1$), а затем второго и третьего столбцов матрицы $P_{23} L_1$:

$$\tilde{L}_1 = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 1 & 0 \\ -1/2 & 0 & 1 \end{pmatrix}.$$

Как видим, матрица \tilde{L}_1 является нижней треугольной, имеющей обратную.

С учетом свойства $P_{23} = P_{23}^{-1}$ из (27) имеем соотношение

$$\tilde{L}_1 P_{23} = P_{23} L_1. \quad (28)$$

Подставляя (28) в (26), получим систему уравнений

$$L_3 L_2 \tilde{L}_1 P_{23} P_{12} A x = L_3 L_2 \tilde{L}_1 P_{23} P_{12} f$$

с верхней унитретугольной матрицей

$$U = L_3 L_2 \tilde{L}_1 P_{23} P_{12} A. \quad (29)$$

Из (29) непосредственно следует разложение

$$PA = LU$$

с матрицами P и L вида

$$P = P_{23} P_{12} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad L = \tilde{L}_1^{-1} L_2^{-1} L_3^{-1} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -1/2 & 3/2 \end{pmatrix}.$$

Таким образом, показана эквивалентность метода Гаусса с выбором главного элемента по столбцу схеме единственного деления, примененной к системе $PAx = Pf$.

5. Вычисление определителей и обращение матриц. Выполнение преобразований матрицы A в процессе решения системы (1) позволяет без дополнительных затрат вычислить ее определитель и найти обратную матрицу. Пусть, например, имеет место разложение (19). Тогда

$$\det(PA) = \det L \det U = l_{11} l_{22} \dots l_{nn},$$

т. е. определитель матрицы PA равен произведению диагональных элементов матрицы L . Поскольку матрицы PA и A отличаются только перестановкой строк, их определители могут отличаться только знаками. Поэтому

$$\det A = (-1)^p a_{11} a_{22}^{(1)} \dots a_{nn}^{(n-1)},$$

где p — количество перестановок, осуществляемых в процессе исключения, $a_{kk}^{(k-1)}$, $k = 1, 2, \dots, n$, — ведущие элементы метода Гаусса. Если матрица A является вырожденной, то при использовании метода Гаусса с выбором главного элемента по столбцу на некотором этапе исключения k все элементы k -го столбца укороченной системы окажутся равными нулю и дальнейшее исключение становится невозможным. Для вычисления определителя матрицы, как и

в случае решения системы уравнений, требуется примерно $(2/3)n^3$ арифметических действий.

Задача нахождения матрицы, обратной матрице A , эквивалентна задаче решения матричного уравнения

$$AX = E,$$

где X — искомая матрица. Обозначим через x_1, x_2, \dots, x_n векторы-столбцы матрицы A^{-1} . Тогда вектор $x_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T$ является решением СЛАУ вида

$$Ax_j = e_j, \quad (30)$$

где e_j — j -й столбец единичной матрицы $E = [e_1, e_2, \dots, e_n]$. Поэтому для нахождения обратной матрицы необходимо решить n систем уравнений (30) с одной и той же матрицей A и различными правыми частями.

Несмотря на то что обращение матрицы с помощью метода Гаусса сводится к решению n систем уравнений, оно требует приблизительно $2n^3$ арифметических действий, т. е. лишь в три раза больше, чем решение одной СЛАУ. Это объясняется тем, что в методе Гаусса большая часть вычислений связана с приведением матрицы к верхнему треугольному виду (прямой ход), а при обращении матрицы эта процедура делается только один раз.

Помимо изложенного подхода, для обращения матрицы может быть эффективно использован метод Гаусса — Жордана, смысл которого заключается в следующем. Обозначим через N_1 матрицу, отличающуюся от единичной недиагональными элементами первого столбца (она совпадает с элементарной нижней треугольной матрицей L_1), и образум матрицу $A_1 = N_1 A$. Далее строим матрицу

$$N_2 = \begin{pmatrix} 1 & -a_{12}^{(1)}/a_{22}^{(1)} & & & \\ & 1/a_{22}^{(1)} & 0 & & \\ 0 & \vdots & 1 & \ddots & \\ & -a_{n2}^{(1)}/a_{22}^{(1)} & & & 1 \end{pmatrix}$$

и умножим ее слева на матрицу A_1 . В полученной таким образом матрице $A_2 = N_2 A_1$ все недиагональные элементы второго столбца окажутся равными нулю.

На k -м этапе исключения приходим к матрице

$$A_k = \begin{pmatrix} 1 & a_{1,k+1}^{(k)} & a_{1,k+2}^{(k)} & \dots & a_{1n}^{(k)} \\ \ddots & a_{2,k+1}^{(k)} & a_{2,k+2}^{(k)} & \dots & a_{2n}^{(k)} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ a_{k+1,k+1}^{(k)} & a_{k+1,k+2}^{(k)} & a_{k+1,k+3}^{(k)} & \dots & a_{k+1,n}^{(k)} \\ 0 & \vdots & \vdots & \ddots & \vdots \\ a_{n,k+1}^{(k)} & a_{n,k+2}^{(k)} & a_{n,k+3}^{(k)} & \dots & a_{nn}^{(k)} \end{pmatrix},$$

связанной с A_{k-1} соотношением $A_k = N_k A_{k-1}$, где

$$N_k = \begin{pmatrix} 1 & -a_{1k}^{(k-1)}/a_{kk}^{(k-1)} & & & & \\ & -a_{2k}^{(k-1)}/a_{kk}^{(k-1)} & 0 & & & \\ & & 1 & a_{kk}^{(k-1)} & & \\ & & & -a_{k+1,k}^{(k-1)}/a_{kk}^{(k-1)} & 1 & \\ 0 & & & & \ddots & \\ & -a_{nk}^{(k-1)}/a_{kk}^{(k-1)} & & & & 1 \end{pmatrix}.$$

В результате после n этапов получим единичную матрицу

$$E = A_n = N_n N_{n-1} \dots N_1 A.$$

Отсюда следует, что

$$A^{-1} = N_n N_{n-1} \dots N_1.$$

Таким образом, приходим к разложению обратной матрицы на элементарные сомножители.

Для вычисления элементов обратной матрицы A^{-1} по методу

Гаусса — Жордана удобно пользоваться формулами

$$b_{ij}^{(k)} = \begin{cases} b_{ij}^{(k-1)} - b_{kj}^{(k-1)} \frac{b_{ik}^{(k-1)}}{b_{kk}^{(k-1)}}, & i, j \neq k, \\ \frac{b_{kj}^{(k-1)}}{b_{kk}^{(k-1)}}, & i = k, j \neq k, \\ -\frac{b_{ik}^{(k-1)}}{b_{kk}^{(k-1)}}, & i \neq k, j = k, \\ \frac{1}{b_{kk}^{(k-1)}}, & i = k, j = k, \end{cases}$$

$b_{ij}^{(0)} = a_{ij}$, $i, j = 1, 2, \dots, n$, $k = 1, 2, \dots, n$, позволяющими последовательно строить матрицы $B_1, B_2, \dots, B_n = A^{-1}$.

6. Диагонально доминирующие матрицы. При решении некоторых классов СЛАУ методом Гаусса необходимость в выборе главного элемента отсутствует. Выявление таких классов систем является важным, поскольку процедура выбора главного элемента заметно усложнит вычислительный процесс.

Определение 4. Матрица $A \in \mathbb{R}^{n \times n}$ называется *строго диагонально доминирующей*, если выполняются условия

$$|a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ij}|, \quad i = 1, 2, \dots, n.$$

Следующая теорема показывает, как свойство диагонального доминирования может гарантировать возможность проведения процесса исключения.

Теорема 5. Пусть матрица A является строго диагонально доминирующей. Тогда существует LU-разложение матрицы A .

Доказательство. Пусть $n = 2$. Так как

$$A_2 = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

является матрицей со строгим диагональным доминированием, то $|a_{11}| > |a_{12}|$, $|a_{22}| > |a_{21}|$ и $a_{11}a_{22} - a_{12}a_{21} \neq 0$, т.е. все угловые миноры матрицы A_2 отличны от нуля и, следовательно, имеет место равенство $A_2 = L_2 U_2$.

Далее, запишем матрицу A в виде

$$A_n = \begin{pmatrix} a_{11} & & z^T \\ v & A_{n-1} \end{pmatrix},$$

$$A_{n-1} = \begin{pmatrix} a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots \\ a_{n2} & \dots & a_{nn} \end{pmatrix}, \quad z = \begin{pmatrix} a_{12} \\ \vdots \\ a_{1n} \end{pmatrix}, \quad v = \begin{pmatrix} a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}.$$

После одного этапа LU-факторизации получим разложение

$$A_n = \begin{pmatrix} a_{11} & 0 \\ v & E \end{pmatrix} \begin{pmatrix} 1 & z^T/a_{11} \\ 0 & B_{n-1} \end{pmatrix}, \quad (31)$$

где $B_{n-1} = A_{n-1} - v z^T / a_{11}$. Покажем, что матрица B_{n-1} является строго диагонально доминирующей. С учетом индуктивного предположения о матрице A_{n-1} имеем

$$\sum_{j=2, j \neq i}^n |b_{ij}| = \sum_{j=2, j \neq i}^n \left| a_{ij} - \frac{v_i z_j}{a_{11}} \right| \leq \sum_{j=2, j \neq i}^n |a_{ij}| + \frac{|v_i|}{|a_{11}|} \sum_{j=2, j \neq i}^n |z_j| <$$

$$< |a_{ii}| - |v_i| + \frac{|v_i|}{|a_{11}|} (|a_{11}| - |z_i|) \leq |a_{ii} - \frac{z_i v_i}{a_{11}}| = |b_{ii}|.$$

Следовательно, $B_{n-1} = L_{n-1} U_{n-1}$. Но тогда из (31) получим

$$A_n = \begin{pmatrix} a_{11} & 0 \\ v & L_{n-1} \end{pmatrix} \begin{pmatrix} 1 & z^T/a_{11} \\ 0 & U_{n-1} \end{pmatrix} \equiv L_n U_n.$$

Теорема доказана.

7. Метод квадратного корня. Одним из наиболее важных классов СЛАУ (1) специального вида являются системы уравнений с симметричной положительно определенной^{*} матрицей. Выясним вначале, какими свойствами обладают элементы такой матрицы.

*Напомним, что матрица A называется положительно определенной ($A > 0$), если $(Ax, x) > 0 \forall x \in \mathbb{R}^n$, $x \neq 0$. Из условия $A > 0$ следует существование константы $\delta > 0$ такой, что $(Ax, x) \geq \delta \|x\|^2 \forall x \in \mathbb{R}^n$, $x \neq 0$.

После того как разложение (32) получено, решение системы (1) сводится к последовательному решению двух систем уравнений с треугольными матрицами $S^T y = f$, $Sx = y$ по формулам

$$y_i = \frac{f_i - \sum_{j=1}^{i-1} s_{ji} y_j}{s_{ii}}, \quad i = 2, 3, \dots, n, \quad y_1 = \frac{f_1}{s_{11}},$$

$$x_i = \frac{\sum_{j=i+1}^n s_{ij} x_j}{s_{ii}}, \quad i = n-1, n-2, \dots, 1, \quad x_n = \frac{y_n}{s_{nn}}.$$

Определитель матрицы A , вычисляемый методом квадратного корня, очевидно, равен

$$\det A = \det S^T \det S = s_{11}^2 s_{22}^2 \dots s_{nn}^2.$$

Разложение Холецкого позволяет построить *кампактную схему* для вычисления элементов обратной матрицы, не использующую процедуру обращения матрицы S . Обозначим $A^{-1} = B = \{b_{ij}\}_{i,j=1}^n$. Из равенства $B = S^{-1}(S^T)^{-1}$ следует соотношение

$$SB = (S^T)^{-1}, \quad (38)$$

в котором нижняя треугольная матрица $(S^T)^{-1}$ имеет вид

$$(S^T)^{-1} = \begin{pmatrix} s_{11}^{-1} & & & 0 \\ \times & s_{22}^{-1} & & \\ \vdots & & \ddots & \\ \times & \times & \dots & s_{nn}^{-1} \end{pmatrix}.$$

Сравнивая между собой верхние треугольные части матриц в (38), приходим к рекуррентным формулам для вычисления элементов обратной матрицы:

$$b_{nn} = \frac{1}{s_{nn}^2}, \quad b_{ij} = b_{ji} = -\frac{\sum_{k=i+1}^{n-1} s_{ik} b_{kj}}{s_{ii}}, \quad j > i, \quad i = n-1, n-2, \dots, 1;$$

$$b_{ii} = \frac{s_{ii}^{-1} - \sum_{k=i+1}^n s_{ik} b_{ik}}{s_{ii}}, \quad i = n-1, n-2, \dots, 1.$$

Пример 2. Пусть дана симметричная положительно определенная матрица

$$A = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 6 \end{pmatrix}.$$

Вычислим ее определитель и найдем обратную матрицу с помощью метода квадратного корня. Используя формулу (37), строим верхнюю треугольную матрицу

$$S = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}.$$

Определитель матрицы A равен

$$\det A = s_{11}^2 s_{22}^2 s_{33}^2 = 64.$$

Далее последовательно находим элементы b_{ij} обратной матрицы:

$$b_{33} = s_{33}^{-2} = \frac{1}{4}, \quad b_{23} = b_{32} = -\frac{s_{23} b_{33}}{s_{22}} = -\frac{1}{8},$$

$$b_{13} = b_{31} = -\frac{s_{12} b_{23} + s_{13} b_{33}}{s_{11}} = -\frac{1}{16}, \quad b_{22} = \frac{s_{22}^{-1} - s_{23} b_{33}}{s_{22}} = \frac{5}{16},$$

$$b_{12} = b_{21} = -\frac{s_{12} b_{22} + s_{13} b_{33}}{s_{11}} = -\frac{3}{32}, \quad b_{11} = \frac{s_{11}^{-1} - s_{12} b_{22} - s_{13} b_{33}}{s_{11}} = \frac{21}{64}.$$

Таким образом, окончательно имеем

$$A^{-1} = \begin{pmatrix} 21/64 & -3/32 & -1/16 \\ -3/32 & 5/16 & -1/8 \\ -1/16 & -1/8 & 1/4 \end{pmatrix}.$$

Замечание 3. Метод квадратного корня при больших n требует примерно в два раза меньше арифметических действий по сравнению с методом Гаусса, что объясняется использованием информации о симметрии матрицы при строении вычислительного алгоритма.

В комплексном случае метод квадратного корня применяется для решения СЛАУ с эрмитовой матрицей $A = A^*$ ($a_{ij} = \bar{a}_{ji}$). При этом вместо разложения (32) реализуется более общее разложение

$$A = S^* D S, \quad (39)$$

где S — верхняя треугольная матрица с положительными элементами на главной диагонали; S^* — эрмитово сопряженная к ней матрица; D — диагональная матрица, на диагонали которой расположены элементы, равные ± 1 . Формулы для вычисления элементов матриц S и D соответственно принимают вид

$$\begin{aligned} d_{ii} &= \operatorname{sign}(a_{ii}) \sum_{k=1}^{i-1} d_{kk} |s_{ki}|^2, \quad s_{ii} = \sqrt{\left| a_{ii} - \sum_{k=1}^{i-1} d_{kk} |s_{ki}|^2 \right|^2}, \\ a_{ij} &= \frac{a_{ij} - \sum_{k=1}^{i-1} d_{kk} \bar{s}_{ki} s_{kj}}{d_{ii} s_{ii}}, \quad j > i, \end{aligned} \quad (40)$$

а искомое решение находится обратным ходом из систем

$$S^* D y = f, \quad S x = y.$$

Отметим, что возможность факторизованного представления (39) обеспечивается отличием от нуля всех угловых миноров матрицы A . В противном случае какой-либо элемент s_{ii} может оказаться равным или близким к нулю (например, $s_{11} = 0$ при $a_{11} = 0$), что делает невозможным использование формулы (40). Этого можно избежать, если применить *симметрическую перестановку* $P_{km} A P_{km}$, позволяющую переупорядочить главную диагональ матрицы A (элементы a_{kk} и a_{mm} меняются местами).

§ 2. Ортогональные разложения

1. Ортогональные матрицы. Как мы уже отмечали, всякая система линейных алгебраических уравнений

$$Ax = f \quad (1)$$

характеризуется числом обусловленности $\kappa(A)$, которое не зависит от метода решения задачи (1). Естественно требовать, чтобы численный алгоритм не вносил дополнительных изменений в степень обусловленности исходной системы, т. е. обеспечивал достаточную вычислительную устойчивость. Рассмотренные выше методы

(за исключением метода квадратного корня) не всегда удовлетворяют указанному требованию. Например, в методе Гаусса с выбором главного элемента по столбцу совершается переход от задачи (1) к задаче с верхней треугольной матрицей

$$U = L_n L_{n-1} P_{n-1} \dots L_1 P_1 A,$$

при этом

$$\kappa_2(U) \leq \kappa_2(L_n) \kappa_2(L_{n-1}) \kappa_2(P_{n-1}) \dots \kappa_2(L_1) \kappa_2(P_1) \kappa_2(A).$$

Поскольку $\kappa_2(P_k) = 1$, а

$$\kappa_2(L_k) = \begin{cases} |a_{kk}^{(k-1)}|, & \text{если } |a_{kk}^{(k-1)}| > 1, \\ |a_{kk}^{(k-1)}|^{-1}, & \text{если } |a_{kk}^{(k-1)}| < 1, \end{cases}$$

то при существенном росте (умножении) элементов матриц L_k обратный ход в методе Гаусса будет происходить с потерей точности. Такого нежелательного явления можно избежать, если осуществлять приведение матрицы A к верхнему треугольному виду с помощью ортогональных преобразований.

Определение 1. Матрица Q с вещественными элементами q_{ij} , $i, j = 1, 2, \dots, n$, называется *ортогональной*, если $Q^{-1} = Q^T$.

Из данного определения вытекают следующие свойства ортогональных матриц:

$$1. Q^T Q = Q Q^T = E.$$

2. Произведение любого числа ортогональных матриц есть ортогональная матрица.

3. Если $B = QA$, где Q — ортогональная матрица, то число обусловленности $\kappa_2(B) = \kappa_2(A)$.

Первые два свойства ортогональных матриц очевидны. Для доказательства свойства 3 воспользуемся тем фактом, что

$$\|Q\|_2 = \sqrt{\lambda_{\max}(Q^T Q)} = \sqrt{\lambda_{\max}(E)} = 1,$$

$$\|Q^{-1}\|_2 = \sqrt{\lambda_{\max}[(Q^{-1})^T Q^{-1}]} = \sqrt{\lambda_{\max}[(QQ^T)^{-1}]} = 1,$$

т. е. $\kappa_2(Q) = 1$. Тогда из свойства 2 числа обусловленности получим

$$\kappa_2(QA) \leq \kappa_2(Q) \kappa_2(A) = \kappa_2(A).$$

С другой стороны,

$$\kappa_2(A) = \kappa_2(Q^{-1}QA) \leq \kappa_2(Q^{-1}) \kappa_2(QA) = \kappa_2(QA).$$

Следовательно, $\kappa_2(QA) = \kappa_2(A)$.

Теорема 1. Всякую невырожденную матрицу A можно представить в виде произведения ортогональной матрицы Q на верхнюю треугольную матрицу U :

$$A = QU. \quad (2)$$

Доказательство. Матрица $A^T A$ является симметричной и положительно определенной. Согласно теореме 7 из § 1, она может быть факторизована следующим образом:

$$A^T A = U^T U, \quad (3)$$

где U — верхняя треугольная матрица. Запишем матрицу A в виде $A = (AU^{-1})U$. С учетом (3) получим

$$\begin{aligned} (AU^{-1})(AU^{-1})^T &= AU^{-1}(U^{-1})^T A^T = A(U^T U)^{-1} A^T = \\ &= A(A^T A)^{-1} A^T = AA^{-1}(A^T)^{-1} A^T = E. \end{aligned}$$

Следовательно, матрица AU^{-1} ортогональна и мы имеем разложение (2). Теорема доказана.

Замечание 1. Аналогично можно показать, что любая невырожденная матрица представима в виде

$$A = LQ, \quad (4)$$

где L — нижняя треугольная, Q — ортогональная матрица.

Таким образом, фактически доказана возможность приведения системы (1) к эквивалентной системе уравнений с треугольной матрицей, при этом устойчивость вычислительного процесса определяется только свойствами исходной матрицы A . В качестве ортогональных матриц Q , с помощью которых осуществляется разложение (2), на практике используются элементарные ортогональные матрицы — матрицы отражения и матрицы вращения.

2. Метод отражений. Определим матрицу отражения (матрицу Хаусхолдера) как матрицу

$$V = E - 2ww^T, \quad (5)$$

где w — некоторый вектор-столбец единичной длины: $(w, w) = 1$. Отметим следующие свойства матрицы отражения (5).

1. Матрица V является симметричной и ортогональной.

Действительно,

$$V^T = (E - 2ww^T)^T = E - 2(w^T)^T w^T = E - 2ww^T = V;$$

$$\begin{aligned} VV^T &= (E - 2ww^T)(E - 2ww^T) = E - 4ww^T + 4ww^T w w^T = \\ &= E - 4ww^T + 4(w, w)ww^T = E. \end{aligned}$$

2. Матрица V оставляет без изменений все векторы, ортогональные w :

$$Vx = x - 2ww^T x = x - 2w(w, x) = x.$$

3. Матрица V меняет на противоположные векторы, коллинеарные w :

$$Vx = x - 2ww^T x = \lambda w - 2w(w, \lambda w) = \lambda w - 2\lambda w = -\lambda w = -x.$$

Используя указанные геометрические свойства матрицы отражения, нетрудно решить задачу построения матрицы V , переводящей заданный вектор $y \neq 0$ в вектор, коллинеарный единичному вектору e , т. е. подобрать w таким образом, чтобы $Vy = \alpha e$. Имеем

$$(E - 2ww^T)y = \alpha e,$$

откуда

$$y - \alpha e = 2ww^T y = 2(y, w)w,$$

т. е. $w = \rho^{-1}(y - \alpha e)$, где $\rho = 2(y, w)$ — нормирующий множитель. Подставляя выражение для w в последнее равенство, получим

$$\frac{1}{\rho^2} (2y, y - \alpha e) = 1, \quad \rho = \sqrt{(2y, y - \alpha e)}.$$

Осталось выбрать число α так, чтобы $(y, y - \alpha e) > 0$. Этому условию, очевидно, удовлетворяет $\alpha = \sqrt{(y, y)}$.

Таким образом, для решения задачи построения матрицы V такой, что $Vy = \alpha e$, где y и e — заданные векторы, необходимо положить

$$w = \frac{1}{\rho} (y - \alpha e), \quad \alpha = \sqrt{(y, y)}, \quad \rho = \sqrt{2(y, y - \alpha e)}. \quad (6)$$

Опишем теперь схему *метода отражений* решения СЛАУ (1). На первом этапе, используя формулы (6), образуем матрицу отражения V_1 по векторам $y^{(1)} = (a_{11}, a_{21}, \dots, a_{n1})^T$, $e = (1, 0, \dots, 0)^T$. Умножив слева (1) на V_1 , приходим к системе

$$A^{(1)}x = V_1 f, \quad (7)$$

где в матрице $A^{(1)} = V_1 A$ все поддиагональные элементы первого столбца равны нулю:

$$\begin{aligned} a_{11}^{(1)} &= \alpha^{(1)}, \quad a_{11}^{(1)} = 0, \quad i = 2, 3, \dots, n, \quad a_{ij}^{(1)} = a_{ij} - 2(y_j^{(1)}, w^{(1)})w_i^{(1)}, \\ y_j^{(1)} &= (a_{1j}, a_{2j}, \dots, a_{nj})^T, \quad i = 1, 2, \dots, n, \quad j = 2, 3, \dots, n. \end{aligned}$$

На втором этапе аналогично образуем матрицу V_2 по векторам $y^{(2)} = (0, a_{22}^{(1)}, a_{32}^{(1)}, \dots, a_{n2}^{(1)})^T$, $e = (0, 1, 0, \dots, 0)^T$ и умножим (7) слева на V_2 :

$$A^{(2)}x = V_2 V_1 f.$$

В матрице $A^{(2)} = V_2 V_1 A$ первая строка совпадает с первой строкой матрицы $A^{(1)}$, а все поддиагональные элементы второго столбца равны нулю:

$$\begin{aligned} a_{ij}^{(2)} &= a_{ij}^{(1)}, \quad j = 1, 2, \dots, n, \quad a_{22}^{(2)} = \alpha^{(2)}, \quad a_{ij}^{(2)} = 0, \\ i = j + 1, j + 2, \dots, n, \quad j &= 1, 2, \quad a_{ij}^{(2)} = a_{ij}^{(1)} - 2(y_j^{(2)}, w^{(2)})w_i^{(2)}, \\ y_j^{(2)} &= (0, a_{2j}^{(1)}, a_{3j}^{(1)}, \dots, a_{nj}^{(1)})^T, \quad i = 2, 3, \dots, n, \quad j = 3, 4, \dots, n. \end{aligned}$$

Предположим, что построена матрица $A^{(k-1)}$, у которой

$$a_{ij}^{(k-1)} = 0, \quad i = j + 1, j + 2, \dots, n, \quad j = 1, 2, \dots, k - 1.$$

Тогда, выбирая векторы $y^{(k)} = (0, 0, \dots, 0, a_{kk}^{(k-1)}, a_{k+1,k}^{(k-1)}, \dots, a_{nk}^{(k-1)})^T$, $e = (0, 0, \dots, 0, 1, 0, 0, \dots, 0)^T$ и образуя V_k , приходим к системе

$$A^{(k)}x = V_k V_{k-1} \dots V_1 f, \quad A^{(k)} = V_k A^{(k-1)},$$

$$\begin{aligned} a_{ij}^{(k)} &= a_{ij}^{(k-1)}, \quad i = 1, 2, \dots, k - 1, \quad j = 1, 2, \dots, n, \\ a_{kk}^{(k)} &= \alpha^{(k)}, \quad a_{ij}^{(k)} = 0, \quad i = j + 1, j + 2, \dots, n, \quad j = 1, 2, \dots, k, \\ a_{ij}^{(k)} &= a_{ij}^{(k-1)} - 2(y_j^{(k)}, w^{(k)})w_i^{(k)}, \\ y_j^{(k)} &= (0, 0, \dots, 0, a_{kj}^{(k-1)}, a_{k+1,j}^{(k-1)}, \dots, a_{nj}^{(k-1)})^T, \\ i = k, k + 1, \dots, n, \quad j &= k + 1, k + 2, \dots, n. \end{aligned}$$

После выполнения $n - 1$ этапов, очевидно, будем иметь

$$A^{(n-1)}x = V_{n-1} V_{n-2} \dots V_1 f,$$

где $A^{(n-1)} = V_{n-1} V_{n-2} \dots V_1 A$ — верхняя треугольная матрица.

Таким образом, получено разложение (2) с ортогональной матрицей $Q = V^T$, $V = V_{n-1} V_{n-2} \dots V_1$.

3. Метод вращений. Разложение матрицы на ортогональный и треугольный множители можно осуществить также с помощью матриц вращения (матриц Гивенса)

$$T_{kl} = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & \cos \varphi & \dots & -\sin \varphi & & 0 \\ & & & \vdots & \ddots & \vdots & & \\ & & & \sin \varphi & \dots & \cos \varphi & & \\ 0 & & & & & & 1 & \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix}, \quad k < l,$$

которые отличаются от единичной матрицы только четырьмя элементами, расположеными на пересечении строк и столбцов с номерами k, l . Соответствующий алгоритм *метода вращений* выглядит следующим образом. Умножим слева (1) на T_{12} :

$$A^{(1,2)}x = T_{12}f, \quad A^{(1,2)} = T_{12}A. \quad (8)$$

Легко убедиться, что матрица $A^{(1,2)}$ отличается от матрицы A только первыми двумя строками:

$$\begin{aligned} a_{1j}^{(1,2)} &= (\cos \varphi)_{12} a_{1j} - (\sin \varphi)_{12} a_{2j}, \\ a_{2j}^{(1,2)} &= (\sin \varphi)_{12} a_{1j} + (\cos \varphi)_{12} a_{2j}, \quad j = 1, 2, \dots, n, \\ a_{ij}^{(1,2)} &= a_{ij}, \quad i = 3, 4, \dots, n, \quad j = 1, 2, \dots, n. \end{aligned} \quad (9)$$

Положив в (9)

$$(\cos \varphi)_{12} = \frac{a_{11}}{\sqrt{a_{11}^2 + a_{21}^2}}, \quad (\sin \varphi)_{12} = -\frac{a_{21}}{\sqrt{a_{11}^2 + a_{21}^2}},$$

получим

$$a_{11}^{(1,2)} = \sqrt{a_{11}^2 + a_{21}^2}, \quad a_{21}^{(1,2)} = 0.$$

Далее рассмотрим матрицу вращения T_{13} . Умножение слева (8) на матрицу T_{13} приводит к системе

$$A^{(1,3)}x = T_{13}T_{12}f, \quad A^{(1,3)} = T_{13}T_{12}A,$$

у которой матрица $A^{(1,3)}$ отличается от матрицы $A^{(1,2)}$ первой и третьей строками:

$$\begin{aligned} a_{ij}^{(1,3)} &= a_{ij}^{(1,2)}, \quad i = 2, \quad i = 4, 5, \dots, n, \quad j = 1, 2, \dots, n, \\ a_{1j}^{(1,3)} &= (\cos \varphi)_{13} a_{1j}^{(1,2)} - (\sin \varphi)_{13} a_{3j}^{(1,2)}, \\ a_{3j}^{(1,3)} &= (\sin \varphi)_{13} a_{1j}^{(1,2)} + (\cos \varphi)_{13} a_{3j}^{(1,2)}, \quad j = 1, 2, \dots, n. \end{aligned}$$

Положим теперь

$$(\cos \varphi)_{13} = \frac{a_{11}^{(1,2)}}{\sqrt{(a_{11}^{(1,2)})^2 + (a_{31}^{(1,2)})^2}}, \quad (\sin \varphi)_{13} = -\frac{a_{31}^{(1,2)}}{\sqrt{(a_{11}^{(1,2)})^2 + (a_{31}^{(1,2)})^2}}$$

Тогда

$$a_{11}^{(1,3)} = \sqrt{(a_{11}^{(1,2)})^2 + (a_{31}^{(1,2)})^2}, \quad a_{31}^{(1,3)} = 0.$$

Дальнейшие преобразования свяжем с матрицами вращения T_{14} , T_{15}, \dots, T_{1n} . При этом

$$(\cos \varphi)_{1k} = \frac{\sqrt{a_{11}^2 + a_{21}^2 + \dots + a_{k-1,1}^2}}{\sqrt{a_{11}^2 + a_{21}^2 + \dots + a_{k1}^2}},$$

$$(\sin \varphi)_{1k} = \frac{a_{k1}}{\sqrt{a_{11}^2 + a_{21}^2 + \dots + a_{k1}^2}},$$

$$a_{ij}^{(1,k)} = (\cos \varphi)_{1k} a_{ij}^{(1,k-1)} - (\sin \varphi)_{1k} a_{kj}^{(1,k-1)},$$

$$a_{kj}^{(1,k)} = (\sin \varphi)_{1k} a_{ij}^{(1,k-1)} + (\cos \varphi)_{1k} a_{kj}^{(1,k-1)},$$

$$a_{ij}^{(1,k)} = a_{ij}^{(1,k-1)}, \quad i \neq 1, \quad i \neq k, \quad j = 1, 2, \dots, n, \quad k = 2, 3, \dots, n.$$

После завершения цикла система приводится к виду

$$A^{(1)}x = T_{11}f, \quad A^{(1)} = T_1 A, \quad T_1 = T_{1n} T_{1,n-1} \dots T_{12}, \quad (10)$$

где

$$a_{11}^{(1)} = \sqrt{a_{11}^2 + a_{21}^2 + \dots + a_{n1}^2}, \quad a_{11}^{(1)} = 0, \quad i = 2, 3, \dots, n.$$

Второй этап связан с матрицами вращения $T_{23}, T_{24}, \dots, T_{2n}$ и равносителем умножению (10) слева на ортогональную матрицу $T_2 = T_{2n} T_{2,n-1} \dots T_{23}$:

$$A^{(2)}x = T_2 T_1 f, \quad A^{(2)} = T_2 T_1 A.$$

После $(n - 1)$ -го этапа получим систему уравнений с треугольной матрицей

$$A^{(n-1)}x = T_{n-1} T_{n-2} \dots T_1 f, \quad A^{(n-1)} = T_{n-1} T_{n-2} \dots T_1 A,$$

$$T_k = T_{kn} T_{k,n-1} \dots T_{k+1,1}, \quad k = 1, 2, \dots, n - 1.$$

Отсюда вытекает, что

$$A = T^T A^{(n-1)}, \quad T = T_{n-1} T_{n-2} \dots T_1,$$

и мы имеем разложение (2). $A = Q U$

$Q = \text{ортогональная матрица}$

$U = \text{трехугольная матрица}$

4. Метод ортогонализации. Рассмотрим метод, позволяющий получить разложение (4) исходной матрицы на треугольный и ортогональный множители с помощью процесса *ортогонализации* системы векторов. Перепишем задачу (1) в виде

$$(4) \quad A \in \mathbb{R}^{n \times n}, \quad (a_i, y) = 0, \quad i = 1, 2, \dots, n, \quad (11)$$

$L_i = \text{линейная форма}$

$Q = \text{ортогональная матрица}$

$a_i = (a_{i1}, a_{i2}, \dots, a_{in}, -f_i)$, $i = 1, 2, \dots, n$, $y = (x_1, x_2, \dots, x_n, 1)^T$. Равенство (11) означает, что решение системы (1) эквивалентно нахождению вектора y , ортогонального ко всем линейно независимым векторам a_1, a_2, \dots, a_n и имеющему единичную последнюю координату. Ортогональность вектора y к векторам a_1, a_2, \dots, a_n означает ортогональность y к любому базису подпространства P_n , натянутого на a_1, a_2, \dots, a_n . Применим к системе векторов a_1, a_2, \dots, a_{n+1} , $a_{n+1} = (0, 0, \dots, 0, 1)$, процесс ортогонализации Шмидта, состоящий в построении ортонормированного базиса b_1, b_2, \dots, b_{n+1} по рекуррентным соотношениям

$$v_1 = a_1, \quad b_1 = \frac{v_1}{\sqrt{(v_1, v_1)}}, \quad v_k = a_k + \sum_{i=1}^{k-1} c_{ki} b_i,$$

$$c_{ki} = -(a_k, b_i), \quad b_k = \frac{v_k}{\sqrt{(v_k, v_k)}}, \quad k = 2, 3, \dots, n+1.$$

Векторы a_1, a_2, \dots, a_n линейно выражаются через b_1, b_2, \dots, b_n , поэтому вектор b_{n+1} ортогонален ко всем векторам a_1, a_2, \dots, a_n .

Таким образом, искомое решение системы уравнений вычисляется по формуле

$$x_i = \frac{z_i}{z_{n+1}}, \quad i = 1, 2, \dots, n,$$

где z_1, z_2, \dots, z_{n+1} — компоненты вектора b_{n+1} .

Рассмотрим процесс ортогонализации с точки зрения матричных преобразований. Нормировка v_1 сводится к умножению матрицы A слева на диагональную матрицу $D_1 = [1/\sqrt{(v_1, v_1)}, 1, \dots, 1]$, что приводит к матрице $B_1 = D_1 A$. Далее, вычисление вектора v_2 эквивалентно умножению матрицы B_1 слева на элементарную нижнюю треугольную матрицу

$$M_2 = \begin{pmatrix} 1 & & & \\ c_{21} & 1 & & 0 \\ & \ddots & & \\ 0 & & & 1 \end{pmatrix}, \quad c_{21} = -(a_2, b_1).$$

Нормировка вектора v_2 состоит в умножении матрицы $M_2 B_1$ слева на диагональную матрицу $D_2 = \text{diag}[1, 1/\sqrt{(v_2, v_2)}, 1, \dots, 1]$ и т. д.

На k -м этапе определяется матрица

$$M_k = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & 0 \\ & & \ddots & & & \\ c_{k1} & \dots & c_{kk-1} & 1 & & \\ & & & & \ddots & \\ 0 & & & & & 1 \end{pmatrix},$$

на которую умножается найденная ранее матрица B_{k-1} слева, и вычисляется матрица $B_k = D_k M_k B_{k-1}$, где D_k отличается от единичной матрицы только диагональным элементом $d_{kk} = 1/\sqrt{(v_k, v_k)}$. После выполнения n этапов мы получим ортогональную матрицу

$$B_n = D_n M_n D_{n-1} M_{n-1} \dots D_2 M_2 D_1 A,$$

откуда следует, что $A = L B_n$, $L = (D_n M_n D_{n-1} M_{n-1} \dots D_2 M_2 D_1)^{-1}$. Матрица L является нижней треугольной, и, таким образом, разложение (4) построено.

Процесс разложения матрицы A по методу ортогонализации является устойчивым. При этом, однако, свойство ортогональности системы векторов b_1, b_2, \dots, b_n может нарушаться, что приводит к накоплению вычислительной погрешности. Для устранения этого недостатка используются различные методы *переортогонализации*, например, видя

$$v_k^{(s+1)} = v_k^{(s)} - \sum_{i=1}^{k-1} (v_k^{(s)}, b_i) b_i,$$

$$s = 0, 1, \dots, \quad v_k^{(0)} = a_k, \quad v_k = \lim_{s \rightarrow \infty} v_k^{(s+1)}, \quad k = 1, 2, \dots, n+1.$$

Отметим, что при практической реализации метода в основном достаточно выполнения двух итераций.

§ 3. Метод прогонки решения СЛАУ с трехдиагональной матрицей

В предыдущих разделах мы изучили методы решения задачи

$$Ax = f, \quad (1)$$

для которых $Q(A) = O(n^3)$. В этом параграфе рассмотрим важнейший для практики случай системы уравнений с трехдиагональной матрицей (они часто встречаются при аппроксимации краевых задач для дифференциальных уравнений второго порядка). Предположение о ленточной структуре матрицы* позволяет нам построить метод решения системы (1) с $Q(A) = O(n)$, т. е. экономичный метод. Таким методом является *метод прогонки*, представляющий собой частный случай метода исключения Гаусса для систем уравнений указанного вида.

1. Алгоритм метода прогонки. В трехдиагональной матрице A ненулевые элементы расположены на главной и двух побочных диагоналях: $a_{ij} \neq 0$, $j = i - 1, i, i + 1$. Для упрощения записи этих элементов введем следующие обозначения: $a_{i,i-1} = -a_i$, $a_{ii} = c_i$, $a_{i,i+1} = -b_i$. Тогда систему (1) можно переписать в виде

$$\begin{cases} c_0 x_0 - b_0 x_1 = f_0, & i = 0, \\ -a_i x_{i-1} + c_i x_i - b_i x_{i+1} = f_i, & i = 1, 2, \dots, n-1, \\ -a_n x_{n-1} + c_n x_n = f_n, & i = n. \end{cases} \quad (2)$$

Заметим, что именно такая нумерация уравнений (начиная с $i = 0$) более естественна в приложениях.

