ДЛЯ ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЙ

В.М. Вержбицкий

Вычислительная линейная алгебра

$$x_1 - 2x_2 + x_3 = 1,$$
$$2x_1 \quad - x_3 = 8,$$
$$2x_1 - x_2 - x_3 = 5$$

$$B = (A \mid b) = \left[\begin{array}{ccc|c} 1 & -2 & 1 & 1 \\ 2 & 0 & -3 & 8 \\ 2 & -1 & -1 & 5 \end{array}\right]$$

$$x_3 = \frac{-1{,}114}{0{,}557} = -2{,}000$$

$$x_2 = \frac{-2{,}785 - 2{,}290(-2)}{-1{,}795} = -1{,}000$$

$$x_1 = \frac{-9 - 1{,}333(-1) - 2{,}333(-2)}{-3} = -1{,}000$$

МАТЕМАТИКА

В.М. Вержбицкий

Вычислительная линейная алгебра

Допущено
УМО по образованию в области прикладной математики и управления качеством
в качестве учебного пособия
для студентов высших учебных заведений,
обучающихся по направлению подготовки
230400 «Прикладная математика»
и специальности 230401 «Прикладная математика»

Москва «Высшая школа» 2009

УДК 512
ББК 22.143
В31

Рецензенты:

кафедра информационных технологий в машиностроении Казанского государственного технического университета им. А.Н. Туполева (зав. кафедрой – д-р физ.-мат. наук, проф. *И.Х. Саитов*); академик РАН, проф. *А.М. Липанов* (Институт прикладной механики УрО РАН)

Вержбицкий В.М.

В31 Вычислительная линейная алгебра: Учеб. пособие для вузов/ В.М. Вержбицкий.–М.: Высш. шк., 2009.– 351 с.: ил.

ISBN 978-5-06-005829-1

Рассмотрены теория и практика получения треугольных, ортогональных и сингулярных разложений вещественных матриц. Показано, как эти разложения и лежащие в их основе преобразования используются для решения систем линейных алгебраических уравнений (в частности, плохо обусловленных и вырожденных), обращения и псевдообращения матриц, вычисления собственных и сингулярных значений, решения линейных задач о наименьших квадратах и некоторых других задач. Изложение материала сопровождается конкретными алгоритмами и числовыми примерами.

Для студентов вузов, обучающихся по математическим и техническим направлениям, а также для всех, кому важно знание современных численных методов линейной алгебры.

УДК 512
ББК 22.143

ISBN 978-5-06-005829-1 © ОАО «Издательство «Высшая школа», 2009

Оригинал-макет данного ивдания является собственностью издательства «Высшая школа», и его репродуцирование (воспроизведение) любым способом без согласия издательства запрещается

ОГЛАВЛЕНИЕ

Предисловие

Эта книга посвящена изучению методов решения основных задач вычислительной линейной алгебры, к каковым относятся:

- решение систем линейных алгебраических уравнений,
- линейная задача о наименьших квадратах,
- алгебраическая проблема собственных значений,
- задача о сингулярных числах матриц.

Варьируются как постановки задач (только с вещественными векторами и матрицами; проблемы больших размерностей не затрагиваются), так и методы их решения.

Многие численные методы опираются на те или иные факторизации матриц, в связи с чем используемым здесь треугольным и ортогональным разложениям квадратных матриц и лежащим в их основе преобразованиям отведена первая глава. Наиболее сложное, а именно сингулярное, разложение прямоугольных матриц рассматривается в шестой главе, а некоторые его применения — в седьмой.

Из соображений полноты и самодостаточности материала в этом учебном пособии автор не мог избежать большого пересечения нового издания с предыдущими своими книгами [11–14]. В первую очередь это относится к наполнению второй, третьей и четвертой глав, где сосредоточены сведения о прямых и итерационных методах решения, в основном хорошо обусловленных систем уравнений и задач на собственные значения.

Большое внимание в этой книге уделяется методам, основанным на ортогональных преобразованиях отражения (Хаусхолдера) и вращения (Гивенса), в частности **QR**-алгоритму. Этот алгоритм, описываемый в главе 5, не только решает проблему нахождения всех собственных чисел в общем случае несимметричных матриц, но также существенно используется в процессе вычисления сингулярных чисел.

Чтобы не затенять идеи и способы реализации рассматриваемых методов, вопросы арифметической сложности, численной устойчивости и улучшения обусловленности отнесены в послед-

нюю, восьмую главу, хотя по ходу изложения местами приходится их затрагивать.

Описание методов сопровождается рассмотрением примеров, все главы заканчиваются небольшими наборами задач для самостоятельного решения. Для успешного усвоения материала всё это желательно дополнить решением задач из специализированных сборников (таких как [5, 37, 60, 63]) и выполнением вычислительных лабораторных работ (см. [12–14, 55] и др.).

В приложении приводятся краткие сведения: 1) о векторных и матричных нормах; 2) об особенностях компьютерной арифметики. Первое — это тот инструмент, без которого немыслимо изучение численных методов, а второе — тому, кто изучает методы, следует постоянно иметь в виду, чтобы понимать разницу между «идеальным» методом и его реальным воплощением.

Для обозначения матриц и векторов в книге используется только жирный шрифт (векторно-матричный стиль): заглавные буквы — для матриц, строчные — для векторов. Возможно, несколько вольно автор распоряжается символом «:=». В разных ситуациях его следует воспринимать либо как «присвоить», «вычислить по формуле», либо как «положить по определению», «обозначить». Например, формула $\Delta(\lambda) := \det(\mathbf{A} - \lambda\mathbf{E}) = 0$ с переменной λ должна быть воспринята как уравнение, левая часть которого обозначена посредством $\Delta(\lambda)$.

Более полное и глубокое отражение затрагиваемых здесь вопросов можно найти в книгах [17, 23 ,26 ,45 ,54 ,57 ,67 ,68]; конкретные ссылки на эти и другие учебные пособия и монографии даются по ходу изложения материала.

Автор выражает искреннюю благодарность заведующему кафедрой «Прикладная математика и информатика» ИжГТУ доценту А. А. Айзиковичу за поддержку и внимание к работе, профессору А. Л. Тептину за прочтение рукописи и ценные замечания, а также всем тем, кто так или иначе причастен к появлению этой книги.

В. М. Вержбицкий

Глава 1

РАЗЛОЖЕНИЯ КВАДРАТНЫХ МАТРИЦ

§ 1.1. ВИДЫ ФАКТОРИЗАЦИЙ

Факторизацией, или ***разложением***, ***матрицы*** будем называть ее мультипликативное представление, т.е. представление в виде произведения нескольких матриц (обычно, двух - трех), обладающих теми или иными заданными свойствами. Процесс факторизации матриц осуществляется на основе различных линейных преобразований в соответствующих пространствах над векторами, отождествляемыми со столбцами или строками исходных матриц, а также матриц промежуточных этапов в применяемых алгоритмах.

Приведем несколько широко известных матричных разложений из тех, которые существенно используются в дальнейшем.

Пусть $\mathbf{A}$ — заданная вещественная квадратная матрица.

1. ***Треугольное разложение*** матрицы $\mathbf{A}$, иначе называемое $\mathbf{LU}$-***разложением*** или $\mathbf{LR}$-***разложением***, — это ее представление в виде $\mathbf{A} = \mathbf{LU}$, где $\mathbf{L}$ и $\mathbf{U}$ — соответственно нижняя (левая) и верхняя (правая) вещественные треугольные матрицы. У одной из матриц $\mathbf{L}$ или $\mathbf{U}$ диагональные элементы обычно принимают равными единице.

Для расширения области применимости таких разложений иногда вводят (подбирают) подходящую матрицу перестановок $\mathbf{P}$ и выполняют треугольную факторизацию матрицы $\mathbf{PA}$. С таким обобщенным толкованием понятия $\mathbf{LU}$-разложения при наличии стратегии построения матрицы $\mathbf{P}$ треугольная факторизация может быть выполнена для любой невырожденной матрицы $\mathbf{A}$ [26]. Подробнее $\mathbf{LU}$-разложение будет рассмотрено в следующем параграфе (см. также § 2.4), а его применения — в § 2.2 и 5.1.