Изложим формальную схему метода прогонки. Будем искать решение задачи (2) в виде

$$x_i = \alpha_{i+1} x_{i+1} + \beta_i, \quad i = n-1, n-2, \dots, 0, \quad (3)$$

где α_i, β_i — неизвестные пока коэффициенты. Подставляя (3) в (2), получим для $i = 1, 2, \dots, n-1$ уравнение

$$-a_i(\alpha_i x_i + \beta_i) + c_i x_i - b_i x_{i+1} = f_i,$$

или

$$x_i = \frac{b_i}{c_i - \alpha_i a_i} x_{i+1} + \frac{f_i + \beta_i a_i}{c_i - \alpha_i a_i}. \quad (4)$$

*Матрица A является ленточной, если ее элементы удовлетворяют соотношениям $a_{ij} = 0$, $i - j > p$, $j - i > q$, для некоторых неограниченных чисел p, q . Величина $p+q+1$ называется шириной ленты. В случае трехдиагональной матрицы $p = q = 1$.

Сравнивая (3) и (4), приходим к выводу, что

$$\alpha_{i+1} = \frac{b_i}{c_i - \alpha_i a_i}, \quad \beta_{i+1} = \frac{f_i + \beta_i a_i}{c_i - \alpha_i a_i}, \quad i = 1, 2, \dots, n-1. \quad (5)$$

Соотношения (5) представляют собой рекуррентные уравнения, для решения которых необходимо задать начальные значения α_1, β_1 . Из первого уравнения (2) имеем

$$x_0 = \frac{b_0}{c_0} x_1 + \frac{f_0}{c_0}.$$

Вместе с уравнением (3) при $i = 0$ это дает

$$\alpha_1 = \frac{b_0}{c_0}, \quad \beta_1 = \frac{f_0}{c_0}. \quad (6)$$

Нахождение коэффициентов α_i, β_i по формулам (5), (6) называется *прямой прогонкой*.

После того как прогоночные коэффициенты α_i, β_i , $i = 1, 2, \dots, n-1$, вычислены, решение системы (2) находится по рекуррентной формуле (3). Для начала расчета необходимо определить x_n . Из последнего уравнения (2) и уравнения (3) при $i = n-1$ имеем

$$-a_n x_{n-1} + c_n x_n = f_n, \quad x_{n-1} - \alpha_n x_n = \beta_n,$$

откуда следует, что

$$x_n = \frac{f_n + \beta_n a_n}{c_n - \alpha_n a_n} = \beta_{n+1}. \quad (7)$$

Вычисление неизвестных x_i по формулам (3), (7) называется *обратной прогонкой*.

Объединяя все формулы, окончательно получим алгоритм решения системы (2):

$$\alpha_{i+1} = \frac{b_i}{c_i - \alpha_i a_i}, \quad i = 1, 2, \dots, n-1, \quad \alpha_1 = \frac{b_0}{c_0}, \quad (8)$$

$$\beta_{i+1} = \frac{f_i + \beta_i a_i}{c_i - \alpha_i a_i}, \quad i = 1, 2, \dots, n, \quad \beta_1 = \frac{f_0}{c_0}, \quad (9)$$

$$x_i = \alpha_{i+1} x_{i+1} + \beta_{i+1}, \quad i = n-1, n-2, \dots, 0, \quad x_n = \beta_{n+1}. \quad (10)$$

Так как значения x_i находятся по направлению убывания индексов, то описанный метод иногда называют *методом правой прогонки*. Несложный подсчет числа арифметических действий в формулах (8)–(10) показывает, что для метода правой прогонки $Q(A) \approx 8n$.

2. Связь метода прогонки с методом Гаусса. Установим связь метода прогонки с методом Гаусса последовательного исключения неизвестных. Первое уравнение (2) с учетом (6) можно переписать в виде

$$x_0 = \alpha_1 x_1 + \beta_1. \quad (11)$$

С помощью этого уравнения исключим неизвестное x_0 из оставшихся уравнений системы (2). Но поскольку специфика матрицы A такова, что x_0 содержит только уравнение

$$-a_1 x_0 + c_1 x_1 - b_1 x_2 = f_1,$$

то только оно с помощью (11) приводится к виду

$$x_1 = \alpha_2 x_2 + \beta_2.$$

Точно так же и результат исключения из i -го уравнения (2)

$$x_{i-1} = \alpha_i x_i + \beta_i$$

приводит к уравнению

$$x_i = \alpha_{i+1} x_{i+1} + \beta_{i+1}.$$

Таким образом, после осуществления прямого хода метода прогонки (формулы (8), (9)) мы приходим к эквивалентной системе уравнений с верхней унитреугольной матрицей

$$Ux = \begin{pmatrix} 1 & -\alpha_1 & & & \\ & 1 & -\alpha_2 & & 0 \\ & & \ddots & \ddots & \\ 0 & & & 1 & -\alpha_n \\ & & & & 1 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \\ \beta_{n+1} \end{pmatrix}. \quad (12)$$

При этом для матрицы A задачи (2) система (12) в точности совпадает с системой (11) из § 1.

3. Обоснование метода прогонки. При изложении метода прогонки предполагалась возможность выполнения всех предписанных алгоритмом действий. Однако мы должны убедиться в том, что эти действия могут быть реально выполнены, т. е. отсутствует деление на нуль в формулах (5), (7). Кроме того, необходимо исследовать вопрос влияния погрешности в задании входных данных на полученное методом прогонки решение. Все это вместе взятое составляет обоснование метода прогонки.

Теорема 1. Пусть коэффициенты системы (2) удовлетворяют условиям

$$a_i \neq 0, \quad b_i \neq 0, \quad |c_i| \geq |a_i| + |b_i|, \quad i = 1, 2, \dots, n-1, \quad (13)$$

$$|c_0| \geq |b_0|, \quad |c_n| \geq |a_n|, \quad (14)$$

причем хотя бы в одном из неравенств (13) или (14) выполняется строгое неравенство. Тогда для алгоритма (8)–(10) метода правой прогонки имеют место неравенства

$$c_i - \alpha_i a_i \neq 0, \quad |\alpha_i| \leq 1, \quad i = 1, 2, \dots, n,$$

обеспечивающие корректность метода.

Доказательство. Доказательство теоремы проведем методом математической индукции. Согласно (6), (14), имеем

$$|\alpha_1| = \frac{|b_0|}{|c_0|} \leq 1.$$

Предположим, что $|\alpha_i| \leq 1$ для некоторого значения i , и покажем, что $|\alpha_{i+1}| \leq 1$. Из оценок

$$|c_i - \alpha_i a_i| \geq ||c_i| - |\alpha_i|||a_i|| \geq ||c_i| - |\alpha_i||$$

и условий (13) получим

$$|c_i - \alpha_i a_i| \geq |b_i| > 0, \quad i = 1, 2, \dots, n-1,$$

т. е. знаменатель выражений (5) не обращается в нуль. Кроме того,

$$|\alpha_{i+1}| = \frac{|b_i|}{|c_i - \alpha_i a_i|} \leq \frac{|b_i|}{|b_i|} = 1.$$

Следовательно, $|\alpha_i| \leq 1$, $i = 1, 2, \dots, n$.

Далее, пусть имеет место строгое неравенство в (14), например,

$$|c_n| > |a_n|.$$

В силу оценки $|\alpha_n| \leq 1$ получим

$$|c_n - \alpha_n a_n| \geq ||c_n| - |\alpha_n|||a_n|| > 0,$$

т. е. не обращается в нуль и знаменатель в выражении (7).

К аналогичному выводу можно прийти и в том случае, когда строгое неравенство имеет место в (13). Тогда из предположения $|\alpha_i| \leq 1$ следует

$$|c_i - \alpha_i a_i| \geq ||c_i| - |\alpha_i|| > |b_i|, \quad |\alpha_{i+1}| < 1,$$

т. е. $|\alpha_i| < 1$, $i = 2, 3, \dots, n$. При этом

$$|c_n - \alpha_n a_n| \geq ||c_n| - |\alpha_n|||a_n|| > 0.$$

Наконец, если выполняется условие $|c_0| > |b_0|$, то $|\alpha_1| < 1$ и по индукции $|\alpha_i| < 1$, $i = 1, 2, \dots, n$, $c_n - \alpha_n a_n \neq 0$.

Таким образом, при выполнении условий (13), (14) задача (2) разрешима. Осталось показать устойчивость расчета по рекуррентным формулам (10).

Пусть в (10) при $i = i_0 + 1$ вместо x_{i_0+1} вычислена величина $\tilde{x}_{i_0+1} = x_{i_0+1} + \delta_{i_0+1}$. Тогда на следующем шаге вычислений, т. е. при $i = i_0$, вместо

$$x_{i_0} = \alpha_{i_0+1} x_{i_0+1} + \beta_{i_0+1}$$

получим величину

$$\tilde{x}_{i_0} = \alpha_{i_0+1} (x_{i_0+1} + \delta_{i_0+1}) + \beta_{i_0+1},$$

и погрешность окажется равной

$$\delta_{i_0} = \tilde{x}_{i_0} - x_{i_0} = \alpha_{i_0+1} \delta_{i_0+1},$$

откуда имеем

$$|\delta_{i_0}| \leq |\alpha_{i_0+1}| |\delta_{i_0+1}| \leq |\delta_{i_0+1}|,$$

т. е. погрешность при переходе к следующему шагу не возрастает.

Теорема доказана.

Заметим, что условия $\Delta_i = c_i - \alpha_i a_i \neq 0$, $i = 1, 2, \dots, n$, гарантируют отличие от нуля определителя системы (2), т. е. единственность решения исходной задачи. Действительно, как показано выше, система уравнений $Ax = f$ в методе прогонки приводится к задаче (12) с верхней унитреугольной матрицей U . Поэтому имеет место факторизованное представление $A = LU$, где

$$L = \begin{pmatrix} c_0 & & & & 0 \\ -a_1 & \Delta_1 & & & \\ & -a_2 & \Delta_2 & & \\ & & \ddots & \ddots & \\ 0 & & & -a_n & \Delta_n \end{pmatrix}.$$

Так как

$$\det A = \det L \cdot \det U = \det L = c_0 \prod_{i=1}^n \Delta_i,$$

то в силу теоремы 1 $c_0 \neq 0$, $\Delta_i \neq 0$, $i = 1, 2, \dots, n$, и $\det A \neq 0$. Поэтому в случае выполнения условий теоремы 1 система (2) имеет единственное решение, которое может быть найдено с помощью метода прогонки (8)–(10).

4. Методы левой и встречной прогонки. В п. 1 были получены формулы правой прогонки для решения системы (2). Аналогично выводятся формулы левой прогонки:

$$\xi_i = \frac{a_i}{c_i - \xi_{i+1} b_i}, \quad i = n-1, n-2, \dots, 1, \quad \xi_n = \frac{a_n}{c_n};$$

$$\eta_i = \frac{f_i + \eta_{i+1} b_i}{c_i - \xi_{i+1} b_i}, \quad i = n-1, n-2, \dots, 0, \quad \eta_n = \frac{f_n}{c_n};$$

$$x_{i+1} = \xi_{i+1} x_i + \eta_{i+1}, \quad i = 0, 1, \dots, n-1, \quad x_0 = \eta_0.$$

Здесь значение x_i находится в направлении возрастания индексов.

Комбинируя правую и левую прогонки, получим так называемый *метод встречных прогонок*:

$$\alpha_{i+1} = \frac{b_i}{c_i - \alpha_i a_i}, \quad i = 1, 2, \dots, i_0, \quad \alpha_1 = \frac{b_0}{c_0};$$

Для вычисления элементов матриц A_k удобно использовать следующие расчетные формулы:

$$\begin{aligned} a_{ij}^{(1)} &= a_{ij}, \quad i, j = 1, 2, \dots, n; \\ b_{i,k+1}^{(k)} &= \frac{a_{i,k+1}^{(k)}}{a_{k,k+1}^{(k)}}, \quad i = 1, 2, \dots, n, \quad b_{ij}^{(k)} = a_{ij}^{(k)} - a_{kj}^{(k)} b_{i,k+1}^{(k)}, \\ i, j &= 1, 2, \dots, n, \quad j \neq k+1, \quad k = 1, 2, \dots, n-1, \\ a_{ij}^{(k+1)} &= b_{ij}^{(k)}, \quad i = 1, 2, \dots, n, \quad i \neq k+1, \quad j = 1, 2, \dots, n, \\ a_{k+1,j}^{(k+1)} &= \sum_{i=1}^n a_{ki}^{(k)} b_{ij}^{(k)}, \quad j = 1, 2, \dots, n, \quad k = 1, 2, \dots, n-1. \end{aligned}$$

Рассмотрим теперь варианты вырождения метода. Пусть на k -м этапе получена матрица A_{k+1} , у которой элемент $a_{k+1,k+2}^{(k+1)} = 0$. Это означает, что матрица A_{k+2} не может быть построена. В дальнейшем реализуется одна из двух возможностей.

Предположим, что в $(k+1)$ -й строке хотя бы один из элементов, расположенных правее $a_{k+1,k+2}^{(k+1)}$, отличен от нуля, т.е. $a_{k+1,j}^{(k+1)} \neq 0$, $j > k+2$. В этом случае с помощью подобного преобразования $P_{k+2,j} A_k P_{k+2,j}$, позволяющего переставить $(k+2)$ -ю и j -ю столбцы и одновременно $(k+2)$ -ю и j -ю строки, задача сводится к невырожденному случаю. С целью улучшения вычислительной устойчивости алгоритма аналогичную процедуру можно использовать на каждом этапе приведения матрицы A к канонической форме Фробениуса.

Если же $a_{k+1,k+2}^{(k+1)} = 0$, $j > k+1$, то матрица A_{k+1} имеет вид

$$A_{k+1} = \begin{pmatrix} 0 & 1 & & & & & 0 \\ \vdots & \vdots & \ddots & & & & \\ 0 & 0 & \dots & 1 & & & \\ a_{k+1,1}^{(k+1)} & a_{k+1,2}^{(k+1)} & \dots & a_{k+1,k+1}^{(k+1)} & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ a_{n,1}^{(k+1)} & a_{n,2}^{(k+1)} & \dots & a_{n,k+1}^{(k+1)} & a_{n,k+2}^{(k+1)} & \dots & a_{nn}^{(k+1)} \end{pmatrix},$$

или в блочной записи

$$\begin{aligned} A_{k+1} &= \begin{pmatrix} \Phi_{k+1} & 0 \\ B & C \end{pmatrix}, \quad \Phi_{k+1} = \begin{pmatrix} 0 & 1 & & & & 0 \\ \vdots & \vdots & \ddots & & & \\ 0 & 0 & \dots & 1 & & \\ a_{k+1,1}^{(k+1)} & a_{k+1,2}^{(k+1)} & \dots & a_{k+1,n}^{(k+1)} & & \\ \vdots & \vdots & & \vdots & & \\ a_{n,k+2}^{(k+1)} & a_{n,k+3}^{(k+1)} & \dots & a_{nn}^{(k+1)} & & \end{pmatrix}, \\ C &= \begin{pmatrix} a_{k+2,k+2}^{(k+1)} & a_{k+2,k+3}^{(k+1)} & \dots & a_{k+2,n}^{(k+1)} \\ a_{k+3,k+2}^{(k+1)} & a_{k+3,k+3}^{(k+1)} & \dots & a_{k+3,n}^{(k+1)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,k+2}^{(k+1)} & a_{n,k+3}^{(k+1)} & \dots & a_{nn}^{(k+1)} \end{pmatrix}. \end{aligned}$$

Подобная ситуация возникает в том случае, когда исходная матрица A является неполной*. Характеристический многочлен матрицы A_{k+1} , очевидно, равен произведению характеристических многочленов матриц Φ_{k+1} и C . Матрица Φ_{k+1} уже является матрицей Фробениуса и коэффициенты ее характеристического многочлена выписываются по элементам последней строки. Поэтому дальнейшие преобразования связанны с приведением к канонической форме Фробениуса только матрицы C .

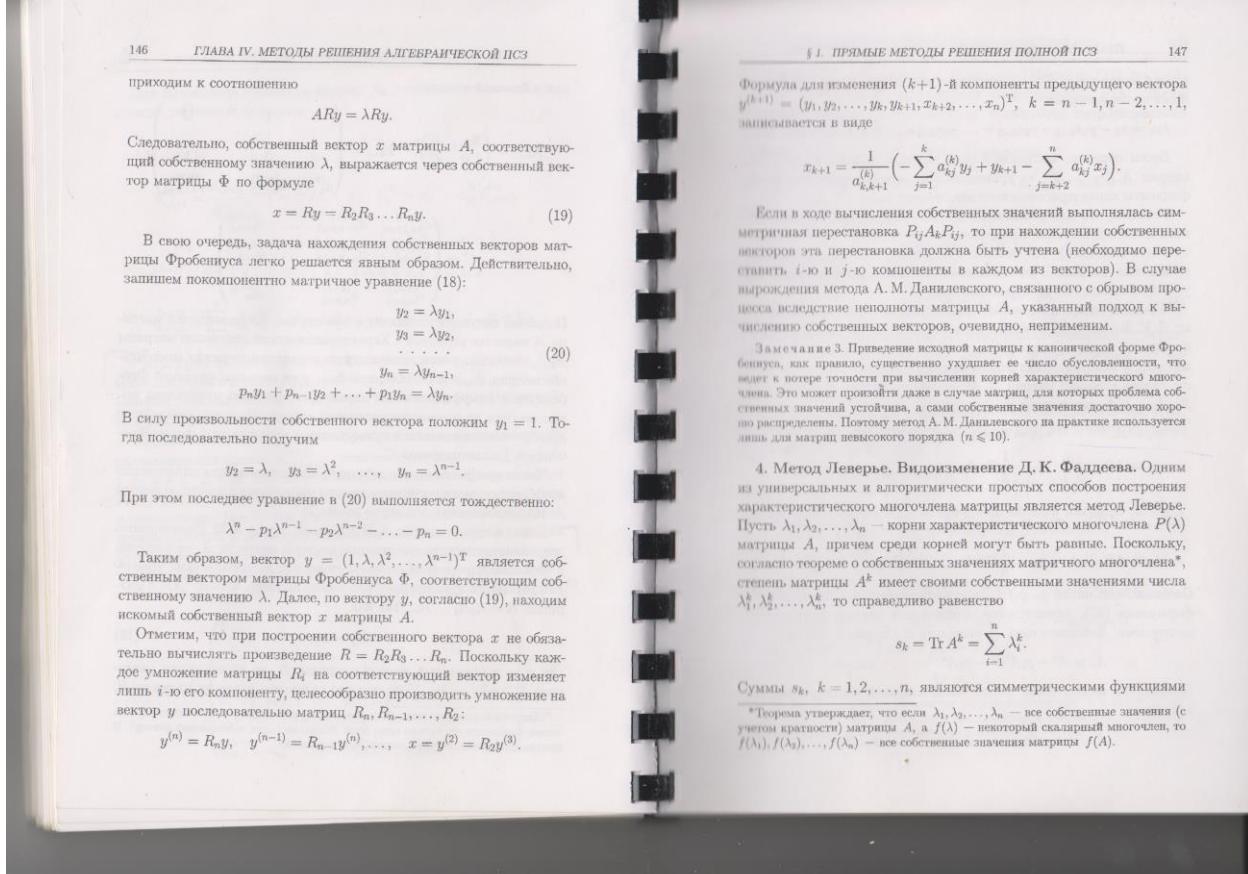
Число арифметических действий, необходимых для определения коэффициентов характеристического многочлена с помощью метода А. М. Данилевского, примерно равно $2n^3$.

Как и в случае метода А. Н. Крылова, метод А. М. Данилевского позволяет вычислять собственные векторы матрицы A , не прибегая к решению однородной системы линейных алгебраических уравнений (4). Пусть имеется преобразование подобия $\Phi = R^{-1}AR$ с матрицей $R = R_2R_3\dots R_n$. Так как

$$Ry = \lambda y, \quad (18)$$

где y — собственный вектор матрицы Φ , то выполняется равенство $R^{-1}ARy = \lambda y$. Умножая это равенство слева на матрицу R ,

*Матрица называется полной, если каждому различному ее собственному значению соответствует только один блок Жордана (один собственный вектор). В противном случае матрица называется неполной.



приходим к соотношению

$$ARy = \lambda Ry.$$

Следовательно, собственный вектор x матрицы A , соответствующий собственному значению λ , выражается через собственный вектор матрицы Φ по формуле

$$x = Ry = R_2R_3\dots R_n y. \quad (19)$$

В свою очередь, задача нахождения собственных векторов матрицы Фробениуса легко решается явным образом. Действительно, запишем покомпонентно матричное уравнение (18):

$$\begin{aligned} y_2 &= \lambda y_1, \\ y_3 &= \lambda y_2, \\ &\dots \\ y_n &= \lambda y_{n-1}, \\ p_n y_1 + p_{n-1} y_2 + \dots + p_1 y_n &= \lambda y_n. \end{aligned} \quad (20)$$

В силу произвольности собственного вектора положим $y_1 = 1$. Тогда последовательно получим

$$y_2 = \lambda, \quad y_3 = \lambda^2, \quad \dots, \quad y_n = \lambda^{n-1}.$$

При этом последнее уравнение в (20) выполняется тождественно:

$$\lambda^n - p_1 \lambda^{n-1} - p_2 \lambda^{n-2} - \dots - p_n = 0.$$

Таким образом, вектор $y = (1, \lambda, \lambda^2, \dots, \lambda^{n-1})^T$ является собственным вектором матрицы Фробениуса Φ , соответствующим собственному значению λ . Далее, по вектору y , согласно (19), находим искомый собственный вектор x матрицы A .

Отметим, что при построении собственного вектора x не обязательно вычислять произведение $R = R_2R_3\dots R_n$. Поскольку каждое умножение матрицы R_i на соответствующий вектор изменяет лишь i -ю его компоненту, целесообразно производить умножение на вектор y последовательно матриц R_n, R_{n-1}, \dots, R_2 :

$$y^{(n)} = R_n y, \quad y^{(n-1)} = R_{n-1} y^{(n)}, \dots, \quad x = y^{(2)} = R_2 y^{(3)}.$$

Формула для изменения $(k+1)$ -й компоненты предыдущего вектора $y^{(k+1)} = (y_1, y_2, \dots, y_k, y_{k+1}, x_{k+2}, \dots, x_n)^T$, $k = n-1, n-2, \dots, 1$, записывается в виде

$$x_{k+1} = \frac{1}{a_{k,k+1}^{(k)}} \left(-\sum_{j=1}^k a_{kj}^{(k)} y_j + y_{k+1} - \sum_{j=k+2}^n a_{kj}^{(k)} x_j \right).$$

Если в ходе вычисления собственных значений выполнялась симметричная перестановка $P_{ij} A_k P_{ij}^T$, то при нахождении собственных векторов эта перестановка должна быть учтена (необходимо переставить i -ю и j -ю компоненты в каждом из векторов). В случае вырождения метода А. М. Данилевского, связанного с обратным процессом вследствие неполноты матрицы A , указанный подход к вычислению собственных векторов, очевидно, неприменим.

З а м е ч а н и е 3. Приведение исходной матрицы к канонической форме Фробениуса, как правило, существенно ухудшает ее число обусловленности, что ведет к потере точности при вычислении корней характеристического многочлена. Это может произойти даже в случае матриц, для которых проблема собственных значений устойчива, а сами собственные значения достаточно хорошо распределены. Поэтому метод А. М. Данилевского на практике используется лишь для матриц невысокого порядка ($n \leq 10$).

4. Метод Леверье. Видоизменение Д. К. Фаддеева. Одним из универсальных и алгоритмически простых способов построения характеристического многочлена матрицы является метод Леверье. Пусть $\lambda_1, \lambda_2, \dots, \lambda_n$ — корни характеристического многочлена $P(\lambda)$ матрицы A , причем среди корней могут быть равные. Поскольку, согласно теореме о собственных значениях матричного многочлена*, степень матрицы A^k имеет своим собственными значениями числа $\lambda_1^k, \lambda_2^k, \dots, \lambda_n^k$, то справедливо равенство

$$s_k = \text{Tr } A^k = \sum_{i=1}^n \lambda_i^k.$$

Суммы s_k , $k = 1, 2, \dots, n$, являются симметрическими функциями

*Теорема утверждает, что если $\lambda_1, \lambda_2, \dots, \lambda_n$ — все собственные значения (с учетом кратности) матрицы A , а $f(\lambda)$ — некоторый скалярный многочлен, то $f(\lambda_1), f(\lambda_2), \dots, f(\lambda_n)$ — все собственные значения матрицы $f(A)$.

корней многочлена $P(\lambda)$ и связаны с его коэффициентами известными соотношениями

$$kp_k = s_k - p_1 s_{k-1} - p_2 s_{k-2} - \dots - p_{k-1} s_1, \quad k = 1, 2, \dots, n. \quad (21)$$

Таким образом, последовательно вычисляя следы s_1, s_2, \dots, s_n матриц A, A^2, \dots, A^n , из уравнений (21) можно определить коэффициенты характеристического многочлена p_1, p_2, \dots, p_n .

Отметим, что для реализации метода Леверье требуется выполнить довольно большой объем вычислений (примерно $2n^4$ арифметических действий), что связано с необходимостью многократного умножения матриц. Этот недостаток метода компенсируется, как уже отмечалось, его универсальностью.

Перейдем теперь к видоизменению метода Леверье, предложеному Д. К. Фаддеевым, которое позволяет, помимо вычисления коэффициентов характеристического многочлена, определять обратную матрицу и находить собственные векторы исходной матрицы. Вместо следов степеней матриц A, A^2, \dots, A^n вычислим следы матриц A_1, A_2, \dots, A_n по следующим формулам:

$$\begin{aligned} A_1 &= A, & p_1 &= \text{Tr } A_1, & B_1 &= A_1 - p_1 E, \\ A_2 &= AB_1, & p_2 &= \frac{1}{2} \text{Tr } A_2, & B_2 &= A_2 - p_2 E, \\ &\dots & &\dots & &\dots \\ A_{n-1} &= AB_{n-2}, & p_{n-1} &= \frac{1}{n-1} \text{Tr } A_{n-1}, & B_{n-1} &= A_{n-1} - p_{n-1} E, \\ A_n &= AB_{n-1}, & p_n &= \frac{1}{n} \text{Tr } A_n, & B_n &= A_n - p_n E = 0. \end{aligned} \quad (22)$$

Покажем, что числа p_1, p_2, \dots, p_n , последовательно определяемые формулами (22), являются коэффициентами характеристического многочлена. Действительно, по построению алгоритма имеем

$$A_k = A^k - p_1 A^{k-1} - p_2 A^{k-2} - \dots - p_{k-1} A.$$

Приведем между собой следы левой и правой части последнего равенства:

$$kp_k = s_k - p_1 s_{k-1} - p_2 s_{k-2} - \dots - p_{k-1} s_1.$$

Но эти формулы совпадают с формулами Ньютона (21), по которым последовательно определяются коэффициенты $P(\lambda)$. Таким образом, числа p_k , $k = 1, 2, \dots, n$, и являются коэффициентами характеристического многочлена.

Далее, в силу теоремы Кели – Гамильтона

$$B_n = A_n - p_n E = A^n - p_1 A^{n-1} - \dots - p_{n-1} A - p_n E = 0. \quad (23)$$

Из равенства (23) получим соотношение

$$A_n = AB_{n-1} = p_n E,$$

откуда имеем

$$A^{-1} = \frac{1}{p_n} B_{n-1}. \quad (24)$$

Формула (24) используется для получения матрицы A^{-1} , обратной исходной матрице A . Если матрица A вырожденная, то метод Д. К. Фаддеева позволяет получить присоединенную матрицу* $B = (-1)^{n-1} B_{n-1}$.

Наконец, рассмотрим вопрос построения собственных векторов. Пусть все собственные значения матрицы A различны. Докажем, что любой столбец матрицы

$$Q_i = \lambda_i^{n-1} E + \lambda_i^{n-2} B_1 + \dots + \lambda_i B_{n-2} + B_{n-1}$$

может принять в качестве собственного вектора матрицы A , соответствующего собственному значению λ_i . В самом деле

$$\begin{aligned} (\lambda_i E - A) Q_i &= (\lambda_i E - A)(\lambda_i^{n-1} E + \lambda_i^{n-2} B_1 + \dots + \lambda_i B_{n-2} + B_{n-1}) = \\ &= \lambda_i^n E + \lambda_i^{n-1}(B_1 - A) + \lambda_i^{n-2}(B_2 - AB_1) + \dots + \lambda_i(B_{n-1} - AB_{n-2}) - \\ &- AB_{n-1} = (\lambda_i^n - p_1 \lambda_i^{n-1} - p_2 \lambda_i^{n-2} - \dots - p_n) E = 0. \end{aligned}$$

Отсюда следует, что $(\lambda_i E - A)x = 0$ или

$$Ax = \lambda_i x,$$

*Матрица $B = \{b_{ij}\}_{i,j=1}^n$ называется присоединенной для $A = \{a_{ij}\}_{i,j=1}^n$, если $b_{ij} = A_{ji}$, где A_{ji} – алгебраическое дополнение элемента a_{ij} . Обратная матрица связана с присоединенной соотношением $A^{-1} = (\det A)^{-1} B$.

где x – любой столбец матрицы Q_i . Последнее равенство и доказывает сделанное утверждение.

Очевидно, что при нахождении собственных векторов указанным способом нет необходимости вычислять все столбцы матрицы Q_i . Достаточно для каждого λ_i , $i = 1, 2, \dots, n$, вычислить лишь один столбец, воспользовавшись рекуррентной формулой

$$x^{(k)} = \lambda_i x^{(k-1)} + b_j^{(k)}, \quad k = 1, 2, \dots, n-1, \quad x_0 = e_j,$$

где $b_j^{(k)}$ – выбранный столбец матрицы B_k , e_j – соответствующий столбец единичной матрицы. Тогда собственный вектор, соответствующий собственному значению λ_i , равен $x = x^{(n-1)}$.

При наличии кратных собственных значений процедура вычисления собственных векторов несколько усложняется. В этом случае наряду с матрицей Q_i необходимо привлекать к рассмотрению также матрицы, полученные дифференцированием Q_i по λ .

5. Применение прямых методов к решению полной ПСЗ.

1. Рассмотрим (3×3) -матрицу:

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

Определим собственные значения и соответствующие им собственные векторы этой матрицы методом А. Н. Крылова. Сначала по вектору $b_0 = (1, 0, 0)^T$ последовательно находим

$$b_1 = Ab_0 = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}, \quad b_2 = Ab_1 = \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}, \quad b_3 = Ab_2 = \begin{pmatrix} 3 \\ 1 \\ 4 \end{pmatrix}.$$

Далее запишем систему для вычисления коэффициентов линейной комбинации $b_3 = q_1 b_2 + q_2 b_1 + q_3 b_0$:

$$\begin{aligned} q_3 + q_2 + q_1 &= 3, \\ q_2 &= 1, \\ 2q_2 + q_1 &= 4. \end{aligned}$$

Решая ее, находим $q_1 = 2$, $q_2 = 1$, $q_3 = -2$. Следовательно, характеристический многочлен равен $P(\lambda) = -\lambda^3 + 2\lambda^2 + \lambda - 2$, а его корни $\lambda_1 = -1$, $\lambda_2 = 1$, $\lambda_3 = 2$.

Используем формулу (16) для вычисления коэффициентов β_{ij} :

$$\beta_{11} = \beta_{21} = \beta_{31} = 1, \quad \beta_{12} = -3, \quad \beta_{13} = 2,$$

$$\beta_{22} = -1, \quad \beta_{23} = -2, \quad \beta_{32} = 0, \quad \beta_{33} = -1.$$

Окончательно по формуле (14) находим собственные векторы

$$x_1 = (-1, 3, 5)^T, \quad x_2 = (1, 1, 1)^T, \quad x_3 = (1, 0, 1)^T.$$

2. В качестве второго примера рассмотрим матрицу

$$A = \begin{pmatrix} 3 & 4 & -6 \\ 1 & 2 & -3 \\ 1 & 3 & -4 \end{pmatrix}$$

и найдем ее собственные значения с помощью метода А. Н. Крылова. Зададим начальный вектор $b_0 = (1, 0, 0)^T$ и вычислим b_1, b_2, b_3 . В результате получим систему линейных алгебраических уравнений

$$q_3 + 3q_2 + 7q_1 = 17,$$

$$q_2 + 2q_1 = 5,$$

$$q_2 + 2q_1 = 5,$$

в которой второе и третье уравнения совпадают, т. е. имеет место вырожденный случай. Соответствующая укороченная система

$$q_2 + 3q_1 = 7,$$

$$q_1 = 2,$$

имеет решение $q_1 = 2$, $q_2 = 1$. Тем самым мы определяем коэффициенты многочлена второй степени $Q(\lambda) = \lambda^2 - 2\lambda - 1$, являющегося делителем характеристического многочлена $P(\lambda)$. Это позволяет найти лишь два собственных значения матрицы $\lambda_{1,2} = 1 \pm \sqrt{2}$.

Для определения собственного значения λ_3 изменим начальный вектор b_0 , полагая $b_0 = (1, 1, 0)^T$. Аналогичные предыдущему вычисления приводят к системе

$$q_3 + 7q_2 + 9q_1 = 31,$$

$$q_3 + 3q_2 + q_1 = 11,$$

$$4q_2 = 12,$$

решая которую методом Гаусса, находим $q_1 = 1$, $q_2 = 3$, $q_3 = 1$.

Таким образом, характеристический многочлен матрицы A имеет вид $P(\lambda) = -\lambda^3 + \lambda^2 + 3\lambda + 1$. Разделив его на многочлен $Q(\lambda)$, получим $Q_1(\lambda) = -\lambda - 1$. Следовательно, недостающее собственное значение равно $\lambda_3 = -1$.

3. Пусть требуется решить полную ПСЗ для матрицы

$$A = \begin{pmatrix} 3 & -2 & -1 \\ 3 & -4 & -3 \\ 2 & -4 & 0 \end{pmatrix},$$

используя метод А. Н. Крылова. Здесь линейно независимы только два первых вектора: $b_0 = (1, 0, 0)^T$ и $b_1 = (3, 3, 2)^T$, а вектор $b_2 = (1, -9, -6)^T$ является их линейной комбинацией. Решением соответствующей укороченной СЛАУ служат числа $q_1 = -3$, $q_2 = 10$. Следовательно, приходим к минимальному анулирующему вектору b_0 многочлену $Q(\lambda) = \lambda^2 + 3\lambda - 10$, который, в отличие от предыдущего примера, является минимальным многочленом матрицы A . Его собственные значения $\lambda_1 = -5$, $\lambda_2 = 2$ совпадают (без учета кратности) с корнями многочлена $P(\lambda)$. Соответствующие им собственные векторы определяются аналогично невырожденному случаю: $x_1 = (1, 3, 2)^T$, $x_2 = (8, 3, 2)^T$.

Наконец, из условия $\lambda_1 + \lambda_2 + \lambda_3 = \text{Tr } A$ получим $\lambda_3 = 2$. Задавая другой начальный вектор, например $b_0 = (1, 1, 0)^T$, с помощью процесса А. Н. Крылова находим еще один линейно независимый собственный вектор $x_3 = (3, 2, -1)^T$, соответствующий кратному собственному значению $\lambda = 2$.

4. Методом А. М. Данилевского вычислим собственные значения и собственные векторы матрицы

$$A = A_1 = \begin{pmatrix} -2 & 0 & 3 \\ 0 & -2 & 3 \\ -2 & -2 & 5 \end{pmatrix}.$$

Как видим, уже на первом этапе имеет место ситуация вырождения: $a_{12}^{(1)} = 0$. Поскольку элемент $a_{13}^{(1)} \neq 0$, можно применить преобразо-

вание подобия

$$P_{23}A_1P_{23} = \begin{pmatrix} -2 & 3 & 0 \\ -2 & 5 & -2 \\ 0 & 3 & -2 \end{pmatrix}, \quad P_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

с помощью которого меняются местами второй и третий столбец и одновременно вторая и третья строки матрицы A_1 . Далее последовательно вычислим матрицы

$$R_2 = \begin{pmatrix} 1 & 0 & 0 \\ 2/3 & 1/3 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad R_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$A_2 = R_2^{-1}P_{23}A_1P_{23}R_2 = \begin{pmatrix} 0 & 1 & 0 \\ 4 & 3 & -6 \\ 2 & 1 & -2 \end{pmatrix}.$$

На втором этапе по аналогичному правилу образуем матрицы

$$R_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2/3 & 1/2 & -1/6 \end{pmatrix}, \quad R_3^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 4 & 3 & -6 \end{pmatrix},$$

$$A_3 = R_3^{-1}R_2^{-1}P_{23}A_1P_{23}R_2R_3 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -4 & 4 & 1 \end{pmatrix}.$$

Матрица A_3 представляет собой матрицу Фробениуса, которая получена с помощью трех последовательных преобразований подобия. Таким образом, коэффициенты характеристического многочлена матрицы A определяются как элементы последней строки матрицы A_3 . Имеем

$$P(\lambda) = -\lambda^3 + \lambda^2 + 4\lambda - 4; \quad \lambda_1 = -2, \quad \lambda_2 = 1, \quad \lambda_3 = 2.$$

Определим собственный вектор, соответствующий собственному значению $\lambda = -2$. Сначала вычислим собственный вектор матрицы Фробениуса A_3 :

$$y = (1, \lambda, \lambda^2)^T = (1, -2, 4)^T.$$

Затем по вектору y последовательно находим

$$y^{(3)} = R_3y = \begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}, \quad y^{(2)} = R_2y^{(3)} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad x = P_{23}y^{(2)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Аналогично можно получить остальные собственные векторы:

$$\lambda_1 = 1, \quad x = (1, 1, 1)^T; \quad \lambda_2 = 2, \quad x = (1, 1, 4/3)^T.$$

5. Рассмотрим матрицу из примера 3:

$$A = A_1 = \begin{pmatrix} 3 & -2 & -1 \\ 3 & -4 & -3 \\ 2 & -4 & 0 \end{pmatrix}$$

и вычислим ее собственные значения с помощью метода А. М. Данилевского. Имеем

$$R_2 = \begin{pmatrix} 1 & 0 & 0 \\ 3/2 & -1/2 & -1/2 \\ 0 & 0 & 1 \end{pmatrix}, \quad R_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & -2 & -1 \\ 0 & 0 & 1 \end{pmatrix},$$

$$A_2 = R_2^{-1}A_1R_2 = \begin{pmatrix} 0 & 1 & 0 \\ 10 & -3 & 0 \\ -4 & 2 & 2 \end{pmatrix} = \begin{pmatrix} \Phi_2 & 0 \\ B & C \end{pmatrix},$$

где

$$\Phi_2 = \begin{pmatrix} 0 & 1 \\ 10 & -3 \end{pmatrix}, \quad 0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad B = (-4, 2), \quad C = (2).$$

В этом случае характеристический многочлен записывается в факторизованном виде:

$$P(\lambda) = \det(\Phi_2 - \lambda E) \cdot (2 - \lambda) = (\lambda^2 + 3\lambda - 10)(2 - \lambda),$$

откуда следует, что $\lambda_1 = -5$, $\lambda_{2,3} = 2$.