2. $\mathbf{U}^T\mathbf{U}$-***разложение*** (а также аналогичное ему $\mathbf{LL}^T$-***разложение***) есть частный случай $\mathbf{LU}$-разложения для симметричных матриц. В процессе выполнения такой факторизации

требуется вычислять квадратные корни, что может повлечь за собой появление мнимых чисел, т.е. может нарушиться вещественность разложения. Поэтому применяют такие разложения, как правило, только к положительно определенным матрицам.

В несколько более широких условиях можно осуществить $\mathbf{U}^T\mathbf{DU}$-*разложение* — представление симметричных матриц в виде $\mathbf{A} = \mathbf{U}^T\mathbf{DU}$, где $\mathbf{D}$ — диагональная матрица, а $\mathbf{U}$ — верхняя треугольная с единичной диагональю.

Реализациям этих разложений симметричных матриц посвящен § 1.3, а применениям — § 2.4, 4.5 (см. также § 8.4).

3. ***Ортогональное разложение*** — представление произвольной квадратной матрицы $\mathbf{A}$ в виде произведения ортогональной матрицы $\mathbf{Q}$ на правую треугольную матрицу $\mathbf{R}$. Отсюда другое его название: $\mathbf{QR}$-***разложение.*** (Иногда при ортогональном разложении вместо правой треугольной матрицы $\mathbf{R}$ берут левую треугольную матрицу $\mathbf{L}$ и в соответствии с этим строят $\mathbf{QL}$-***разложение***.) Подчеркнем факт (который далее будет обоснован), что такое вещественное разложение может быть выполнено для любой вещественной матрицы $\mathbf{A}$.

То, как можно получить $\mathbf{QR}$-представление матрицы $\mathbf{A}$, показывается в § 1.4 (на основе преобразований отражения) и в § 1.5 (на основе преобразований вращения). Такое представление, вообще говоря, — не единственное, но различия в матрицах $\mathbf{Q}$ и $\mathbf{Q}_1$, а также соответственно в матрицах $\mathbf{R}$ и $\mathbf{R}_1$ разложений $\mathbf{A} = \mathbf{QR}$ и $\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1$ предсказуемы (см. замечание 1.2 в § 1.4).

В § 2.5, 2.6 показано, как $\mathbf{QR}$-разложение используется для решения систем линейных алгебраических уравнений, в гл. 5 на нем базируется $\mathbf{QR}$-***алгоритм*** нахождения всех собственных чисел заданной матрицы, а он, в свою очередь, в последующей главе служит составной частью процедуры сингулярного разложения.

4. ***Сингулярное разложение*** (иначе, **SVD-*разложение***) — весьма универсальная факторизация матриц. Это разложение может быть выполнено для любой прямоугольной $m \times n$-матрицы

$\mathbf{A}$ и имеет вид $\mathbf{A} = \mathbf{U\Sigma V}$, где $\mathbf{U}$ и $\mathbf{V}$ — ортогональные матрицы размера $m \times m$ и $n \times n$ соответственно, а $\mathbf{\Sigma}$ — диагональная матрица с диагональю из так называемых ***сингулярных*** (или иначе, ***главных*** [18]) ***чисел***. Эти числа суть квадратные корни из собственных чисел симметричной квадратной матрицы $\mathbf{A}^T\mathbf{A}$ (или, что то же, матрицы $\mathbf{AA}^T$, если иметь в виду только ненулевые собственные числа) и играют роль, аналогичную роли собственных чисел.

Процедура сингулярного разложения является одной из наиболее сложных в вычислительной линейной алгебре. Она описана в гл. 6 (с привлечением сведений из первой и пятой глав). Применениям сингулярных разложений посвящена гл. 7.

Разумеется, упомянутыми мультипликативными представлениями матриц далеко не исчерпывается множество возможных разложений, используемых в разных областях математики и ее приложений.

Зачастую бывает важным иметь не само представление данной матрицы в виде произведения двух треугольных или ортогональной и треугольной матриц, а процесс ее приведения к правой треугольной форме. В таком случае говорят о процедуре ***триангуляризации матрицы***. Название той или иной триангуляризации связывают с конкретными линейными (матричными) преобразованиями, лежащими в основе приведения матрицы к треугольному виду.

Замечание 1.1. Следует обратить внимание на использование здесь термина «ортогональное разложение» фактически только для названия **QR**-разложения квадратной матрицы. В других источниках [29, 45] можно встретить более широкое трактование этого термина. Так, например, в монографии [45] под ортогональным разложением $m \times n$-матрицы $\mathbf{A}$ понимают ее представление в виде $\mathbf{A} = \mathbf{PBS}$, где $\mathbf{P}$ и $\mathbf{S}$ — ортогональные матрицы размеров $m \times m$ и $n \times n$ соответственно. В такой трактовке сингулярное разложение (равно как и распространенное на прямоугольные матрицы **QR**-разложение) можно считать частным случаем ортогонального разложения. Это соподчинение типов разложений используется в дальнейшем при изучении методов решения линейной задачи наименьших квадратов (§ 7.6).

§ 1.2. LU-РАЗЛОЖЕНИЕ

Пусть $\mathbf{A} := (a_{ij})_{i,\,j=1}^{n}$ — данная $n \times n$-матрица, а $\mathbf{L} := (l_{ij})_{i,\,j=1}^{n}$ и $\mathbf{U} := (u_{ij})_{i,\,j=1}^{n}$ — соответственно нижняя (левая) и верхняя (правая) треугольные матрицы*.

Будем искать мультипликативное представление матрицы $\mathbf{A}$ с указанными множителями $\mathbf{L}$ и $\mathbf{U}$, осуществляя непосредственное перемножение этих матриц с неизвестными (искомыми) элементами l_{ij} и u_{ij} и приравнивая полученные элементы матрицы-результата соответствующим элементам a_{ij} данной матрицы. Ясно, что таких равенств — уравнений относительно l_{ij} и u_{ij} — получится столько, сколько элементов в матрице $\mathbf{A}$, т.е. n^2, в то время как суммарное число искомых элементов $n^2 + n$. Чтобы иметь возможность найти однозначное решение поставленной задачи, нужно наложить n дополнительных условий. Это можно сделать, например, полагая элементы диагонали одной из матриц $\mathbf{L}$ или $\mathbf{U}$ равными заданным наперед числам. Обычно принимают или $l_{ii} := 1$, или $u_{ii} := 1$ (треугольные матрицы с единицами на диагонали иногда называют ***унитреугольными*** [23, 26]).

Итак, ищем такие значения l_{ij} (при $i > j$) и u_{ij} (при $i \le j$), с которыми справедливо равенство

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ l_{21} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ l_{n1} & l_{n2} & \dots & 1 \end{pmatrix} \cdot \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & u_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}.$$

Выполнив перемножение матриц, на основе поэлементного приравнивания левых и правых частей приходим к $n \times n$-матрице

---

* Общепринятые обозначения **L** и **U** связаны с английскими словами *lower* (нижний) и *upper* (верхний). Существует другой стандарт обозначения: **L** и **R**, определяемый словами *left* (левый) и *right* (правый).

уравнений

$$\begin{array}{llll} u_{11}=a_{11}, & u_{12}=a_{12}, & \dots, & u_{1n}=a_{1n}, \\ l_{21}u_{11}=a_{21}, & l_{21}u_{12}+u_{22}=a_{22}, & \dots, & l_{21}u_{1n}+u_{2n}=a_{2n}, \\ \dots\dots & \dots\dots\dots\dots & \dots & \dots\dots\dots\dots \\ l_{n1}u_{11}=a_{n1}, & l_{n1}u_{12}+l_{n2}u_{22}=a_{n2}, & \dots, & l_{n1}u_{1n}+\dots+u_{nn}=a_{nn} \end{array}$$

относительно $n\times n$-матрицы неизвестных

$$\mathbf{L}+\mathbf{U}-\mathbf{E}:=\begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ l_{21} & u_{22} & \dots & u_{2n} \\ \dots & \dots & \dots & \dots \\ l_{n1} & l_{n2} & \dots & u_{nn} \end{pmatrix}. \qquad (1.1)$$

Специфика этой системы позволяет вычислять фигурирующие в (1.1) неизвестные одно за другим в следующем порядке.

Из первой строки уравнений имеем

$$u_{1j}=a_{1j} \quad (j=1,\dots,n) \ ;$$

из оставшейся части первого столбца уравнений находим

$$l_{i1}=\frac{a_{i1}}{u_{11}} \quad (i=2,\dots,n) \ ;$$

из оставшейся части второй строки —

$$u_{2j}=a_{2j}-l_{21}u_{1j} \quad (j=2,\dots,n) \ ;$$

из оставшейся части второго столбца —

$$l_{i2}=\frac{a_{i\,2}-l_{i1}u_{12}}{u_{22}} \quad (i=3,\dots,n) \ ;$$

и т.д. Наконец, последним вычисляем элемент

$$u_{nn}=a_{nn}-\sum_{k=1}^{n-1} l_{nk}u_{kn} \ .$$

Легко видеть, что все отличные от фиксированных заранее

значений 0 и 1 элементы матриц **L** и **U** можно получить, применяя всего две формулы:

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj} \quad (\text{где } i \le j), \tag{1.2}$$

$$l_{ij} = \frac{1}{u_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj} \right) \quad (\text{где } i > j). \tag{1.3}$$

При практическом выполнении **LU**-разложения матрицы **A** нужно иметь в виду следующие два обстоятельства.

Во-первых, организация вычислений по формулам (1.2), (1.3) должна предусматривать переключение счета с одной формулы на другую в соответствии с показанным выше процессом получения неизвестных, приведшим к этим формулам. Это удобно делать, ориентируясь на матрицу $\mathbf{L}+\mathbf{U}-\mathbf{E}$ неизвестных (1.1) (ее, кстати, можно интерпретировать как n^2-мерный массив для компактного хранения в памяти компьютера матрицы полученного разложения $\mathbf{L}+\mathbf{U}-\mathbf{E}$, например, на месте «затираемой» матрицы **A**). А именно, сначала, полагая $i := 1$, $j := 1, 2, \ldots, n$ в формуле (1.2), заполняем первую строку матрицы (1.1), затем по формуле (1.3) при $j := 1$, $i := 2, \ldots, n$ получаем первый столбец матрицы (1.1) (без первого элемента) и т.д.

Во-вторых, препятствием для осуществимости описанного процесса **LU**-разложения матрицы **A** может оказаться равенство нулю диагональных элементов матрицы **U**, поскольку на них выполняется деление в формуле (1.3). Как показывает детальный анализ рассматриваемой ситуации, деления на нуль не будет происходить в том случае, если главные миноры данной матрицы отличны от нуля*. Пусть это требование выполнено. Так как $u_{11} = a_{11}$, т.е. первый диагональный элемент матрицы **U** совпадает с первым главным минором матрицы **A**, то он должен быть от-

---

* Напомним, что ***главными минорами*** матрицы $\mathbf{A} = (a_{ij})_{i,j=1}^{n}$ называются определители подматриц $\mathbf{A}_k := (a_{ij})_{i,j=1}^{k}$, где $k = 1, 2, \ldots, n-1$.

личным от нуля. Тогда второй диагональный элемент матрицы **U** может быть представлен так:

$$u_{22} = a_{22} - l_{21}u_{12} = a_{22} - \frac{a_{21}}{a_{11}}a_{12} = \frac{1}{a_{11}}\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}.$$

Следовательно, он не равен нулю, если отличен от нуля второй главный минор, и т.д. Ясно, что вместо проверки на равенство нулю главных миноров данной матрицы удобнее делать такую проверку для элементов u_{jj} в процессе их вычисления, причем, чтобы уменьшить влияние погрешностей округлений, лучше сравнивать модули u_{jj} с малой положительной константой (допуском).

Из однозначности арифметических операций, выполняемых в процессе треугольной факторизации матрицы **A** по формулам (1.2), (1.3), при упомянутом требовании к главным минорам следует ее однозначная разложимость. Факт треугольной разложимости матрицы фиксируется следующей теоремой [17, 61, 71].

Теорема 1.1. *Если все главные миноры квадратной матрицы* **A** *отличны от нуля, то существуют такие нижняя* **L** *и верхняя* **U** *треугольные матрицы, что* $\mathbf{A} = \mathbf{LU}$. *Если элементы диагонали одной из матриц* **L** *или* **U** *фиксированы* (*ненулевые*), *то такое разложение единственно.*

Для определенных классов матриц указанное в теореме требование заведомо выполняется. Это относится, например, к ***матрицам с диагональным преобладанием***, т.е. к таким, для которых

$$|a_{i\,i}| > \sum_{\substack{j=1 \\ (j \neq i)}}^{n} |a_{i\,j}| \qquad \forall i \in \{1, 2, \ldots, n\}. \tag{1.4}$$

Пример 1.1. Выполнить LU-разложение матрицы $\mathbf{A} := \begin{pmatrix} 2 & -1 & 1 \\ 4 & 3 & 1 \\ 6 & -13 & 6 \end{pmatrix}$.

По формулам (1.2), (1.3) последовательно вычисляем:

$$u_{11} := a_{11} = 2, \quad u_{12} := a_{12} = -1, \quad u_{13} := a_{13} = 1;$$

$$l_{21} := \frac{a_{21}}{u_{11}} = \frac{4}{2} = 2, \quad l_{31} := \frac{a_{31}}{u_{11}} = \frac{6}{2} = 3;$$

$$u_{22} := a_{22} - l_{21}u_{12} = 3 - 2(-1) = 5, \quad u_{23} := a_{23} - l_{21}u_{13} = 1 - 2 \cdot 1 = -1;$$

$$l_{32} := \frac{1}{u_{22}}(a_{32} - l_{31}u_{12}) = \frac{1}{5}(-13 - 3(-1)) = -2;$$

$$u_{33} := a_{33} - l_{31}u_{13} - l_{32}u_{23} = 6 - 3 \cdot 1 - (-2)(-1) = 1.$$

Таким образом, равенство $\mathbf{A} = \mathbf{LU}$ в данном случае выглядит так:

$$\begin{pmatrix} 2 & -1 & 1 \\ 4 & 3 & 1 \\ 6 & -13 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & -2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & -1 & 1 \\ 0 & 5 & -1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Хранить это разложение можно в виде матрицы

$$\mathbf{L} - \mathbf{E} + \mathbf{U} = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 5 & -1 \\ 3 & -2 & 1 \end{pmatrix}.$$

Показанные выше поэлементные преобразования, приводящие удовлетворяющую условиям теоремы 1.1 матрицу $\mathbf{A}$ к мультипликативному виду $\mathbf{A} = \mathbf{LU}$, можно отождествить с выполнением некоторой последовательности простых матричных преобразований. Матрицы, с помощью которых совершаются эти преобразования, имеют следующую структуру [1, 71]: на первом этапе — это матрица вида

$$\mathbf{M}_1 := \begin{pmatrix} 1 & 0 & \dots & 0 \\ -l_{21} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ -l_{n1} & 0 & \dots & 1 \end{pmatrix}, \quad \text{где} \quad l_{i1} = \frac{a_{i1}}{a_{11}} \quad (i = 2, \dots, n),$$

на втором этапе — матрица

$$\mathbf{M}_2 := \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & -l_{32} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & -l_{n2} & 0 & \dots & 1 \end{pmatrix},$$

где элементы l_{i2} при $i = 3, \ldots, n$ определяются аналогично элементам l_{i1}, но через элементы матрицы $\mathbf{A}_1 := \mathbf{M}_1\mathbf{A}$, и т.д.