6. Вычислим собственные значения и соответствующие им собственные векторы методом Д. К. Фаддеева для матрицы

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

Последовательно применим формулы (22):

$$p_1 = \text{Tr } A = 1, \quad B_1 = A_1 - p_1E = \begin{pmatrix} 0 & -2 & -1 \\ -1 & 0 & 1 \\ 1 & 0 & -2 \end{pmatrix},$$

$$A_2 = AB_1 = \begin{pmatrix} 1 & -2 & -1 \\ 0 & 2 & 0 \\ -1 & -2 & 1 \end{pmatrix}, \quad p_2 = \frac{1}{2} \text{Tr } A_2 = 2,$$

$$B_2 = A_2 - 2E = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix}, \quad A_3 = 0, \quad p_3 = 0, \quad B_3 = 0,$$

$$P(\lambda) = -\lambda^3 + \lambda^2 + 2\lambda; \quad \lambda_1 = -1, \quad \lambda_2 = 0, \quad \lambda_3 = 2.$$

Поскольку $\det A = 0$, то процесс Д. К. Фаддеева дает присоединенную для A матрицу $B = B_2$.

Определим собственный вектор, соответствующий собственному значению $\lambda_1 = -1$. С этой целью построим матрицу

$$Q_1 = \lambda_1^2 E + \lambda_1 B_1 + B_2 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & -1 \\ -2 & -2 & 2 \end{pmatrix}.$$

Следовательно, $x_1 = (0, 1, -2)^T$. Аналогично имеем

$$Q_2 = \lambda_2^2 E + \lambda_2 B_1 + B_2 = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix}, \quad x_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix},$$

$$Q_3 = \lambda_3^2 E + \lambda_3 B_1 + B_2 = \begin{pmatrix} 3 & -6 & -3 \\ -2 & 4 & 2 \\ 1 & -2 & -1 \end{pmatrix}, \quad x_3 = \begin{pmatrix} 3 \\ -2 \\ 1 \end{pmatrix}.$$

6. Методы, использующие другие канонические формы.

В п. 3 мы рассмотрели метод А. М. Данилевского построения характеристического многочлена $P(\lambda)$ матрицы A . В результате матрица была преобразована к канонической форме Фробениуса, по виду которой выписывались коэффициенты $P(\lambda)$. В то же время большая

группа методов решения полной проблемы собственных значений основана на приведении матрицы A подобными ортогональными преобразованиями к почти треугольному виду, а в случае $A = A^T$ к трехдиагональному виду. Для таких матриц существует способ быстрого вычисления определителя $\det(A - \lambda E)$ при фиксированных значениях λ без нахождения явного выражения характеристического многочлена.

Пусть матрица A является верхней почти треугольной: $a_{ij} = 0$ при $i > j + 1$. Вычисление $\det A$ удобно проводить методом исключения Гаусса. С учетом структуры матрицы A каждый этап исключения сводится к преобразованию только двух строк. Поэтому расчетные формулы принимают вид (см. гл. II, § 1)

$$c_{kj} = \frac{a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}}, \quad a_{k+1,j}^{(k)} = a_{k+1,j}^{(k-1)} - a_{k+1,k}^{(k-1)}c_{kj}, \quad i = k+1, k+2, \dots, n,$$

$$k = 1, 2, \dots, n-1, \quad \det A = a_{11}a_{22}^{(1)} \dots a_{nn}^{(n-1)}.$$

Поскольку нас интересует определитель $\det(A - \lambda E)$, то для его вычисления достаточно в указанных формулах заменить элементы a_{ii} на $a_{ii} - \lambda$. Этот способ позволяет вычислить определитель за n^2 арифметических действий.

Еще проще может быть вычислено значение характеристического многочлена при фиксированном λ для трехдиагональной матрицы. Обозначим через $D_k(\lambda)$ главный минор k -го порядка матрицы $A - \lambda E$. Последняя строка такого минора содержит лишь два ненулевых элемента $a_{k,k-1}$ и $a_{kk} - \lambda$. Проведя разложение по элементам последней строки, получим

$$D_k(\lambda) = \begin{vmatrix} D_{k-2}(\lambda) & 0 & 0 \\ \vdots & \vdots & \vdots \\ a_{k-2,k-1} & 0 & 0 \\ \hline 0 & \dots & a_{k-1,k-2} & a_{k-1,k-1} - \lambda & a_{k-1,k} \\ 0 & \dots & 0 & a_{k,k-1} & a_{kk} - \lambda \end{vmatrix} =$$

$$= (a_{kk} - \lambda)D_{k-1}(\lambda) - a_{k,k-1}B_{k,k-1}(\lambda),$$

где $B_{k,k-1}(\lambda)$ — минор, дополняющий элемент $a_{k,k-1}$. Но у $B_{k,k-1}(\lambda)$ в последнем столбце содержится только один ненулевой элемент $a_{k-1,k}$. Поэтому справедливо равенство

$$B_{k,k-1}(\lambda) = a_{k-1,k}D_{k-2}(\lambda).$$

Тем самым имеем рекуррентное соотношение для вычисления миноров k -го порядка матрицы $A - \lambda E$:

$$D_k(\lambda) = (a_{kk} - \lambda)D_{k-1}(\lambda) - a_{k,k-1}a_{k-1,k}D_{k-2}(\lambda). \quad (25)$$

Для начала расчета достаточно положить

$$D_0(\lambda) = 1, \quad D_1(\lambda) = a_{11} - \lambda.$$

Вычисление определителя по формуле (25) требует всего $5n$ арифметических действий.

Далее, утром, что многочлен n -й степени однозначно определяется своими значениями в $n+1$ точках $\lambda_1^*, \lambda_2^*, \dots, \lambda_{n+1}^*$. Поэтому для вычисления коэффициентов p_1, p_2, \dots, p_n характеристического многочлена получим следующую СЛАУ:

$$\begin{aligned} (-1)^n((\lambda_1^*)^n - p_1(\lambda_1^*)^{n-1} - p_2(\lambda_1^*)^{n-2} - \dots - p_n) &= D(\lambda_1^*), \\ (-1)^n((\lambda_2^*)^n - p_1(\lambda_2^*)^{n-1} - p_2(\lambda_2^*)^{n-2} - \dots - p_n) &= D(\lambda_2^*), \\ \dots \\ (-1)^n((\lambda_n^*)^n - p_1(\lambda_n^*)^{n-1} - p_2(\lambda_n^*)^{n-2} - \dots - p_n) &= D(\lambda_n^*). \end{aligned} \quad (26)$$

Для различных $\lambda_i^*, i = 1, 2, \dots, n$, определитель системы (26) отличен от нуля, следовательно, коэффициенты характеристического многочлена $P(\lambda)$ находятся единственным образом. Задачу решения системы (26) иногда заменяют задачей построения интерполяционного многочлена по значениям $(\lambda_i^*, D(\lambda_i^*))$, который в силу единственности совпадает с характеристическим многочленом.

Описанный метод получил название *метода интерполяции*. Вычисление коэффициентов характеристического многочлена методом интерполяции требует примерно $(2/3)n^4$ арифметических действий, поэтому на практике его применяют лишь при $n \leq 10$.

Перейдем теперь к простейшим алгоритмам приведения матрицы A подобными преобразованиями к верхнему треугольному (трехдиагональному) виду. Рассмотрим сначала *метод отражений*, являющийся одним из самых эффективных методов решения полной ПСЗ. Напомним (см. гл. II, § 2), что матрица отражения имеет вид

$$V = E - 2uw^T,$$

где w — вектор-столбец единичной длины. При этом задача построения матрицы V , переводящей произвольный ненулевой вектор y в заданный единичный вектор e , решается по формулам

$$(E - 2uw^T)y = \alpha e, \quad w = \rho^{-1}(y - \alpha e),$$

$$\alpha = \sqrt{(y, y)}, \quad \rho = 2(y, y - \alpha e). \quad (27)$$

Покажем, как для произвольной матрицы A подобрать конечную последовательность отражений, приводящую ее к верхнему почти треугольному виду. Как и в случае решения СЛАУ, каждое последующее отражение должно уничтожать элементы самого длинного ненулевого столбца в нижней части матрицы. Предположим, что получена матрица A , у которой уничтожены соответствующие элементы первых $k-1$ столбцов. Представим ее в блочном виде:

$$\left(\begin{array}{cccc|ccccc} a_{11} & a_{12} & \dots & a_{1,k-1} & a_{1k} & a_{1,k+1} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2,k-1} & a_{2k} & a_{2,k+1} & \dots & a_{2n} \\ 0 & a_{32} & \dots & a_{3,k-1} & a_{3k} & a_{3,k+1} & \dots & a_{3n} \\ 0 & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \ddots & a_{k,k-1} & a_{kk} & a_{k,k+1} & \dots & a_{kn} & & \\ \hline 0 & a_{k+1,k} & a_{k+1,k+1} & \dots & a_{k+1,n} \\ 0 & a_{k+2,k} & a_{k+2,k+1} & \dots & a_{k+2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{n,k} & a_{n,k+1} & \dots & a_{nn} \end{array} \right) = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix}.$$

Отметим, что матрица $A_1 \in \mathbb{R}^{k \times k}$ является верхней почти треугольной, а в матрице $A_3 \in \mathbb{R}^{(n-k) \times k}$ только последний столбец отличен от нуля. Выполним преобразование с помощью вектора $w^{(k)}$, у которого первые k компоненты равны нулю:

$$w^{(k)} = (0, 0, \dots, 0, w_{k+1}, w_{k+2}, \dots, w_n)^T.$$

По вектору $w^{(k)}$ образуем матрицу отражения V_k и разобьем ее на блоки того же размера, что у матрицы A :

$$V_k = \begin{pmatrix} E & 0 \\ 0 & W \end{pmatrix}, \quad w_{ij} = \delta_{ij} - 2w_i^{(k)}w_j^{(k)}, \quad i, j = k+1, k+2, \dots, n.$$

Здесь δ_{ij} — символ Кронекера. Тогда искомое преобразование подобно можно записать в виде

$$B = V_k^{-1}AV_k = \begin{pmatrix} E & 0 \\ 0 & W \end{pmatrix} \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix} \begin{pmatrix} E & 0 \\ 0 & W \end{pmatrix} = \begin{pmatrix} A_1 & B_2 \\ B_3 & A_4 \end{pmatrix}. \quad (28)$$

П о блочном представлении (28) элементы последнего столбца матрицы $B_3 = WA_3$ отличны от нуля. Для их уничтожения в (27) положим $y = (a_{k+1,k}, a_{k+2,k}, \dots, a_{nk})^T$, $e = (1, 0, \dots, 0)^T$. Имеем

$$\alpha = b_{k+1,k} = \left(\sum_{i=k+1}^n a_{ik}^2 \right)^{1/2}, \quad (y, y - \alpha e) = b_{k+1,k}(b_{k+1,k} - a_{k+1,k}).$$

Окончательно формулы для вычисления компонент вектора $w^{(k)}$ принимают вид

$$\alpha = \alpha^{(k)} = b_{k+1,k} = \sqrt{\sum_{i=k+1}^n a_{ik}^2}, \quad \rho = \sqrt{2b_{k+1,k}(b_{k+1,k} - a_{k+1,k})},$$

$$w_{k+1}^{(k)} = \rho^{-1}(a_{k+1,k} - \alpha^{(k)}), \quad w_i^{(k)} = \rho^{-1}a_{ik}, \quad i = k+2, k+3, \dots, n.$$

Последовательно определяя векторы $w^{(k)}$ и осуществляя подобные преобразования отражения, после $n-2$ этапов приходим к верхней почти треугольной матрице (на это затрачивается примерно $(10/3)n^3$ арифметических действий). Если исходная матрица A симметрична, то алгоритм отражений преобразует ее к трехдиагональному виду за $(4/3)n^3$ арифметических действий.

Помимо метода отражений, для приведения матрицы к почти треугольному (трехдиагональному) виду может быть также использован *прямой метод брашнелей*. Этот метод несколько уступает по скорости методу отражений, однако формулы для преобразования

элементов в нем пропаде. В гл. II, § 2 была определена матрица вращения T_{kl} , $k < l$, которая отличается от единичной матрицы только четырьмя элементами, расположеными на пересечении строк и столбцов с номерами k и l . Одно подобное преобразование вращения состоит из двух преобразований. Первое из них заключается в построении матрицы

$$B = AT_{kl}.$$

Столбцы матрицы B совпадают со столбцами матрицы A , за исключением столбцов с номерами k и l :

$$B_k = \cos \varphi A_k + \sin \varphi A_l, \quad B_l = -\sin \varphi A_k + \cos \varphi A_l. \quad (29)$$

Второе преобразование связано с построением матрицы

$$C = T_{kl}^{-1}B = T_{kl}^{-1}AT_{kl}.$$

Строки матрицы C совпадают со строками матрицы B , за исключением k -й и l -й строк:

$$C^k = \cos \varphi B^k + \sin \varphi B^l, \quad C^l = -\sin \varphi B^k + \cos \varphi B^l. \quad (30)$$

Очевидно, что всегда можно подобрать угол поворота в матрице вращения T_{kl} так, чтобы уничтожить элемент, находящийся в позиции $(l, k-1)$. Для этого достаточно положить

$$\cos \varphi = \frac{a_{k,k-1}}{\sqrt{a_{k,k-1}^2 + a_{l,k-1}^2}}, \quad \sin \varphi = \frac{a_{l,k-1}}{\sqrt{a_{k,k-1}^2 + a_{l,k-1}^2}}.$$

На основании данного факта выберем следующую стратегию уничтожения элементов столбцов матрицы A . На первом этапе уничтожим элементы первого столбца, начиная с третьего, т. е. элементы в позициях $(3, 1), (4, 1), \dots, (n, 1)$. Для этого используем подобные преобразования $T_{23}, T_{24}, \dots, T_{2n}$. Очевидно, что ранее уничтоженные элементы первого столбца в позициях $(i, 1)$ при дальнейших преобразованиях $T_{2j}, j > i$, изменяться не будут. На втором этапе с помощью подобных преобразований вращения $T_{34}, T_{35}, \dots, T_{3n}$ уничтожим элементы второго столбца, начиная с четвертого. При этом легко убедиться, что преобразования второго этапа не изменяют нулевых элементов в позициях $(i, 1)$, $i = 3, 4, \dots, n$, полученных

§ 2. ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ ПОЛНОЙ ПСЗ

161

после первого этапа. Действительно, преобразование (29) не изменяет первый столбец, а из преобразований (30) следует, что элементы c_{kl} и c_{ll} равны нулю, поскольку являются линейной комбинацией нулевых элементов b_{kl} и b_{ll} . Далее переходим к третьему этапу, который заключается в уничтожении элементов третьего столбца, начиная с пятого, и т. д. Таким образом, после $0,5(n-1)(n-2)$ подобных преобразований вращения матрица A будет приведена к верхнему почти треугольному (трехдиагональному) виду.

§ 2. Итерационные методы решения полной ПСЗ

Рассмотренные в § 1 прямые методы решения полной ПСЗ не всегда удовлетворительны, в первую очередь из-за плохой устойчивости при больших размерах матриц и неоднородности вычислительного процесса. Кроме того, расчетные формулы этих методов не упрощаются для некоторых специальных видов матриц, наиболее важных в приложениях. Поэтому интерес представляют итерационные методы решения полной ПСЗ, более трудоемкие по сравнению с прямыми методами, однако позволяющие вычислить собственные значения матрицы без предварительного нахождения корней характеристического многочлена.

1. Метод Якоби (итерационный метод вращений). Для симметричной матрицы A рассмотрим метод Якоби, или итерационный метод вращений, в котором с помощью подобных преобразований вращения $T^{-1}AT$ строится последовательность матриц $A_0 = A, A_1, \dots, A_m, \dots$, такая, что $A_m \rightarrow \Lambda$, $\Lambda = \text{diag}[\lambda_1, \lambda_2, \dots, \lambda_n]$. В основе метода Якоби лежит следующая теорема.

Теорема 1. Сферическая норма матрицы не изменяется при подобном преобразовании вращения.

Доказательство. По определению сферической нормы имеем

$$\|A\|_E^2 = \sum_{i,j=1}^n a_{ij}^2. \quad (1)$$

Рассмотрим матрицу вращения T_{kl} и образуем матрицу $B = AT_{kl}$.

Исходная матрица A отличается от матрицы B лишь столбцами с номерами k и l , при этом

$$\begin{aligned} b_{ik} &= a_{ik} \cos \varphi + a_{il} \sin \varphi, & b_{il} &= -a_{ik} \sin \varphi + a_{il} \cos \varphi, \\ b_{ij} &= a_{ij}, \quad j \neq k, \quad j \neq l, & i &= 1, 2, \dots, n. \end{aligned} \quad (2)$$

Тогда получим соотношение

$$b_{ik}^2 + b_{il}^2 = a_{ik}^2 + a_{il}^2,$$

а поскольку элементы остальных столбцов не изменяются, то

$$\|B\| = \|AT_{kl}\| = \|A\|.$$

Далее образуем матрицу $C = T_{kl}^{-1}AT_{kl}$, отличающуюся от матрицы B только строками с номерами k и l :

$$\begin{aligned} c_{ki} &= b_{ki} \cos \varphi + b_{li} \sin \varphi, & c_{ii} &= -b_{ki} \sin \varphi + b_{li} \cos \varphi, \\ c_{ji} &= b_{ji}, \quad j \neq k, \quad j \neq l, & i &= 1, 2, \dots, n. \end{aligned} \quad (3)$$

Так как

$$c_{ki}^2 + c_{li}^2 = b_{ki}^2 + b_{li}^2,$$

то, очевидно, имеем

$$\|C\| = \|T_{kl}^*B\| = \|B\| = \|A\|.$$

Теорема доказана.

Разобьем сумму, входящую в норму (1), на две части:

$$\sum_{i,j=1}^n a_{ij}^2 = \sum_{i=1}^n a_{ii}^2 + \sum_{i,j=1, i \neq j}^n a_{ij}^2 = S(A) + t^2(A).$$

Как показано в теореме 1, при подобном преобразовании вращения внедиагональные элементы $a_{ik}, a_{il}, a_{ki}, a_{li}$ меняются так, что сохраняются неизменными величины

$$a_{ik}^2 + a_{il}^2, \quad a_{ki}^2 + a_{li}^2, \quad i \neq k, \quad i \neq l.$$

Кроме этих элементов, внедиагональным является также элемент a_{kl} . Поэтому изменение величины $t^2(A)$ полностью определяется

§ 2. ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ ПОЛНОЙ ПСЗ

163

изменением a_{kl}^2 . Для того чтобы максимально уменьшить $t^2(A)$ за одно вращение, выберем угол поворота φ таким образом, чтобы уничтожить элемент a_{kl} . Из формул (2), (3) следует, что

$$\begin{aligned} c_{kl} &= b_{kl} \cos \varphi + b_{ll} \sin \varphi = (-a_{kk} \sin \varphi + a_{kl} \cos \varphi) \cos \varphi + \\ &+ (-a_{lk} \sin \varphi + a_{ll} \cos \varphi) \sin \varphi = a_{kl} \cos 2\varphi + \frac{1}{2}(a_{ll} - a_{kk}) \sin 2\varphi. \end{aligned} \quad (4)$$

Полагая $c_{kl} = 0$, из равенства (4) получим

$$\tan 2\varphi = \frac{2a_{kl}}{a_{kk} - a_{ll}}, \quad |\varphi| \leq \frac{\pi}{4},$$

или в эквивалентном виде

$$\cos \varphi = \sqrt{\frac{1}{2}\left(1 + \frac{1}{\sqrt{1 + \mu^2}}\right)}, \quad \sin \varphi = \operatorname{sign} \mu \sqrt{\frac{1}{2}\left(1 - \frac{1}{\sqrt{1 + \mu^2}}\right)},$$

где $\mu = \frac{2a_{kl}}{a_{kk} - a_{ll}}$. Нетрудно убедиться, что при таком вращении максимально уменьшается сумма квадратов внедиагональных элементов матрицы A . Действительно,

$$\begin{aligned} \sum_{i,j=1, i \neq j}^n c_{ij}^2 &= \sum_{i,j=1, i \neq j}^n a_{ij}^2 + \\ &+ \frac{1}{2}((a_{ll} - a_{kk}) \sin 2\varphi + 2a_{kl} \cos 2\varphi)^2 = \sum_{i \neq j} a_{ij}^2 - 2a_{kl}^2. \end{aligned} \quad (5)$$

Если a_{kl} — максимальный по модулю внедиагональный элемент матрицы A , то из (5) следует, что преобразование подобия с матрицей T_{kl} максимально уменьшает величину $t^2(A)$.

Таким образом, выбирая последовательность подобных преобразований, уничтожающую внедиагональные элементы, мы тем самым строим монотонные последовательности

$$t^2(A_0) > t^2(A_1) > \dots > t^2(A_m) \rightarrow 0,$$

$$S(A_0) < S(A_1) < \dots < S(A_m) \rightarrow \|A\|^2.$$

Это означает, что последовательность матриц A_m , $m = 0, 1, \dots$, сходится к диагональной матрице Λ :

$$a_{ii}^{(m)} \rightarrow \lambda_i, \quad i = 1, 2, \dots, n.$$

Поскольку собственными векторами диагональной матрицы являются единичные векторы e_i и $A_m \approx \Lambda$, то хорошим приближением к собственным векторам матрицы A будут служить столбцы результирующей матрицы

$$T = T_{k_0 l_0}^{(0)} T_{k_1 l_1}^{(1)} \cdots T_{k_m l_m}^{(m)}.$$

Выберем теперь стратегию уничтожения внедиагональных элементов. Если каждый раз уничтожать максимальный по модулю внедиагональный элемент, то хотя скорость убывания $t^2(A_m)$ при этом будет наибольшей, его нахождение требует $n(n-1)$ переборов всех внедиагональных элементов, что существенно снижает эффективность алгоритма. Существуют более экономичные стратегии, основанные на уничтожении оптимального элемента.

Составим суммы квадратов внедиагональных элементов строк:

$$\sigma_i = \sum_{j=1, j \neq i}^n a_{ij}^2, \quad i = 1, 2, \dots, n.$$

Выберем из этих сумм наибольшую, а в ней — наибольший по модулю элемент, подлежащий уничтожению. Его нахождение требует всего $2n-1$ перебора. Если оптимальным является элемент a_{kl} , то при подобном преобразовании вращения изменятся лишь σ_k и σ_l , а именно:

$$\sigma_k^{(1)} = \sigma_k + (a_{kk}^{(1)})^2 - a_{kk}^2 - a_{kl}^2, \quad \sigma_l^{(1)} = \sigma_l + \sigma_k - \sigma_k^{(1)},$$

т. е. для нахождения оптимального элемента на следующем этапе необходимо пересчитывать только σ_k , σ_l .

Описанная стратегия выбора оптимального элемента позволяет судить о скорости сходимости метода Якоби. В силу равенства (5)

$$t^2(A_{m+1}) = t^2(A_m) - 2(a_{kl}^{(m)})^2.$$

По определению оптимального элемента имеем

$$(a_{kl}^{(m)})^2 \geq \frac{1}{n-1} \sigma_l^{(m)} \geq \frac{1}{n(n-1)} t^2(A_m),$$

откуда следует оценка скорости сходимости

$$t^2(A_{m+1}) \leq \left(1 - \frac{2}{n(n-1)}\right) t^2(A_m) \leq \dots \leq \left(1 - \frac{2}{n(n-1)}\right)^{m+1} t^2(A_0).$$

Таким образом, итерационный метод вращений с выбором оптимального элемента сходится:

$$\lim_{m \rightarrow \infty} t^2(A_m) = 0.$$

2. QR-алгоритм. Представление матрицы в верхнем треугольном либо трехдиагональном виде является одним из существенных элементов и в ряде других методов решения полной проблемы собственных значений. Например, в рассмотренном методе Якоби для $A = A^T$ предпочтительно сначала привести матрицу A к трехдиагональному виду, а лишь затем перейти на обычную стратегию уничтожения оптимального элемента. Еще большую роль играет предварительное приведение матрицы к верхнему почти треугольному виду в так называемом QR-алгоритме, идея которого заключается в следующем.

Пусть $\det A \neq 0$. Тогда, согласно теореме 1 из гл. II, § 2, матрица A может быть представлена в виде

$$A = QR, \quad (6)$$

где Q — ортогональная матрица, R — верхняя треугольная матрица. Образуем последовательность матриц

$$\begin{aligned} A &= Q_1 R_1, \quad A_1 = R_1 Q_1 = Q_1^{-1} A Q_1, \\ A_1 &= Q_2 R_2, \quad A_2 = R_2 Q_2 = Q_2^{-1} Q_1^{-1} A Q_1 Q_2, \end{aligned} \quad (7)$$

$$A_{k-1} = Q_k R_k, \quad A_k = R_k Q_k = Q_k^{-1} \dots Q_1^{-1} A Q_1 \dots Q_k.$$

Имеет место утверждение.

Теорема 2. Пусть невырожденная матрица A имеет различные вещественные собственные значения

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_n| \quad (8)$$

и пусть существует LU-разложение матрицы Q^{-1} :

$$Q^{-1} = LU, \quad A = Q \Lambda Q^{-1}, \quad \Lambda = \text{diag}[\lambda_1, \lambda_2, \dots, \lambda_n]. \quad (9)$$

Тогда последовательность (7) QR-алгоритма сходится к верхней треугольной матрице.

Доказательство. Пусть имеют место представления

$$P_k = Q_1 Q_2 \dots Q_k, \quad U_k = R_k R_{k-1} \dots R_1,$$

где Q_k, R_k определены в (7). Тогда

$$\begin{aligned} P_k U_k &= P_{k-1} Q_k U_{k-1} = P_{k-1} A_k U_{k-1} = \\ &= P_{k-1} P_{k-1}^{-1} A P_{k-1} U_{k-1} = A(P_{k-1} U_{k-1}) = \dots = A^k. \end{aligned}$$

Следовательно, справедливо соотношение

$$A^k = P_k U_k.$$

С другой стороны,

$$A^k = Q \Lambda^k Q^{-1} = Q \Lambda^k L U.$$

Поэтому

$$P_k U_k = (Q \Lambda^k L \Lambda^{-k}) \Lambda^k U = G_k \Lambda^k U.$$

Из последнего равенства следует, что матрица

$$H_k = P_k^{-1} Q \Lambda^k L \Lambda^{-k} = U_k U^{-1} \Lambda^{-k}$$

верхняя треугольная. Далее находим

$$P_k = G_k \Lambda^k U U_k^{-1} = G_k H_k^{-1}, \quad P_k^{-1} = U_k U^{-1} \Lambda^{-k} G_k^{-1} = H_k G_k^{-1}. \quad (10)$$

Подставим (10) в (7) и воспользуемся разложением (9):

$$\begin{aligned} A_k &= P_k^{-1} Q \Lambda Q^{-1} P_k = H_k (G_k^{-1} Q \Lambda Q^{-1} G_k) H_k^{-1} = \\ &= H_k (\Lambda^k L^{-1} \Lambda L \Lambda^{-k}) H_k^{-1}. \end{aligned} \quad (11)$$

Матрица $B = L^{-1} A L$ в (11) является нижней треугольной, причем $b_{ii} = \lambda_i$. Матрица $C_k = \Lambda^k B \Lambda^{-k}$ также нижняя треугольная с элементами

$$c_{ij}^{(k)} = b_{ij} \left(\frac{\lambda_i}{\lambda_j}\right)^k, \quad i \geq j.$$

Но в силу условия (8)

$$\lim_{k \rightarrow \infty} c_{ij}^{(k)} = \begin{cases} 0, & i > j, \\ \lambda_i, & i = j. \end{cases}$$

Поэтому

$$\lim_{k \rightarrow \infty} C_k = \Lambda.$$

Остается изучить поведение матриц H_k, H_k^{-1} при $k \rightarrow \infty$. Для этого рассмотрим их составляющие. Матрицы P_k и P_k^{-1} — ограниченные для всех k , так как они являются ортогональными, матрицы Q и Q^{-1} — постоянные. В силу теоремы об LU-разложении можно считать, что $I_n = 1$. Поэтому

$$\lim_{k \rightarrow \infty} \Lambda^k L \Lambda^{-k} = \lim_{k \rightarrow \infty} \Lambda^k L^{-1} \Lambda^{-k} = E.$$

Таким образом, элементы матриц H_k, H_k^{-1} ограничены при любом значении k . Теорема доказана.

Практическое применение QR-алгоритма неразрывно связано со следующими необходимыми элементами.

а) Приведение исходной матрицы A к верхней почти треугольной. Основным фактором здесь является инвариантность процесса (7) к верхнему почти треугольному виду матрицы. При этом для осуществления одного этапа QR-алгоритма требуется лишь $O(n^2)$ арифметических действий, т. е. на порядок меньше, чем в случае матрицы общего вида.

б) Ускорение сходимости метода. Пусть λ_m^* — некоторое приближение к собственному значению матрицы A . Тогда для $k \geq m$ в (7) положим

$$A_k - \lambda_m^* E = Q_k R_k, \quad A_{k+1} = R_k Q_k + \lambda_m^* E.$$

Эти формулы задают *QR-алгоритм со сдвигом*. При этом собственные значения матриц A_k , A_{k+1} совпадают, поскольку, как и в обычном *QR-алгоритме*, матрица A_{k+1} подобна матрице A_k :

$$\begin{aligned} A_{k+1} &= R_k Q_k + \lambda_m^* E = Q_k^{-1} Q_k R_k Q_k + Q_k^{-1} Q_k \lambda_m^* E = \\ &= Q_k^{-1} (Q_k R_k + \lambda_m^* E) Q_k = Q_k^{-1} A_k Q_k. \end{aligned}$$

Выбирая число λ_m^* подходящим образом, можно существенно ускорить сходимость *QR-алгоритма*. На практике после m этапов обычного алгоритма в качестве λ_m^* берут элемент в позиции (n, n) . Тогда через m_0 последующих этапов *QR-алгоритма со сдвигом*, в этой позиции будет находиться достаточно хорошее приближение наименьшего по модулю собственного значения. Затем обычный алгоритм применяют к матрице $(n - 1)$ -го порядка, полученной из A_{m+m_0} вычеркиванием последней строки и последнего столбца, и т. д.

3. Метод биссекций. В этом пункте рассмотрим *метод биссекций* решения полной проблемы собственных значений, являющихся наряду с методом Якоби одним из основных для симметрических матриц с вещественными элементами. Будем считать, что матрица A приведена подобными ортогональными преобразованиями к трехдиагональному виду

$$A = \begin{pmatrix} c_1 & -b_2 & & & 0 \\ -b_2 & c_2 & -b_3 & & \\ & \ddots & \ddots & \ddots & \\ 0 & & -b_{n-1} & c_{n-1} & -b_n \\ & & -b_n & c_n & \end{pmatrix}.$$

Найдем интервал, содержащий все собственные значения задачи

$$Ax = \lambda x. \quad (12)$$

Специфика матрицы A позволяет достаточно точно определить такой интервал. Из уравнения (12) имеем

$$\lambda = \frac{(Ax, x)}{(x, x)}. \quad (13)$$

По определению запишем

$$(Ax, x) = \sum_{i=1}^n c_i x_i^2 - 2 \sum_{i=1}^{n-1} b_{i+1} x_i x_{i+1}.$$

Для оценки величины (13) воспользуемся формулами $-u^2 - v^2 \leq \pm 2uv \leq u^2 + v^2$. Тогда получим

$$\begin{aligned} \sum_{i=1}^n c_i x_i^2 - \sum_{i=1}^{n-1} |b_{i+1}| x_i^2 - \sum_{i=1}^{n-1} |b_{i+1}| x_{i+1}^2 &\leq (Ax, x) \leq \\ &\leq \sum_{i=1}^n c_i x_i^2 + \sum_{i=1}^{n-1} |b_{i+1}| x_i^2 + \sum_{i=1}^{n-1} |b_{i+1}| x_{i+1}^2 \end{aligned}$$

или после преобразований

$$\begin{aligned} (c_1 - |b_2|) x_1^2 + (c_n - |b_n|) x_n^2 + \sum_{i=2}^{n-1} (c_i - |b_i| - |b_{i+1}|) x_i^2 &\leq (Ax, x) \leq \\ (c_1 + |b_2|) x_1^2 + (c_n + |b_n|) x_n^2 + \sum_{i=2}^{n-1} (c_i + |b_i| + |b_{i+1}|) x_i^2. \end{aligned}$$

Таким образом, в качестве искомого интервала достаточно взять промежуток (γ_1, γ_2) , где

$$\gamma_1 = \min \left\{ c_1 - |b_2|, \min_{2 \leq i \leq n-1} (c_i - |b_i| - |b_{i+1}|), c_n - |b_n| \right\},$$

$$\gamma_2 = \max \left\{ c_1 + |b_2|, \max_{2 \leq i \leq n-1} (c_i + |b_i| + |b_{i+1}|), c_n + |b_n| \right\}.$$

Второй этап алгоритма биссекций состоит в делении интервала (γ_1, γ_2) пополам и в определении числа корней характеристического многочлена $A(\lambda) = \det(A - \lambda E)$, содержащихся в каждом из полученных таким образом интервалов. Теоретическую основу этого этапа составляет закон инерции.

Теорема 3. Пусть вещественные симметрические матрицы A и B подобны:

$$A = S^T B S, \quad \det S \neq 0.$$

Тогда матрицы A и B имеют одинаковое число положительных, отрицательных и разных нуля собственных значений.

Теорема 4. Пусть A – вещественная симметричная матрица. Тогда число положительных, отрицательных и разных нуля ее собственных значений равно числу положительных, отрицательных и разных нулях *ведущих* элементов в методе Гаусса.

Доказательство. Известно, что вещественная симметричная матрица подобна диагональному (разложение Шурра): $A = Q^T \Lambda Q$, где Q – ортогональная матрица, $\Lambda = \text{diag}[\lambda_1, \lambda_2, \dots, \lambda_n]$. С другой стороны, согласно теореме об *LDU*-разложении (см. гл. II, § 1), $A = U^T D U$, где U – верхняя треугольная матрица с единичной главной диагональю, D – диагональная матрица, на диагонали которой расположены ведущие элементы метода исключения Гаусса. Тогда имеем

$$D = (U^T)^{-1} A U^{-1} = (U^T)^{-1} Q^T \Lambda Q U^{-1} = (Q U^{-1})^T \Lambda (Q U^{-1}).$$

Полагая в последнем равенстве $S = Q U^{-1}$, получим искомое утверждение. Теорема доказана.

Следующая теорема позволяет локализовать собственные значения в методе биссекций.

Теорема 5. Пусть λ^* не является собственным значением вещественной симметрической матрицы A . Тогда число положительных (отрицательных) ведущих элементов метода Гаусса для матрицы $A - \lambda^* E$ равно числу собственных значений матрицы A , больших (меньших) λ^* .

Доказательство очевидно в силу теоремы 4, так как собственные значения μ_i матрицы $A - \lambda^* E$ связаны с собственными значениями λ_i матрицы A соотношением $\mu_i = \lambda_i - \lambda^*$.

Таким образом, приходим к следующей вычислительной схеме метода биссекций. Находим интервал (γ_1, γ_2) такой, что $\gamma_1 \leq \lambda_1 \leq \gamma_2$. Затем с помощью ведущих элементов матрицы $A - [(\gamma_1 + \gamma_2)/2] E$ определяем число собственных значений в каждом из промежутков

$$\left(\gamma_1, \frac{\gamma_1 + \gamma_2}{2} \right), \quad \left(\frac{\gamma_1 + \gamma_2}{2}, \gamma_2 \right)$$

и повторяем указанную процедуру. Не более чем за $2^m - 1$ итераций все собственные значения матрицы A будут локализованы в

интервалах длины $(\gamma_2 - \gamma_1) \cdot 2^{-m}$. Для вычисления ведущих элементов матрицы $A - \lambda^* E$ используем рекуррентные соотношения (см. формулы метода прогонки, гл. II, § 3):

$$\Delta_i(\lambda^*) = c_i - \lambda^* - \frac{b_i^2}{\Delta_{i-1}(\lambda^*)}, \quad i = 2, 3, \dots, n, \quad \Delta_1(\lambda^*) = c_1 - \lambda^*. \quad (14)$$

Наряду с последовательностью (14) алгоритм метода биссекций может быть построен и на основе последовательности угловых миноров матрицы $A - \lambda^* E$. Из формулы (25) § 1 имеем

$$\begin{aligned} \Delta_i(\lambda^*) &= (c_i - \lambda^*) \Delta_{i-1}(\lambda^*) - b_i^2 \Delta_{i-2}(\lambda^*), \quad i = 2, 3, \dots, n, \\ A_0(\lambda^*) &= 1, \quad A_1(\lambda^*) = c_1 - \lambda^*. \end{aligned} \quad (15)$$

Обоснованность применения формул (15) для локализации собственных значений гарантируется уже упомянутой теоремой об *LDU*-разложении, в силу которой

$$\Delta_i(\lambda^*) = \frac{A_i(\lambda^*)}{A_{i-1}(\lambda^*)}, \quad i = 1, 2, \dots, n,$$

и следующими очевидными свойствами последовательности (15).

1. Никакие два соседних угловых минора не могут одновременно равняться нулю.

2. Если $A_i(\lambda^*) = 0$, $i = 2, 3, \dots, n - 1$, то $A_{i-1}(\lambda^*) A_{i+1}(\lambda^*) < 0$.

Потому число собственных значений матрицы A , больших или меньших λ^* , можно определять также и по знакам угловых миноров матрицы $A - \lambda^* E$.

Формулы (14), (15) в прямом виде нельзя использовать для вычислений на ЭВМ по ряду причин. Например, равенство нулю любого из ведущих элементов означает, что λ^* – собственное значение матрицы A . На самом деле этот вывод не всегда будет справедлив из-за того, что малое ненулевое число на конкретной ЭВМ может восприниматься как машинный нуль. Опасность появления близких к машинному нулю чисел заключается также и в том, что в этих формулах должны быть правильно вычислены знаки $\Delta_i(\lambda^*)$, $A_i(\lambda^*)$, что достаточно непросто в условиях влияния погрешностей округления. Сказанное означает, что при реализации на ЭВМ алгоритмов (14), (15) эти вопросы должны быть заранее исследованы.