Легко видеть, что матрица $\mathbf{A}_1$ имеет нулевые поддиагональные элементы первого столбца, матрица $\mathbf{A}_2 := \mathbf{M}_2\mathbf{A}_1$ — поддиагональные элементы второго столбца,... В итоге, произведение $\mathbf{M}_{n-1}\ldots\mathbf{M}_2\mathbf{M}_1\mathbf{A}$ при подобных преобразованиях оказывается верхней треугольной матрицей, т.е. оно может быть принято за матрицу $\mathbf{U}$. Найдя произведение $\mathbf{M}_1^{-1}\mathbf{M}_2^{-1}\ldots\mathbf{M}_{n-1}^{-1}$, убеждаемся, что это нижняя треугольная матрица (с единичной диагональю). При этом получение последней не требует выполнения действий, связанных с обращением и перемножением матриц, поскольку она может быть явно выписана через определенные выше элементы l_{ij} (проверьте!):

$$\mathbf{L} := \mathbf{M}_1^{-1}\mathbf{M}_2^{-1}\ldots\mathbf{M}_{n-1}^{-1} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ l_{21} & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ l_{n1} & l_{n2} & \ldots & 1 \end{pmatrix}.$$

В силу единственности LU-разложения при фиксировании диагонали треугольной матрицы найденные таким способом матрицы $\mathbf{L} := \mathbf{M}_1^{-1}\mathbf{M}_2^{-1}\ldots\mathbf{M}_{n-1}^{-1}$ и $\mathbf{U} := \mathbf{M}_{n-1}\ldots\mathbf{M}_2\mathbf{M}_1\mathbf{A}$ будут иметь те же самые элементы, что и вычисляемые по формулам (1.2), (1.3).

Заметим, что так же употребительно фиксирование единичной диагонали у правой треугольной матрицы, т.е. представление матрицы $\mathbf{A}$ в виде:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{pmatrix} \cdot \begin{pmatrix} 1 & u_{12} & \cdots & u_{1n} \\ 0 & 1 & \cdots & u_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

В этом случае элементы l_{ij} и u_{ij} находят по формулам, аналогичным формулам (1.2) и (1.3):

$$\begin{aligned} l_{ij} &= a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj} \qquad (i \geq j); \\ u_{ij} &= \frac{1}{l_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}\right) \qquad (i < j)\,, \end{aligned} \tag{1.5}$$

где индексы фиксируются так, чтобы вычислялись поочередно: столбец $(l_{i1})_{i=1}^{n}$, затем строка $(u_{1j})_{j=2}^{n}$, столбец $(l_{i2})_{i=2}^{n}$, строка $(u_{2j})_{j=3}^{n}$ и т.д.

Для матриц трехдиагональной структуры, как показано далее в § 2.4, процесс подобного разложения существенно упрощается.

§ 1.3. $\mathbf{U}^T\mathbf{U}$- И $\mathbf{U}^T\mathbf{D}\mathbf{U}$-РАЗЛОЖЕНИЯ

Объем вычислений, требующихся для решения линейных алгебраических задач с симметричными матрицами, можно сократить почти вдвое, если учитывать симметрию при треугольной факторизации матриц.

Пусть $\mathbf{A} := (a_{ij})_{i,\,j=1}^{n}$ — данная симметричная матрица, т.е. $a_{ij} = a_{ji}$. Будем строить ее представление в виде $\mathbf{A} = \mathbf{U}^T\mathbf{U}$, где

$$\mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & u_{nn} \end{pmatrix}, \quad \mathbf{U}^T = \begin{pmatrix} u_{11} & 0 & \dots & 0 \\ u_{12} & u_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ u_{1n} & u_{2n} & \dots & u_{nn} \end{pmatrix}.$$

Аналогично тому, как это делалось в предыдущем параграфе, составим систему $\dfrac{n(n+1)}{2}$ уравнений относительно такого же

числа неизвестных (элементов матрицы **U**):

$$
\begin{aligned}
u_{11}^2 = a_{11}, \quad u_{12}u_{11} &= a_{12}, \ \ldots, & u_{11}u_{1\,n} &= a_{1\,n},\\
u_{12}^2 + u_{22}^2 &= a_{22}, \ \ldots, & u_{12}u_{1\,n} + u_{22}u_{2\,n} &= a_{2\,n},\\
&\ldots\ldots\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & &\ldots\\
& & u_{1\,n}^2 + u_{2\,n}^2 + \ldots + u_{n\,n}^2 &= a_{n\,n}.
\end{aligned}
$$

Из первой строки уравнений находим сначала $u_{11} = \sqrt{a_{11}}$, затем $u_{1j} = \dfrac{a_{1j}}{u_{11}}$ при $j = 2, \ldots, n$. Из второй — $u_{22} = \sqrt{a_{22} - u_{12}^2}$, затем $u_{2j} = \dfrac{a_{2j} - u_{12}u_{1j}}{u_{22}}$ при $j = 3, \ldots, n$, и т.д. Завершается процесс вычислением

$$u_{n\,n} = \sqrt{a_{n\,n} - \sum_{k=1}^{n-1} u_{k\,n}^2}\,.$$

Таким образом, матрица **U** может быть получена с помощью следующей совокупности формул:

$$u_{i\,i} = \sqrt{a_{i\,i} - \sum_{k=1}^{i-1} u_{k\,i}^2} \quad \text{при} \quad i = 1,\ 2, \ldots,\ n\ ; \tag{1.6}$$

$$u_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} u_{ki}u_{kj}}{u_{ii}} \quad \text{при} \quad j = 2, \ldots, n; \quad j > i \tag{1.7}$$

$$(u_{ij} := 0 \quad \text{при} \quad j < i\ )\,.$$

Осуществимости определяемого формулами (1.6) – (1.7) вещественного $\mathbf{U}^T\mathbf{U}$-***разложения*** вещественной симметричной матрицы **A** по этим формулам (называемого также ***разложени-***

*ем Холецкого**) могут помешать два обстоятельства: обращение в нуль элемента u_{ii} при каком-либо значении $i \in \{1, 2, \ldots, n\}$ и отрицательность подкоренного выражения. Так как $\mathbf{U}^T\mathbf{U}$-разложение можно считать частным случаем **LU**-разложения, то для его осуществимости достаточно потребовать неравенство нулю главных миноров данной матрицы. Для симметричных матриц это условие будет выполнено в случае их положительной определенности, что для многих участвующих в приложениях матриц имеет место.

Замечание 1.2. Формально, находя u_{11} из равенства $u_{11}^2 = a_{11}$, как и последующие диагональные элементы u_{ii} из аналогичных соответствующих равенств, мы берем только один корень из двух — для простоты положительный. С таким же успехом можно брать отрицательные или, например, чередующиеся при изменении i положительные и отрицательные корни — все это не противоречит теореме 1.1 об **LU**-разложении. Поскольку диагональ треугольной матрицы не фиксирована, здесь нет единственности; единственность разложения появляется с фиксированием знака перед арифметическим корнем.

Более универсальным, чем $\mathbf{U}^T\mathbf{U}$-разложение Холецкого, является $\mathbf{U}^*\mathbf{DU}$-*разложение*, пригодное для эрмитовых матриц, частным случаем которых являются вещественные симметричные матрицы ([3, 18, 61]). Для матриц с вещественными элементами эрмитово сопряжение равносильно транспонированию, и $\mathbf{U}^*\mathbf{DU}$-разложение реализуется в виде $\mathbf{U}^T\mathbf{DU}$-*разложения*. Под такой модификацией разложения Холецкого понимают следующее мультипликативное представление вещественной симметричной матрицы $\mathbf{A}$:

$$\mathbf{A} = \mathbf{U}^T\mathbf{DU}\,, \tag{1.8}$$

где $\mathbf{U}$ — верхняя треугольная матрица с элементами, равными

---

* Холе́цкий Андре-Луи (1875–1918) — французский военный геодезист. Часто можно встретить в литературе другое русскоязычное написание его фамилии Cholesky: *Холесский*.

единице на главной диагонали; $\mathbf{U}^T$ — транспонированная к ней матрица; $\mathbf{D}$ — диагональная матрица.

Вывод формул для выполнения данной факторизации можно осуществить аналогично предыдущему, рассматривая матричное равенство (1.8) в поэлементном представлении. Имеем:

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ u_{12} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ u_{1n} & u_{2n} & \dots & 1 \end{pmatrix} \cdot \begin{pmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_{nn} \end{pmatrix} \cdot \begin{pmatrix} 1 & u_{12} & \dots & u_{1n} \\ 0 & 1 & \dots & u_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{12} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{pmatrix}.$$

Учитывая, что в искомом разложении неизвестных величин d_{ii} и u_{ij} всего $0{,}5n(n+1)$, из получающихся при перемножении матриц и поэлементном приравнивании левых и правых частей уравнений выбираем такое же число несовпадающих:

$$d_{11} = a_{11}, \quad u_{12}d_{11} = a_{12}, \quad \dots, \quad u_{1n}d_{11} = a_{1n},$$

$$u_{12}^2 d_{11} + d_{22} = a_{22}, \dots, \quad u_{12}u_{1n}d_{11} + u_{2n}d_{22} = a_{2n},$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

$$u_{1n}^2 d_{11} + u_{2n}^2 d_{22} + \dots + d_{nn} = a_{nn}.$$

Отсюда последовательно находим:

$$d_{11} = a_{11} \quad \text{и} \quad u_{1j} = \frac{a_{1j}}{d_{11}} \quad (j = 2, \dots, n)$$

— из первой строки уравнений,

$$d_{22} = a_{22} - u_{12}^2 d_{11} \quad \text{и} \quad u_{2j} = \frac{a_{2j} - u_{12}u_{1j}d_{11}}{d_{22}} \quad (j = 3, \dots, n)$$

— из второй строки уравнений, и т.д., пока не будет найдено последнее значение

$$d_{nn} = a_{nn} - \sum_{k=1}^{n-1} u_{kn}^2 d_{kk}.$$

В результате приходим к следующей совокупности формул, по которым, циклически переключаясь с одной на другую, можно найти $\mathbf{U}^T\mathbf{DU}$-разложение симметричной матрицы $\mathbf{A} := (a_{ij})_{i,\,j=1}^{n}$:

$$d_{ii} = a_{ii} - \sum_{k=1}^{i-1} u_{ki}^2 d_{kk} \qquad \text{при} \quad i = 1, 2, \ldots, n;$$

$$u_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} u_{ki} u_{kj} d_{kk}}{d_{ii}} \qquad \text{при} \quad j = 2, \ldots, n, \quad j > i; \qquad (1.9)$$

$$u_{ij} := 0 \qquad \text{при} \quad j < i.$$

Как видим, реализация вещественного $\mathbf{U}^T\mathbf{DU}$-разложения не требует извлечения квадратных корней, что расширяет границы его применимости (факторизуемая симметричная матрица не обязана быть положительно определенной).

Пример 1.2. Дана симметричная матрица $\mathbf{A} := \begin{pmatrix} 25 & 5 & 5 \\ 5 & 10 & 4 \\ 5 & 4 & 1 \end{pmatrix}$.

Найдем ее $\mathbf{U}^T\mathbf{DU}$-разложение.

Пользуясь формулами (1.9), последовательно вычисляем ненулевые элементы матриц $\mathbf{D}$ и $\mathbf{U}$:

$$d_{11} := a_{11} = 25, \qquad u_{12} := \frac{a_{12}}{d_{11}} = \frac{5}{25} = \frac{1}{5}, \qquad u_{13} := \frac{a_{13}}{d_{11}} = \frac{5}{25} = \frac{1}{5};$$

$$d_{22} := a_{22} - u_{12}^2 d_{11} = 10 - \frac{1}{25} \cdot 25 = 9,$$

$$u_{23} := \frac{a_{23} - u_{12} u_{13} d_{11}}{d_{22}} = \frac{4 - \frac{1}{5} \cdot \frac{1}{5} \cdot 25}{9} = \frac{1}{3};$$

$$d_{33} := a_{33} - u_{13}^2 d_{11} - u_{23}^2 d_{22} = 1 - \frac{1}{25} \cdot 25 - \frac{1}{9} \cdot 9 = -1.$$

Таким образом, $\mathbf{U}^T\mathbf{DU}$-разложение матрицы $\mathbf{A}$ имеет вид

$$\begin{pmatrix} 25 & 5 & 5 \\ 5 & 10 & 4 \\ 5 & 4 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0{,}2 & 1 & 0 \\ 0{,}2 & 0{,}333 & 1 \end{pmatrix} \cdot \begin{pmatrix} 25 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0{,}2 & 0{,}2 \\ 0 & 1 & 0{,}333 \\ 0 & 0 & 1 \end{pmatrix}.$$

Формулы, по которым можно выполнить $\mathbf{U}^T\mathbf{D}\mathbf{U}$-разложение симметричной матрицы, значительно упрощаются в очень важном для дальнейшего (§ 4.5) случае, когда эта матрица — трехдиагональная.

Пусть дана матрица $\mathbf{A}$ следующего вида:

$$\mathbf{A} := \begin{pmatrix} \alpha_1 & \beta_1 & 0 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \mathbf{0} & \\ 0 & \beta_2 & \alpha_3 & & & \\ & & & \ddots & & \\ & \mathbf{0} & & \beta_{n-2} & \alpha_{n-1} & \beta_{n-1} \\ & & & 0 & \beta_{n-1} & \alpha_n \end{pmatrix}.$$

Ищем такие матрицы

$$\mathbf{U} := \begin{pmatrix} 1 & u_1 & 0 & & & \\ 0 & 1 & u_2 & & \mathbf{0} & \\ & & & \ddots & & \\ & \mathbf{0} & & & 1 & u_{n-1} \\ & & & & 0 & 1 \end{pmatrix} \quad \text{и} \quad \mathbf{D} := \begin{pmatrix} d_1 & 0 & & & \\ 0 & d_2 & & \mathbf{0} & \\ & & \ddots & & \\ & \mathbf{0} & & d_{n-1} & 0 \\ & & & 0 & d_n \end{pmatrix},$$

что $\mathbf{U}^T\mathbf{D}\mathbf{U} = \mathbf{A}$.

Выполнив соответствующие перемножения и приравнивания, приходим к системе уравнений

$$\begin{cases} d_1 = \alpha_1; \quad u_1 d_1 = \beta_1; \\ u_1^2 d_1 + d_2 = \alpha_2; \quad u_2 d_2 = \beta_2; \\ u_2^2 d_2 + d_3 = \alpha_3; \quad u_3 d_3 = \beta_3; \\ \qquad\qquad \ddots \\ u_{n-2}^2 d_{n-2} + d_{n-1} = \alpha_{n-1}; \quad u_{n-1} d_{n-1} = \beta_{n-1}; \\ u_{n-1}^2 d_{n-1} + d_n = \alpha_n \end{cases}$$

относительно искомых n элементов d_i и $n-1$ элементов u_i. Легко видеть, что эти элементы могут быть последовательно

вычислены с помощью совокупности рекуррентных формул

$$d_1 := \alpha_1, \quad u_i = \frac{\beta_i}{d_i} \ (i = 1, 2, ..., n-1),$$
$$d_i = \alpha_i - u_{i-1}^2 d_{i-1} \quad (i = 2, 3, ..., n) \ . \tag{1.10}$$

Как и в случае **LU**-разложения, при $\mathbf{U}^T\mathbf{DU}$-разложении также должно быть выполнено условие разложимости, сформулированное в теореме 1.1: главные миноры симметричной матрицы **A** должны быть отличными от нуля.

Замечание 1.3. Иногда бывает полезной ***неполная факторизация Холецкого*** [26, 50]. Под этим понимют представление симметричной матрицы **A** в виде

$$\mathbf{A} = \mathbf{U}^T\mathbf{U} + \mathbf{B},$$

где **U** — верхняя треугольная матрица, часть элементов которой не вычисляется, а фиксируется определенным образом, например, так, чтобы сохранить разреженную структуру матриц. Это уменьшает объем работы, требуемой для выполнения разложения, но порождает матрицу-остаток **B**.