Одним из способов, которые могут гарантировать правильность знаков вычисляемых величин, является их нормировка. В частности, соотношения (15) представляют собой линейные однородные уравнения относительно A_1, A_2, \dots, A_n . Если перед каждым вычислением A_i умножать A_{i-2} и A_{i-1} на произвольное положительное число ρ_i , то знаки новых величин $\rho A_1, \rho A_2, \dots, \rho A_n$ будут совпадать со знаками угловых миноров. Следовательно, число нулевых, положительных и отрицательных собственных значений исходной матрицы будет определено корректно. Нормирующие множители ρ выбираются таким образом, чтобы обеспечить требования точности.

§ 3. Методы решения частичной ПСЗ

Как было сказано выше, частичная проблема собственных значений подразумевает нахождение одного или нескольких собственных значений и соответствующих им собственных векторов матрицы. Все методы решения этой проблемы являются итерационными и основаны на использовании структурных свойств матриц.

1. Степенной метод вычисления наибольшего по модулю собственного значения матрицы. В этом пункте рассмотрим простейший итерационный метод вычисления наибольшего по модулю собственного значения и соответствующего ему собственного вектора — *степенной метод*. Его идея заключается в построении такой итерационной последовательности векторов, чтобы в ней доминировала одна составляющая в разложении по базису из собственных векторов. Тем самым достигается сходимость по направлению к выделенному собственному вектору. Эффективность метода определяется тем, какходит наибольшее по модулю собственное значение матрицы в ее каноническую форму Жордана. Для упрощения изложения мы ограничимся рассмотрением матриц, имеющих линейные элементарные делители*.

Расположим собственные значения матрицы A в порядке невоз-

*Элементарным делителем матрицы A назовем определитель блока канонической формы $J - \lambda E$, где J — жорданова каноническая форма матрицы A .

растания их модулей:

$$|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|.$$

С учетом сделанного предположения матрица A имеет базис из собственных векторов x_1, x_2, \dots, x_n . Пусть наибольшее по модулю собственное значение вещественное и простое:

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_n|.$$

В этом случае говорят, что λ_1 является *доминирующим собственным значением*, а соответствующий ему собственный вектор — *доминирующим собственным вектором*. Возьмем произвольный начальный вектор y^0 и образуем итерационную последовательность:

$$y^k = Ay^{k-1} = \dots = A^ky^0, \quad k = 1, 2, \dots \quad (1)$$

Разложим y^0 по системе собственных векторов $\{x_s\}$:

$$y^0 = c_1x_1 + c_2x_2 + \dots + c_nx_n, \quad (2)$$

где c_s — некоторые постоянные. Предположим, что $c_1 \neq 0$ (это требование будет выполнено, если вектор y^0 не ортогонален доминирующему собственному вектору транспонированной матрицы A^T). Принимая во внимание, что $A^kx_s = \lambda_s^kx_s$, $s = 1, 2, \dots, n$, из соотношений (1), (2) получим

$$y^k = c_1\lambda_1^kx_1 + c_2\lambda_2^kx_2 + \dots + c_n\lambda_n^kx_n. \quad (3)$$

Запишем равенство (3) в покомпонентной форме:

$$\begin{aligned} y_i^{(k)} &= c_1\lambda_1^kx_{i1} + c_2\lambda_2^kx_{i2} + \dots + c_n\lambda_n^kx_{in} = \\ &= c_1\lambda_1^kx_{i1} \left[1 + \frac{c_2}{c_1} \left(\frac{\lambda_2}{\lambda_1} \right)^k x_{i2} + \dots + \frac{c_n}{c_1} \left(\frac{\lambda_n}{\lambda_1} \right)^k x_{in} \right]. \end{aligned}$$

Для соседней итерации аналогично получим

$$y_i^{(k+1)} = c_1\lambda_1^{k+1}x_{i1} \left[1 + \frac{c_2}{c_1x_{i1}} \left(\frac{\lambda_2}{\lambda_1} \right)^{k+1} x_{i2} + \dots + \frac{c_n}{c_1x_{i1}} \left(\frac{\lambda_n}{\lambda_1} \right)^{k+1} x_{in} \right].$$

Так как λ_1 — доминирующее собственное значение, то величины $\left| \frac{\lambda_s}{\lambda_1} \right| < 1$, $s = 2, 3, \dots, n$, и при больших k

$$\frac{y_i^{(k+1)}}{y_i^{(k)}} = \lambda_1 + O\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k\right), \quad \lim_{k \rightarrow \infty} \frac{y_i^{(k+1)}}{y_i^{(k)}} = \lambda_1, \quad i = 1, 2, \dots, n.$$

Таким образом, для вычисления доминирующего собственного значения λ_1 можно взять отношение $\frac{y_i^{(k+1)}}{y_i^{(k)}}$ любых компонент до статочно высоких итераций.

Определим теперь собственный вектор x_1 , соответствующий доминирующему собственному значению λ_1 . Рассмотрим сначала общий подход к получению собственного вектора, соответствующего любому собственному значению. Пусть простое собственное значение λ_m матрицы A вычислено с какой-либо точностью, т. е. $\tilde{\lambda}_m = \lambda_m + O(\varepsilon)$, при этом $\tilde{\lambda}_m \neq \lambda_s$, $s = 1, 2, \dots, n$. Тогда

$$\det(A - \tilde{\lambda}_m E) \neq 0$$

и для любого произвольного вектора f задача

$$(A - \tilde{\lambda}_m E)y = f \quad (4)$$

имеет единственное решение. Поэтому в качестве собственного вектора x_m , соответствующего собственному значению λ_m , можно выбрать решение задачи (4). В самом деле, поскольку

$$y = \sum_{s=1}^n c_s x_s, \quad f = \sum_{s=1}^n b_s x_s,$$

из (4) следует

$$\sum_{s=1}^n [c_s(\lambda_s - \tilde{\lambda}_m) - b_s] x_s = 0,$$

или

$$c_s = \frac{b_s}{\lambda_s - \tilde{\lambda}_m}.$$

Так как по предположению собственное значение λ_m простое, то в случае $b_s = O(1)$ коэффициент $c_m = O(1/\varepsilon)$ намного больше всех остальных и, следовательно,

$$y \approx c_m x_m. \quad (5)$$

Указанныя процедура может быть всегда применена в случае наличия простого собственного значения λ_m , в том числе и для нахождения собственного вектора x_1 . В то же время если λ_1 находится с помощью степенного метода, то наряду с вектором y из соотношения (5) хорошим приближением для x_1 будет служить вектор y^k последовательности (1). Действительно, запишем (3) в виде

$$y^k = \sum_{s=1}^n c_s \lambda_s^k x_s = c_1 \lambda_1^k \left[x_1 + \sum_{s=2}^n \left(\frac{c_s}{c_1} \right) \left(\frac{\lambda_s}{\lambda_1} \right)^k x_s \right], \quad (6)$$

откуда, отбрасывая бесконечно малые слагаемые, получим

$$y^k \approx c_1 \lambda_1^k x_1. \quad (7)$$

Входящая в выражение (7) константа λ_1^k может быть как достаточно большой, так и достаточно малой величиной. Поэтому в реальных вычислениях необходимо на каждой итерации проводить нормировку найденного вектора y^k .

На практике эффективность степенного метода зависит от отношения $\left| \frac{\lambda_2}{\lambda_1} \right|$, определяющего скорость сходимости. Если это отношение близко к единице, она будет достаточно медленной. Тот факт, что начальный вектор y^0 может не содержать x_1 ($c_1 = 0$), не должен вызывать затруднений. Ошибки округления, появляющиеся при выполнении итераций, обычно приводят к тому, что последующие y^k имеют компоненту в направлении доминирующего вектора.

Если доминирующее собственное значение вещественное и кратное, но соответствующие ему элементарные делители линейны, это не влияет на сходимость степенного метода. Пусть

$$\lambda_1 = \lambda_2 = \dots = \lambda_r, \quad |\lambda_1| > |\lambda_{r+1}| \geq |\lambda_{r+2}| \geq \dots \geq |\lambda_n|.$$

В этом случае разложение (3) принимает вид

$$y^k = \lambda_1^k (c_1 x_1 + c_2 x_2 + \dots + c_r x_r) + c_{r+1} \lambda_{r+1}^k x_{r+1} + \dots + c_n \lambda_n^k x_n =$$

$$= \lambda_1^k \left[c_1 x_1 + c_2 x_2 + \dots + c_r x_r + c_{r+1} \left(\frac{\lambda_{r+1}}{\lambda_1} \right)^k x_{r+1} + \dots + c_n \left(\frac{\lambda_n}{\lambda_1} \right)^k x_n \right] = \\ = \lambda_1^k (c_1 x_1 + c_2 x_2 + \dots + c_r x_r) \left[1 + O \left(\left| \frac{\lambda_{r+1}}{\lambda_1} \right|^k \right) \right].$$

Построенная последовательность $(c_1 x_1 + c_2 x_2 + \dots + c_r x_r \neq 0)$ вновь ведет себя как геометрическая прогрессия, на этот раз со знаменателем $|\lambda_{r+1}/\lambda_1|$. Имеем

$$\frac{y_i^{(k+1)}}{y_i^{(k)}} = \lambda_1 + O \left(\left| \frac{\lambda_{r+1}}{\lambda_1} \right|^k \right), \quad i = 1, 2, \dots, n.$$

Как и в предыдущем случае, в качестве доминирующего собственного вектора x_1 можно взять вектор y^k . Выбирая другие начальные приближения y^0 , приходим, вообще говоря, к другим доминирующими собственным векторам x_2, x_3, \dots, x_r .

Далее рассмотрим ситуацию, когда матрица A имеет два доминирующих собственных значения, которые вещественны и противоположны по знаку:

$$\lambda_1 = -\lambda_2, \quad |\lambda_1| = |\lambda_2| > |\lambda_3| \geq |\lambda_4| \geq \dots \geq |\lambda_n|.$$

Тогда разложение (3) приводит к равенствам

$$y^{2k} = \lambda_1^{2k} (c_1 x_1 + c_2 x_2) + c_3 \lambda_3^{2k} x_3 + \dots + c_n \lambda_n^{2k} x_n, \\ y^{2k+1} = \lambda_1^{2k+1} (c_1 x_1 - c_2 x_2) + c_3 \lambda_3^{2k+1} x_3 + \dots + c_n \lambda_n^{2k+1} x_n.$$

Отсюда видно, что y^{2k} и y^{2k+1} не могут быть использованы для вычисления λ_1 в силу различия у них главных частей. Однако у векторов только с четными либо только с нечетными степенями главные части уже одинаковы. Следовательно,

$$\frac{y_i^{(2k+2)}}{y_i^{(2k)}} = \lambda_1^2 + O \left(\left| \frac{\lambda_3}{\lambda_1} \right|^{2k} \right), \quad \text{или} \quad \frac{y_i^{(2k+1)}}{y_i^{(2k-1)}} = \lambda_1^2 + O \left(\left| \frac{\lambda_3}{\lambda_1} \right|^{2k-1} \right).$$

Хорошим приближением для доминирующего собственного вектора, соответствующего λ_1 , является вектор $y^k + \lambda_1 y^{k-1}$, а для доминирующего собственного вектора, соответствующего $\lambda_2 = -\lambda_1$, вектор $y^k - \lambda_1 y^{k-1}$. В самом деле

$$y^k + \lambda_1 y^{k-1} = 2c_1 \lambda_1^k x_1 + c_3 \lambda_3^{k-1} (\lambda_3 + \lambda_1) x_3 + \dots$$

$$\dots + c_n \lambda_n^{k-1} (\lambda_n + \lambda_1) x_n = 2c_1 \lambda_1^k \left[x_1 + O \left(\left| \frac{\lambda_3}{\lambda_1} \right|^k \right) \right], \\ y^k - \lambda_1 y^{k-1} = 2c_2 (-\lambda_1)^k x_2 + c_3 \lambda_3^{k-1} (\lambda_3 - \lambda_1) x_3 + \dots \\ \dots + c_n \lambda_n^{k-1} (\lambda_n - \lambda_1) x_n = 2c_2 (-\lambda_1)^k \left[x_2 + O \left(\left| \frac{\lambda_3}{\lambda_1} \right|^k \right) \right].$$

Предположим теперь, что доминирующие собственные значения λ_1 и λ_2 образуют комплексно-сопряженную пару:

$$\lambda_1 = re^{i\varphi}, \quad \lambda_2 = re^{-i\varphi}, \quad \lambda_1 = \bar{\lambda}_2,$$

$$|\lambda_1| = |\lambda_2| > |\lambda_3| \geq |\lambda_4| \geq \dots \geq |\lambda_n|.$$

Тогда произвольный вещественный вектор представим в виде

$$y^0 = c_1 x_1 + \bar{c}_1 \bar{x}_1 + c_3 x_3 + \dots + c_n x_n, \quad c_1 = \rho e^{i\varphi}, \quad \bar{c}_1 = \rho e^{-i\varphi}.$$

Отсюда получим

$$y^k = r^k \rho e^{i(\varphi+k\theta)} x_1 + r^k \rho e^{-i(\varphi+k\theta)} \bar{x}_1 + c_3 \lambda_3^k x_3 + \dots + c_n \lambda_n^k x_n = \\ = r^k \left[2\rho \cos(\varphi + k\theta) x_1 + \sum_{s=3}^n c_s \left(\frac{\lambda_s}{r} \right)^k x_s \right]. \quad (8)$$

Как видим, хотя составляющие x_3, x_4, \dots, x_n исчезают при $k \rightarrow \infty$, выражение (8) не стремится к пределу. Наличие множителя $\cos(\varphi + k\theta)$ является причиной осцилляций и перемен знаков компонент $y_i^{(k)}$ последовательности векторов y^k .

Получим формулы для нахождения чисел r и θ , обеспечивающие сходимость степенного метода. Именем место разложения

$$y^k = 2\rho r^k \cos(\varphi + k\theta) + O(|\lambda_3|^k), \\ y^{k+1} = 2\rho r^{k+1} \cos(\varphi + (k+1)\theta) + O(|\lambda_3|^{k+1}), \\ y^{k+2} = 2\rho r^{k+2} \cos(\varphi + (k+2)\theta) + O(|\lambda_3|^{k+2}). \quad (9)$$

С помощью (9) вычислим выражение

$$I^k = y^k y^{k+2} - (y^{k+1})^2 = 4\rho^2 r^{2k+2} [\cos(\varphi + (k+2)\theta) \cos(\varphi + k\theta) - \\ - \cos^2(\varphi + (k+1)\theta)] + r^k O(|\lambda_3|^k) = -4\rho^2 r^{2k+2} \sin^2 \theta + r^k O(|\lambda_3|^k).$$

Аналогично получим

$$I^{k-1} = -4\rho^2 r^{2k} \sin^2 \theta + r^{k-1} O(|\lambda_3|^k).$$

Следовательно, для модуля комплексного доминирующего собственного значения справедлива формула

$$r^2 = \frac{I^{(k)}}{I^{(k-1)}} + O \left(\left| \frac{\lambda_3}{\lambda_2} \right|^k \right) = \frac{y_i^{(k)} y_i^{(k+2)} - (y_i^{(k+1)})^2}{y_i^{(k-1)} y_i^{(k+1)} - (y_i^{(k)})^2} + O \left(\left| \frac{\lambda_3}{\lambda_2} \right|^k \right). \quad (10)$$

Далее находим

$$y^{k+2} + r^2 y^k = 2\rho r^{k+2} [\cos(\varphi + (k+2)\theta) + \cos(\varphi + k\theta)] + O(|\lambda_3|^k) = \\ = 4\rho r^{k+2} \cos(\varphi + (k+1)\theta) \cos \theta + O(|\lambda_3|^k) = 2ry^{k+1} \cos \theta + O(|\lambda_3|^k),$$

откуда вытекает соотношение

$$\cos \theta = \frac{y_i^{(k+2)} + r^2 y_i^{(k)}}{2ry_i^{(k+1)}} + O \left(\left| \frac{\lambda_3}{\lambda_2} \right|^k \right). \quad (11)$$

Окончательно для вычисления комплексной пары собственных значений получим формулы

$$\lambda_1 = r(\cos \theta + i \sin \theta), \quad \lambda_2 = r(\cos \theta - i \sin \theta),$$

где величины r и $\cos \theta$ определяются, согласно (10), (11), с точностью до малых слагаемых.

По найденным собственным значениям λ_1 и λ_2 нетрудно определить соответствующие им собственные векторы. В силу (3) имеем

$$y^k = c_1 \lambda_1^k x_1 + c_2 \lambda_2^k x_2 + O(|\lambda_3|^k), \\ y^{k+1} = c_1 \lambda_1^{k+1} x_1 + c_2 \lambda_2^{k+1} x_2 + O(|\lambda_3|^{k+1}),$$

откуда получим равенства

$$y^{k+1} - \lambda_2 y^k = c_1 \lambda_1^k (\lambda_1 - \lambda_2) \left[x_1 + O \left(\left| \frac{\lambda_3}{\lambda_1} \right|^k \right) \right],$$

$$y^{k+1} - \lambda_1 y^k = c_2 \lambda_2^k (\lambda_2 - \lambda_1) \left[x_2 + O \left(\left| \frac{\lambda_3}{\lambda_1} \right|^k \right) \right].$$

Таким образом, векторы $y^{k+1} - \lambda_2 y^k$ и $y^{k+1} - \lambda_1 y^k$ при больших k можно взять в качестве собственных векторов, соответствующих комплексно-сопряженным собственным значениям λ_1 и λ_2 .

До сих пор при конструировании различных вариантов степенного метода мы предполагали, что исходная матрица имеет линейные элементарные делители. В заключение этого пункта на простейшем примере рассмотрим изменения, возникающие в том случае, когда доминирующему собственному значению соответствует нелинейный элементарный делитель.

Пусть λ_1 вещественно и принадлежит в канонической форме блоку Жордана $J_2 = \begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_1 \end{pmatrix}$, а остальными собственными значениями $\lambda_2, \lambda_3, \dots, \lambda_n$ соответствуют линейные элементарные делители, причем

$$|\lambda_1| > |\lambda_3| \geq |\lambda_4| \geq \dots \geq |\lambda_n|.$$

Возьмем корневой базис Жордана x_1, x_2, \dots, x_n , для векторов которого справедливы соотношения

$$Ax_1 = \lambda_1 x_1, \\ Ax_2 = \lambda_1 x_2 + x_1, \\ Ax_3 = \lambda_3 x_3, \\ \dots \\ Ax_n = \lambda_n x_n.$$

Отсюда вытекают равенства

$$A^k x_1 = \lambda_1^k x_1, \\ A^k x_2 = \lambda_1^k x_2 + k \lambda_1^{k-1} x_1, \\ A^k x_3 = \lambda_3^k x_3, \\ \dots \\ A^k x_n = \lambda_n^k x_n. \quad (12)$$

Расложим вектор y^0 по векторам корневого базиса Жордана:

$$y^0 = c_1 x_1 + c_2 x_2 + \dots + c_n x_n. \quad (13)$$

Из (13) с учетом (12) получим выражение

$$y^k = \lambda_1^k (c_1 x_1 + c_2 x_2) + c_2 k \lambda_1^{k-1} x_1 + c_3 \lambda_3^k x_3 + \dots + c_n \lambda_n^k x_n. \quad (14)$$

Следовательно, для отношения компонент двух соседних итераций справедливо равенство

$$\frac{y_i^{(k+1)}}{y_i^{(k)}} = \lambda_1 \left[1 + O\left(\frac{1}{k}\right) \right]. \quad (15)$$

Формула (15) показывает, что отношение $\frac{y_i^{(k+1)}}{y_i^{(k)}}$ стремится к собственному значению λ_1 при $k \rightarrow \infty$. Однако эта сходимость намного медленнее, чем сходимость, соответствующая случаю линейных элементарных делителей, из-за наличия множителя k во втором слагаемом (14). Вычислить доминирующее собственное значение из (15) практически не представляется возможным.

Видоизменим алгоритм таким образом, чтобы искать не само собственное значение λ_1 , а коэффициенты $p = -2\lambda_1$ и $q = \lambda_1^2$ квадратного уравнения

$$\lambda^2 + p\lambda + q = 0,$$

кратным корнем которого является λ_1 . Запишем разложения

$$\begin{aligned} y^{k-1} &= \lambda_1^{k-1}(c_1x_1 + c_2x_2) + c_2(k-1)\lambda_1^{k-2}x_1 + \sum_{s=3}^n c_s\lambda_s^{k-1}x_s, \\ y^{k+1} &= \lambda_1^{k+1}(c_1x_1 + c_2x_2) + c_2(k+1)\lambda_1^kx_1 + \sum_{s=3}^n c_s\lambda_s^{k+1}x_s. \end{aligned}$$

Тогда имеем

$$\begin{aligned} y^{k+1} + py^k + qy^{k-1} &= (c_1x_1 + c_2x_2)\lambda_1^{k-1}(\lambda_1^2 + p\lambda_1 + q) + \\ &+ c_2x_1\lambda_1^{k-2}[(k+1)\lambda_1^2 + pk\lambda_1 + q(k-1)] + O(|\lambda_3|^k) = O(|\lambda_3|^k). \end{aligned}$$

Таким образом, справедливы приближенные равенства

$$\begin{aligned} y_i^{(k+1)} + py_i^{(k)} + qy_i^{(k-1)} &= O(|\lambda_3|^k), \\ y_j^{(k+1)} + py_j^{(k)} + qy_j^{(k-1)} &= O(|\lambda_3|^k). \end{aligned}$$

Отсюда получим формулы для коэффициентов p и q :

$$p = -\frac{y_i^{(k-1)}y_j^{(k+1)} - y_j^{(k-1)}y_i^{(k+1)}}{y_i^{(k-1)}y_j^{(k)} - y_j^{(k-1)}y_i^{(k)}} + O\left(\frac{|\lambda_3|^k}{|\lambda_1|}\right), \quad (16)$$

$$q = \frac{y_i^{(k)}y_j^{(k+1)} - y_j^{(k)}y_i^{(k+1)}}{y_i^{(k-1)}y_j^{(k)} - y_j^{(k-1)}y_i^{(k)}} + O\left(\frac{|\lambda_3|^k}{|\lambda_1|}\right). \quad (17)$$

Очевидно, что для нахождения доминирующего собственного значения достаточно определить один из коэффициентов p или q . Вычисление второго коэффициента необходимо для контроля справедливости предположения о вхождении λ_1 в жорданов блок J_2 .

В качестве собственного вектора x_1 , соответствующего собственному значению λ_1 , можно приблизенно взять вектор $y^{k+1} - \lambda_1y^k$. Действительно, в силу (14) имеем

$$y^{k+1} - \lambda_1y^k = c_2\lambda_1^kx_1 + O\left(|\lambda_3|^k\right) = c_2\lambda_1^k[x_1 + O\left(\frac{|\lambda_3|^k}{|\lambda_1|}\right)].$$

Рассмотрим теперь вектор $x = c_1x_1 + c_2x_2$, являющийся проекцией начального вектора y^0 на корневое подпространство векторов, соответствующее собственному значению λ_1 . Для этого вектора аналогично получим соотношение

$$(k+1)\lambda_1y^k - ky^{k+1} = \lambda_1^{k+1}[x + O\left(k\left|\frac{\lambda_3}{\lambda_1}\right|^{k+1}\right)],$$

откуда следует, что с точностью до постоянного множителя

$$x \approx (k+1)\lambda_1y^k - ky^{k+1}.$$

Таким образом, по найденному вектору x можно определить соответствующий λ_1 корневой вектор x_2 с точностью до слагаемого, пропорционального собственному вектору x_1 :

$$c_2x_2 = x - c_1x_1.$$

Замечание 1. Если доминирующее собственное значение λ_1 вещественное и простое, то отношение $\left|\frac{\lambda_2}{\lambda_1}\right|$ близко к единице, то, как уже отмечалось, формула

$$\frac{y_i^{(k+1)}}{y_i^{(k)}} = \lambda_1 + O\left(\frac{|\lambda_3|^k}{|\lambda_1|}\right)$$

становится практически неприменимой вследствие низкой скорости сходимости итерационного процесса. В этом случае также может быть использован аналогичный подход, связанный с вычислением коэффициентов p и q по формулам

(16), (17). Корень соответствующего квадратного уравнения будут уже вещественными и близкими по модулю. Решая это уравнение, определим λ_1 и λ_2 с точностью до величин порядка $O\left(\frac{|\lambda_3|^k}{|\lambda_2|}\right)$.

Замечание 2. Рассмотренный выше метод может быть обобщен для нахождения нескольких вещественных или комплексных собственных значений. Однако следует учесть, что при этом собственные значения определяются путем решения алгебраических (в простейшем случае квадратных) уравнений. Если эти собственные значения одного знака и близки по модулю друг к другу (а именно в таких условиях обычный метод сходится медленно), то потеря точности при их вычислении неизбежна вследствие неустойчивости процесса. Эта неустойчивость связана не только с характером чувствительности многочлена к малым возмущениям коэффициентов, но также и с возможной плохой обусловленностью уравнений, по которым определяются сами коэффициенты. Поэтому для одновременного вычисления нескольких близких собственных значений необходимо применять более устойчивые итерационные методы.

2. Ускорение сходимости степенного метода. Изложим некоторые способы ускорения сходимости степенного метода, весьма полезные в приложениях. Сначала рассмотрим случай использования скалярных произведений, особенно эффективный в применении к симметричным матрицам.

Пусть $A = A^T$ и все собственные значения матрицы A простые:

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|.$$

Зададим произвольный пеенуловый вектор y^0 и построим итерационную последовательность

$$y^k = Ay^{k-1} = \dots = A^ky^0, \quad k = 1, 2, \dots \quad (18)$$

Наряду с (18) образуем также последовательности скалярных произведений (y^k, y^k) и (y^k, y^{k-1}) . Тогда можно утверждать, что справедливо соотношение

$$\frac{(y^k, y^k)}{(y^k, y^{k-1})} = \lambda_1 + O\left(\frac{|\lambda_2|^k}{|\lambda_1|}\right) \quad (19)$$

и, следовательно,

$$\lim_{k \rightarrow \infty} \frac{(y^k, y^k)}{(y^k, y^{k-1})} = \lambda_1.$$

Действительно, в силу сделанных предположений система собственных векторов матрицы A образует ортонормированный базис в пространстве \mathbb{R}^n . С учетом равенства $A^T x_s = \lambda_s^k x_s$, $s = 1, 2, \dots, n$, запишем разложения по векторам этого базиса:

$$y^0 = c_1x_1 + c_2x_2 + \dots + c_nx_n,$$

$$y^{k-1} = c_1\lambda_1^{k-1}x_1 + c_2\lambda_2^{k-1}x_2 + \dots + c_n\lambda_n^{k-1}x_n,$$

$$y^k = c_1\lambda_1^kx_1 + c_2\lambda_2^kx_2 + \dots + c_n\lambda_n^kx_n.$$

Тогда для скалярных произведений получим

$$(y^k, y^k) = \sum_{s=1}^n c_s^2 \lambda_s^{2k} = c_1^2 \lambda_1^{2k} \left[1 + \sum_{s=2}^n \left(\frac{c_s}{c_1} \right)^2 \left(\frac{\lambda_s}{\lambda_1} \right)^{2k} \right],$$

$$(y^k, y^{k-1}) = \sum_{s=1}^n c_s^2 \lambda_s^{2k-1} = c_1^2 \lambda_1^{2k-1} \left[1 + \sum_{s=2}^n \left(\frac{c_s}{c_1} \right)^2 \left(\frac{\lambda_s}{\lambda_1} \right)^{2k-1} \right],$$

откуда сразу следует искомое утверждение.

Из формулы (19) видно, что использование скалярных произведений для вычисления доминирующего собственного значения симметричной матрицы позволяет почти вдвое уменьшить число итераций по сравнению с обычным степенным методом.

Еще одним способом ускорения сходимости последовательностей, возникающих при использовании степенного метода, является так называемый δ^2 -процесс Эйткена. Этот процесс может применяться не только при решении задач линейной алгебры. Основное назначение метода связано с улучшением сходимости любых итерационных последовательностей, в которых погрешность убывает по закону, близкому к геометрическому.

Сначала изложим некоторые теоретические предпосылки метода. Пусть какой-либо итерационный процесс сходится линейно со скоростью геометрической прогрессии к решению S :

$$s_k \approx S + aq^k, \quad q \in (0, 1), \quad a \neq 0, \quad k = 1, 2, \dots \quad (20)$$

Составим уравнения для трех последовательных итераций:

$$s_{k-1} = S + aq^{k-1}, \quad s_k = S + aq^k, \quad s_{k+1} = S + aq^{k+1}$$

(эти равенства, очевидно, следует рассматривать как приближенные) и вычислим соотношения

$$\begin{aligned}s_{k+1} &= Saq^{k-1}(q-1)^2, \\ s_{k+1} - 2s_k + s_{k-1} &= aq^{k-1}(q-1)^2 \neq 0,\end{aligned}$$

откуда находим

$$S = \frac{s_{k+1}s_{k-1} - s_k^2}{s_{k+1} - 2s_k + s_{k-1}}. \quad (21)$$

Метод Эйткена ускорения сходимости заключается в вычислении по значениям x_{k-1}, x_k, x_{k+1} нового приближения

$$x_k = \frac{s_{k+1}s_{k-1} - s_k^2}{s_{k+1} - 2s_k + s_{k-1}}. \quad (22)$$

Если бы равенство (21) выполнялось точно, то σ_k совпало бы с точным решением S . В общем случае σ_k не совпадает с S , но дает существенно лучшее приближение к S , чем очередная итерация s_{k+1} последовательности $\{s_k\}$. Заметим, что главным предположением при использовании преобразования Эйткена (22) является требование геометрической скорости сходимости основного итерационного процесса. В случае методов, имеющих более высокую скорость сходимости, данный подход малоэффективен.

Рассмотрим теперь применение δ^2 -процесса Эйткена к решению задачи ускорения сходимости степенного метода. Предположим, что

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \dots \geq |\lambda_n|.$$

В предыдущем пункте была получена формула для вычисления собственного значения λ_1 :

$$\frac{y_i^{(k+1)}}{y_i^{(k)}} = \lambda_1^{(k)} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right), \quad (23)$$

из которой следует, что последовательность приближений к λ_1 изменяется по закону, близкому к геометрической прогрессии. Этот факт позволяет для последовательности $\{\lambda_1^{(k)}\}$ применить процедуру ускорения сходимости

$$u_k = \frac{\lambda_1^{(k+1)}\lambda_1^{(k-1)} - (\lambda_1^{(k)})^2}{\lambda_1^{(k+1)} - 2\lambda_1^{(k)} + \lambda_1^{(k-1)}}. \quad (24)$$

Можно показать, что справедливо соотношение

$$u_k = \lambda_1 + O\left(\left|\frac{\lambda_3}{\lambda_1}\right|^k\right) + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right),$$

т.е. погрешность в определении доминирующего собственного значения по формуле (24) может быть существенно меньше, чем при вычислении его непосредственно из последовательности (23).

В предположении, что собственное значение λ_1 вычислено с хорошей точностью, применим δ^2 -процесс Эйткена к определению доминирующего собственного вектора. Правило уточнения собственного вектора построено таким образом, чтобы каждая компонента уточнялась отдельно. С этой целью умножим i -ю компоненту векторов y^{k-1} , y^k и y^{k+1} на λ_1 , 1 и λ_1^{-1} соответственно:

$$\begin{aligned}\lambda_1 y_i^{(k-1)} &= c_1 x_{i1} \lambda_1^k + c_2 x_{i2} \lambda_2^{k-1} \lambda_1 + \dots + c_n x_{in} \lambda_n^{k-1} \lambda_1, \\ y_i^{(k)} &= c_1 x_{i1} \lambda_1^k + c_2 x_{i2} \lambda_2^k + \dots + c_n x_{in} \lambda_n^k, \\ \frac{y_i^{(k+1)}}{\lambda_1} &= c_1 x_{i1} \lambda_1^k + c_2 x_{i2} \frac{\lambda_2^{k+1}}{\lambda_1} + \dots + c_n x_{in} \frac{\lambda_n^{k+1}}{\lambda_1}.\end{aligned}$$

Введем обозначения $b_s = c_s x_{i1}$, $s = 1, 2, \dots, n$. Образуем величины

$$\begin{aligned}\lambda_1^{-1} y_i^{(k+1)} \lambda_1 y_i^{(k-1)} - (y_i^{(k)})^2 &= [b_1 b_2 \lambda_1^{k-1} \lambda_2^{k-1} (\lambda_1 - \lambda_2)^2 + \dots \\ &\quad + b_1 b_n \lambda_1^{k-1} \lambda_n^{k-1} (\lambda_1 - \lambda_n)^2] \left[1 + O\left(\left|\frac{\lambda_3}{\lambda_1}\right|^k\right)\right], \\ \lambda_1^{-1} y_i^{(k+1)} - 2y_i^{(k)} + \lambda_1 y_i^{(k-1)} &= \\ &= b_2 \lambda_1^{-1} \lambda_2^{k-1} (\lambda_1 - \lambda_2)^2 + \dots + b_n \lambda_1^{-1} \lambda_n^{k-1} (\lambda_1 - \lambda_n)^2\end{aligned}$$

и используем их для преобразования Эйткена. Имеем

$$v_i^{(k)} = \frac{\lambda_1^{-1} y_i^{(k+1)} \lambda_1 y_i^{(k-1)} - (y_i^{(k)})^2}{\lambda_1^{-1} y_i^{(k+1)} - 2y_i^{(k)} + \lambda_1 y_i^{(k-1)}} = c_1 \lambda_1^k x_{i1} \left[1 + O\left(\left|\frac{\lambda_3}{\lambda_1}\right|^k\right)\right]. \quad (25)$$

Сравнивая (25) с формулой (6) для i -й компоненты вектора y^k :

$$y_i^{(k)} = c_1 \lambda_1^k x_{i1} \left[1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right],$$

видим, что при сделанном предположении $|\lambda_3| < |\lambda_2|$ скорость сходимости последовательности $v_i^{(k)}$ к пределу x_{i1} выше, чем последовательности $y_i^{(k)}$.

3. Метод λ -разности. Рассмотрим теперь общую идею, позволяющую в предположении

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \dots \geq |\lambda_n|$$

находить собственное значение λ_2 после вычисления λ_1 . Наряду с последовательностью

$$y^k = A^k y_0 \quad (26)$$

образуем последовательность

$$\Delta y^m = y^{m+1} - \lambda_1 y^m. \quad (27)$$

Тогда для собственного значения λ_2 справедливы формулы

$$\frac{\Delta y_i^{(m)}}{\Delta y_i^{(m-1)}} = \lambda_2 + O\left(\left|\frac{\lambda_3}{\lambda_2}\right|^m\right), \quad \lim_{m \rightarrow \infty} \frac{\Delta y_i^{(m)}}{\Delta y_i^{(m-1)}} = \lambda_2. \quad (28)$$

Соотношения (28) непосредственно вытекают из разложения

$$\begin{aligned}\Delta y^m &= y^{m+1} - \lambda_1 y^m = c_2 \lambda_2^m (\lambda_2 - \lambda_1) x_2 + c_3 \lambda_3^m (\lambda_3 - \lambda_1) x_3 + \dots \\ &\dots + c_n \lambda_n^m (\lambda_n - \lambda_1) x_n = c_2 \lambda_2^m (\lambda_2 - \lambda_1) x_2 \left[1 + O\left(\left|\frac{\lambda_3}{\lambda_2}\right|^m\right)\right].\end{aligned}$$

В качестве собственного вектора x_2 , соответствующего λ_2 , очевидно, можно взять вектор Δy^m .

Теоретически данный подход можно применять и к вычислению следующих собственных значений матрицы, однако при этом необходимо иметь в виду одно существенное обстоятельство. Пусть доминирующее собственное значение найдено степенным методом. Тогда число k в равенстве (26) следует выбирать таким образом, чтобы можно было пренебречь слагаемыми

$$\left(\frac{\lambda_2}{\lambda_1}\right)^k, \quad \left(\frac{\lambda_3}{\lambda_1}\right)^k, \dots, \quad \left(\frac{\lambda_n}{\lambda_1}\right)^k.$$

Однако как раз слагаемые отброшенного порядка малости используются при нахождении λ_2 с помощью последовательности (27). Например, если $m = k$ и $\left|\frac{\lambda_2}{\lambda_1}\right|^k \leq \varepsilon$, то

$$\left|\frac{\lambda_3}{\lambda_2}\right|^k \leq \varepsilon \left|\frac{\lambda_1}{\lambda_2}\right|^k.$$

Это означает, что чем с большей точностью будет найдено λ_1 , тем меньшая точность гарантируется при вычислении λ_2 . Последующие собственные значения, очевидно, будут вычисляться с еще меньшей точностью.

4. Метод обратных итераций. В тех же условиях относительно матрицы A (наличие базиса из собственных векторов) и расположения собственных значений

$$|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_{n-1}| > |\lambda_n|$$

образуем последовательные итерации

$$z^k = A^{-1} z^{k-1} = \dots = (A^{-1})^k z^0, \quad k = 1, 2, \dots \quad (29)$$

Так как собственные значения ν_s матрицы A^{-1} связаны с собственными значениями λ_s матрицы A очевидным соотношением $\nu_s = \frac{1}{\lambda_s}$, то имеет место равенство

$$\nu_{\max}(A^{-1}) = \frac{1}{\lambda_{\min}(A)} = \frac{1}{\lambda_n}.$$

Поэтому для последовательности (29) получим

$$\frac{z_i^{(k+1)}}{z_i^{(k)}} = \nu_1 + O\left(\left|\frac{\nu_2}{\nu_1}\right|^k\right) \approx \frac{1}{\lambda_n}. \quad (30)$$

Алгоритм (30) вычисления наименьшего по модулю собственного значения матрицы получил название *метода обратных итераций*. Как и в случае степенного метода вычисление простого собственного значения λ_1 , метод обратных итераций сходится со скоростью геометрической прогрессии со знаменателем $\left|\frac{\nu_2}{\nu_1}\right| = \left|\frac{\lambda_2}{\lambda_{n-1}}\right|$.

Для собственного вектора x_n , соответствующего λ_n , достаточно хорошим приближением является вектор

$$z^k = \sum_{s=1}^n \frac{a_s}{\lambda_s^k} x_s = \frac{a_n}{\lambda_n^k} \left[x_n + \sum_{s=2}^n \left(\frac{a_s}{\lambda_n} \right) \left(\frac{\lambda_n}{\lambda_s} \right)^k x_s \right] \approx \frac{a_n}{\lambda_n^k} x_n,$$

где a_i — координаты вектора x^0 в базисе $\{x_s\}$.