Примером неполной факторизации Холецкого может служить матричное равенство

$$\begin{pmatrix} 9 & 3 & 0 & 3 \\ 3 & 5 & 2 & 0 \\ 0 & 2 & 2 & 0 \\ 3 & 0 & 0 & 5 \end{pmatrix} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} 3 & 1 & 0 & 1 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix},$$

при получении которого элементы u_{13}, u_{24}, u_{34} треугольной матрицы **U**, соответствующие нулевым элементам данной матрицы **A**, также полагались равными нулю, а остальные элементы вычислялись, как обычно, по формулам (1.6) – (1.7); матрица **B** находилась как невязка $\mathbf{A} - \mathbf{U}^T\mathbf{U}$.

Одной из сфер приложения неполной $\mathbf{U}^T\mathbf{U}$-факторизации и аналогичной ей неполной $\mathbf{U}^T\mathbf{DU}$-факторизации является процедура предобусловливания при решении систем линейных алгебраических уравнений методом сопряженных градиентов с целью повышения численной устойчивости метода (см. § 8.4).

§ 1.4. ПРЕОБРАЗОВАНИЕ ХАУСХОЛДЕРА И QR-РАЗЛОЖЕНИЕ

Пусть **w** — некоторый фиксированный вектор (столбец) евклидова пространства $\mathbb{R}_n$ со скалярным произведением $(\cdot, \cdot)$, такой

что

$$\|\mathbf{w}\|_2 = \sqrt{(\mathbf{w}, \mathbf{w})} = \sqrt{\mathbf{w}^T \mathbf{w}} = 1. \qquad (1.11)$$

Образованная с его помощью $n \times n$-матрица

$$\mathbf{H} := \mathbf{E} - 2\mathbf{w}\ \mathbf{w}^T \qquad (1.12)$$

называется ***матрицей Хаусхолдера****.

Чтобы выявить простейшие геометрические свойства ***преобразования Хаусхолдера***, осуществляемого посредством матрицы $\mathbf{H}$, посмотрим, что представляет собой вектор $\mathbf{y}$, служащий при этом преобразовании образом произвольного вектора $\mathbf{x} \in \mathbb{R}_n$:

$$\mathbf{y} = \mathbf{H}\mathbf{x} = \mathbf{x} - 2\mathbf{w}\mathbf{w}^T\mathbf{x} = \mathbf{x} - 2(\mathbf{x}, \mathbf{w})\mathbf{w}. \qquad (1.13)$$

Если взять $\mathbf{x}$ коллинеарным $\mathbf{w}$, т.е. $\mathbf{x} := \alpha\mathbf{w}$, где $\alpha \neq 0$ — const, то, в силу (1.11), имеем $(\mathbf{x}, \mathbf{w}) = \alpha(\mathbf{w}, \mathbf{w}) = \alpha$. В таком случае, согласно (1.13), получаем

$$\mathbf{y} = \mathbf{x} - 2\alpha\mathbf{w} = \mathbf{x} - 2\mathbf{x} = -\mathbf{x}.$$

Если же вектор $\mathbf{x}$ ортогонален вектору $\mathbf{w}$, то $(\mathbf{x}, \mathbf{w}) = 0$ и, значит, из (1.13) следует $\mathbf{y} = \mathbf{x}$.

Итак, преобразование Хаусхолдера действует на векторы n-мерного евклидова пространства следующим образом: *векторы, ортогональные определяющему матрицу Хаусхолдера* (1.12) *вектору* $\mathbf{w}$, *оно оставляет неизменными, а векторы, коллинеарные* $\mathbf{w}$, *переводит в противоположные — отражает.* Отсюда проистекают другие названия матрицы $\mathbf{H}$ и соответствующего ей преобразования — ***матрица отражения*** и ***преобразование отражения***.

Непосредственным перемножением вектора $\mathbf{w}$ на вектор $\mathbf{w}^T$

---

*Ха́усхо́лдер Олстон (1904—1993) — американский математик. Занимался вопросами вариационного исчисления, математической биологии, численного анализа.

находим:

$$\mathbf{w}\mathbf{w}^T := \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \cdot \left(w_1; w_2; \ldots; w_n\right) = \begin{pmatrix} w_1^2 & w_1 w_2 & \ldots & w_1 w_n \\ w_2 w_1 & w_2^2 & \ldots & w_2 w_n \\ \ldots & \ldots & \ldots & \ldots \\ w_n w_1 & w_n w_2 & \ldots & w_n^2 \end{pmatrix}.$$

Очевидная симметричность матрицы $\mathbf{w}\mathbf{w}^T$ влечет симметричность матрицы $\mathbf{H}$. Пользуясь этим, имеем

$$\mathbf{H}\mathbf{H}^T = \mathbf{H}^2 = \mathbf{E} - 4\mathbf{w}\mathbf{w}^T + 4\mathbf{w}\left(\mathbf{w}^T\mathbf{w}\right)\mathbf{w}^T = \mathbf{E}$$

[поскольку $\mathbf{w}^T\mathbf{w} = 1$, в силу (1.11)]. Полученное в итоге равенство $\mathbf{H}\mathbf{H}^T = \mathbf{E}$ означает, что *матрица Хаусхолдера — ортогональная.*

Одним из важнейших свойств ортогональных преобразований является *сохранение длин преобразуемых векторов.* Благодаря этому свойству, согласно вышепоказанному, можно утверждать, что

$$\|\mathbf{y}\|_2 = \|\mathbf{H}\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \quad \forall\, \mathbf{x} \in \mathbb{R}_n. \tag{1.14}$$

Равенство (1.14) играет существенную роль для конкретизации векторов $\mathbf{w}$ при построении таких матриц Хаусхолдера, чтобы преобразованиями с их помощью достичь определенных целей.

Пусть требуется подобрать такую матрицу Хаусхолдера, которая преобразовывала бы заданный ненулевой вектор $\mathbf{x} := \left(x_1; x_2; \ldots; x_n\right)^T$ в вектор $\mathbf{y}$, коллинеарный орту $\mathbf{e}_1$, т.е. в вектор $\mathbf{y} = y_1\mathbf{e}_1 = \left(y_1; 0; \ldots; 0\right)^T$.

Согласно (1.14), должно быть

$$\|\mathbf{y}\|_2 = |y_1| = \|\mathbf{x}\|_2,$$

откуда следует, что $y_1 = \pm\|\mathbf{x}\|_2$ и, значит, $\mathbf{y} = \pm\|\mathbf{x}\|_2\,\mathbf{e}_1$. Подставляя это выражение вектора $\mathbf{y}$ в равенство (1.13), имеем

$$\pm\|\mathbf{x}\|_2\,\mathbf{e}_1 = \mathbf{x} - 2\left(\mathbf{x}, \mathbf{w}\right)\mathbf{w},$$

или иначе

$$2\left(\mathbf{x}, \mathbf{w}\right)\mathbf{w} = \mathbf{x} \mp \|\mathbf{x}\|_2\,\mathbf{e}_1.$$

Так как в последнем равенстве величина $2(\mathbf{x}, \mathbf{w})$ — скалярная, то можно сказать, что вектор

$$\tilde{\mathbf{w}} := \mathbf{x} \mp \|\mathbf{x}\|_2 \mathbf{e}_1 \tag{1.15}$$

(точнее, любой из двух фигурирующих здесь векторов) задает нужный вектор по направлению. Остается привести его к единичной длине, что требуется «стартовым» условием (1.11). Таким образом, можно принять

$$\mathbf{w} := \frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|_2} = \frac{\tilde{\mathbf{w}}}{\sqrt{(\tilde{\mathbf{w}}, \tilde{\mathbf{w}})}}. \tag{1.16}$$

Свободой выбора одного из двух векторов в выражении (1.15) распоряжаются так, чтобы процесс вычислений был более устойчивым. С этой целью знак в формуле для $\tilde{\mathbf{w}}$ выбирают таким, при котором не будет происходить вычитания, т.е. исключается возможность пропадания значащих цифр при вычитании близких чисел. Достигают этого, полагая

$$\tilde{\mathbf{w}} := (x_1 - \beta; x_2; \ldots; x_n)^T, \tag{1.17}$$

$$\beta := \operatorname{sgn}_+(-x_1)\|\mathbf{x}\|_2 = \operatorname{sgn}_+(-x_1)\sqrt{\sum_{i=1}^{n} x_i^2}, \tag{1.18}$$

где $\operatorname{sgn}_+(x) := \begin{cases} x, & \text{если } x \geq 0, \\ -x, & \text{если } x < 0. \end{cases}$

Для сокращения объема вычислений при подсчете нормы вектора $\tilde{\mathbf{w}}$, требуемой выражением (1.16), можно воспользоваться уже подсчитанным значением β. Действительно, раскрывая скалярный квадрат вектора $\tilde{\mathbf{w}}$ из (1.17), имеем:

$$\|\tilde{\mathbf{w}}\|_2^2 = (\tilde{\mathbf{w}}, \tilde{\mathbf{w}}) = (x_1 - \beta)^2 + x_2^2 + \ldots + x_n^2 = \sum_{i=1}^{n} x_i^2 - 2\beta x_1 + \beta^2. \tag{1.19}$$

Но, согласно (1.18), $\sum_{i=1}^{n} x_i^2 = \beta^2$. Поэтому из (1.19) следует, что

$$\|\tilde{\mathbf{w}}\|_2^2 = 2\beta^2 - 2\beta x_1. \tag{1.20}$$

Итогом проведенных рассуждений является следующее заклю-

чение:

если матрицу Хаусхолдера $\mathbf{H} := \mathbf{E} - 2\,\mathbf{w}\,\mathbf{w}^T$ *строить с помощью вектора* $\mathbf{w} := \mu\left(x_1 - \beta; x_2; \ldots; x_n\right)^T$, *где* $\beta := \operatorname{sgn}_+(-x_1)\sqrt{\sum_{i=1}^{n} x_i^2}$, *а* $\mu := \dfrac{1}{\sqrt{2\beta^2 - 2\beta x_1}}$, *то заданный вектор* $\mathbf{x} := \left(x_1; x_2; \ldots; x_n\right)^T$ *(отличный от нулевого) преобразуется матрицей* $\mathbf{H}$ *в вектор, все компоненты которого нули, кроме первой, причем первую компоненту, очевидно, можно считать равной* β *(т.е. в вектор* $\mathbf{Hx} = \left(\beta; 0; \ldots; 0\right)^T$*).*

Пример 1.3. На векторе $\mathbf{x} := \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix}$ проследим за действиями, преобразующими его в вектор $\mathbf{y}$ с нулями во второй и в третьей позициях.

Имеем:

$$\|\mathbf{x}\|_2 = \sqrt{2^2 + (-2)^2 + 1} = 3; \quad \beta := \operatorname{sgn}_+(-x_1)\|\mathbf{x}\|_2 = -1 \cdot 3 = -3;$$

$$\mu^2 := \frac{1}{2\beta^2 - 2\beta x_1} = \frac{1}{2 \cdot 9 - 2 \cdot (-3) \cdot 2} = \frac{1}{30}; \quad \mathbf{w} := \mu \begin{pmatrix} x_1 - \beta \\ x_2 \\ x_3 \end{pmatrix} = \frac{1}{\sqrt{30}} \begin{pmatrix} 5 \\ -2 \\ 1 \end{pmatrix};$$

$$\mathbf{w}^T = \frac{1}{\sqrt{30}}(5; -2; 1); \quad 2\,\mathbf{w}\,\mathbf{w}^T = \frac{1}{15} \begin{pmatrix} 25 & -10 & 5 \\ -10 & 4 & -2 \\ 5 & -2 & 1 \end{pmatrix};$$

$$\mathbf{H} := \mathbf{E} - 2\,\mathbf{w}\,\mathbf{w}^T = \frac{1}{15} \begin{pmatrix} -10 & 10 & -5 \\ 10 & 11 & 2 \\ -5 & 2 & 14 \end{pmatrix}$$

и

$$\mathbf{y} := \mathbf{Hx} = \frac{1}{15} \begin{pmatrix} -10 & 10 & -5 \\ 10 & 11 & 2 \\ -5 & 2 & 14 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix} = \frac{1}{15} \begin{pmatrix} -45 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -3 \\ 0 \\ 0 \end{pmatrix}.$$

Заметим, что той же матрицей $\mathbf{H}$, ввиду ее ортогональности, вектор $\mathbf{y}$ можно обратно преобразовать в $\mathbf{x}$:

$$\mathbf{x} := \mathbf{Hy} = \frac{1}{15} \begin{pmatrix} -10 & 10 & -5 \\ 10 & 11 & 2 \\ -5 & 2 & 14 \end{pmatrix} \cdot \begin{pmatrix} -3 \\ 0 \\ 0 \end{pmatrix} = \frac{1}{15} \begin{pmatrix} 30 \\ -30 \\ 15 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix}.$$

Поставим теперь следующую задачу: ортогональными преобразованиями привести $n \times n$-матрицу $\mathbf{A} := (a_{ij})$ к треугольному виду. Иначе, осуществить ***QR-разложение*** матрицы $\mathbf{A}$, т.е. описать процедуру, посредством которой получается равенство

$$\mathbf{A} = \mathbf{QR}, \tag{1.21}$$

где $\mathbf{Q}$ — ортогональная матрица, а $\mathbf{R}$ — правая треугольная.

Будем решать эту задачу поэтапно.

На первом этапе отдадим роль преобразуемого вектора $\mathbf{x}$ в предыдущих рассуждениях и формулах первому столбцу $(a_{11}; a_{21}; \ldots; a_{n1})^T$ матрицы $\mathbf{A}$. Согласно им, построив матрицу Хаусхолдера $\mathbf{H}_1 := \mathbf{E} - 2\mathbf{w}_1\mathbf{w}_1^T$ с помощью вектора

$$\mathbf{w}_1 := \mu_1 (a_{11} - \beta_1; a_{21}; \ldots; a_{n1})^T$$

и скаляров

$$\beta_1 := \operatorname{sgn}_+(-a_{11}) \sqrt{\sum_{k=1}^{n} a_{k1}^2}, \quad \mu_1 := \frac{1}{\sqrt{2\beta_1^2 - 2\beta_1 a_{11}}}$$

и применив ее к матрице $\mathbf{A}$, получим матрицу

$$\mathbf{A}_1 = \mathbf{H}_1\mathbf{A}$$

со столбцом нулей под первым диагональным элементом, т.е. матрицу вида

$$\mathbf{A}_1 := \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \ldots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & \ldots & a_{2n}^{(1)} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & a_{n2}^{(1)} & \ldots & a_{nn}^{(1)} \end{pmatrix} \quad (\text{где } a_{11}^{(1)} := \beta_1).$$

На втором этапе нужно поступить таким же образом с $(n-1) \times (n-1)$-подматрицей матрицы $\mathbf{A}_1$, которая получается вычеркиванием в $\mathbf{A}_1$ первой строки и первого столбца. Легко проверить, что это равносильно применению ко всей матрице $\mathbf{A}_1$

преобразования Хаусхолдера, определяемого формулами

$$\mathbf{H}_2 := \mathbf{E} - 2\mathbf{w}_2\mathbf{w}_2^T, \quad \mathbf{w}_2 := \mu_2\left(0; a_{22}^{(1)} - \beta_2; a_{32}^{(1)}; \ldots; a_{n2}^{(1)}\right)^T,$$

$$\beta_2 := \operatorname{sgn}_+\left(-a_{22}^{(1)}\right)\sqrt{\sum_{k=2}^{n}\left(a_{k2}^{(1)}\right)^2}, \quad \mu_2 := \frac{1}{\sqrt{2\beta_2^2 - 2\beta_2 a_{22}^{(1)}}}.$$