Отметим, что при вычислении векторов z^k нет необходимости определять обратную матрицу A^{-1} . Эти векторы могут быть найдены путем последовательного решения СЛАУ вида

$$A z^k = z^{k-1}, \quad k = 1, 2, \dots$$

Задачи к главе IV

1. Пусть $A \in \mathbb{R}^{n \times n}$ — симметричная матрица. Доказать, что для ее наибольшего и наименьшего собственных значений справедливы оценки

$$-\lambda_{\min}(A) \leq \min_{1 \leq i \leq n} a_{ii}, \quad \lambda_{\max}(A) \geq \max_{1 \leq i \leq n} a_{ii}.$$

2. Показать, что для модулей $|\lambda_i|$ собственных значений матрицы $A \in \mathbb{C}^{n \times n}$ имеют место неравенства

$$\min_{x \neq 0} \left| \frac{(Ax, x)}{(x, x)} \right| \leq |\lambda_i| \leq \max_{x \neq 0} \left| \frac{(Ax, x)}{(x, x)} \right|, \quad i = 1, 2, \dots, n$$

(матрица A не обязательно кримитова).

3. Доказать, что для любых квадратных матриц A и B матрицы AB и BA имеют одинаковые характеристические многочлены.

4. Доказать, что если матрицы A и B перестановочны, т.е. $AB = BA$, то существует собственное значение $\lambda(AB)$, равное произведению собственных значений $\lambda(A)\lambda(B)$.

5. Доказать, что если A и B — перестановочные симметричные положительноподопределеные матрицы, то матрица AB положительно определена.

6. Доказать, что если A — симметричная положительно определенная матрица, B — положительно определенная матрица, то система собственных векторов матрицы AB является полной.

7. Пусть A — симметризуемая матрица, т.е. существует невырожденная матрица T такая, что $T^{-1}AT$ — симметричная матрица. Доказать, что система собственных векторов матрицы A является полной.

8. Пусть $A, B \in \mathbb{R}^{n \times n}$ — симметричные матрицы. Показать, что необходимым и достаточным условием равенства $AB = BA$ является существование базиса в \mathbb{R}^n , составленного из общих собственных векторов матриц A и B .

9. Пусть A — симметричная положительно определенная матрица. Доказать, что справедливы соотношения

$$\lambda_{\min}(A) = \min_{x \neq 0} \frac{(Ax, x)}{(x, x)}, \quad \lambda_{\max}(A) = \max_{x \neq 0} \frac{(Ax, x)}{(x, x)}.$$

10. Доказать положительную определенность матрицы

$$A = \begin{pmatrix} 12 & -6 & 3 & -2 \\ -6 & 18 & -6 & 6 \\ 3 & -6 & 24 & 15 \\ -2 & 6 & 15 & 20 \end{pmatrix}.$$

11. Пусть $A \in \mathbb{R}^{n \times n}$ — симметричная положительно определенная матрица. Показать, что если $\lambda_{\max}(A) = a_{kk}$ при некотором $1 \leq k \leq n$, то $a_{ik} = a_{kj} = 0$ при всех $i \neq k, j \neq k$.

12. Пусть матрица $A \in \mathbb{C}^{n \times n}$ является эрмитовой, $A = B + iC$. Доказать, что матрица $T = \begin{pmatrix} B & -C \\ C & B \end{pmatrix}$ симметричная. Сравнить собственные значения и собственные векторы матриц A и T .

13. Доказать, что у трехдиагональной матрицы

$$A = \begin{pmatrix} a_1 & b_1 & & & & & & 0 \\ a_2 & c_2 & b_2 & & & & & \\ a_3 & c_3 & b_3 & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & & & \ddots & & & \\ & & & & & \ddots & & \\ 0 & & & & & & a_{n-1} & c_{n-1} \\ & & & & & & a_n & c_n \end{pmatrix}$$

все собственные значения $\lambda_i(A)$ вещественны, если выполняются условия

$$a_{i+1}b_i > 0, \quad i = 1, 2, \dots, n-1.$$

14. Доказать, что любое собственное значение матрицы $A \in \mathbb{C}^{n \times n}$ лежит по крайней мере в одной из областей

$$|\lambda - a_{ii}| |\lambda - a_{jj}| \leq \sum_{k=1, k \neq i}^n |a_{ik}| \sum_{k=1, k \neq j}^n |a_{jk}|, \quad i \neq j$$

(эти области называются *областями Кассини*).

15. Пусть собственные значения симметричной матрицы $A \in \mathbb{R}^{n \times n}$ удовлетворяют соотношениям $\lambda_1 \approx 1, 1 \leq \lambda_i \leq 3, i = 2, 3, \dots, n$. Построить итерационный процесс вида

$$x^{k+1} = (A + cE)x^k, \quad c = \text{const},$$

для получения собственного значения λ_1 с наилучшей при данной информации скоростью сходимости.

ЛИТЕРАТУРА

- Бахвалов, Н. С. Численные методы / Н. С. Бахвалов, Н. П. Жидков, Г. М. Ко-бельков. М.: Наука, 1987. 600 с.
 Бахвалов, Н. С. Численные методы в задачах и упражнениях / Н. С. Бахвалов, А. В. Лапин, Е. В. Чижиков. М.: Выш. пис., 2000. 190 с.
 Воеводин, В. В. Вычислительные основы линейной алгебры / В. В. Воеводин. М.: Наука, 1977. 304 с.
 Воеводин, В. В. Матрицы и вычисления / В. В. Воеводин, Ю. А. Кузнецов. М.: Наука, 1984. 320 с.
 Гантмахер, Ф. Р. Теория матриц / Ф. Р. Гантмахер. М.: Наука, 1967. 576 с.
 Годунов, С. К. Решение систем линейных уравнений / С. К. Годунов. Новосибирск: Наука, 1980.
 Голуб, Дж. Матричные вычисления / Дж. Голуб, Ч. Ван Лоуп. М.: Мир, 1999. 548 с.
 Егоров, А. А. Вычислительные методы алгебры / А. А. Егоров. Мн.: БГУ, 1998. 74 с.
 Колапкин, И. Н. Численные методы / И. Н. Колапкин. М.: Наука, 1978. 512 с.
 Коновалов, А. И. Введение в вычислительные методы линейной алгебры / А. И. Коновалов. Новосибирск: НГУ, 1983. 84 с.
 Крылов, В. И. Вычислительные методы высшей математики. В 2 т. Т.1 / В. И. Крылов, В. В. Бобков, П. И. Монастырский. Мн.: Выш. пис., 1972. 584 с.
 Крылов, В. И. Вычислительные методы. В 2 т. Т. 1 / В. И. Крылов, В. В. Бобков, П. И. Монастырский. М.: Наука, 1976. 304 с.
 Паретт, Б. Симметрическая проблема собственных значений. Численные методы / Б. Паретт. М.: Мир, 1983. 384 с.
 Самарский, А. А. Введение в численные методы / А. А. Самарский. М.: Наука, 1987. 288 с.
 Самарский, А. А. Численные методы / А. А. Самарский, А. В. Гудин. М.: Наука, 1989. 432 с.
 Самарский, А. А. Методы решения сеточных уравнений / А. А. Самарский, Е. С. Николаев. М.: Наука, 1978. 592 с.
 Стрепка, Г. Линейная алгебра и ее применение / Г. Стрепка. М.: Мир, 1980.
 Уилькисон, Дж. Х. Алгебраическая проблема собственных значений / Дж. Х. Уилькисон. М.: Наука, 1970.
 Фаддеев, Д. К. Вычислительные методы линейной алгебры / Д. К. Фаддеев, В. Н. Фаддеева. М.; Л.: Физматлит, 1963.

БИБЛИОТЕКА
ВГУ

1824386

Учебное издание

Егоров Андрей Александрович

ВЫЧИСЛИТЕЛЬНЫЕ АЛГОРИТМЫ
ЛИНЕЙНОЙ АЛГЕБРЫ

Учебное пособие

Редактор Н. Ф. Акулич
Художник обложки Е. П. Протасеня
Технический редактор Т. К. Раманович
Корректор Г. М. Добыши

86747р.

Подписано в печать 06.10.2005. Формат 60×84/16. Бумага офсетная,
Гарнитура Roman. Печать офсетная. Усл. печ. л. 11,16. Уч.-изд. л. 9,27.

Тираж 250 экз. Зак.959

Белорусский государственный университет.
Лицензия на осуществление издательской деятельности № 02330/0056804 от 02.03.2004.

220050, Минск, проспект Независимости, 4.

Отпечатано с оригиналами-макетами заказчика
Республиканское унитарное предприятие
«Издательский центр Белорусского государственного университета».

Лицензия на осуществление полиграфической деятельности № 02330/0056850 от 30.04.2004.

220030, Минск, ул. Красноречьевская, 6.

$$\begin{aligned} \beta_{i+1} &= \frac{f_i + \beta_i a_i}{c_i - \alpha_i a_i}, \quad i = 1, 2, \dots, i_0, \quad \beta_1 = \frac{f_0}{c_0}; \\ \xi_i &= \frac{a_i}{c_i - \xi_{i+1} b_i}, \quad i = n-1, n-2, \dots, i_0+1, \quad \xi_n = \frac{a_n}{c_n}; \\ \eta_i &= \frac{f_i + \eta_{i+1} b_i}{c_i - \xi_{i+1} b_i}, \quad i = n-1, n-2, \dots, i_0+1, \quad \eta_0 = \frac{f_n}{c_n}; \\ x_i &= \alpha_{i+1} x_{i+1} + \beta_{i+1}, \quad i = i_0-1, i_0-2, \dots, 0; \\ x_{i+1} &= \xi_{i+1} x_i + \eta_{i+1}, \quad i = i_0, i_0+1, \dots, n-1; \\ x_{i_0} &= \frac{\beta_{i_0+1} + \alpha_{i_0+1} \eta_{i_0+1}}{1 - \alpha_{i_0+1} \xi_{i_0+1}}, \quad i_0 = 1, 2, \dots, n-1. \end{aligned}$$

Этот метод, очевидно, выгодно применять в том случае, когда необходимо найти только одно неизвестное x_{i_0} либо группу следующих друг за другом неизвестных $x_{i_0}, x_{i_0+1}, \dots, x_{i_0+k}$.

Если матрица исходной системы не обладает свойством диагонального доминирования, то для ее решения обычно применяется метод *немонотонной прогонки*, базирующейся на использовании схемы Гаусса с выбором главного элемента по строке.

5. Метод циклической прогонки. Остановимся более подробно на еще одной модификации алгоритма прогонки — *методе циклической прогонки*. Рассмотрим СЛАУ

$$-a_i x_{i-1} + c_i x_i - b_i x_{i+1} = f_i, \quad i = 0, \pm 1, \pm 2, \dots, \quad (15)$$

коэффициенты и правая часть которой периодичны с периодом n :

$$a_i = a_{i+n}, \quad b_i = b_{i+n}, \quad c_i = c_{i+n}, \quad f_i = f_{i+n}. \quad (16)$$

Если решение системы (15) существует, то в силу (16) оно также будет периодичным с периодом n :

$$x_i = x_{i+n}. \quad (17)$$

Задача (15) содержит бесконечное число неизвестных, однако, учитывая (17), достаточно найти решение x_i при $i = 0, 1, \dots, n-1$. В этом случае (15) можно записать следующим образом:

$$\begin{cases} -a_0 x_{n-1} + c_0 x_0 - b_0 z_1 = f_0, & i = 0, \\ -a_i x_{i-1} + c_i x_i - b_i x_{i+1} = f_i, & i = 1, 2, \dots, n-1, \\ x_n = x_0. \end{cases} \quad (18)$$

Второе из уравнений (18) при $i = n-1$ содержит неизвестное x_n . Чтобы в дальнейшем не менять структуру этого уравнения, мы добавили к системе условие $x_n = x_0$.

Получим теперь формулы метода циклической прогонки. Сразу же отметим, что матрица системы (18)

$$A = \begin{pmatrix} c_0 & -b_0 & 0 & \dots & \dots & \dots & -a_0 \\ -a_1 & c_1 & -b_1 & & & & \vdots \\ \vdots & -a_2 & c_2 & -b_2 & & 0 & \vdots \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ & 0 & & & & & \vdots \\ -b_{n-1} & \dots & \dots & \dots & \dots & -a_{n-2} & -b_{n-2} \\ & & & & & & c_{n-1} \end{pmatrix}$$

не является трехдиагональной ввиду наличия ненулевых элементов $-a_0, -b_{n-1}$, что не позволяет решать эту задачу обычным методом прогонки, описанным в п. 1.

Обозначим

$$Pv_i = -a_i v_{i-1} + c_i v_i - b_i v_{i+1}, \quad i = 1, 2, \dots, n-1.$$

Пусть y_i — решение задачи

$$Py_i = f_i, \quad y_0 = 0, \quad y_n = 0, \quad (19)$$

а z_i есть решение задачи

$$Pz_i = 0, \quad z_0 = 1, \quad z_n = 1. \quad (20)$$

Будем искать теперь решение системы (18) в виде

$$x_i = y_i + x_0 z_i, \quad i = 0, 1, \dots, n. \quad (21)$$

Из (19)–(21) имеем

$$Px_i = Py_i + x_0 Pz_i = f_i,$$

и поэтому выполнено второе уравнение (18). Нетрудно убедиться, что выполнено также и третье уравнение. Осталось удовлетворить первому из уравнений (18). Подставляя в него (21), получим

$$-a_0 y_{n-1} - a_0 x_0 z_{n-1} + c_0 x_0 - b_0 y_1 = f_0,$$

откуда

$$x_0 = \frac{f_0 + a_0 y_{n-1} + b_0 y_1}{c_0 - a_0 z_{n-1} - b_0 z_1}. \quad (22)$$

Таким образом, алгоритм метода циклической прогонки заключается в следующем. Сначала решаются вспомогательные задачи (19), (20), а затем из (22) определяется x_0 , после чего из (21) находятся x_1, x_2, \dots, x_{n-1} . Поскольку матрица каждой из систем (19), (20) является трехдиагональной, для их решения может быть применен метод правой прогонки:

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}, \quad i = n-1, n-2, \dots, 1, \quad y_n = 0,$$

$$z_i = \alpha_{i+1} z_{i+1} + \gamma_{i+1}, \quad i = n-1, n-2, \dots, 1, \quad z_n = 1,$$

где прогоночные коэффициенты α_i, β_i и γ_i находятся по формулам

$$\begin{aligned} \alpha_{i+1} &= \frac{b_i}{c_i - \alpha_i a_i}, \quad i = 1, 2, \dots, n, \quad \alpha_1 = 0; \\ \beta_{i+1} &= \frac{f_i + \beta_i a_i}{c_i - \alpha_i a_i}, \quad i = 1, 2, \dots, n, \quad \beta_1 = 0; \\ \gamma_{i+1} &= \frac{\gamma_i a_i}{c_i - \alpha_i a_i}, \quad i = 1, 2, \dots, n, \quad \gamma_1 = 1. \end{aligned}$$

При этом формула (22) для определения x_0 преобразуется к виду

$$x_0 = \frac{\beta_{n+1} + \alpha_{n+1} y_1}{1 - \gamma_{n+1} - \alpha_{n+1} z_1}.$$

Теорема 2. Пусть коэффициенты задачи (18) удовлетворяют условиям

$$a_i \neq 0, \quad b_i \neq 0, \quad |c_i| \geq |a_i| + |b_i|, \quad i = 1, 2, \dots, n,$$

и существует хотя бы одно i_0 такое, что $|c_{i_0}| > |a_{i_0}| + |b_{i_0}|$. Тогда имеют место соотношения

$$c_i - \alpha_i a_i \neq 0, \quad |\alpha_i| \leq 1, \quad |\alpha_i| + |\gamma_i| \leq 1, \quad i = 2, 3, \dots, n, \quad (23)$$

$$1 - \gamma_{n+1} - \alpha_{n+1} z_1 \neq 0.$$

Доказательство. Доказательство теоремы 2 проведем только для двух последних соотношений (23), поскольку справедливость

остальных была установлена при обосновании метода правой прогонки. Имеем

$$\begin{aligned} |\alpha_{i+1}| + |\gamma_{i+1}| &= \frac{|b_i| + |a_i||\gamma_i|}{|c_i - \alpha_i a_i|} \leq \frac{|a_i| + |b_i| - |a_i|(1 - |\gamma_i|)}{|c_i - |a_i||a_i|} \leq \\ &\leq \frac{|a_i| + |b_i| - |a_i|(1 - |\gamma_i|)}{|a_i| + |b_i| - |a_i||\alpha_i|}. \end{aligned} \quad (24)$$

Если предположить выполнение $|\alpha_i| + |\gamma_i| \leq 1$, то из (24) по индукции будет следовать, что и $|\alpha_{i+1}| + |\gamma_{i+1}| \leq 1$. Так как $|\alpha_1| + |\gamma_1| = 1$, то третье из неравенств (23) доказано.

Остается показать, что $1 - \gamma_{n+1} - \alpha_{n+1} z_1 \neq 0$. Так как

$$z_{n-1} = \alpha_n z_n + \gamma_n,$$

то

$$|z_{n-1}| \leq |\alpha_n| + |\gamma_n| \leq 1,$$

и по индукции легко устанавливается справедливость неравенства

$$|z_i| \leq |\alpha_{i+1}| |z_{i+1}| + |\gamma_{i+1}| \leq |\alpha_{i+1}| + |\gamma_{i+1}| \leq 1.$$

В частности, $|z_1| \leq 1$. Далее

$$|1 - \gamma_{n+1} - \alpha_{n+1} z_1| \geq 1 - |\gamma_{n+1}| - |\alpha_{n+1}| |z_1| \geq 1 - |\gamma_{n+1}| - |\alpha_{n+1}| \geq 0.$$

Для получения строгого неравенства в последнем соотношении заметим, что для некоторого i_0 $|c_{i_0}| > |a_{i_0}| + |b_{i_0}|$ и, следовательно, $|\alpha_{i_0}| + |\gamma_{i_0}| < 1$, т. е. для $i \geq i_0$ имеет место строгое неравенство $|\alpha_i| + |\gamma_i| < 1$. Значит, $1 - |\gamma_{n+1}| - |\alpha_{n+1}| > 0$. Теорема доказана.

6. Связь метода циклической прогонки с методом окаймления. Изложенный метод циклической прогонки является одним из вариантов *метода окаймления*, основная идея которого заключается в следующем. Пусть A_n — матрица порядка $n \times n$:

$$A_n = \begin{pmatrix} A_{n-1} & u_n \\ v_n & a_{nn} \end{pmatrix}, \quad (25)$$

Задачи к главе II

1. Пусть $A \in \mathbb{R}^{n \times n}$ — строго треугольная матрица: $a_{ij} = 0$, $i \leq j$ ($i \geq j$). Доказать, что справедливо равенство $A^T = 0$.
2. Обозначим через A_k , L_k и U_k матрицы угловых миноров k -го порядка для матриц A , L и U соответственно в разложении $A = LU$. Доказать, что для всех k справедливо равенство $A_k = L_k U_k$.
3. Описать вариант исключения Гаусса, основанный на разложении $A = UL$, где U — верхняя треугольная матрица, L — нижняя унитретугольная матрица. Сформулировать условия, при которых это разложение возможно.
4. Пусть матрица $A + iB$ эрмитова и положительно определенная, причем $A, B \in \mathbb{R}^{n \times n}$. Доказать, что матрица $C \in \mathbb{R}^{2n \times 2n}$ виду

$$C = \begin{pmatrix} A & -B \\ B & A \end{pmatrix}$$

является симметричной и положительно определенной.

5. Для симметричной положительно определенной матрицы A построить алгоритм вычисления верхней треугольной матрицы S такой, что $A = SS^T$.
6. Пусть A — положительно определенная матрица, $A_0 = (A + A^T)/2$. Доказать справедливость неравенства

$$\|A^{-1}\|_2 \leq \|A_0^{-1}\|_2, \quad (A^{-1}x, x) \leq (A_0^{-1}x, x) \quad \forall x \in \mathbb{R}^n.$$

7. Привести пример матрицы $A \in \mathbb{R}^{2 \times 2}$, обладающей свойством $(Ax, x) > 0$ для $x \in \mathbb{R}^2$, $x \neq 0$, но которая не является положительно определенной для комплексных векторов $x \in \mathbb{C}^2$.

8. Доказать, что если матрицы A и A^T строго диагонально доминирующие, то A положительно определена, причем диагональные элементы матрицы A положительны.

9. Доказать, что функция $f(x) = (Ax, x)^{1/2}$ является нормой вектора в \mathbb{R}^n тогда и только тогда, когда A положительно определена.

10. Для матрицы $S \in \mathbb{R}^{n \times n}$ в разложении Холецкого $A = S^T S$ проверить справедливость соотношения

$$\|S\|_2 = \|A\|_2^{1/2}, \quad \|A\|_F^{1/2} \leq \|S\|_F \leq n^{1/4} \|A\|_F^{1/2}.$$

11. Пусть матрица $A \in \mathbb{R}^{n \times n}$ имеет вид $A = E + uu^T$, $\|u\|_2 = 1$. Дать точное описание матрицы S в разложении Холецкого.

12. Адаптировать алгоритм метода квадратного корня для случая симметричной положительно определенной трехдиагональной матрицы.

13. Доказать, что первые k векторов-столбцов матрицы Q в разложении $A = QU$ § 2 для всех k представляют собой ортонормированный базис подпространства, наложенного на первые k векторов-столбцов матрицы A .

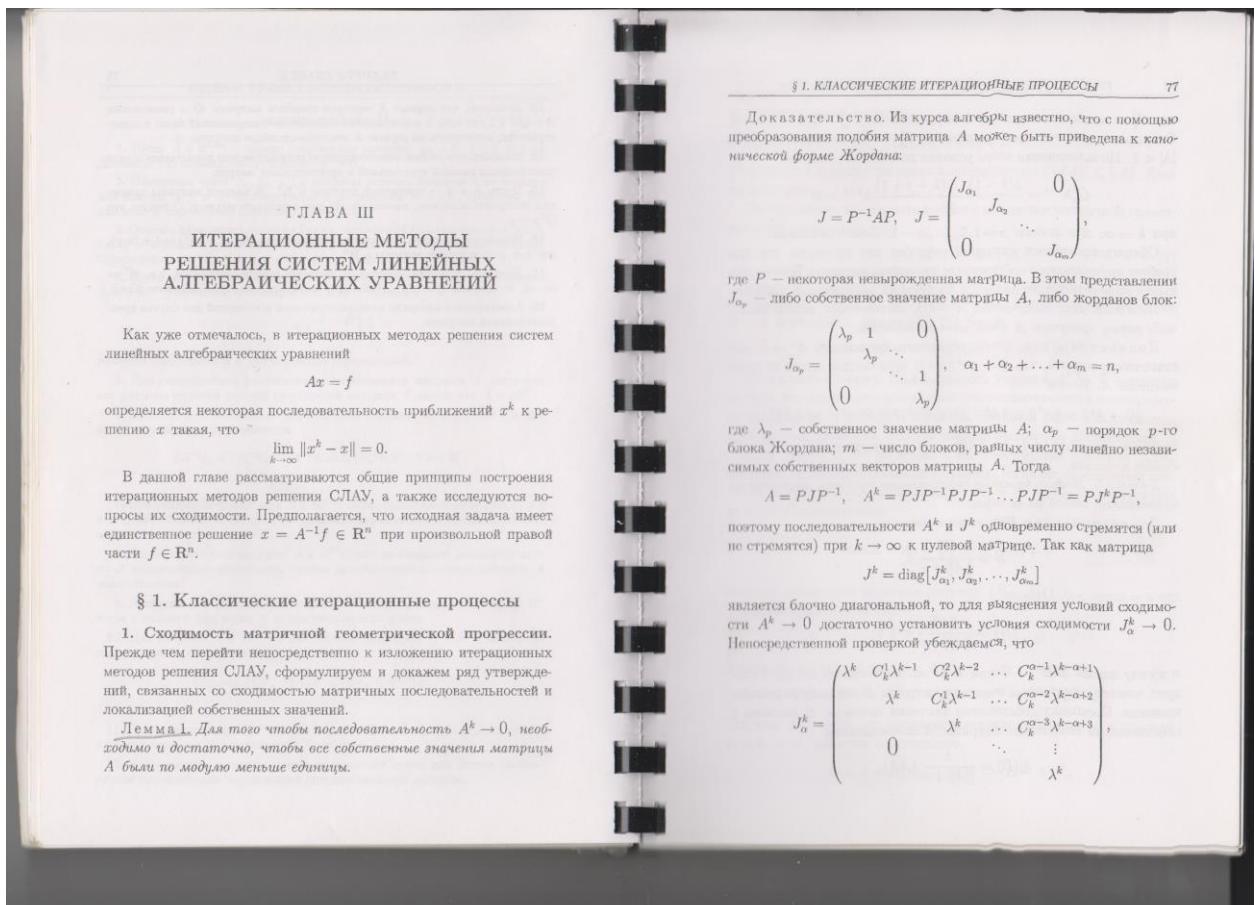
14. Доказать, что любую невырожденную матрицу можно представить в виде произведения нижней треугольной и ортогональной матриц.

15. Пусть x и y — единичные векторы в \mathbb{R}^n . Используя матрицы вращения, построить алгоритм, вычисляющий ортогональную матрицу Q такую, что $Q^T x = y$.

16. Используя матрицы отражений, показать, что $\det(E + xy^T) = 1 + (x, y)$, где x и y — заданные векторы в \mathbb{R}^n .

17. Доказать, что если матрица A является кососимметричной, т. е. $A^T = -A$, то матрица $Q = (E + A)(E - A)^{-1}$ ортогональна (преобразование Кэли).

18. Адаптировать алгоритм методов отражений и вращений для случая трехдиагональной матрицы.



§ I. КЛАССИЧЕСКИЕ ИТЕРАЦИОННЫЕ ПРОЦЕССЫ

Доказательство. Из курса алгебры известно, что с помощью преобразования подобия матрица A может быть приведена к канонической форме Жордана:

$$J = P^{-1}AP, \quad J = \begin{pmatrix} J_{\alpha_1} & & & 0 \\ & J_{\alpha_2} & & \cdot \\ & & \ddots & \\ 0 & & & J_{\alpha_m} \end{pmatrix},$$

где P — некоторая невырожденная матрица. В этом представлении J_{α_p} — либо собственное значение матрицы A , либо жорданов блок:

$$J_{\alpha_p} = \begin{pmatrix} \lambda_p & 1 & & 0 \\ & \lambda_p & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_p \end{pmatrix}, \quad \alpha_1 + \alpha_2 + \dots + \alpha_m = n,$$

где λ_p — собственное значение матрицы A ; α_p — порядок p -го блока Жордана; m — число блоков, равных числу линейно независимых собственных векторов матрицы A . Тогда

$$A = PJP^{-1}, \quad A^k = PJP^{-1}PJP^{-1} \dots PJP^{-1} = PJ^kP^{-1},$$

поэтому последовательности A^k и J^k одновременно стремятся (или не стремятся) при $k \rightarrow \infty$ к нулевой матрице. Так как матрица

$$J^k = \text{diag}[J_{\alpha_1}^k, J_{\alpha_2}^k, \dots, J_{\alpha_m}^k]$$

является блочно диагональной, то для выяснения условий сходимости $A^k \rightarrow 0$ достаточно установить условия сходимости $J_{\alpha_i}^k \rightarrow 0$. Непосредственной проверкой убеждаемся, что

$$J_{\alpha_i}^k = \begin{pmatrix} \lambda^k & C_1^1 \lambda^{k-1} & C_2^1 \lambda^{k-2} & \dots & C_{\alpha_i}^{k-1} \lambda^{k-\alpha+1} \\ & \lambda^k & C_1^2 \lambda^{k-1} & \dots & C_{\alpha_i}^2 \lambda^{k-\alpha+2} \\ & & \lambda^k & \dots & C_{\alpha_i}^{\alpha_i-2} \lambda^{k-\alpha+3} \\ & & & \ddots & \vdots \\ 0 & & & & \lambda^k \end{pmatrix},$$

из которого и следует искомое утверждение. Теорема доказана.

С учетом лемм 1, 2 и теоремы 1 критерий сходимости матричной геометрической прогрессии можно сформулировать в другом виде.

Теорема 2. Для сходимости ряда (1) необходимо и достаточно, чтобы все собственные значения матрицы A были по модулю меньшие единицы.

Теорема 3. Если какая-либо норма матрицы A меньше единицы, то ряд (1) сходится.

При выполнении последнего признака нетрудно получить оценку скорости сходимости ряда $E + A + A^2 + \dots + A^k + \dots$

Теорема 4. Если $\|A\| < 1$, то справедливо неравенство

$$\|(E - A)^{-1} - (E + A + A^2 + \dots + A^k)\| \leq \frac{\|A\|^{k+1}}{1 - \|A\|}.$$

Доказательство. В силу теоремы 3 имеем

$$(E - A)^{-1} - (E + A + A^2 + \dots + A^k) = A^{k+1} + A^{k+2} + \dots$$

Следовательно,

$$\begin{aligned} \|(E - A)^{-1} - (E + A + A^2 + \dots + A^k)\| &\leq \|A^{k+1}\| + \|A^{k+2}\| + \dots \leq \\ &\leq \|A\|^{k+1} + \|A\|^{k+2} + \dots = \frac{\|A\|^{k+1}}{1 - \|A\|}. \end{aligned}$$

Теорема доказана.

В лемме 3 установлен тот факт, что все собственные значения матрицы A по модулю не превосходят любой ее нормы. Отметим, что легко проверяемыми являются оценки при помощи C и 1-норм. Сформулируем теорему, позволяющую более точно оценить расположение собственных значений матрицы.

Теорема 5. Любое собственное значение матрицы $A \in \mathbb{C}^{n \times n}$ находится по крайней мере в одном из кругов:

$$|z - a_{ii}| \leq \sum_{j=1, j \neq i}^n |a_{ij}|, \quad i = 1, 2, \dots, n. \quad (2)$$

Доказательство. Пусть λ — любое собственное значение матрицы A , $x = (x_1, x_2, \dots, x_n)^T$ — соответствующий ему собственный

вектор. Тогда имеем систему уравнений

$$\sum_{j=1}^n a_{ij} x_j = a_i x_i, \quad i = 1, 2, \dots, n. \quad (3)$$

Предположим, что x_s является наибольшей по модулю компонентой вектора x . Запишем s -е уравнение (3) в виде

$$\lambda - a_{ss} = \sum_{j=1, j \neq s}^n a_{sj} \frac{x_j}{x_s},$$

откуда следует, что

$$|\lambda - a_{ss}| \leq \sum_{j=1, j \neq s}^n |a_{sj}| \left| \frac{x_j}{x_s} \right| \leq \sum_{j=1, j \neq s}^n |a_{sj}|.$$

Таким образом, собственное значение λ лежит в одном из кругов вида (2). Теорема доказана.

Области (2) называются *кругами Гершгорина*. Они применяются в исследованиях, связанных с собственными значениями матриц. Приведем без доказательства теорему, дающую более детальную информацию относительно распределения собственных значений по кругам Гершгорина.

Теорема 6. Если r кругов Гершгорина образуют связную область, изолированную от остальных кругов, то в этой связной области находится ровно r собственных значений матрицы A .

2. Градиент функционала. Функционалом (вообще говоря, нелинейным), определенным в линейном пространстве L , называется вещественная числовая функция $F(x)$, аргументами которой являются векторы этого пространства*. Функционалы широко используются в линейной алгебре. В частности, функционалы являются

*Функционал $F(x)$ называется линейным, если для любых векторов $x, y \in L$ и для любых чисел $\alpha, \beta \in \mathbb{R}$ имеет место соотношение

$$F(\alpha x + \beta y) = \alpha F(x) + \beta F(y).$$

Функционал, не являющийся линейным, называется нелинейным.

норма, скалярное произведение, билинейная и квадратичная формы, определитель Грама и т. п. Введем понятие градиента функционала, необходимое при изучении вариационных методов решения систем линейных алгебраических уравнений.

Пусть вектор $x \in \mathbb{R}^n$ задан своими координатами x_1, x_2, \dots, x_n в некотором ортонормированном базисе. Тогда функционал $F(x)$ можно определить как функцию $F(x_1, x_2, \dots, x_n)$ от n переменных. Будем предполагать, что функционал $F(x)$ дифференцируем, т. е. функция $F(x_1, x_2, \dots, x_n)$ имеет непрерывные частные производные по всем аргументам. Пусть y — произвольный вектор единичной длины с координатами y_1, y_2, \dots, y_n .

Определение 1. Производной функционала $F(x)$ в точке x по направлению y называется выражение

$$\frac{\partial F(x)}{\partial y} = \lim_{t \rightarrow 0} \frac{F(x + ty) - F(x)}{t} = \frac{d}{dt} F(x + ty) \Big|_{t=0}. \quad (4)$$

Так как

$$F(x + ty) = F(x_1 + ty_1, x_2 + ty_2, \dots, x_n + ty_n),$$

то выражение (4) можно записать в виде

$$\begin{aligned} \frac{\partial F(x)}{\partial y} &= \frac{d}{dt} F(x_1 + ty_1, x_2 + ty_2, \dots, x_n + ty_n) \Big|_{t=0} = \\ &= \frac{\partial F}{\partial x_1} y_1 + \frac{\partial F}{\partial x_2} y_2 + \dots + \frac{\partial F}{\partial x_n} y_n = (z, y). \end{aligned} \quad (5)$$

Таким образом, производная функционала $F(x)$ по направлению y является линейным функционалом как функция от y .

Определение 2. Вектор z с координатами $\frac{\partial F}{\partial x_1}, \frac{\partial F}{\partial x_2}, \dots, \frac{\partial F}{\partial x_n}$ называется градиентом функционала $F(x)$ в точке x и обозначается $\text{grad } F(x)$.

Из равенства (5) в силу условия $\|y\| = 1$ получим

$$\frac{\partial F(x)}{\partial y} = \|z\| \cos(z, y)$$

откуда вытекают неравенства

$$-\|z\| \leq \frac{\partial F(x)}{\partial y} \leq \|z\|.$$

Следовательно, наибольшая скорость роста функционала $F(x)$ в точке x происходит в направлении $\text{grad } F(x)$, а наибольшая скорость убывания соответственно в направлении $-\text{grad } F(x)$.

Пример 1. Вычислим градиенты некоторых наиболее часто используемых функционалов. Для симметричной матрицы A рассмотрим

$$e(x) = (Ax, x) - 2(f, x) + c \quad (\text{функционал ошибки}).$$

По определению имеем

$$\begin{aligned} \frac{\partial e(x)}{\partial y} &= \frac{d}{dt} e(x + ty) \Big|_{t=0} = \frac{d}{dt} ((Ax, x) - 2(f, x) + c + \\ &+ 2t(Ax - f, y) + t^2(Ay, y)) \Big|_{t=0} = 2(Ax - f, y), \end{aligned}$$

откуда следует, что градиент функционала ошибки $e(x)$ равен

$$\text{grad } ((Ax, x) - 2(f, x) + c) = 2(Ax - f).$$

Далее в том же предположении относительно матрицы A вычислим градиент функционала

$$\mu(x) = \frac{(Ax, x)}{(x, x)} \quad (\text{отношение Роджерса}).$$

Используя формулу производной частного двух величин, получим

$$\begin{aligned} \frac{\partial \mu(x)}{\partial y} &= \frac{(x, x) \frac{\partial}{\partial y} (Ax, x) - (Ax, x) \frac{\partial}{\partial y} (x, x)}{(x, x)^2} = \\ &= \frac{2(x, x)(Ax, y) - 2(Ax, x)(x, y)}{(x, x)^2} = \frac{2}{(x, x)} \left(Ax - \frac{(Ax, x)}{(x, x)} x, y \right). \end{aligned}$$

Таким образом, градиент функционала $\mu(x)$ равен

$$\text{grad } \frac{(Ax, x)}{(x, x)} = \frac{2}{(x, x)} \left(Ax - \frac{(Ax, x)}{(x, x)} x \right).$$

3. Классификация итерационных процессов. Рассмотрим систему линейных алгебраических уравнений

$$Ax = f.$$

При конструировании итерационных методов решения (6) часто исходную систему приводят к эквивалентному виду:

$$x = Bx + g. \quad (7)$$

Тогда последовательность приближений x^k к решению x задачи (6) можно построить, например, по формуле

$$x^{k+1} = Bx^k + g, \quad k = 0, 1, \dots, \quad (8)$$

при этом начальное приближение x^0 выбирается, вообще говоря, произвольным.

Приведение системы (6) к виду (7) осуществляется различными способами. Пусть C — некоторая невырожденная матрица. Тогда можно записать

$$x = x + C(f - Ax).$$

В этом случае $B = E - CA$, $g = Cf$ и метод (8) принимает вид

$$x^{k+1} = x^k + C(f - Ax^k), \quad k = 0, 1, \dots$$

Если подобные преобразования проводить для каждой итерации с новой матрицей C , то приходим к итерационному методу

$$x^{k+1} = x^k + C_k(f - Ax^k), \quad k = 0, 1, \dots,$$

или

$$x^{k+1} = B_k x^k + g^k, \quad k = 0, 1, \dots \quad (9)$$

Итерационные процессы (8), когда матрица B не зависит от номера итерации, называют *стационарными*, а итерационные процессы (9) соответственно *нестационарными*.

При построении итерационных методов решения системы (6) ее можно предварительно привести также к виду

$$Px + Qx = g,$$

где $P + Q = CA$, $g = Cf$, $\det C \neq 0$. Аналогично предыдущему здесь можно построить два типа итерационных процессов:

$$Px^{k+1} + Qx^k = g, \quad k = 0, 1, \dots, \quad (10)$$

$$P_k x^{k+1} + Q_k x^k = g^k, \quad k = 0, 1, \dots \quad (11)$$

В отличие от языков итерационных методов (8) и (9), итерационные процессы (10) и (11) являются *нелинейными*, поскольку для нахождения очередного приближения x^{k+1} необходимо обращать матрицу P или P_k . Естественно, эта задача должна быть более простой, чем задача нахождения матрицы A^{-1} . В качестве матрицы P выбирается одна из простейших матриц (диагональная, треугольная, трехдиагональная, факторизованная и т. п.).

Все указанные выше итерационные методы являются, во-первых, *линейными* и, во-вторых, *одношаговыми* (или *двухшаговыми*). Именно такие методы будут предметом нашего рассмотрения. Отметим, что в общем случае можно строить нелинейные и многошаговые итерационные процессы.

4. Метод простой итерации. Будем считать, что исходная система (6) приведена к виду (7), т. е. к виду, удобному для итераций. Тогда итерационный метод (8)

$$x^{k+1} = Bx^k + g, \quad k = 0, 1, \dots,$$

обычно и называют *методом простой итерации*. Так как процедура вычисления приближения x^{k+1} линейна, то последовательность приближений x^k , $k = 1, 2, \dots$, всегда может быть построена. При этом если последовательность сходится, то она сходится к решению задачи (7). Действительно, если $x^k \rightarrow x$, то, переходя в (8) к пределу при $k \rightarrow \infty$, получим равенство $x = Bx + g$.

Выясним условия сходимости последовательности приближений, получаемых методом простой итерации.

Теорема 7. Для сходимости метода простой итерации (8) при любом начальном приближении x^0 необходимо и достаточно, чтобы все собственные значения матрицы B были по модулю меньше единицы.

Доказательство. Достаточность. Выразим произвольное приближение x^{k+1} через начальный вектор x^0 :

$$\begin{aligned} x^{k+1} &= Bx^k + g = B(Bx^{k-1} + g) + g = B^2x^{k-1} + \\ &+ (E + B)g = \dots = B^{k+1}x^0 + (E + B + B^2 + \dots + B^k)g. \end{aligned} \quad (12)$$

В силу леммы 1 имеем $B^{k+1} \rightarrow 0$, а по теореме 1

$$E + B + B^2 + \dots + B^k \rightarrow (E - B)^{-1}.$$

Тогда из равенства (12) получим

$$x^{k+1} \rightarrow (E - B)^{-1}g = x.$$

Необходимость. Пусть $x^k \rightarrow x$ при любом x^0 . Поскольку имеет место точное соотношение $x = Bx + g$, то

$$x - x^{k+1} = B(x - x^k) = B^2(x - x^{k-1}) = \dots = B^{k+1}(x - x^0).$$

Переходя в этом равенстве к пределу при $k \rightarrow \infty$ и учитывая производность вектора x^0 , получим, что $B^{k+1} \rightarrow 0$. Следовательно, согласно лемме 1, все собственные значения матрицы B по модулю меньше единицы. Теорема доказана.

Использование теоремы 7 для проверки сходимости метода простой итерации довольно затруднительно, так как требует знания спектра матрицы B . Судить о сходимости метода можно при помощи следующего утверждения.

Теорема 8. Для того чтобы метод простой итерации (8) сходился, достаточно, чтобы какая-либо норма матрицы B была меньше единицы.

Доказательство очевидно, так как если $\|B\| < 1$, то в силу леммы 3 справедливы неравенства $|\lambda_i(B)| < 1$ и по теореме 7 метод простой итерации (8) сходится.

Введенные в гл. I нормы матриц

$$\|B\|_C = \max_{1 \leq i \leq n} \sum_{j=1}^n |b_{ij}|, \quad \|B\|_1 = \max_{1 \leq i \leq n} \sum_{j=1}^n |b_{ij}|, \quad \|B\|_E = \sqrt{\sum_{i,j=1}^n b_{ij}^2}$$

позволяют сформулировать легкое проверяемые достаточные признаки сходимости, непосредственно вытекающие из теоремы 8.

Теорема 9. Метод простой итерации (8) сходится, если для элементов b_{ij} , $i, j = 1, 2, \dots, n$, матрицы B выполняется одно из следующих условий:

$$1) \sum_{j=1}^n |b_{ij}| < 1, \quad i = 1, 2, \dots, n;$$

$$2) \sum_{i=1}^n |b_{ij}| < 1, \quad j = 1, 2, \dots, n;$$

$$3) \sum_{i,j=1}^n b_{ij}^2 < 1.$$

При практическом применении итерационных методов недостаточно просто установить факт сходимости того или иного итерационного процесса. Весьма важным является вопрос о его скорости сходимости. Определим *погрешность метода* на k -й итерации $x^k = x^0$.

Теорема 10. Если какая-либо норма матрицы B , согласованная с данной нормой вектора, меньше единицы, то для погрешности метода простой итерации (8) справедлива оценка

$$\|x^k - x\| \leq \|B\|^k \|x^0\| + \frac{\|B\|^k}{1 - \|B\|} \|g\|. \quad (13)$$

Доказательство. Имеем

$$x^k = Bx^{k-1} + g = \dots = B^k x^0 + (E + B + B^2 + \dots + B^{k-1})g.$$

С другой стороны, так как $\|B\| < 1$, то

$$x = (E - B)^{-1}g = (E + B + B^2 + \dots + B^{k-1} + \dots)g.$$

и, следовательно, $x^k - x = B^k x^0 - (B^k + B^{k+1} + \dots)g$. Отсюда, переходя к нормам, получим

$$\begin{aligned} \|x^k - x\| &\leq \|B\|^k \|x^0\| + (\|B\|^k + \|B\|^{k+1} + \dots) \|g\| = \\ &= \|B\|^k \|x^0\| + \frac{\|B\|^k}{1 - \|B\|} \|g\|, \end{aligned}$$

что и требовалось доказать.

Замечание 1. Если в качестве начального приближения x^0 выбрать вектор g , оценка (13) упрощается и принимает вид

$$\|x^k\| = \|(B^{k+1} + B^{k+2} + \dots)g\| \leq \frac{\|B\|^{k+1}}{1 - \|B\|} \|g\|. \quad (14)$$

Из соотношений $x^k = Bx^{k-1} + g$, $x = Bx + g$, получим уравнение для погрешности метода

$$x^k - x = B(x^{k-1} - x). \quad (15)$$

В дальнейшем при анализе сходимости итерационных методов матрицу B будем называть *матрицей перехода от $(k-1)$ -й итерации к k -й*. В общем случае нестационарного итерационного процесса матрица перехода зависит от номера итерации.

Из (15) вытекает оценка, снимающая погрешность метода на k -й итерации с начальной погрешностью:

$$\|x^k - x\| \leq \|B\| \|x^{k-1} - x\| \leq \dots \leq \|B\|^k \|x^0 - x\|. \quad (16)$$

Неравенство (16) позволяет сделать вывод, что метод простой итерации (8) сходится со скоростью геометрической прогрессии со знаменателем $\|B\| < 1$.

Одной из важных характеристик итерационного процесса является число итераций $k_0(\varepsilon)$, обеспечивающее выполнение неравенства

$$\|x^k - x\| \leq \varepsilon \|x^0 - x\|. \quad (17)$$

Используя соотношение (16), нетрудно оценить $k_0(\varepsilon)$ в методе простой итерации (8). В самом деле условие (17) будет выполнено, если

$$\|B\|^k \leq \varepsilon.$$

Последнее требование приводит к оценке минимального числа итераций, необходимых для достижения заданной точности ε :

$$k \geq k_0(\varepsilon) = \frac{\ln(1/\varepsilon)}{\ln(1/\|B\|)}. \quad (18)$$

Величина $\ln(1/\|B\|)$ называется *скоростью сходимости итерационного метода*. Отметим, что скорость сходимости полностью определяется свойствами матрицы перехода B и не зависит ни от начального приближения x_0 , ни от заданной точности ε .

5. Способы приведения СЛАУ к виду, удобному для итераций. Как уже отмечалось, один из способов приведения исходной

системы к виду, удобному для итераций, состоит в преобразовании задачи (6) к эквивалентной системе уравнений $x = x + C(f - Ax)$, где C – некоторая невырожденная матрица. Остановимся на наиболее часто встречающихся способах выбора матрицы C . Отметим, что если бы $C = A^{-1}$, то мы сразу бы получили точное решение задачи (6). Поэтому иногда подбор матрицы C осуществляют путем “грубого” обращения матрицы A (например, с помощью метода Гаусса). Такой подход очевидно, связан с большим объемом вычислительной работы. На практике исходную матрицу A представляют в виде суммы двух матриц $A = P + Q$, где матрица P обращается сравнительно легко. Тогда система (6) принимает вид

$$x = x + P^{-1}(f - (P + Q)x) = -P^{-1}Qx + P^{-1}f,$$

а метод простой итерации записывается следующим образом:

$$x^{k+1} = -P^{-1}Qx^k + P^{-1}f. \quad (19)$$

Чаше всего указанный способ используют для систем с матрицей A , имеющей строгое диагональное доминирование. Разделив каждое уравнение системы (6) на диагональный элемент, получим

$$\begin{aligned} x_1 + \frac{a_{12}}{a_{11}} x_2 + \dots + \frac{a_{1n}}{a_{11}} x_n &= \frac{f_1}{a_{11}}, \\ \frac{a_{21}}{a_{22}} x_1 + \dots + \frac{a_{2n}}{a_{22}} x_n &= \frac{f_2}{a_{22}}, \\ \vdots & \\ \frac{a_{n1}}{a_{nn}} x_1 + \dots + \frac{a_{nn}}{a_{nn}} x_n &= \frac{f_n}{a_{nn}}. \end{aligned} \quad (20)$$

Система (20), очевидно, записывается в виде $x = Bx + g$, где

$$B = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \dots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \dots & -\frac{a_{2n}}{a_{22}} \\ \vdots & & & \vdots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \dots & 0 \end{pmatrix}, \quad g = \begin{pmatrix} \frac{f_1}{a_{11}} \\ \frac{f_2}{a_{22}} \\ \vdots \\ \frac{f_n}{a_{nn}} \end{pmatrix}.$$

Соответствующий итерационный процесс $x^{k+1} = Bx^k + g$ называется *методом Якоби*. Этот метод является частным случаем метода (19) с матрицей $P = D$, где $D = \text{diag}[a_{11}, a_{22}, \dots, a_{nn}]$. Нетрудно сформулировать критерий сходимости метода Якоби, вытекающий из теоремы сходимости метода простой итерации.

Теорема 11. *Итерационный метод Якоби сходится при любом начальном приближении x^0 тогда и только тогда, когда все собственные значения матрицы $B = E - D^{-1}A$ по модулю меньше единицы.*

Таким образом, сходимость метода Якоби связана с расположением корней уравнения $\det(B - \lambda E) = 0$. Несложные преобразования приводят к следующей эквивалентной форме этого уравнения:

$$\det(A - D + \lambda D) = \begin{vmatrix} \lambda a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & \lambda a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & \lambda a_{nn} \end{vmatrix} = 0.$$

В случае симметричной матрицы A , имеющей положительные диагональные элементы, критерии сходимости метода Якоби можно придать легко проверяемую форму. Определим матрицу $2D - A$, отличающуюся от матрицы A знаками недиагональных элементов.

Теорема 12. *Для того чтобы метод Якоби решения системы $Ax = f$ с симметричной матрицей A , имеющей положительные диагональные элементы, сходился при любом начальном приближении x^0 , необходимо и достаточно, чтобы матрицы A и $2D - A$ были положительно определенными.*

Доказательство. Рассмотрим матрицу

$$D^{-1}A = D^{-1/2}(D^{-1/2}AD^{-1/2})D^{1/2}.$$

Собственные значения матрицы $D^{-1}A$, очевидно, совпадают с собственными значениями симметричной матрицы $D^{-1/2}AD^{-1/2}$. Значит, все собственные значения матрицы $D^{-1}A$ вещественны. Согласно теореме 11, для сходимости метода Якоби необходимо и достаточно выполнения условия

$$|\lambda_i(E - D^{-1}A)| = |1 - \lambda_i(D^{-1}A)| < 1,$$

откуда следует, что

$$0 < \lambda_i(D^{-1}A) < 2, \quad 0 < \lambda_i(2E - D^{-1}A) < 2.$$

Таким образом, собственные значения матриц $D^{-1}A$ и $2E - D^{-1}A$ положительны. Последнее равносильно положительной определенности матриц A и $2D - A$. Теорема доказана.

Сформулированный в предыдущем пункте достаточный признак сходимости метода простой итерации в случае метода Якоби принимает более конкретный вид.

Теорема 13. *Метод Якоби для системы $Ax = f$ сходится, если элементы a_{ij} , $i, j = 1, 2, \dots, n$, матрицы A удовлетворяют одному из следующих условий:*

- 1) $\sum_{j=1, j \neq i}^n \left| \frac{a_{ij}}{a_{ii}} \right| < 1, \quad i = 1, 2, \dots, n;$
- 2) $\sum_{i=1, i \neq j}^n \left| \frac{a_{ij}}{a_{ii}} \right| < 1, \quad j = 1, 2, \dots, n;$
- 3) $\sum_{i,j=1}^n \left(\frac{a_{ij}}{a_{ii}} \right)^2 < 1.$

Указанные в теореме 13 условия фактически означают, что матрица A должна быть строго диагонально доминирующей. Это требование можно несколько ослабить.

Определение 3. Матрица $A \in \mathbb{R}^{n \times n}$ называется матрицей со *слабым диагональным доминированием*, если выполняются условия

$$|a_{ii}| \geq \sum_{j=1, j \neq i}^n |a_{ij}|, \quad i = 1, 2, \dots, n,$$

причем хотя бы для одного i имеет место строгое неравенство.

Теорема 14. *Пусть A – неразложимая матрица⁸ со слабым диагональным доминированием. Тогда метод Якоби решения системы $Ax = f$ сходится.*

⁸Матрица называется неразложимой, если симметричной перестановкой (перестановкой строк и столбцов с одинаковыми номерами) она не может быть приведена к блочно треугольному виду.

Быстрая сходимость итерационного метода Якоби, как видим, обеспечивается наличием у матрицы A существенного диагонального доминирования. Если это условие не выполняется, то в качестве матрицы P целесообразно выделять не чисто диагональную, а блочно диагональную матрицу, например, вида

$$P = \begin{pmatrix} a_{11} & a_{12} & & & & 0 \\ a_{21} & a_{22} & & & & \\ & & a_{33} & a_{34} & & \\ & & a_{43} & a_{44} & & \\ & & & & \ddots & \\ 0 & & & & & a_{n-1,n-1} & a_{n-1,n} \\ & & & & & a_{n,n-1} & a_{nn} \end{pmatrix}.$$

Обращение матрицы P в этом случае не представляет сложности, так как сводится к обращению матрицы второго порядка.

С целью еще большего упрощения преобразований в (19) в качестве матрицы P можно выделить скалярную матрицу $P = \frac{1}{\tau} E$, $\tau = \text{const} \neq 0$. В этом случае

$$C = P^{-1} = \tau E, \quad B = E - \tau A.$$

Если матрица A положительно определена, то можно указать такое τ , что метод простой итерации (19) будет заведомо сходящимся. Для этого достаточно положить $\tau = \|A\|^{-1}$. В самом деле тогда

$$B = E - \frac{1}{\|A\|} A, \quad \lambda_i(B) = 1 - \frac{\lambda_i(A)}{\|A\|} > 0,$$

а поскольку $|\lambda_i(A)| \leq \|A\|$, то $0 \leq \lambda_i(B) < 1$ и выполнены условия теоремы 7 о сходимости метода простой итерации. Более детальный анализ последнего метода с точки зрения общей теории одншаговых итерационных процессов будет дан в следующем параграфе.

6. Метод Зейделя. Пусть исходная система (6) приведена к виду, удобному для итераций, и для нее покомпонентно записан метод

простой итерации

$$x_i^{(k+1)} = \sum_{j=1}^n b_{ij} x_j^{(k)} + g_i, \quad i = 1, 2, \dots, n.$$

Очевидно, что компоненты вектора x^{k+1} можно находить в любом порядке и независимо друг от друга, т. е. метод простой итерации является методом параллельного типа. Изменим алгоритм так, чтобы он позволял использовать при вычислении последующих компонент вектора x^{k+1} уже найденные компоненты этого вектора. Будем вести вычисление последовательных приближений по формулам

$$x_i^{(k+1)} = \sum_{j=1}^{i-1} b_{ij} x_j^{(k+1)} + \sum_{j=i}^n b_{ij} x_j^{(k)} + g_i, \quad i = 1, 2, \dots, n. \quad (21)$$

Построенный итерационный процесс называется *методом Зейделя* и его реализация состоит в следующем. Пусть известен вектор $x^k = (x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)})^T$. Из первого уравнения

$$x_1^{(k+1)} = \sum_{j=1}^n b_{1j} x_j^{(k)} + g_1$$

находим $x_1^{(k+1)}$ и подставим его во все оставшиеся уравнения. Затем из преобразованного второго уравнения

$$x_2^{(k+1)} = b_{21} x_1^{(k+1)} + \sum_{j=2}^n b_{2j} x_j^{(k)} + g_2$$

определим $x_2^{(k+1)}$ и т. д. Такая организация вычислений позволяет отнести метод Зейделя к методам последовательного типа.

Введем в рассмотрение треугольные матрицы

$$F = \begin{pmatrix} 0 & & & 0 \\ b_{21} & 0 & & \\ \vdots & & \ddots & \\ b_{n1} & \dots & b_{n,n-1} & 0 \end{pmatrix}, \quad R = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{22} & & & b_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & & & b_{nn} \end{pmatrix}.$$

В этих обозначениях метод (21) можно записать в матричной форме:

$$x^{k+1} = Fx^{k+1} + Rx^k + g, \quad k = 0, 1, \dots \quad (22)$$

Отсюда следует, что

$$(E - F)x^{k+1} = Rx^k + g, \quad k = 0, 1, \dots,$$

т. е. метод Зейделя представляет собой неявный одншаговый итерационный метод (10) с матрицей

$$P = E - F = \begin{pmatrix} 1 & & & 0 \\ -b_{21} & 1 & & \\ \vdots & & \ddots & \\ -b_{n1} & \dots & -b_{n,n-1} & 1 \end{pmatrix}.$$

Так как P — невырожденная матрица ($\det P = 1$), то метод Зейделя принимает вид

$$x^{k+1} = (E - F)^{-1} Rx^k + (E - F)^{-1} g.$$

Таким образом, метод (22) эквивалентен методу простой итерации, примененному к системе

$$x = (E - F)^{-1} Rx + (E - F)^{-1} g.$$

Этот факт позволяет сформулировать следующий критерий сходимости метода Зейделя.

Теорема 15. Для сходимости метода Зейделя (22) при любом начальном приближении x^0 необходимо и достаточно, чтобы все собственные значения матрицы $S = (E - F)^{-1} R$ были по модулю меньше единицы.

Нетрудно указать уравнение, корни которого совпадают с корнями уравнения $\det(S - \lambda E) = 0$. Имеем

$$\det[(E - F)^{-1} R - \lambda E] = \det[(E - F)^{-1}(R - \lambda(E - F))] =$$

$$= \det(E - F)^{-1} \cdot \det[R - \lambda(E - F)] = \det[R + \lambda F - \lambda E].$$

Таким образом, справедливо следующее утверждение.

Теорема 16. Для сходимости метода Зейделя (22) при любом начальном приближении x^0 необходимо и достаточно, чтобы все корни уравнения

$$\det(R + \lambda F - \lambda E) = 0 \quad (23)$$

были по модулю меньше единицы.

Сравнение полученного уравнения $\det(R + \lambda F - \lambda E) = 0$ с аналогичным уравнением $\det(B - \lambda E) = 0$ для метода простой итерации свидетельствует о том, что области сходимости методов Зейделя и простой итерации, вообще говоря, различны.

Как и в случае метода простой итерации, для метода Зейделя имеет место достаточный признак, условия сходимости которого формулируются непосредственно через элементы матрицы B .

Теорема 17. Для сходимости метода Зейделя (22) достаточно выполнения одного из условий:

$$1) \sum_{j=1}^n |b_{ij}| < 1, \quad i = 1, 2, \dots, n;$$

$$2) \sum_{i=1}^n |b_{ij}| < 1, \quad j = 1, 2, \dots, n.$$

Доказательство. Пусть, например, выполняется условие 1). Возьмем значение λ , для которого $|\lambda| \geq 1$. Рассмотрим при таком λ сумму модулей недиагональных элементов любой строки матрицы

$$G = R + \lambda F - \lambda E = \begin{pmatrix} b_{11} - \lambda & b_{12} & \dots & b_{1n} \\ \lambda b_{21} & b_{22} - \lambda & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda b_{n1} & \lambda b_{n2} & \dots & b_{nn} - \lambda \end{pmatrix}.$$

Для $i = 1, 2, \dots, n$ имеем

$$|\lambda||b_{1i}| + \dots + |\lambda||b_{i,i-1}| + |b_{i,i+1}| + |b_{in}| \leq |\lambda| \left(\sum_{j=1}^n |b_{ij}| - |b_{ii}| \right) <$$

$$< |\lambda|(1 - |b_{ii}|) = |\lambda| - |\lambda||b_{ii}| \leq |\lambda| - |b_{ii}| \leq |\lambda - b_{ii}|.$$

Полученные неравенства являются условиями строгого диагонального доминирования элементов матрицы G . Следовательно (см. тео-

ремя 5, гл. II, § 1), $\det G \neq 0$, т. е. выбранное λ не может быть корнем уравнения (23).

Таким образом, при выполнении условия 1) все корни уравнения (23) по модулю меньше единицы и в силу теоремы 16 метод Зейделя сходится. Достаточность условия 2) для сходимости метода доказывается аналогично. Теорема доказана.

Способ редукции (1) к системе уравнений, удобной для применения метода Зейделя, определяет конкретный вид того или иного итерационного процесса. Рассмотрим итерационный метод, который, как и метод Якоби, базируется на эквивалентности исходной задачи системе (20). Запишем ее в покомпонентной форме:

$$x_i = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j - \sum_{j=i+1}^n \frac{a_{ij}}{a_{ii}} x_j + \frac{f_i}{a_{ii}}, \quad i = 1, 2, \dots, n.$$

Одношаговый итерационный процесс

$$x_i^{(k+1)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(k+1)} - \sum_{j=i+1}^n \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \frac{f_i}{a_{ii}}, \quad i = 1, 2, \dots, n, \quad (24)$$

называется *методом Гаусса – Зейделя* (иногда этот метод называют *методом Некрасова*).

Получим условия сходимости метода Гаусса – Зейделя. Для этого представим матрицу A в виде $A = L + D + U$, где

$$L = \begin{pmatrix} 0 & & & 0 \\ a_{21} & 0 & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \dots & a_{n,n-1} & 0 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & a_{12} & \dots & a_{1n} \\ 0 & 0 & \ddots & a_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix},$$

$D = \text{diag}[a_{11}, a_{22}, \dots, a_{nn}]$. Тогда в прежних обозначениях

$$F = -D^{-1}L, \quad R = -D^{-1}U, \quad g = D^{-1}f.$$

Поэтому

$$(E - F)^{-1}R = -(E + D^{-1}L)^{-1}D^{-1}U = -(D + L)^{-1}U$$

и, таким образом, приходим к матричной форме записи итерационного метода (24):

$$x^{k+1} = -(D + L)^{-1}Ux^k + (D + L)^{-1}f.$$

Поскольку уравнение

$$\det[-(D + L)^{-1}U - \lambda E] = 0,$$

очевидно, равносильно уравнению $\det[U + \lambda(D + L)] = 0$, то справедлив следующий критерий сходимости метода Гаусса – Зейделя.

Теорема 18. *Метод Гаусса – Зейделя (24) сходится при любом начальном приближении x^0 в том и только в том случае, когда все корни уравнения*

$$\det[U + \lambda(D + L)] = \begin{vmatrix} \lambda a_{11} & a_{12} & \dots & a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \dots & a_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ \lambda a_{n1} & \lambda a_{n2} & \dots & \lambda a_{nn} \end{vmatrix} = 0$$

по модулю меньше единицы.

Из большого числа достаточных признаков сходимости метода Гаусса – Зейделя приведем две теоремы, представляющие интерес для приложений.

Теорема 19. *Пусть A – неразложимая матрица со слабым диагональным доминированием. Тогда метод Гаусса – Зейделя (24) сходится.*

Теорема 20. *Если матрица A симметричная и положительно определенная, то метод Гаусса – Зейделя (24) сходится.*

Доказательство. Запишем матрицу A в виде

$$A = L + D + L^T,$$

где D – диагональная матрица, образованная из диагональных элементов матрицы A ; L – нижняя треугольная часть матрицы A .

Покажем, что все собственные значения матрицы перехода $S = -(D + L)^{-1}L^T$ по модулю меньше единицы. Рассмотрим матрицу

$$S_1 = D^{1/2}SD^{-1/2} = -(E + L_1)^{-1}L_1^T,$$

где $L_1 = D^{-1/2}LD^{-1/2}$ (существование матрицы S_1 гарантируется условием положительной определенности матрицы A). Поскольку матрицы S и S_1 имеют одинаковые собственные значения, достаточно убедиться в том, что $|\lambda_i(S_1)| < 1$. Пусть $x \in \mathbb{C}^n$ – собственный вектор единичной длины ($(x, x) = 1$), соответствующий некоторому собственному значению λ матрицы S_1 . Тогда справедливо соотношение

$$-L_1^T x = \lambda(E + L_1)x,$$

или в терминах скалярных произведений

$$-(L_1^T x, x) = \lambda[1 + (L_1 x, x)].$$

Полагая $(L_1 x, x) = a + ib$, $(L_1^T x, x) = a - ib$, получим

$$|\lambda|^2 = \left| \frac{-a + ib}{1 + a + ib} \right|^2 = \frac{a^2 + b^2}{1 + 2a + a^2 + b^2}.$$

В силу положительной определенности матрицы

$$E + L_1 + L_1^T = D^{-1/2}AD^{-1/2}$$

имеем

$$0 < ((E + L_1 + L_1^T)x, x) = 1 + (L_1 x, x) + (L_1^T x, x) = 1 + 2a,$$

откуда и следует неравенство $|\lambda| < 1$.

Таким образом, все собственные значения матрицы S_1 (а значит, и матрицы S) по модулю меньше единицы. Согласно теореме 15, метод Гаусса – Зейделя сходится. Теорема доказана.

§ 2. Элементы общей теории одношаговых итерационных методов

1. Каноническая форма одношаговых итерационных методов. Основная теорема сходимости. В этом разделе мы рассмотрим некоторые элементы общей теории одношаговых итерационных методов для нахождения решения систем линейных алгебраических уравнений

$$Ax = f$$

(1)

с невырожденной матрицей $A \in \mathbb{R}^{n \times n}$. Важную роль в разработке и исследовании итерационных процессов играет их запись в единой обобщенной форме. В предыдущем параграфе отмечалось, что любой линейный одношаговый итерационный метод решения системы (1) можно записать в виде

$$P_k x^{k+1} + Q_k x^k = g^k,$$

где P_k и Q_k – заданные матрицы, записанные, вообще говоря, от номера итерации, причем существует P_k^{-1} , g^k – заданный вектор.

Естественно требовать, чтобы точное решение x задачи (1) тождественно удовлетворяло уравнению

$$(P_k + Q_k)x = g^k. \quad (2)$$

Выполнение (2) возможно только при условии

$$(P_k + Q_k)A^{-1}f = g^k.$$

Следовательно, правая часть исходной системы (1) принимает вид $f = A(P_k + Q_k)^{-1}g^k$. Положим

$$P_k + Q_k = \tau_{k+1}A, \quad g^k = \tau_{k+1}f, \quad B_k = P_k.$$

В результате приходим к канонической форме одношагового итерационного метода решения системы (1):

$$B_k \frac{x^{k+1} - x^k}{\tau_{k+1}} + Ax^k = f, \quad k = 0, 1, \dots \quad (3)$$

Здесь $B_k \in \mathbb{R}^{n \times n}$ – невырожденная матрица, задающая тот или иной итерационный метод; τ_{k+1} – последовательность итерационных параметров. Начальное приближение $x^0 \in \mathbb{R}^n$ предполагается произвольным. В общем случае при $B \neq E$ метод (3) является невынужденным, и для определения x^{k+1} по известным значениям f и x^k необходимо решить систему линейных алгебраических уравнений

$$B_k x^{k+1} = (B_k - \tau_{k+1}A)x^k + \tau_{k+1}f, \quad k = 0, 1, \dots \quad (4)$$

Введем обозначения $r^k = Ax^k - f$, $w^k = B_k^{-1}r^k$. Тогда система уравнений (4) преобразуется к виду

$$x^{k+1} = x^k - \tau_{k+1}B_k^{-1}r^k = x^k - \tau_{k+1}w^k. \quad (5)$$

Векторы r^k и w^k будем называть *невязкой* и *поправкой метода на k-й итерации* соответственно.

Обычно задают некоторую относительную погрешность $\varepsilon > 0$ (точность итерационного процесса), с которой требуется найти приближенное решение задачи (1). Вычисления прекращают, когда выполнено условие (см. § 1)

$$\|x^k - x\| \leq \varepsilon \|x^0 - x\|.$$

Это условие неудобно для проверки, так как неизвестно точное решение x . На практике критерием окончания итерационного процесса выступает аналогичное требование:

$$\|Ax^k - f\| \leq \varepsilon \|Ax^0 - f\|$$

для невязки $r^k = Ax^k - f$, которая вычисляется непосредственно.

Если $k = k_0(\varepsilon)$ — минимальное число итераций, необходимых для достижения заданной точности ε , то общий объем вычислений, которые затрачиваются для нахождения приближенного решения системы (1), характеризуется числом

$$Q(\varepsilon) = \sum_{m=1}^{k_0(\varepsilon)} Q_m,$$

где Q_m — число арифметических действий для вычисления x^m . Тем самым задача построения экономичного итерационного процесса заключается в минимизации Q_m и $k_0(\varepsilon)$. Минимизация Q_m осуществляется за счет выбора соответствующей матрицы B_k , а минимизация $k_0(\varepsilon)$ как за счет выбора B_k , так и за счет выбора τ_{k+1} .

Запишем задачу для погрешности метода $z^k = x^k - x$:

$$B_k \frac{z^{k+1} - z^k}{\tau_{k+1}} + Az^k = 0, \quad k = 0, 1, \dots, \quad z^0 = x^0 - x. \quad (6)$$

Покажем, что если матрица $B_k = B$ не зависит от номера итерации, то поправка метода $w^k = B^{-1}r^k$ является решением однородной системы уравнений

$$B_k \frac{w^{k+1} - w^k}{\tau_{k+1}} + Aw^k = 0, \quad k = 0, 1, \dots, \quad w^0 = B^{-1}r^0. \quad (7)$$

Действительно, из (5) получим

$$x^{k+1} - x^k = -\tau_{k+1} B^{-1} r^k = -\tau_{k+1} w^k.$$

Умножая это равенство на матрицу A и учитывая соотношения

$$Ax^{k+1} - Ax^k = (Ax^{k+1} - f) - (Ax^k - f) = r^{k+1} - r^k,$$

$$r^{k+1} - r^k = B(B^{-1}r^{k+1} - B^{-1}r^k) = B(w^{k+1} - w^k),$$

приходим к однородной системе (7).

Разрешим уравнение (6) относительно z^{k+1} :

$$z^{k+1} = (E - \tau_{k+1} B^{-1} A) z^k = S_k z^k,$$

где S_k — матрица перехода. Исключая z^k, z^{k-1}, \dots, z^1 , получим

$$z^k = S_{k-1} S_{k-2} \dots S_0 z^0 = T_{k,0} z^0.$$

Матрица $T_{k,0}$ называется *разрешающей матрицей*.

Из последнего равенства находим

$$\|z^k\| = \|T_{k,0} z^0\| \leq \|T_{k,0}\| \|z^0\|.$$

Таким образом, для выяснения вопроса сходимости итерационного процесса (3) необходимо оценить норму разрешающей матрицы $T_{k,0}$.

Далее будем рассматривать несимметричные стационарные итерационные

методы с $B_k = B$, $\tau_{k+1} = \tau$:

$$B \frac{x^{k+1} - x^k}{\tau} + Ax^k = f, \quad k = 0, 1, \dots \quad (8)$$

Определим норму, порожденную матрицей $A > 0$:

$$\|x\|_A = \sqrt{(Ax, x)}.$$

Будем говорить, что итерационный процесс (8) *сходится в норме* $\|\cdot\|_A$, порожденной матрицей A , если

$$\lim_{k \rightarrow \infty} \|z^k\|_A = \lim_{k \rightarrow \infty} \|x^k - x\|_A = 0.$$

Теорема 1. Пусть A — симметричная положительно определенная матрица. Тогда при условии

$$B > \frac{\tau}{2} A \quad \text{или} \quad (Bx, x) > \frac{\tau}{2} (Ax, x) \quad \forall x \in \mathbb{R}^n, \quad x \neq 0, \quad (9)$$

итерационный метод (8) сходится.

Доказательство. Рассмотрим задачу для погрешности

$$B \frac{z^{k+1} - z^k}{\tau} + Az^k = 0. \quad (10)$$

Представим погрешность z^k в виде

$$z^k = \frac{1}{2}(z^{k+1} + z^k) - \frac{\tau}{2} \frac{z^{k+1} - z^k}{\tau}.$$

Тогда получим равенство

$$\left(B - \frac{\tau}{2} A \right) \frac{z^{k+1} - z^k}{\tau} + \frac{1}{2} A(z^{k+1} + z^k) = 0. \quad (11)$$

Умножим (11) скалярно на $2(z^{k+1} - z^k)$ и используем справедливое для $A = A^T$ равенство

$$(A(z^{k+1} + z^k), z^{k+1} - z^k) = \|z^{k+1}\|_A^2 - \|z^k\|_A^2.$$

В результате приходим к так называемому *основному тождеству*:

$$2\tau \left(\left(B - \frac{\tau}{2} A \right) \frac{z^{k+1} - z^k}{\tau}, \frac{z^{k+1} - z^k}{\tau} \right) + \|z^{k+1}\|_A^2 = \|z^k\|_A^2. \quad (12)$$

В силу условия (9) и $\tau > 0$ из (12) следует

$$0 \leq \|z^{k+1}\|_A^2 \leq \|z^k\|_A^2 \leq \dots \leq \|z^0\|_A^2.$$

Таким образом, последовательность $\{(Az^k, z^k)\}$ является невозрастающей и ограниченной снизу нулем. Значит, по теореме Вейерштрасса эта последовательность сходится.

Из положительной определенности матрицы $B - 0,5\tau A$ следует существование константы $\delta > 0$ такой, что

$$((B - 0,5\tau A)x, x) \geq \delta \|x\|^2 \quad \forall x \in \mathbb{R}^n, \quad x \neq 0.$$

Поэтому из основного тождества (12) получим неравенство

$$\frac{2\delta}{\tau} \|z^{k+1} - z^k\|^2 + \|z^{k+1}\|_A^2 \leq \|z^k\|_A^2 \quad (13)$$

и в силу сходимости последовательности $\{(Az^k, z^k)\}$

$$\lim_{k \rightarrow \infty} \|z^{k+1} - z^k\|^2 = 0.$$

Далее из уравнения для погрешности (10) находим

$$Az^k = -\frac{B(z^{k+1} - z^k)}{\tau}, \quad z^k = -A^{-1}B \frac{z^{k+1} - z^k}{\tau},$$

откуда вытекает оценка

$$\|z^k\|_A^2 \leq \|A^{-1}\| \|B\|^2 \frac{\|z^{k+1} - z^k\|^2}{\tau^2}. \quad (14)$$

Переходя в (14) к пределу при $k \rightarrow \infty$, получим $\lim_{k \rightarrow \infty} \|z^k\|_A = 0$. Теорема доказана.

Замечание 1. При условиях теоремы 1 итерационный процесс (8) сходится также и в евклидовой норме. Действительно, так как $z^k = -\tau^{-1}A^{-1}B(z^{k+1} - z^k)$, то имеет место оценка

$$\|z^k\|^2 \leq \|A^{-1}\|^2 \|B\|^2 \frac{\|z^{k+1} - z^k\|^2}{\tau^2}.$$

и, следовательно,

$$\lim_{k \rightarrow \infty} \|z^k\| = 0.$$

Замечание 2. Из неравенств (13) и (14) следует, что итерационный метод (8) сходится со скоростью геометрической прогрессии, и для погрешности метода справедлива оценка

$$\|z^k\|_A \leq \rho^k \|z^0\|_A, \quad \rho = \sqrt{1 - \frac{2\delta\tau}{\|A^{-1}\| \|B\|^2}} < 1.$$

Теорема 1 при достаточно минимальных требованиях на матрицы A и B устанавливает легко проверяемое условие сходимости итерационного процесса (8) и потому имеет значительную практическую ценность. Ниже мы рассмотрим конкретные примеры итерационных процессов (некоторые из них изучались в § 1) и с помощью доказанной теоремы исследуем вопросы их сходимости.

2. Явный итерационный метод. В предположении симметричности и положительной определенности матрицы A рассмотрим явный итерационный процесс ($B = E$) с постоянным итерационным параметром $\tau > 0$:

$$\frac{z^{k+1} - z^k}{\tau} + Ax^k = f. \quad (15)$$

В покомпонентной записи метод (15) имеет вид

$$x_i^{(k+1)} = x_i^{(k)} - \tau \left(\sum_{j=1}^n a_{ij} x_j^{(k)} - f_i \right), \quad i = 1, 2, \dots, n.$$

Именно этот итерационный процесс часто называют методом простой итерации. Согласно (9), метод сходится при условии

$$E - \frac{\tau}{2} A > 0. \quad (16)$$

Выясним, какие ограничения на параметр τ накладывает условие (16). С учетом неравенства $E \geq A/\|A\|$ имеем

$$E - \frac{\tau}{2} A \geq \left(\frac{1}{\|A\|} - \frac{\tau}{2} \right) A > 0,$$

если $\frac{1}{\|A\|} - \frac{\tau}{2} > 0$. Следовательно, явный итерационный метод (15) сходится при выполнении неравенства

$$\tau < \frac{2}{\|A\|} = \frac{2}{\lambda_{\max}(A)}, \quad (17)$$

где $\lambda_{\max}(A)$ — наибольшее собственное значение матрицы A .

Покажем, что условие (17) является и необходимым для сходимости метода (15). В качестве начального приближения выберем вектор $x^0 = x + \mu$, где x — точное решение задачи (6), а μ — собственный вектор, соответствующий собственному значению $\lambda_{\max}(A)$. При таком выборе начального приближения $z^0 = \mu$. Тогда получим

$$z^k = (E - \tau A)^k z^0 = (E - \tau A)^k \mu,$$

откуда следует, что

$$\|z^k\| = |1 - \tau \lambda_{\max}(A)|^k \|\mu\|.$$

Если $\tau = \frac{2}{\lambda_{\max}(A)}$, то $\|z^k\| = \|\mu\| \rightarrow 0$ при $k \rightarrow \infty$. Если же

$\tau > \frac{2}{\lambda_{\max}(A)}$, то $|1 - \tau \lambda_{\max}(A)| > 1$ и $\|z^k\| \rightarrow \infty$ при $k \rightarrow \infty$. Таким образом, условие (17) необходимо и достаточно для сходимости явного итерационного метода (15).

Вопросы, связанные с оптимизацией итерационного параметра τ в методе (15), будут рассмотрены ниже.

3. Методы Якоби и Гаусса — Зейделя. Рассмотрим итерационный метод Якоби:

$$x_i^{(k+1)} = - \sum_{j=1, j \neq i}^n \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \frac{f_i}{a_{ii}}, \quad i = 1, 2, \dots, n. \quad (18)$$

Если обозначить $D = \text{diag}[a_{11}, a_{22}, \dots, a_{nn}]$, то метод (18) можно записать в каноническом виде:

$$D(x^{k+1} - x^k) + Ax^k = f.$$

Для вектора погрешности имеем

$$D(z^{k+1} - z^k) + Az^k = 0.$$

При $A = A^T > 0$ достаточное условие сходимости метода Якоби принимает вид

$$2D - A > 0. \quad (19)$$

Но поскольку в рассматриваемом случае $D = D^T > 0$, то условие (19) будет являться и необходимым. Действительно, матрица перехода имеет вид

$$S = S^T = E - D^{-1}A.$$

Поэтому из соотношений $-E < E - D^{-1}A < E$ вытекает неравенство (19). Отметим, что этот результат полностью согласуется с выводами теоремы 12 § 1.

Перейдем теперь к итерационному методу Гаусса — Зейделя:

$$x_i^{(k+1)} = - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(k+1)} - \sum_{j=i+1}^n \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \frac{f_i}{a_{ii}}, \quad i = 1, 2, \dots, n. \quad (20)$$

Реализация метода (20) осуществляется последовательно, начиная с первого уравнения. Приведем (20) к каноническому виду (8). Представим матрицу A в виде

$$A = L + D + U,$$

где $D = \text{diag}[a_{11}, a_{22}, \dots, a_{nn}]$; L — строго нижняя треугольная матрица; U — строго верхняя треугольная матрица. Тогда для метода Гаусса — Зейделя (20) имеем

$$(D + L)(x^{k+1} - x^k) + Ax^k = f.$$

Теорема 2. Если A — симметричная положительно определенная матрица ($A = A^T > 0$), то метод Гаусса — Зейделя (20) сходится.

Доказательство. Согласно теореме 1, достаточно проверить выполнение скалярного неравенства

$$((D + L - 0,5A)x, x) > 0.$$

Преобразуем левую часть этого неравенства:

$$\begin{aligned} ((D + L - 0,5A)x, x) &= (Dx, x) + (Lx, x) - 0,5(Ax, x) = \\ &= (Dx, x) + (Lx, x) - 0,5(Dx, x) - 0,5(Ux, x) = \\ &= 0,5(Dx, x) + 0,5(Lx, x) - 0,5(Ux, x). \end{aligned}$$

Так как $A = A^T > 0$, то $D = D^T > 0$, $L^T = U$ и, следовательно,

$$((D + L - 0,5A)x, x) = 0,5(Dx, x) > 0.$$

Теорема доказана.

Сформулируем и докажем теорему о скорости сходимости итерационного метода (20).

Теорема 3. Пусть матрица A является строго диагонально доминирующими:

$$\sum_{j=1, j \neq i}^n |a_{ij}| \leq q |a_{ii}|, \quad 0 < q < 1, \quad i = 1, 2, \dots, n. \quad (21)$$

Тогда итерационный метод Гаусса — Зейделя (20) сходится со скоростью геометрической прогрессии, и для погрешности метода справедлива оценка

$$\|z^k\|_C \leq q^k \|z^0\|_C.$$

Доказательство. Из (20) имеем

$$|z_i^{(k+1)}| \leq \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| |z_j^{(k+1)}| + \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right| |z_j^{(k)}|.$$

Пусть максимум величины $|z_i^{(k+1)}|$ достигается при $i = i_0$. В силу условия (21) получим

$$\sum_{j=i_0+1}^n \left| \frac{a_{ij}}{a_{ii_0}} \right| \leq q - \sum_{j=1}^{i_0-1} \left| \frac{a_{ij}}{a_{ii_0}} \right| \leq q \left(1 - \sum_{j=1}^{i_0-1} \left| \frac{a_{ij}}{a_{ii_0}} \right| \right).$$

Поэтому можно записать

$$|z_{i_0}^{(k+1)}| \leq \sum_{j=1}^{i_0-1} \left| \frac{a_{ij}}{a_{ii_0}} \right| |z_{i_0}^{(k+1)}| + q \left(1 - \sum_{j=1}^{i_0-1} \left| \frac{a_{ij}}{a_{ii_0}} \right| \right) \|z^k\|_C.$$

Отсюда следует неравенство

$$\left(1 - \sum_{j=1}^{i_0-1} \left| \frac{a_{ij}}{a_{ii_0}} \right| \right) |z_{i_0}^{(k+1)}| \leq q \left(1 - \sum_{j=1}^{i_0-1} \left| \frac{a_{ij}}{a_{ii_0}} \right| \right) \|z^k\|_C,$$

которое и доказывает теорему.

4. Метод релаксации. Пусть при вычислении какой-либо итерации в (8) построено приближение $x^s = (x_1^{(s)}, \dots, x_n^{(s)})^T$, а следующее приближение x^{s+1} таково, что изменяет лишь i -ю компоненту вектора x^s , т. е. $x^{s+1} = (x_1^{(s)}, \dots, x_i^{(s+1)}, \dots, x_n^{(s)})^T$. Подобный способ построения приближений характерен для рассмотренного выше метода Гаусса — Зейделя. К такому же типу методов относятся и итерационный метод релаксации. Положим

$$z^{s+1} = z^s - \alpha e_i, \quad \alpha = \text{const},$$

где e_i — вектор-столбец, i -я компонента которого равна единице, а остальные нули. Вычислим норму вектора погрешности:

$$\begin{aligned} \|z^{s+1}\|_A^2 &= (A z^{s+1}, z^{s+1}) = (A(z^s - \alpha e_i), z^s - \alpha e_i) = \\ &= \|z^s\|_A^2 + \alpha^2 a_{ii} - 2\alpha r_i^{(s)} = \|z^s\|_A^2 + \frac{1}{a_{ii}} (a_{ii}\alpha - r_i^{(s)})^2 - \frac{(r_i^{(s)})^2}{a_{ii}}. \end{aligned} \quad (22)$$

В этом выражении

$$r_i^{(s)} = (Ax^s, e_i) = (A(x^s - x), e_i) = (Ax^s - f, e_i) = (r^s, e_i).$$

Как следует из (22), минимум величины $\|z^{s+1}\|_A^2$ достигается при

$$\alpha = \frac{r_i^{(s)}}{a_{ii}},$$

Именно при этом значении параметра α

$$r_i^{(s+1)} = (A(z^s - \alpha e_i), e_i) = (Az^s, e_i) - \alpha a_{ii} = r_i^{(s)} - \alpha a_{ii} = 0.$$

Последнее означает, что i -я компонента вектора невязки r^{s+1} обращается в нуль. Это свойство, в частности, характеризует метод Гаусса — Зейделя.

Однако при построении очередного приближения не обязатель но каждый раз минимизировать норму $\|z^{s+1}\|_A^2$, т.е. осуществлять полную релаксацию. Достаточно потребовать, чтобы

$$\|z^{s+1}\|_A^2 < \|z^s\|_A^2. \quad (23)$$

Поскольку справедливо соотношение

$$\|z^{s+1}\|_A^2 - \|z^s\|_A^2 = \frac{1}{a_{ii}} (a_{ii}\alpha - r_i^{(s)})^2 - \frac{(r_i^{(s)})^2}{a_{ii}},$$

то (23) будет выполнено, если $|a_{ii}\alpha - r_i^{(s)}| < |r_i^{(s)}|$. Поэтому

$$\alpha = \omega \frac{r_i^{(s)}}{a_{ii}}, \quad 0 < \omega < 2.$$

При $\omega \in (1, 2)$ имеет место верхняя релаксация, при $\omega \in (0, 1)$ нижняя релаксация. Значение $\omega = 1$ соответствует полной релаксации или методу Гаусса — Зейделя. Для неполной релаксации

$$r_i^{(s+1)} = (Az^{s+1}, e_i) = (1 - \omega)r_i^{(s)}. \quad (24)$$

Соотношение (24) позволяет записать расчетные формулы метода релаксации для системы (6). Пусть

$$x^s = (x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}, x_i^{(k)}, x_{i+1}^{(k)}, \dots, x_n^{(k)})^T,$$

$$x^{s+1} = (x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}, x_i^{(k+1)}, x_{i+1}^{(k)}, \dots, x_n^{(k)})^T.$$

Тогда получим

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} - \omega \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(k+1)} - \omega \sum_{j=i+1}^n \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \omega \frac{f_i}{a_{ii}}, \quad i = 1, 2, \dots, n, \quad k = 0, 1, \dots$$

Для приведения (25) к каноническому виду используем уже известное расложение $A = L + D + U$. Тогда из (25) имеем

$$(D + \omega L) \frac{x^{k+1} - x^k}{\omega} + Ax^k = f.$$

Теорема 4. Пусть A — симметричная положительно определенная матрица ($A = A^T > 0$), $0 < \omega < 2$. Тогда метод релаксации (25) сходится.

Доказательство. Воспользуемся теоремой 1 и покажем, что справедливо неравенство

$$D + \omega L > \frac{\omega}{2} A.$$

В силу $L^T = U$, $D > 0$ имеем

$$\begin{aligned} ((D + \omega L - 0,5\omega A)x, x) &= (1 - 0,5\omega)(Dx, x) + 0,5\omega(Lx, x) - \\ &- 0,5\omega(Ux, x) = (1 - 0,5\omega)(Dx, x) > 0. \end{aligned}$$

Теорема доказана.

5. Попеременно-треугольный метод. В настоящее время методы такого типа являются одними из самых эффективных итерационных методов для решения системы (6), поскольку сочетают высокую скорость сходимости и простоту обращения матрицы B в (8).

Основная идея метода заключается в следующем. Представим матрицу A в виде суммы двух треугольных матриц:

$$A = \left(L + \frac{D}{2} \right) + \left(\frac{D}{2} + U \right) = R_1 + R_2.$$

Тогда матрицу B определим как произведение матриц

$$B = B_1 B_2 = (E + \omega R_1)(E + \omega R_2), \quad \omega > 0. \quad (26)$$

110 ГЛАВА III. ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ СЛАУ

Одношаговый итерационный метод теперь запишется в виде

$$(E + \omega R_1)(E + \omega R_2) \frac{x^{k+1} - x^k}{\tau} + Ax^k = f. \quad (27)$$

Реализация метода (27) сводится к последовательному обращению треугольных матриц B_1 и B_2 :

$$B_1 u = f - Ax^k, \quad B_2 \frac{x^{k+1} - x^k}{\tau} = u,$$

или в покомпонентном виде

$$u_i = 2 \left(- \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} u_j - \sum_{j=i+1}^n \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \frac{f_i}{a_{ii}} \right), \quad i = 1, 2, \dots, n,$$

$$x_i^{(k+1)} = x_i^{(k)} + 2 \left(- \sum_{j=n}^{i+1} \frac{a_{ij}}{a_{ii}} (x_j^{(k+1)} - x_j^{(k)}) + \tau \frac{u_i}{a_{ii}} \right), \quad i = n, n-1, \dots, 1.$$

Достоинство построения матрицы (26), помимо простоты реализации итерационного процесса (27), состоит также в том, что если $A = A^T > 0$, то и $B = B^T > 0$. В самом деле так как $R_1^T = R_2$, то

$$B^T = (E + \omega R_2)^T (E + \omega R_1)^T = (E + \omega R_1)(E + \omega R_2) = B,$$

$$(Bx, x) = ((E + \omega A + \omega^2 R_1 R_2)x, x) = \|x\|^2 + \omega \|x\|_A^2 + \omega^2 (R_2 x, R_2 x) > 0. \quad (28)$$

Теорема 5. Пусть A — симметричная положительно определенная матрица ($A = A^T > 0$). Тогда попеременно-треугольный метод (27) сходится при условии $\omega > \tau/2$.

Доказательство. Согласно теореме 1, проверим выполнение неравенства

$$B = (E + \omega R_1)(E + \omega R_2) > \frac{\tau}{2} A. \quad (29)$$

В силу (28) имеем

$$(Bx, x) = \|x\|^2 + \omega \|x\|_A^2 + \omega^2 (R_2 x, R_2 x),$$

а это вместе с предположением $\omega > \tau/2$ делает очевидным (29). Теорема доказана.

§ 2. ЭЛЕМЕНТЫ ОБЩЕЙ ТЕОРИИ

Замечание 3. Возможны варианты попеременно-треугольного метода, когда в расложении

$$A = A_1 + A_2, \quad A_1 > 0, \quad A_2 > 0$$

матрицы A_1, A_2 не являются треугольными, однако матрицы

$$B_1 = (E + \omega A_1), \quad B_2 = (E + \omega A_2)$$

легко обращаются. В этом случае факторизованная конструкция матрицы $B = B_1 B_2$ также является достаточно эффективной при построении итерационного процесса.

6. Оптимизация скорости сходимости итерационных процессов. Рассмотрим явный итерационный метод (15):

$$\frac{x^{k+1} - x^k}{\tau} + Ax^k = f.$$

Предположим, что

$$A = A^T > 0, \quad \gamma_1 E \leqslant A \leqslant \gamma_2 E, \quad \gamma_1 > 0. \quad (30)$$

Второе условие в (30) означает, что

$$\gamma_1 \|x\|^2 \leqslant (Ax, x) \leqslant \gamma_2 \|x\|^2 \quad \forall x \in \mathbb{R}^n, \quad x \neq 0.$$

Постоянные γ_1 и γ_2 называются константами энергетической эквивалентности матриц A и E . Для явного метода имеем

$$z^{k+1} = (E - \tau A)z^k = Sz^k.$$

Как уже отмечалось, чем меньше норма матрицы перехода $\|S\|$, тем меньше число итераций $k_0(\varepsilon)$, необходимых для достижения заданной точности ε , а следовательно, выше скорость сходимости итерационного процесса. Поскольку $\|S\|$ зависит от параметра τ , то естественно выбирать τ , минимизирующими $\|S\|$. Так как $A = A^T$, то $S = S^T$ и поэтому

$$\|S\| = \max_{1 \leq i \leq n} |s_i| = \max_{1 \leq i \leq n} |1 - \tau \lambda_i|,$$

где λ_i — собственные значения матрицы A . В силу (30)

$$\gamma_1 \leq \lambda_i \leq \gamma_2,$$

и условие сходимости явного метода $|s_i| < 1$ выполняется при

$$\tau < \frac{2}{\gamma_2}.$$

Минимизация нормы $\|S\|$ сводится к отысканию такого $\tau > 0$, при котором достигается

$$\min_{\tau > 0} \max_{\gamma_1 \leq \lambda_i \leq \gamma_2} |1 - \tau \lambda_i|. \quad (31)$$

Рассмотрим на отрезке $[\gamma_1, \gamma_2]$, при этом найдется τ_0 такое, что $g_{\tau_0}(\gamma_1) = -g_{\tau_0}(\gamma_2)$, которое и дает решение задачи (31). В самом деле если $\tau < \tau_0$, то $|g_{\tau}(\gamma_2)| > |g_{\tau_0}(\gamma_2)|$, а если $\tau > \tau_0$, то $|g_{\tau}(\gamma_1)| > |g_{\tau_0}(\gamma_1)|$. Таким образом, оптимальное значение τ_0 равно

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2} < \frac{2}{\gamma_2}.$$

При этом справедливо равенство

$$\|S\|_{\min} = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1} = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \quad (32)$$

и для явного итерационного метода (15) имеет место следующая оценка скорости сходимости:

$$\|z^k\| \leq \left(\frac{1 - \xi}{1 + \xi} \right)^k \|z^0\|. \quad (33)$$

Если $\gamma_1 = \lambda_{\min}(A)$, $\gamma_2 = \lambda_{\max}(A)$, то в формулах (32), (33) ξ есть величина, обратная числу обусловленности $\kappa(A)$ матрицы A . Поэтому чем меньше число обусловленности матрицы системы (6), тем выше скорость сходимости явного итерационного метода.

Исследуем теперь вопрос о применении последовательности итерационных параметров в явном итерационном методе. Этую задачу

будем решать в предположениях (30). Кроме того, будем предполагать, что система собственных векторов $\{\varphi_i\}$ матрицы A образует ортонормированный базис, т. е. $(\varphi_i, \varphi_j) = \delta_{ij}$, где δ_{ij} — символ Кронекера. Запишем уравнение для погрешности метода

$$z^{k+1} = (E - \tau_{k+1} A) z^k. \quad (34)$$

Разложим z^k по системе собственных векторов матрицы A :

$$z^k = \sum_{i=1}^n c_i^k \varphi_i,$$

Тогда из (34) получим

$$c_i^k = \left[\prod_{m=1}^k (1 - \tau_m \lambda_i) \right] c_i^0 = P_k(\lambda_i) c_i^0,$$

откуда следует, что

$$\begin{aligned} \|z^k\| &= \sqrt{\sum_{i=1}^n (c_i^k)^2} = \sqrt{\sum_{i=1}^n (P_k(\lambda_i))^2 (c_i^0)^2} \leq \\ &\leq \max_{\gamma_1 \leq \lambda_i \leq \gamma_2} |P_k(\lambda_i)| \sqrt{\sum_{i=1}^n (c_i^0)^2} \leq \max_{\gamma_1 \leq \lambda_i \leq \gamma_2} |P_k(\lambda_i)| \|z^0\|. \end{aligned}$$

Таким образом, получена оценка для нормы разрешающей матрицы

$$\|T_{k,0}\| \leq \max_{\gamma_1 \leq \lambda_i \leq \gamma_2} |P_k(\lambda_i)|.$$

Дальнейшее заключается в отыскании такой последовательности итерационных параметров $\tau_1, \tau_2, \dots, \tau_k$, при которой достигается

$$\min_{\tau_1, \tau_2, \dots, \tau_k} \max_{\gamma_1 \leq \lambda_i \leq \gamma_2} |P_k(\lambda_i)|. \quad (35)$$

Другими словами, среди всех многочленов k -й степени, удовлетворяющих условию $P_k(0) = 1$, требуется найти многочлен $\tilde{P}_k(\lambda)$, наименее уклоняющийся от нуля на отрезке $\gamma_1 \leq \lambda \leq \gamma_2$.

Искомый многочлен $\tilde{P}_k(\lambda)$ связан с многочленом Чебышева

$$T_k(x) = 2^{1-k} \cos(k \arccos x), \quad |x| \leq 1.$$

При $|x| > 1$ многочлен $T_k(x)$ определяется формулой

$$T_k(x) = 2^{-k} [(x + \sqrt{x^2 - 1})^k + (x - \sqrt{x^2 - 1})^k].$$

Многочлен $T_k(x)$ обладает тем свойством, что среди всех многочленов степени k со старшим коэффициентом, равным единице, он наименее уклоняется от нуля на отрезке $[-1, 1]$. Поэтому для решения задачи (35) достаточно привести замену переменных

$$x = \frac{2\lambda - (\gamma_2 + \gamma_1)}{\gamma_2 - \gamma_1}, \quad (36)$$

с помощью которой отрезок $[-1, 1]$ переводится в отрезок $[\gamma_1, \gamma_2]$. Из условия нормировки $P_k(0) = 1$ при $\lambda = 0$ находим

$$x_0 = -\frac{\gamma_2 + \gamma_1}{\gamma_2 - \gamma_1}.$$

Тогда имеем

$$\tilde{P}_k(\lambda) = \frac{T_k(x)}{T_k(x_0)} = T_k\left(\frac{2\lambda - (\gamma_2 + \gamma_1)}{\gamma_2 - \gamma_1}\right) / T_k\left(-\frac{\gamma_2 + \gamma_1}{\gamma_2 - \gamma_1}\right). \quad (37)$$

Чтобы найти набор итерационных параметров $\tau_1, \tau_2, \dots, \tau_k$, потребуем совпадения нулей многочленов $P_k(\lambda)$ и $T_k\left(\frac{2\lambda - (\gamma_2 + \gamma_1)}{\gamma_2 - \gamma_1}\right)$.

Нули многочлена $P_k(\lambda)$ суть $\lambda_m = \frac{1}{\tau_m}$, $m = 1, 2, \dots, k$, а нули многочлена Чебышева $T_k(x)$ $x_m = \cos \frac{2m-1}{2k} \pi$, $m = 1, 2, \dots, k$.

Учитывая связь (36) между x и λ , получим

$$x_m = \frac{2\lambda_m - (\gamma_2 + \gamma_1)}{\gamma_2 - \gamma_1}, \quad m = 1, 2, \dots, k,$$

и, следовательно,

$$\tau_m = \frac{2}{\gamma_2 + \gamma_1 + (\gamma_2 - \gamma_1) \cos \frac{2m-1}{2k} \pi}, \quad m = 1, 2, \dots, k. \quad (38)$$

Из (37) вытекает соотношение

$$\max_{\gamma_1 \leq \lambda_i \leq \gamma_2} |\tilde{P}_k(\lambda_i)| = \frac{1}{|T_k(x_0)|} = q_k,$$

и, таким образом, оценка для погрешности принимает вид

$$\|z^k\| \leq q_k \|z^0\|.$$

Осталось найти выражение для q_k . Обозначим

$$\rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

Поскольку $|x_0| > 1$, то

$$q_k = \frac{1}{|T_k\left(-\frac{1}{\rho_0}\right)|} = \frac{2}{\left(\frac{1}{\rho_1}\right)^k + \rho_1^k} = \frac{2\rho_1^k}{1 + \rho_1^{2k}} < 1.$$

Определим $k_0(\varepsilon)$ так, чтобы $q_k \leq \varepsilon$. Для этого достаточно, чтобы $\rho_1^k \leq \varepsilon/2$, или

$$k \geq k_0(\varepsilon) = \frac{\ln(2/\varepsilon)}{\ln(1/\rho_1)}. \quad (39)$$

Изложенный итерационный метод получил название *явного метода с чебышевским набором параметров*. Обычно этот метод применяют в циклическом варианте. Длина цикла определяется из (39), затем из (38) находятся параметры τ_m и проводится k итераций, после чего описанный процесс повторяется.

При практическом применении метода обнаружилось, что порядок выбора итерационных параметров существенно влияет на его вычислительную устойчивость. Оказалось, что использование параметров в произвольном порядке может привести к недопустимо сильному возрастанию погрешностей. Переход от итерации к итерации осуществляется с помощью матрицы перехода $S_m = E - \tau_m A$. В силу (30)

$$(1 - \tau_m \gamma_2) E \leq E - \tau_m A \leq (1 - \tau_m \gamma_1) E. \quad (40)$$

Подставим в (40) τ_m из (38) и учтем равенство

$$\tau_m = \frac{\tau_0}{1 + \rho_0 x_m}, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}.$$

Тогда получим

$$\frac{\rho_0(1 - x_m)}{1 + \rho_0 x_m} E \leq S_m \leq \frac{\rho_0(1 + x_m)}{1 + \rho_0 x_m} E. \quad (41)$$

Поскольку при сделанных предположениях $S_m = S_m^T$, то из (41) следуют неравенства

$$\begin{aligned} \|S_m\| &= \frac{\rho_0(1 + x_m)}{1 + \rho_0 x_m} < 1, \quad x_m \geq 0, \\ \|S_m\| &= \frac{\rho_0(1 + |x_m|)}{1 - \rho_0 |x_m|}, \quad x_m < 0. \end{aligned} \quad (42)$$

Из (42) получим

$$\|S_m\| > 1, \quad \rho_0(1 + 2|x_m|) > 1. \quad (43)$$

Таким образом, хотя и показано, что

$$\|T_{k,0}\| = \|S_k S_{k-1} \dots S_m \dots S_2 S_1\| < 1,$$

но среди матриц S_1, S_2, \dots, S_k имеются матрицы S_m из (43). Если число k достаточно велико, то найдется достаточно большое p , таковое, что для всех $m_0 + \beta$, $\beta = 1, 2, \dots, p$, $m_0 + p < k$

$$\|S_{m_0+\beta}\| > 1.$$

Тогда

$$z^{m_0+p} = T_{m_0+p, m_0} z^{m_0} = S_{m_0+p} S_{m_0+p-1} \dots S_{m_0+1} z^{m_0},$$

и при любом ограниченном z^{m_0} величина z^{m_0+p} может превзойти максимально допустимую для данной ЭВМ.

Естественно попытаться переворотить параметры τ_m , для которых $\|S_m\| < 1$, с параметрами, для которых $\|S_m\| > 1$. На этом пути и проводится построение такой последовательности параметров $\{\tau_m\}$, для которой сходимость итераций носит монотонный характер

и вычислительная неустойчивость отсутствует. Существует правило такого упорядочения нулей $x_m = -\cos \frac{2m-1}{2k} \pi$ многочлена Чебышева (а тем самым и параметров $\{\tau_m\}$) для любого k , при котором имеет место вычислительная устойчивость.

Приведем это правило для случая $k = 2^p$, где $p > 0$ — целое число. Обозначим множество нулей многочлена Чебышева через

$$\mathfrak{M}_k^* = \left\{ -\cos \beta_i, \beta_i = \frac{\pi}{2k} \theta_i^{(k)}, i = 1, 2, \dots, k \right\}, \quad k = 2^p,$$

где $\theta_i^{(k)}$ — одно из нечетных чисел $1, 3, 5, \dots, 2k-1$. Задача сводится к упорядочению множества нечетных чисел

$$\theta_k = \{\theta_1^{(k)}, \theta_2^{(k)}, \dots, \theta_k^{(k)}\}, \quad 1 \leq \theta_i^{(k)} \leq 2k-1, \quad i = 1, 2, \dots, k.$$

Исходя из множества $\theta_1 = \{1\}$, построим множество $\theta_k^* = \theta_{2^p}^*$ по формулам перехода от θ_s к θ_{2s} :

$$\theta_{2s-1}^{(2s)} = \theta_s^{(s)}, \quad \theta_{2s}^{(2s)} = 4s - \theta_{2s-1}^{(2s)}, \quad i = 1, 2, \dots, s. \quad (44)$$

Соответствующую последовательность параметров $\{\tau_m\}$ будем называть *устойчивым набором*.

Пример 1. Построим устойчивый набор параметров для случая $k = 16 = 2^4$. Полагая $\theta_1 = \{1\}$, последовательно находим

$$\theta_2 = \{1, 3\}, \quad \theta_4 = \{1, 7, 3, 5\}, \quad \theta_8 = \{1, 15, 7, 9, 3, 13, 5, 11\},$$

$$\theta_{16} = \{1, 31, 15, 17, 7, 25, 9, 23, 3, 29, 13, 19, 5, 27, 11, 21\}.$$

При переходе от θ_s к θ_{2s} , согласно (44), достаточно после каждого $\theta_{2s-1}^{(2s)} = \theta_s^{(s)}$ поставить число $\theta_{2s}^{(2s)} = 4s - \theta_{2s-1}^{(2s)}$.

§ 3. Итерационные методы вариационного типа

Методы этого параграфа предназначены для решения системы линейных алгебраических уравнений

$$Ax = f \quad (1)$$

с симметричной положительно определенной матрицей $A \in \mathbf{R}^{n \times n}$. Основная идея итерационных методов вариационного типа состоит в

минимизации выбранного функционала в направлении вектора градиента функционала. Такой выбор направления минимизации функционала связан с тем фактом, что наибольшая скорость убывания функционала $F(x)$ в точке x происходит в направлении $-\text{grad } F(x)$ (см. § 1, п. 2). Поэтому некоторые итерационные методы вариационного типа также называют методами наискорейшего спуска.

1. Метод скорейшего (градиентного) спуска. В методе скорейшего спуска нахождение решения системы (1) связано с задачей минимизации квадратичного функционала

$$F(x) = (Ax, x) - 2(f, x). \quad (2)$$

Покажем, что решение $x^* = A^{-1}f$ системы (1) доставляет минимум функционалу (2) на множестве векторов из пространства \mathbf{R}^n . Действительно, имеем

$$\begin{aligned} F(x) - F(x^*) &= (Ax, x) - 2(f, x) - (Ax^*, x^*) + 2(f, x^*) = (Ax, x) - \\ &- 2(Ax^*, x) - (Ax^*, x^*) + 2(Ax^*, x^*) = (A(x - x^*), x - x^*) \geq 0 \end{aligned}$$

в силу положительной определенности матрицы A . Знак равенства в этом выражении возможен только при $x = x^*$, т. е. задачи (1) и (2) эквивалентны.

Выберем произвольный вектор x^0 и вычислим направление, противоположное градиенту функционала $\text{grad } F(x^0)$ (это направление с точностью до постоянного множителя противоположно направлению вектора невязки $r^0 = Ax^0 - f$). Из точки x^0 будем теперь двигаться в направлении вектора $-r^0$ до точки $x^1 = x^0 - \tau r^0$, в которой функционал $F(x)$ достигнет своего минимального значения:

$$\begin{aligned} F(x^0 - \tau r^0) &= (Ax^0 - \tau Ar^0, x^0 - \tau r^0) - 2(f, x^0 - \tau r^0) = \\ &= F(x^0) - 2\tau(r^0, r^0) + \tau^2(Ar^0, r^0). \end{aligned}$$

Очевидно, что это выражение достигает минимума при

$$\tau = \tau_1 = \frac{(r^0, r^0)}{(Ar^0, r^0)}.$$

В качестве нового приближения к решению (1) принимается вектор

$$x^1 = x^0 - \tau_1 r^0,$$

при этом справедливо соотношение

$$F(x^1) = F(x^0) - \frac{(r^0, r^0)^2}{(Ar^0, r^0)}.$$

Далее находим приближение

$$x^2 = x^1 - \tau_2 r^1,$$

где невязка r^1 и параметр τ_2 находятся по формулам

$$r^1 = Ax^1 - f, \quad \tau_2 = \frac{(r^1, r^1)}{(Ar^1, r^1)}.$$

Для произвольной итерации получим

$$\begin{aligned} x^{k+1} &= x^k - \tau_{k+1} r^k, \\ r^k &= Ax^k - f = r^{k-1} - \tau_k Ar^{k-1}, \end{aligned} \quad (3)$$

где итерационный параметр τ_{k+1} выбирается из условия минимума функционала $F(x^k - \tau_{k+1} r^k)$:

$$\tau_{k+1} = \frac{(r^k, r^k)}{(Ar^k, r^k)}. \quad (4)$$

Таким образом, методом скорейшего или градиентного спуска называется явный итерационный метод (3), в котором параметр τ_{k+1} вычисляется по формуле (4).

Отметим, что невязки r^k и r^{k+1} двух последовательных приближений метода скорейшего спуска ортогональны друг другу. Действительно, имеем

$$(r^{k+1}, r^k) = (r^k - \tau_{k+1} Ar^k, r^k) = (r^k, r^k) - \tau_{k+1}(Ar^k, r^k) = 0$$

в силу выбора итерационного параметра τ_{k+1} .

Для исследования сходимости метода скорейшего спуска удобно применить общую теорию односторонних итерационных методов. Запишем итерационный процесс (3) в канонической форме:

$$\frac{x^{k+1} - x^k}{\tau_{k+1}} + Ax^k = f. \quad (5)$$

Теорема 1. Пусть A — симметричная положительно определенная матрица ($A = A^T > 0$). Тогда метод скорейшего спуска сходится, и для погрешности метода справедлива оценка

$$\|x^k - x\|_A \leq \rho_0 \|x^0 - x\|_A, \quad k = 0, 1, \dots, \quad (6)$$

где

$$\rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}. \quad (7)$$

Доказательство. Как следует из (5), погрешность метода скорейшего спуска удовлетворяет уравнению

$$z^{k+1} = z^k - \tau_{k+1} A z^k = z^k - \tau_{k+1} r^k.$$

Отсюда получим

$$\|z^{k+1}\|_A^2 = \|z^k\|_A^2 - 2\tau_{k+1}(r^k, r^k) + \tau_{k+1}^2(Ar^k, r^k). \quad (8)$$

При заданном векторе r^k правая часть равенства (8) достигает минимума, если τ_{k+1} выбрать согласно (4). При любом другом значении τ_{k+1} правая часть соотношения (8) может только увеличиться. Поэтому, полагая в (8) $\tau_{k+1} = \tau_0$, где

$$\tau_0 = \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)},$$

приходим к неравенству

$$\|z^{k+1}\|_A^2 \leq \|(E - \tau_0 A)\| \|z^k\|_A^2$$

или

$$\|z^{k+1}\|_A \leq \|E - \tau_0 A\| \|z^k\|_A. \quad (9)$$

Так как $\|E - \tau_0 A\| = \rho_0$ (см. § 2, п. 6), то из (9) получим

$$\|x^{k+1} - x\|_A \leq \rho_0 \|x^k - x\|_A,$$

откуда следует оценка (6). Теорема доказана.

Замечание 1. Доказательство теоремы 1 можно провести также с использованием неравенства

$$(y, y)^2 \leq (Ay, y)(A^{-1}y, y) \leq \frac{1}{4}(\sqrt{\xi} + \frac{1}{\sqrt{\xi}})^2(y, y)^2,$$

справедливого для симметрических положительно определенных матриц A и произвольного неподвластного вектора y .

Замечание 2. Как видно из теоремы 1, метод скорейшего спуска (3) имеет ту же скорость сходимости, что и явный итерационный метод с оптимальным параметром $t = t_0$, рассмотренный в предыдущем параграфе. Однако преимущество метода (3) состоит в том, что он не требует информации о границах спектра матрицы A .

2. Методы минимальных невязок и минимальных поправок. Рассмотрим систему (1) с симметрической положительно определенной матрицей A . Зададим начальное приближение x^0 и, как и в методе скорейшего спуска, будем искать очередное приближение в виде $x^1 = x^0 - \tau_1 r^0$. Параметр τ_1 при этом выберем так, чтобы минимизировать функционал невязки $(r^0, r^0) = \|r^0\|^2$. Соответствующий метод носит название *метода минимальных невязок*.

Получим формулы метода для произвольной итерации. Пусть

$$x^{k+1} = x^k - \tau_{k+1} r^k. \quad (10)$$

Тогда для невязки r^k справедливо равенство

$$r^{k+1} = r^k - \tau_{k+1} A r^k. \quad (11)$$

Возведем обе части уравнения (11) скалярно в квадрат:

$$\|r^{k+1}\|^2 = \|r^k\|^2 - 2\tau_{k+1}(Ar^k, r^k) + \tau_{k+1}^2(Ar^k, Ar^k).$$

Отсюда видно, что норма $\|r^{k+1}\|$ достигает минимума при

$$\tau_{k+1} = \frac{(Ar^k, r^k)}{(Ar^k, Ar^k)}. \quad (12)$$

Для сходимости метода минимальных невязок справедлива теорема, аналогичная теореме 1.

Теорема 2. Пусть A — симметричная положительно определенная матрица ($A = A^T > 0$). Тогда метод минимальных невязок сходится, и для погрешности метода справедлива оценка

$$\|A(x^k - x)\| \leq \rho_0^k \|A(x^0 - x)\|, \quad k = 0, 1, \dots,$$

где константа ρ_0 определена согласно (7).

Далее рассмотрим неявный итерационный процесс:

$$B \frac{x^{k+1} - x^k}{\tau_{k+1}} + Ax^k = f, \quad (13)$$

где B — симметричная положительно определенная матрица. Запишем метод (13) в виде

$$x^{k+1} = x^k - \tau_{k+1} w^k, \quad (14)$$

где $w^k = B^{-1}r^k$ — поправка метода на k -й итерации.

Метод минимальных поправок называется итерационный метод (14), в котором параметр τ_{k+1} выбирается из условия минимума нормы $\|w^{k+1}\|_B = \sqrt{(Bw^{k+1}, w^{k+1})}$ при заданном векторе w^k . В случае $B = E$ метод минимальных поправок совпадает с методом минимальных невязок.

Вычислим значение итерационного параметра τ_{k+1} , минимизирующего норму $\|w^{k+1}\|_B$. В § 2 было показано, что поправка w^k удовлетворяет однородному уравнению

$$B \frac{w^{k+1} - w^k}{\tau_{k+1}} + Aw^k = 0. \quad (15)$$

Перепишем уравнение (15) в виде

$$w^{k+1} = w^k - \tau_{k+1} B^{-1}Aw^k.$$

Отсюда следует, что

$$\|w^{k+1}\|_B^2 = \|w^k\|_B^2 - 2\tau_{k+1}(Aw^k, w^k) + \tau_{k+1}^2(B^{-1}Aw^k, Aw^k)$$

и норма $\|w^{k+1}\|_B^2$ будет минимальной, если взять

$$\tau_{k+1} = \frac{(Aw^k, w^k)}{(B^{-1}Aw^k, Aw^k)}. \quad (16)$$

Реализация метода минимальных поправок состоит в следующем. На каждой итерации решается система уравнений $Bw^k = r^k$ и находится поправка w^k . Затем решается система уравнений $Bv^k = Aw^k$, откуда определяется вектор $v^k = B^{-1}Aw^k$, необходимый для вычисления итерационного параметра τ_{k+1} .

Для изучения сходимости метода минимальных поправок рассмотрим обобщенную задачу на собственные значения

$$Ax = \lambda Bx. \quad (17)$$

Теорема 3. Пусть A и B — симметрические положительно определенные матрицы, $\lambda_{\min}(B^{-1}A)$, $\lambda_{\max}(B^{-1}A)$ — наименьшее и наибольшее собственные значения задачи (17). Тогда метод минимальных поправок сходится, и для погрешности метода справедлива оценка

$$\|A(x^k - x)\|_{B^{-1}} \leq \rho_0^k \|A(x^0 - x)\|_{B^{-1}}, \quad k = 0, 1, \dots, \quad (18)$$

где

$$\rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\lambda_{\min}(B^{-1}A)}{\lambda_{\max}(B^{-1}A)}.$$

Доказательство. Обозначим $v^k = B^{1/2}w^k$, $C = B^{-1/2}AB^{-1/2}$ и перепишем уравнение (15) относительно v^k :

$$\frac{v^{k+1} - v^k}{\tau_{k+1}} + Cv^k = 0. \quad (19)$$

Тогда выражение (16) для параметра τ_{k+1} принимает вид

$$\tau_{k+1} = \frac{(Cv^k, v^k)}{(Cv^k, Cv^k)}. \quad (20)$$

Равенства (19) и (20) для вектора v^k в точности совпадают с равенствами (11) и (12) в методе минимальных невязок. Следовательно, к методу минимальных поправок применима теорема 2, согласно которой имеет место оценка

$$\|v^k\| \leq \rho_0^k \|v^0\|, \quad (21)$$

где $\rho_0 = \frac{1 - \xi}{1 + \xi}$, $\xi = \frac{\lambda_{\min}(C)}{\lambda_{\max}(C)}$. Принимая во внимание соотношения

$$\lambda_{\min}(C) = \lambda_{\min}(B^{-1}A), \quad \lambda_{\max}(C) = \lambda_{\max}(B^{-1}A),$$

$\|v^k\| = \|B^{1/2}w^k\| = \|B^{-1/2}r^k\| = \|r^k\|_{B^{-1}} = \|A(x^k - x)\|_{B^{-1}}$, из неравенства (21) получим оценку (18). Теорема доказана.

3. Метод сопряженных градиентов. Как и предыдущие методы, метод сопряженных градиентов предназначен для решения системы (1) с симметричной положительно определенной матрицей A . Зададим начальное приближение x^0 и образуем вектор

$$x^k = x^0 + \sum_{i=0}^{k-1} a_i A^i r^0, \quad r^0 = Ax^0 - f.$$

Поставим задачу выбора параметров a_i , минимизирующих квадратичный функционал

$$F(x^k) = (Ax^k, x^k) - 2(f, x^k).$$

Так как справедливо равенство

$$\frac{\partial F(x^k)}{\partial a_j} = 2\left(Ax^k - f, \frac{\partial x^k}{\partial a_j}\right) = 2(Ax^k - f, A^j r^0),$$

то условия $F'_{a_j} = 0$ приводят к соотношениям

$$(Ax^k - f, A^j r^0) = 0, \quad j = 0, 1, \dots, k-1.$$

Подставляя сюда выражение

$$Ax^k - f = r^0 + \sum_{i=0}^{k-1} a_i A^{i+1} r^0,$$

приходим к системе линейных алгебраических уравнений относительно параметров a_1, a_2, \dots, a_{k-1} :

$$(A^j r^0, r^0) + \sum_{i=0}^{k-1} a_i (A^{i+1} r^0, A^j r^0) = 0, \quad j = 0, 1, \dots, k-1.$$

Идея метода сопряженных градиентов состоит в последовательном построении векторов

$$x^k = x^0 + \sum_{i=0}^{k-1} a_i^{(k)} A^i (Ax^0 - f),$$

минимизирующих функционал $F(x)$ при каждом k . Эта постановка задачи принципиально отличается от задачи построения оптимального чебышевского набора параметров (см. § 2), где требуется при фиксированном k найти параметры, минимизирующие $\|z^k\|$.

Приведем без вывода рекуррентные соотношения, которые связывают векторы x^k , минимизирующие $F(x)$ при каждом k :

$$x^{k+1} = x^k - (1 - \alpha_{k+1})(x^k - x^{k-1}) - \tau_{k+1} \alpha_{k+1} r^k, \quad (22)$$

где параметры τ_{k+1} и α_{k+1} вычисляются по формулам

$$\tau_{k+1} = \frac{(r^k, r^k)}{(A r^k, r^k)}, \quad k = 0, 1, \dots,$$

$$\alpha_{k+1} = \left[1 - \frac{\tau_{k+1}}{\tau_k \alpha_k} \frac{(r^k, r^k)}{(A r^{k-1}, r^{k-1})} \right]^{-1}, \quad k = 1, 2, \dots, \quad \alpha_1 = 1. \quad (23)$$

Итерационный метод (22), (23) получил название *метода сопряженных градиентов*. Как видим, метод сопряженных градиентов является двухшаговым итерационным процессом, т. е. для нахождения очередной итерации x^{k+1} используются две предыдущие итерации x^k и x^{k-1} . Начальное приближение x^0 выбирается произвольным, а вектор x^1 совпадает с первым приближением, полученным по методу скорейшего спуска.

Свойства итерационного процесса (22), (23) существенно зависят от минимального числа s такого, что

$$r^0 = \varphi_1 + \varphi_2 + \dots + \varphi_s, \quad \varphi_i \neq 0, \quad i = 1, 2, \dots, s,$$

где $\varphi_1, \varphi_2, \dots, \varphi_s$ — собственные векторы матрицы A , соответствующие различным собственным значениям $\lambda_1, \lambda_2, \dots, \lambda_s$. В предложении отсутствия погрешностей округлений вектор x^s совпадает с точным решением задачи (1), т. е. метод сопряженных градиентов сходится за конечное число итераций.

Для погрешности метода сопряженных градиентов справедлива оценка, аналогичная оценке скорости сходимости для явного итерационного метода с чебышевским набором параметров:

$$\|x^k - x\| \leq q_k \|x^0 - x\|,$$

где

$$q_k = \frac{2\rho_1^k}{1 + \rho_1^{2k}}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

Задачи к главе III

1. Привести систему уравнений $Ax = f$

$$A = \begin{pmatrix} 2 & -4 & 3 \\ 1 & 3 & 2 \\ 3 & -5 & 4 \end{pmatrix}, \quad f = \begin{pmatrix} 1 \\ 4 \\ 1 \end{pmatrix}$$

к виду $x = Bx + g$, чтобы выполнялось необходимое и достаточное условие сходимости метода простой итерации $x^{k+1} = Bx^k + g$.

2. При каких значениях α и β метод простой итерации и Зейделя для задачи $x = Bx + g$ являются сходящимися:

$$B = \begin{pmatrix} \alpha & 4 \\ \beta & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix}; \quad B = \begin{pmatrix} \alpha & 0 & \beta \\ 0 & \alpha & 0 \\ \beta & 0 & \alpha \end{pmatrix}.$$

3. Привести пример задачи $x = Bx + g$ такой, что матрица B имеет собственное значение $|\lambda| > 1$, однако метод простой итерации $x^{k+1} = Bx^k + g$ сходится при некотором начальном приближении.

4. Пусть матрица B в методе простой итерации $x^{k+1} = Bx^k + g$ имеет вид

$$B = \begin{pmatrix} \alpha & 4 \\ 0 & \alpha \end{pmatrix}, \quad 0 < \alpha < 1.$$

Показать, что норма погрешности $\|x^k - x\|_G$ в этом итерационном процессе начинет монотонно убывать лишь с некоторого номера итерации k_0 . Оценить k_0 при $\alpha \approx 1$.

5. Пусть все собственные значения матрицы A принадлежат промежутку $\delta \leq \lambda_i(A) \leq \Delta$, $\delta > 0$. Доказать, что итерационный метод

$$x^{k+1} = x^k - \tau Ax^k + \tau f$$

для решения системы $Ax = f$ сходится при $0 < \tau < 2/\Delta$.

6. Определить число итераций, необходимых для решения системы

$$x = Bx + g, \quad B = \begin{pmatrix} 7 & -4 & 1 \\ 3 & 8 & 2 \\ 1 & -4 & 9 \end{pmatrix}, \quad g = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix},$$

методом простой итерации $x^{k+1} = Bx^k + g$ с точностью $\varepsilon = 10^{-3}$.

7. Пусть все собственные значения $\lambda_i(B) = 0$, $i = 1, 2, \dots, n$. Доказать, что итерационный процесс $x^{k+1} = Bx^k + g$ для решения системы $x = Bx + g$ сойдетесь не более чем за n итераций.

8. Докажите, что итерационные процессы

$$x^{k+1} = (E - \tau AB)x^k + \tau f, \quad x^{k+1} = (E - \tau BA)x^k + \tau f$$

для решения системы $Ax = f$ сходятся или расходятся одновременно.

9. Найти условия сходимости итерационного метода

$$x^{k+1} = (2A^2 - E)x^k + 2(A - E)f$$

для решения системы $Ax = f$.

10. Исследовать сходимость итерационного процесса

$$x^{k+1} = \left(E - \frac{2}{\|A\|} A\right)x^k + \frac{2}{\|A\|} f$$

для системы $Ax = f$ с симметричной положительно определенной матрицей.

11. Найти необходимое и достаточное условие сходимости итерационного процесса

$$x^{k+1} = (E + D)z^k$$

для решения системы $Ax = f$, где D — диагональная невырожденная матрица.

12. Пусть решается система $Ax = f$ с симметричной положительно определенной матрицей A . Доказать, что итерационный процесс

$$x^{k+1} = (E - \tau A)x^k + \tau f, \quad \tau = \text{const} > 0,$$

сходится при условии

$$\tau < \frac{2}{\|A\|}, \quad \|A\| = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|.$$

13. Найти области сходимости методов Якоби и Гаусса — Зейделя для системы $Ax = f$ с матрицей

$$A = \begin{pmatrix} \alpha & \beta & 0 \\ \beta & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix}.$$

14. Система $Ax = f$ с матрицей $A = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$ решается методом Гаусса — Зейделя. Доказать, что:

- 1) если $|\alpha| > 1$, то для некоторого начального приближения $x^0 \in \mathbb{R}^n$ итерационный процесс расходится;
- 2) если $|\alpha| < 1$, то итерационный процесс сходится при любом начальном приближении $x^0 \in \mathbb{R}^n$.

15. Найти значения α и β , при которых метод Гаусса — Зейделя сходится для системы $Ax = f$ с матрицами A вида

$$A = \begin{pmatrix} \alpha & 0 & \beta \\ 0 & \alpha & 0 \\ \beta & 0 & \alpha \end{pmatrix}; \quad A = \begin{pmatrix} \alpha & \beta & 0 \\ \beta & \alpha & 0 \\ 0 & 0 & \alpha \end{pmatrix}; \quad A = \begin{pmatrix} \alpha & \alpha & 0 \\ \alpha & \beta & \beta \\ 0 & \beta & \alpha \end{pmatrix}.$$

16. Доказать, что для (2×2) -систем уравнений с симметричной положительно определенной матрицей метод Якоби сходится.

17. Доказать, что для (2×2) -систем уравнений методы Якоби и Гаусса — Зейделя сходятся или расходятся одновременно.

18. Привести пример (3×3) -системы уравнений, для которой метод Якоби сходится, а метод Гаусса — Зейделя расходится.

19. Привести пример (3×3) -системы уравнений, для которой метод Якоби расходится, а метод Гаусса — Зейделя сходится.

20. Пусть даны матрицы

$$A_1 = \begin{pmatrix} 1 & -1/2 \\ -1/2 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & -3/4 \\ -1/12 & 1 \end{pmatrix}.$$

Пусть B_1 и B_2 — соответствующие им матрицы перехода в итерационном методе Якоби. Показать, что $\rho(B_1) > \rho(B_2)$, тем самым опровергнув утверждение, что усиление диагонального доминирования влечет за собой более быструю сходимость метода Якоби.

21. Доказать, что выполнение неравенства $0 < \tau < 2$ является необходимым для сходимости метода релаксации.

22. Пусть для решения системы $Ax = f$ с симметричной положительно определенной матрицей A используется итерационный метод

$$\frac{x^{k+1} - x^k}{\tau} + A\left(\frac{x^{k+1} + x^k}{2}\right) = f.$$

Доказать, что при $\tau > 0$ этот метод сходится.

23. Пусть система $Ax = f$ с симметричной положительно определенной матрицей решается итерационным методом

$$\frac{x^{k+1} - x^k}{\tau} + A(\sigma x^{k+1} + (1 - \sigma)x^k) = f, \quad 0 \leq \sigma \leq 1.$$

Указать достаточные условия сходимости этого метода.

24. Пусть $A = A_1 + A_2$, $A_i = A_i^\top > 0$, $i = 1, 2$, $A_1 A_2 = A_2 A_1$. Доказать, что итерационный метод переменных направлений

$$\frac{x^{k+1/2} - x^k}{\tau} + A_1 x^{k+1/2} + A_2 x^k = f,$$

$$\frac{x^{k+1} - x^{k+1/2}}{\tau} + A_1 x^{k+1/2} + A_2 x^{k+1} = f$$

для решения системы $Ax = f$ сходится при любом $\tau > 0$.

25. Пусть $A = A_1 + A_2$, $A_i = A_i^\top > 0$, $i = 1, 2$, $A_1 A_2 = A_2 A_1$. Доказать, что итерационный метод стабилизирующей поправки

$$\frac{x^{k+1/2} - x^k}{\tau} + A_1 x^{k+1/2} + A_2 x^k = f,$$

$$\frac{x^{k+1} - x^{k+1/2}}{\tau} + A_2(x^{k+1} - x^k) = 0$$

для решения системы $Ax = f$ сходится при любом $\tau > 0$.

Отметим, что непосредственное развертывание определителя в равенстве (2) сопряжено с существенными вычислительными проблемами. Известно, что коэффициент p_k , $k = 1, 2, \dots, n$, характеристического многочлена равен сумме всех главных миноров* матрицы A порядка k , взятой со знаком $(-1)^{k-1}$. В частности,

$$p_1 = a_{11} + a_{22} + \dots + a_{nn} = \text{Tr } A, \quad p_n = (-1)^{n-1} \det A.$$

Величину $\text{Tr } A$ называется следом матрицы A . Число главных миноров для каждого k равно числу сочетаний C_n^k и, следовательно, определение всех коэффициентов характеристического многочлена сводится к вычислению

$$C_n^1 + C_n^2 + \dots + C_n^n = 2^n - 1$$

миноров различных порядков. Для матриц большой размерности эта задача может оказаться практически неразрешимой вследствие большого объема вычислений.

Иногда вместо характеристического многочлена (3) рассматривается нормированный характеристический многочлен (собственный многочлен) со старшим коэффициентом, равным единице. Сумма квадратов всех корней характеристического многочлена матрицы A называется спектром этой матрицы.

Предположим, что каким-то образом найдены корни характеристического многочлена $\lambda_1, \lambda_2, \dots, \lambda_n$. Пусть λ_i — простой корень уравнения (2). Подставляя λ_i в (1), можно найти соответствующий ему собственный вектор $x_i = (x_{1i}, x_{2i}, \dots, x_{ni})^\top$ как решение однородной системы линейных алгебраических уравнений

$$(A - \lambda_i E)x_i = 0. \quad (4)$$

В случае, если $\lambda_i \neq \lambda_j$ и λ_i, λ_j — простые корни, то векторы x_i, x_j являются линейно независимыми. Поэтому если все корни уравнения (2) простые, то система собственных векторов x_1, x_2, \dots, x_n линейно независима и, следовательно, образует базис пространства \mathbb{R}^n . Корни λ_i кратности m могут соответствовать от одного до

*Главным называется минор, расположенный в строках и столбцах с одинаковыми номерами.

ГЛАВА IV МЕТОДЫ РЕШЕНИЯ АЛГЕБРАИЧЕСКОЙ ПРОБЛЕМЫ СОБСТВЕННЫХ ЗНАЧЕНИЙ

В предыдущих главах мы подробно изучили круг вопросов, связанных с задачей численного решения систем линейных алгебраических уравнений. В этой главе рассмотрим другой важный класс задач, порожденный так называемой алгебраической проблемой собственных значений (ПСЗ).

По-прежнему считается, что A — квадратная невырожденная матрица с вещественными элементами a_{ij} , $i, j = 1, 2, \dots, n$. Как уже отмечалось, число λ — собственное значение матрицы A , если существует некулов вектор $x \in \mathbb{R}^n$, для которого имеет место равенство

$$Ax = \lambda x. \quad (1)$$

Вектор x называется собственным вектором матрицы A , соответствующим данному собственному значению λ .

Известно, что однородная система (1) имеет нетривиальное решение в случае, когда определитель ее матрицы равен нулю:

$$\det(A - \lambda E) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{vmatrix} = 0. \quad (2)$$

Равенство (2) называется характеристическим или якорным уравнением матрицы A , а левая часть этого уравнения

$$P(\lambda) = (-1)^n (\lambda^n - p_1 \lambda^{n-1} - p_2 \lambda^{n-2} - \dots - p_n) = \quad (3)$$

характеристическим многочленом матрицы A .

т линейно независимых собственных векторов. Собственные векторы, соответствующие различным кратным собственным значениям λ_i, λ_j , как и в случае простых корней, линейно независимы, однако система собственных векторов матрицы A при наличии кратных корней характеристического многочлена в общем случае не является базисом пространства \mathbb{R}^n .

Как и в задачах решения систем линейных алгебраических уравнений, методы решения ПСЗ подразделяются на прямые и итерационные. Методы первой группы используются для решения так называемой *полной проблемы собственных значений*, когда требуется найти все собственные значения и соответствующие им собственные векторы. Большинство прямых методов позволяют вычислять коэффициенты p_1, p_2, \dots, p_n характеристического многочлена, минуя процедуру вычисления главных миноров матрицы A . Далее с помощью каких-либо методов приближенного вычисления корней многочлена находятся собственные значения матрицы, а по ним определяются соответствующие собственные векторы. При этом во многих случаях собственные векторы матрицы удается получить путем использования промежуточных результатов вычислений, не прибегая к решению однородных СЛАУ вида (4).

В итерационных методах собственные значения матрицы λ_i находятся непосредственно (без построения характеристического многочлена) как пределы некоторых числовых последовательностей $\lambda_i^{(k)}$, а соответствующие им собственные векторы x_i — как пределы последовательностей векторов $x_i^{(k)}$. Итерационные методы применяются к решению как полной, так и *частичной проблемы собственных значений*, подразумевающей нахождение одного (чаще всего наибольшего по модулю) или нескольких собственных значений и соответствующих им собственных векторов.

§ 1. Прямые методы решения полной ПСЗ

1. Устойчивость задачи на собственные значения. Исследуем сначала вопрос устойчивости задачи

$$Ax = \lambda x. \quad (1)$$

Пусть A — матрица с комплексными элементами $\{a_{ij}\}_{i,j=1}^n$. Наряду с уравнением (1) рассмотрим *сопряженную* к (1) задачу на собственные значения

$$A^*y = \bar{\lambda}y, \quad y \neq 0. \quad (2)$$

Докажем свойство *взаимной ортогональности* собственных векторов x_i матрицы A и y_i сопряженной матрицы A^* :

$$(x_i, y_j) = 0, \quad i \neq j. \quad (3)$$

Действительно, имеем

$$Ax_i = \lambda_i x_i, \quad A^*y_j = \bar{\lambda}_j y_j. \quad (4)$$

Умножим скалярно* первое равенство (4) справа на y_j , а второе слева на x_i и вычтем одно из другого:

$$(Ax_i, y_j) - (x_i, A^*y_j) = (\lambda_i x_i, y_j) - (x_i, \bar{\lambda}_j y_j).$$

Отсюда в силу сопряженности матриц A и A^* получим

$$0 = (\lambda_i x_i, y_j) - (x_i, \bar{\lambda}_j y_j) = (\lambda_i - \bar{\lambda}_j)(x_i, y_j),$$

т. е. $(x_i, y_j) = 0$ при $\lambda_i \neq \bar{\lambda}_j$, что и требовалось доказать.

Из (3), в частности, следует, что у эрмитовых матриц собственные значения вещественны, а собственные векторы, соответствующие различным собственным значениям, ортогональны.

Пусть элементы a_{ij} заданы с некоторой погрешностью. Для простоты изложения ограничимся случаем, когда все собственные значения матрицы A попарно различны. Вместо (1) получим задачу

$$(A + \delta A)(x + \delta x) = (\lambda + \delta\lambda)(x + \delta x),$$

или с точностью до величин второго порядка малости

$$\delta Ax + \delta Ax = \lambda \delta x + \delta \lambda x.$$

Рассмотрим два случая: $\delta x = 0$, $\delta \lambda \neq 0$ и $\delta x \neq 0$, $\delta \lambda = 0$. В первом случае имеем

$$\delta Ax_i = \delta \lambda_i x_i,$$

*В случае комплексных n -векторов скалярное произведение определяется по правилу $(u, v) = u_1 \bar{v}_1 + u_2 \bar{v}_2 + \dots + u_n \bar{v}_n$.

или после скалярного умножения справа на y_i

$$(\delta Ax_i, y_i) = \delta \lambda_i (x_i, y_i).$$

Из последнего равенства следует оценка

$$|\delta \lambda_i| \leq \frac{\|x_i\| \|y_i\|}{|(x_i, y_i)|} \|\delta A\| \quad (5)$$

(под нормами вектора и матрицы здесь понимаются 2-нормы).

Величина

$$\omega_i = \frac{\|x_i\| \|y_i\|}{|(x_i, y_i)|} \quad (6)$$

называется *i-м коэффициентом перекоса* матрицы A . Очевидно, что для любых матриц $\omega_i \geq 1$ и по определению скалярного произведения векторов

$$\omega_i = \frac{1}{|\cos \alpha_i|},$$

где α_i — угол между собственным вектором x_i задачи (1) и собственным вектором y_i сопряженной задачи (2).

Пусть теперь $\delta x \neq 0$, $\delta \lambda = 0$. Тогда

$$\delta Ax_i = \lambda_i \delta x_i. \quad (7)$$

Разложим вектор δx по системе линейно независимых собственных векторов задачи (1):

$$\delta x_i = \sum_{j=1}^n \beta_{ij} x_j. \quad (8)$$

Поскольку вектор δx_i определен с произволом, потребуем, чтобы в (8) коэффициент $\beta_{ii} = 0$. Из равенств (7), (8) получим

$$\delta Ax_i = \sum_{j=1}^n (\lambda_i - \lambda_j) \beta_{ij} x_j, \quad (9)$$

или в силу свойства биортогональности (3)

$$\beta_{ij} = \frac{(\delta Ax_i, y_j)}{(\lambda_i - \lambda_j)(x_j, y_j)}.$$

Таким образом, окончательно приходим к неравенству

$$\|\delta x_i\| \leq \|x_i\| \left(\sum_{j=1}^n \frac{\omega_j}{|\lambda_i - \lambda_j|} \right) \|\delta A\|, \quad i \neq j. \quad (10)$$

Полученные оценки (5), (10) позволяют сделать следующие выводы об устойчивости задачи (1).

1. Если погрешность определения матричных элементов мала и мал i -й коэффициент перекоса, то мала и погрешность определения i -го собственного значения.

2. Если погрешность определения матричных элементов мала, малы все коэффициенты перекоса и i -е собственное значение простое, то мала и погрешность определения соответствующего i -го собственного вектора.

В частности, для эрмитовой (в вещественном случае симметричной) матрицы $A = A^*$ коэффициенты перекоса равны единице, поэтому задача нахождения собственных значений является устойчивой. Если матрица эрмитова и собственные значения простые, то устойчива и задача нахождения собственных векторов.

2. Метод А. Н. Крылова. Изложение методов вычисления коэффициентов характеристического многочлена начнем с метода А. Н. Крылова. Рассмотрим цепулевый вектор $b_0 \in \mathbb{R}^n$, например $b_0 = (1, 0, \dots, 0)^T$. По этому вектору образуем последовательность векторов $b_1 = Ab_0$, $b_2 = A^2 b_0 = A^2 b_0$ и т. д. до тех пор, пока не обнаружим первый вектор $b_m = A^m b_0$, $m \leq n$, являющийся линейной комбинацией предыдущих линейно независимых векторов:

$$b_m = q_1 b_{m-1} + q_2 b_{m-2} + \dots + q_m b_0, \quad (11)$$

$\sum_{i=1}^m q_i^2 > 0$. Для определения коэффициентов q_1, q_2, \dots, q_m соотношение (11) запишем предельно возможную линейную комбинацию

$$b_n = q_1 b_{n-1} + q_2 b_{n-2} + \dots + q_n b_0, \quad (12)$$

которую удобно записать в покомпонентном виде:

$$\begin{aligned} q_n b_{10} + q_{n-1} b_{11} + \dots + q_1 b_{1,n-1} &= b_{1n}, \\ q_n b_{20} + q_{n-1} b_{21} + \dots + q_1 b_{2,n-1} &= b_{2n}, \\ \dots &\dots \\ q_n b_{n0} + q_{n-1} b_{n1} + \dots + q_1 b_{n,n-1} &= b_{nn}, \end{aligned} \quad (13)$$

где $b_i = (b_{i1}, b_{i2}, \dots, b_{in})^T$, $i = 0, 1, \dots, n$. Определитель неоднородной системы (13) будет отличен от нуля в случае линейной независимости векторов b_0, b_1, \dots, b_{n-1} , поскольку столбцы определителя состоят из компонент этих векторов. Для выяснения факта отличия от нуля определителя построенной систему можно решать, например, методом Гаусса с выбором главного элемента по столбцу. Если после n этапов прямого хода метода исключения Гаусса система приводится к верхнему треугольному виду:

$$\begin{aligned} q_n + c_{12}q_{n-1} + c_{13}q_{n-2} + \dots + c_{1n}q_1 &= g_1, \\ \dots &\dots \\ q_{n-1} + c_{23}q_{n-2} + \dots + c_{2n}q_1 &= g_2, \\ \dots &\dots \\ q_1 &= g_n, \end{aligned}$$

это означает, что векторы b_0, b_1, \dots, b_{n-1} линейно независимы. При этом задача (13) имеет единственное решение, компоненты которого q_1, q_2, \dots, q_n последовательно вычисляются обратным ходом.

Если же после m этапов прямого хода метода Гаусса дальнейшее исключение станет невозможным (последние $n - m$ уравнений обраются в тождество $0 = 0$), в этом случае линейно независимыми будут только m первых векторов b_0, b_1, \dots, b_{m-1} . Тогда, отбрасывая последние $n - m$ столбцов и решая укороченную систему:

$$\begin{aligned} q_m + c_{12}q_{m-1} + c_{13}q_{m-2} + \dots + c_{1m}q_1 &= c_{1,m+1}, \\ q_{m-1} + c_{23}q_{m-2} + \dots + c_{2m}q_1 &= c_{2,m+1}, \\ \dots &\dots \\ q_1 &= c_{m,m+1}, \end{aligned}$$

находим коэффициенты q_1, q_2, \dots, q_m линейной комбинации

$$b_m = q_1 b_{m-1} + q_2 b_{m-2} + \dots + q_m b_0.$$

Покажем, что при $m = n$ (невырожденный случай) коэффициенты q_1, q_2, \dots, q_n соотношения

$$b_n = q_1 b_{n-1} + q_2 b_{n-2} + \dots + q_n b_0$$

являются соответствующими коэффициентами p_1, p_2, \dots, p_n характеристического многочлена

$$P(\lambda) = (-1)^n (\lambda^n - p_1 \lambda^{n-1} - p_2 \lambda^{n-2} - \dots - p_n).$$

В самом деле, согласно теореме Кели — Гамильтона (любая квадратная матрица является корнем своего характеристического многочлена), имеем

$$P(A) \equiv (-1)^n (A^n - p_1 A^{n-1} - p_2 A^{n-2} - \dots - p_n E) = 0.$$

Умножим это равенство на вектор b_0 и примем во внимание, что $A^i b_0 = b_i$, $i = 1, 2, \dots, n$:

$$b_n = p_1 b_{n-1} + p_2 b_{n-2} + \dots + p_n b_0.$$

Сравнивая последнее соотношение с (12) и учитывая линейную независимость векторов b_0, b_1, \dots, b_{n-1} , получим

$$p_1 = q_1, \quad p_2 = q_2, \dots, \quad p_n = q_n.$$

Таким образом, в невырожденном случае метод А. Н. Крылова позволяет по виду построенной линейной комбинации определить коэффициенты характеристического многочлена $P(\lambda)$ матрицы A . Для их вычисления требуется произвести порядка $5n^3$ арифметических действий.

При $m < n$ линейная комбинация имеет вид

$$b_m = q_1 b_{m-1} + q_2 b_{m-2} + \dots + q_m b_0,$$

откуда в силу условий $A^i b_0 = b_i$, $i = 1, 2, \dots, m$, вытекает равенство

$$(A^m - q_1 A^{m-1} - q_2 A^{m-2} - \dots - q_m E) b_0 = 0.$$

Последнее соотношение можно переписать в виде $Q(A) b_0 = 0$, где

$$Q(\lambda) = \lambda^m - q_1 \lambda^{m-1} - q_2 \lambda^{m-2} - \dots - q_m.$$

+

Следовательно, в случае вырождения с помощью указанной выше процедуры вычисляются коэффициенты q_1, q_2, \dots, q_m многочлена на наименьшей степени $Q(\lambda)$ такого, что $Q(A) b_0 = 0$, т. е. коэффициенты минимального анулирующего вектора b_0 многочлена матрицы A . Этот многочлен является делителем любого другого анулирующего b_0 многочлена и, в частности, характеристического многочлена $P(\lambda)$.

Таким образом, решая уравнение $Q(\lambda) = 0$, мы находим часть спектра собственных значений матрицы A . Выбирая другой начальный вектор b_0 , можно найти и остальные собственные значения. Если же изменением вектора b_0 не удастся избежать вырождения (причем система (13) дает одно и то же решение), это означает, что построен минимальный многочлен* матрицы A .

Покажем, как можно использовать промежуточные результаты вычислений собственных значений матрицы A для нахождения ее собственных векторов. Пусть с помощью метода А. Н. Крылова определены коэффициенты характеристического многочлена $P(\lambda)$, корнями которого являются попарно различные собственные значения $\lambda_1, \lambda_2, \dots, \lambda_n$. Через x_1, x_2, \dots, x_n обозначим соответствующие им линейно независимые собственные векторы**.

Будем искать собственный вектор x_i , соответствующий собственному значению λ_i , в виде линейной комбинации

$$x_i = \beta_{1i} b_{n-1} + \beta_{2i} b_{n-2} + \dots + \beta_{ni} b_0. \quad (14)$$

где векторы b_0, b_1, \dots, b_{n-1} уже найдены в процессе построения характеристического многочлена. Умножая (14) на матрицу A и учитывая, что $Ax_i = \lambda_i x_i$, $b_j = Ab_{j-1}$, $j = 1, 2, \dots, n$, получим

$$\lambda_i (\beta_{1i} b_{n-1} + \beta_{2i} b_{n-2} + \dots + \beta_{ni} b_0) = \beta_{1i} b_n + \beta_{2i} b_{n-1} + \dots + \beta_{ni} b_0.$$

Подставляя сюда соотношение $b_n = q_1 b_{n-1} + q_2 b_{n-2} + \dots + q_n b_0$ и

* Минимальным многочленом матрицы A называется многочлен наименьшей степени $Q(\lambda)$, обладающий свойством $Q(A) = 0$. Его корнями служат все различные между собой корни характеристического многочлена $P(\lambda)$.

** Дальнейшие рассуждения справедливы и для вырожденного случая, когда определены коэффициенты некоторого делителя характеристического многочлена.

собирая коэффициенты при $b_{n-1}, b_{n-2}, \dots, b_0$, приходим к равенству

$$\begin{aligned} (q_1 \beta_{11} + \beta_{21} - \lambda_i \beta_{11}) b_{n-1} + (q_2 \beta_{11} + \beta_{31} - \lambda_i \beta_{21}) b_{n-2} + \dots \\ \dots + (q_{n-1} \beta_{11} + \beta_{nn} - \lambda_i \beta_{n-1, n-1}) b_1 + (q_n \beta_{11} - \lambda_i \beta_{nn}) b_0 = 0. \end{aligned}$$

Так как векторы b_0, b_1, \dots, b_n линейно независимы, каждое из выражений в скобках должно обращаться в нуль, что приводит к требованию выполнения следующих соотношений:

$$\begin{aligned} q_1 \beta_{11} + \beta_{21} - \lambda_i \beta_{11} &= 0, \\ q_2 \beta_{11} + \beta_{31} - \lambda_i \beta_{21} &= 0, \\ \dots &\dots \\ q_{n-1} \beta_{11} + \beta_{nn} - \lambda_i \beta_{n-1, n-1} &= 0, \\ q_n \beta_{11} - \lambda_i \beta_{nn} &= 0. \end{aligned} \quad (15)$$

Последнее уравнение этой системы справедливо для любых конечных значений β_{11} в силу того, что

$$q_n \beta_{11} - \lambda_i \beta_{nn} = (\lambda_i^n - q_1 \lambda_i^{n-1} - q_2 \lambda_i^{n-2} - \dots - q_n) \beta_{11} = 0.$$

Полагая, например, $\beta_{11} = 1$, из равенств (15) получим рекуррентные формулы для вычисления искомых коэффициентов линейной комбинации (14):

$$\beta_{11} = 1, \quad \beta_{ij} = \lambda_i \beta_{i,j-1} - q_{j-1}, \quad j = 2, 3, \dots, n. \quad (16)$$

Если минимальный многочлен матрицы A не совпадает с характеристическим многочленом, то собственному значению λ_i может соответствовать несколько линейно независимых векторов. Для их отыскания необходимо выбрать другие начальные векторы b_0 и повторить описанную процедуру.

Замечание 1. Пусть в матрице A элементы a_{ij} подчинены условиям $a_{ij} = 0$, $i > j + 1$:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1,n-1} & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2,n-1} & a_{2n} \\ a_{32} & a_{33} & \dots & a_{3,n-1} & a_{3n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & & & a_{n,n-1} & a_{nn} \end{pmatrix}.$$

Матрицы такого вида называют *диагональными* или *триангулярными* или *матрицами Лессенберга*. В этом случае построенная исходя из начального вектора

$b_0 = (1, 0, \dots, 0)^T$ система уравнений (13) будет верхней треугольной, что позволяет сразу определить ее решение обратным ходом. В частности, указанная ситуация имеет место для трехдиагональной матрицы.

Замечание 2. При больших (а в ряде случаев и умеренных) размерах матрицы A вычисление коэффициентов характеристического многочлена методом А.Н.Крылова может происходить с существенной потерей точности, что объясняется возможной плохой обусловленностью матрицы системы (13).

3. Метод А.М. Данилевского. Одним из простых и весьма эффективных способов вычисления коэффициентов характеристического многочлена является метод А.М. Данилевского. Идея метода опирается на известный факт, что преобразование подобия $R^{-1}AR$ не изменяет характеристический многочлен матрицы A . Поэтому естественно попытаться привести исходную матрицу A преобразованиям подобия к такой матрице, по виду которой сразу можно было бы записать ее характеристический многочлен. Для этой цели оказалась удобной каноническая форма Фробениуса:

$$\Phi = \begin{pmatrix} 0 & 1 & & & 0 \\ 0 & 0 & 1 & & 0 \\ 0 & & \ddots & \ddots & \\ & & & 0 & 1 \\ p_n & p_{n-1} & \cdots & p_2 & p_1 \end{pmatrix}. \quad (17)$$

Непосредственная проверка показывает, что числа p_n, p_{n-1}, \dots, p_1 , составляющие последнюю строку матрицы Φ , являются коэффициентами ее характеристического многочлена, а значит, и характеристического многочлена матрицы A :

$$\det(\Phi - \lambda E) = (-1)^n(\lambda^n - p_1\lambda^{n-1} - p_2\lambda^{n-2} - \dots - p_n) = P(\lambda).$$

Следовательно, основная задача сводится к построению подобной матрицы R . Эту задачу осуществим с помощью преобразований подобия, поэтапно переводящих строки матрицы A в соответствующие строки матрицы Φ . Как и в предыдущем пункте, сначала рассмотрим невырожденный случай, предполагающий корректное выполнение всех предписанных алгоритмом действий.

Обратимся к элементарным матрицам вида

$$R_i = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & r_{i1} & \cdots & r_{i,i-1} & r_{ii} & r_{i,i+1} & \cdots & r_{in} \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix},$$

отличающимся от единичной матрицы элементами i -й строки (неортогональные преобразования с матрицами подобного типа L_i , N_i и M_i были рассмотрены в главе I при изучении прямых методов решения систем линейных алгебраических уравнений). Обратная к R_i матрица является матрицей такого же вида, у которой недиагональные элементы r_{ij} заменены на $-r_{ij}r_{ii}^{-1}$, а диагональный элемент i -й строки равен r_{ii}^{-1} :

$$R_i^{-1} = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & -r_{i1} & \cdots & -r_{i,i-1} & \frac{1}{r_{ii}} & -r_{i,i+1} & \cdots & -r_{in} \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix}.$$

Преобразование подобия с матрицей R_i заключается сначала в сложении i -го столбца, умноженного на соответствующие коэффициенты r_{ij} , со всеми остальными столбцами (при этом образуется матрица $A R_i$), а затем в вычитании из i -й строки всех остальных строк, умноженных на коэффициенты $-r_{ij}r_{ii}^{-1}$ (образуется матрица $R_i^{-1} A R_i$). Указанные свойства матриц R_i позволяют сформулировать следующий алгоритм приведения матрицы A к виду (17).

Обозначим через A_1 исходную матрицу A . На первом этапе ме-

тода А.М. Данилевского определим матрицу

$$R_2 = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\frac{a_{11}^{(1)}}{a_{12}^{(1)}} & 1 & -\frac{a_{13}^{(1)}}{a_{12}^{(1)}} & \cdots & -\frac{a_{1n}^{(1)}}{a_{12}^{(1)}} \\ a_{12}^{(1)} & a_{12}^{(1)} & a_{12}^{(1)} & \cdots & a_{12}^{(1)} \\ 0 & & 1 & \ddots & \\ & & & & 1 \end{pmatrix},$$

обратная к которой имеет вид

$$R_2^{-1} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} \\ & 1 & & & \\ 0 & & \ddots & & \\ & & & 1 & \end{pmatrix}.$$

Петрбург убедиться, что в матрице $A_2 = R_2^{-1} A_1 R_2$ первая строка совпадает с первой строкой матрицы Φ . Очевидно, что такое преобразование возможно лишь в случае, когда $a_{12}^{(1)} \neq 0$. Далее в предположении $a_{23}^{(2)} \neq 0$ аналогично образуем матрицу

$$A_3 = R_3^{-1} A_2 R_3 = R_3^{-1} R_2^{-1} A_1 R_2 R_3,$$

у которой уже первые две строки совпадают с соответствующими строками матрицы Φ . На k -м этапе определим матрицу

$$A_{k+1} = \begin{pmatrix} 0 & 1 & & & & & 0 \\ \vdots & \vdots & \ddots & & & & \\ 0 & 0 & \cdots & 1 & & & \\ a_{k+1,1}^{(k+1)} & a_{k+1,2}^{(k+1)} & \cdots & a_{k+1,k+1}^{(k+1)} & a_{k+1,k+2}^{(k+1)} & \cdots & a_{k+1,n}^{(k+1)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ a_{n,1}^{(k+1)} & a_{n,2}^{(k+1)} & \cdots & a_{n,k+1}^{(k+1)} & a_{n,k+2}^{(k+1)} & \cdots & a_{nn}^{(k+1)} \end{pmatrix},$$

рекуррентно связанные с A_k соотношением $A_{k+1} = R_{k+1}^{-1} A_k R_{k+1}$ ($a_{k,k+1}^{(k)} \neq 0$), где

$$R_{k+1} = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & -\frac{a_{kk}^{(k)}}{a_{k,k+1}^{(k)}} & \cdots & -\frac{a_{kk}^{(k)}}{a_{k,k+1}^{(k)}} & \frac{1}{a_{k,k+1}^{(k)}} & -\frac{a_{k,k+2}^{(k)}}{a_{k,k+1}^{(k)}} & \cdots & -\frac{a_{kn}^{(k)}}{a_{k,k+1}^{(k)}} \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix},$$

$$R_{k+1}^{-1} = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & a_{k1}^{(k)} & \cdots & a_{kk}^{(k)} & a_{k,k+1}^{(k)} & a_{k,k+2}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix}.$$

В результате после выполнения $n-1$ этапов приходим к матрице

$$A_n = R_n^{-1} R_{n-1}^{-1} \cdots R_2^{-1} A_1 R_2 \cdots R_{n-1} R_n = R^{-1} AR =$$

$$= \Phi = \begin{pmatrix} 0 & 1 & & & & & 0 \\ 0 & 0 & 1 & & & & \\ 0 & & \ddots & \ddots & & & \\ & & & 0 & 1 & & \\ & & & & a_{n1}^{(n)} & a_{n2}^{(n)} & \cdots & a_{n,n-1}^{(n)} & a_{nn}^{(n)} \end{pmatrix}.$$

Таким образом, исходная матрица A с помощью преобразования подобия $R^{-1}AR$ с матрицей $R = R_2 R_3 \cdots R_n$ приведена к канонической форме Фробениуса, из вида которой явно выписываются коэффициенты характеристического многочлена:

$$p_1 = a_{nn}^{(n)}; \quad p_2 = a_{n,n-1}^{(n)}, \dots, \quad p_n = a_{n1}^{(n)}.$$