Для этого достаточно лишь убедиться, что матрица $\mathbf{H}_2$ имеет структуру вида

$$\begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & * & * & \ldots & * \\ 0 & * & * & \ldots & * \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & * & * & \ldots & * \end{pmatrix},$$

означающую неизменность первых строки и столбца при выполнении преобразования

$$\mathbf{A}_2 = \mathbf{H}_2\mathbf{A}_1.$$

Результат первых двух этапов — это матрица

$$\mathbf{A}_2 := \mathbf{H}_2\mathbf{H}_1\mathbf{A} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \ldots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \ldots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(2)} & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & a_{n3}^{(2)} & \ldots & a_{nn}^{(2)} \end{pmatrix}.$$

Ясно, что для приведения данной $n \times n$-матрицы $\mathbf{A}$ к треугольному виду потребуется $n-1$ таких этапов, причем i-й этап определяется формулами

$$\mathbf{H}_i := \mathbf{E} - 2\mathbf{w}_i\mathbf{w}_i^T,$$

$$\mathbf{w}_i := \mu_i\left(0; \ldots; a_{ii}^{(i-1)} - \beta_i; a_{i+1,i}^{(i-1)}; \ldots; a_{ni}^{(i-1)}\right)^T, \qquad (1.22)$$

$$\beta_i := \operatorname{sgn}_+\left(-a_{ii}^{(i-1)}\right)\sqrt{\sum_{k=i}^{n}\left(a_{ki}^{(i-1)}\right)^2}\,, \quad \mu_i := \frac{1}{\sqrt{2\beta_i^2 - 2\beta_i a_{ii}^{(i-1)}}}. \quad (1.23)$$

Итог всей процедуры из $n-1$ этапов — матрица треугольного вида

$$\mathbf{R} := \mathbf{A}_{n-1} = \mathbf{H}_{n-1}\ldots\mathbf{H}_2\mathbf{H}_1\mathbf{A} = \mathbf{Q}^T\mathbf{A},$$

где через $\mathbf{Q}^T$ обозначена матрица, представляющая собой произведение $n-1$ ортогональных матриц Хаусхолдера $\mathbf{H}_{n-1}\ldots\mathbf{H}_2\mathbf{H}_1$. Так как произведением ортогональных матриц является тоже ортогональная матрица *, то равенство $\mathbf{R} = \mathbf{Q}^T\mathbf{A}$ можно обратить умножением слева на $\left(\mathbf{Q}^T\right)^{-1} = \left(\mathbf{Q}^T\right)^T = \mathbf{Q}$. В результате приходим к искомой факторизации (1.21) с $\mathbf{Q} := \mathbf{H}_1\mathbf{H}_2\ldots\mathbf{H}_{n-1}$.

Единственным препятствием при выполнении i-го этапа описанных преобразований может оказаться невозможность вычислить величину μ_i^2 по причине обращения в нуль знаменателя в ее выражении. Но этот знаменатель есть квадрат нормы вектора

$$\tilde{\mathbf{w}}_i := \left(0;\ldots;0;\ a_{ii}^{(i-1)} - \beta_i;\ a_{i+1,i}^{(i-1)};\ldots;a_{ni}^{(i-1)}\right)^T$$

(см. (1.20)), и его равенство нулю равносильно тому, что $\tilde{\mathbf{w}}_i$ — нуль-вектор. Следовательно, если в процедуре **QR**-факторизации матрицы **A** предусмотреть игнорирование обработки столбцов, которые на соответствующем этапе уже имеют нужный вид, т.е. у них под диагональным элементом (а возможно, и сам диагональный элемент) — нули, то ситуации с делением на нуль никогда не возникнет.

---

* Действительно, если квадратные матрицы **H** и **G** таковы, что $\mathbf{H}\mathbf{H}^T = \mathbf{H}^T\mathbf{H} = \mathbf{E}$ и $\mathbf{G}\mathbf{G}^T = \mathbf{G}^T\mathbf{G} = \mathbf{E}$, то $(\mathbf{H}\mathbf{G})(\mathbf{H}\mathbf{G})^T = \mathbf{H}\mathbf{G}\mathbf{G}^T\mathbf{H}^T = \mathbf{E}$ и $(\mathbf{H}\mathbf{G})^T(\mathbf{H}\mathbf{G}) = \mathbf{G}^T\mathbf{H}^T\mathbf{H}\mathbf{G} = \mathbf{E}$.

Таким образом, справедлива следующая теорема.

Теорема 1.2 (о **QR-разложении**). *Преобразованиями Хаусхолдера любая квадратная матрица с вещественными элементами может быть представлена в виде произведения вещественных ортогональной и правой треугольной матриц.*

Замечание 1.4. Уже из самого процесса построения матриц Хаусхолдера ясно, что **QR**-разложение, вообще говоря, не единственно (обратим внимание на двузначность в равенстве (1.15)). Если учесть, что кроме преобразований отражения в процессе **QR**-факторизации можно использовать и другие ортогональные преобразования, например плоские вращения (см. § 1.5), встает вопрос о том, насколько существенны отличия в разных **QR**-представлениях квадратных матриц. Ответом на него служит следующее утверждение.

Если для невырожденной вещественной матрицы $\mathbf{A}$ *имеют место два* **QR**-*разложения:* $\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1$ *и* $\mathbf{A} = \mathbf{Q}_2\mathbf{R}_2$, *то ортогональные матрицы* $\mathbf{Q}_1$ *и* $\mathbf{Q}_2$ *могут различаться только знаками столбцов, а правые треугольные матрицы* $\mathbf{R}_1$ *и* $\mathbf{R}_2$ – *только знаками строк.*

Доказательство этого утверждения (см., например, [68]) опирается на очевидное равенство $\mathbf{Q}_2^T\mathbf{Q}_1 = \mathbf{R}_2\mathbf{R}_1^{-1}$, из которого следует, что правая треугольная матрица $\mathbf{R} := \mathbf{R}_2\mathbf{R}_1^{-1}$ тоже является ортогональной; отсюда имеем равенство $\mathbf{R}^T\mathbf{R} = \mathbf{E}$, означающее, что $\mathbf{R} = \mathrm{diag}(\pm 1, \pm 1, ..., \pm 1)$.

Замечание 1.5. Как видим, в отличие от **LU**-разложения, описанного в § 1.2, **QR**-разложение можно применять даже к вырожденным матрицам. Наличие нуля на диагонали i -го столбца на i -м этапе **QR**-факторизации в случае, когда в нем имеются ненулевые элементы под диагональю, очевидно, означает подсчет скаляров β_i и μ_i по формулам

$$\beta_i := \sqrt{\sum_{k=i}^{n}\left(a_{ki}^{(i-1)}\right)^2}\,, \quad \mu_i := \frac{1}{\beta_i\sqrt{2}}\,.$$

Пример 1.4. Выполним **QR**-разложение заведомо вырожденной матрицы

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

С учетом замечания 1.5 имеем:

$$\mathbf{H}_1 := \mathbf{E} \quad \Rightarrow \quad \mathbf{A}_1 := \mathbf{H}_1\mathbf{A} = \mathbf{A};$$

$$\beta_2 := \left\|(0;0;-1)^T\right\|_2 = 1, \quad \mu_2 := \frac{1}{\beta_2\sqrt{2}} = \frac{1}{\sqrt{2}}, \quad \mathbf{w}_2 = \frac{1}{\sqrt{2}}(0;-1;-1)^T;$$

$$\mathbf{H}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - 2\cdot\frac{1}{2}\cdot\begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}\cdot(0;-1;-1) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix};$$

$$\mathbf{R} := \mathbf{A}_2 = \mathbf{H}_2\mathbf{A}_1 = \begin{pmatrix} 0 & 2 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & -1 \end{pmatrix}, \quad \mathbf{Q} := \mathbf{H}_1\mathbf{H}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}.$$

Фактическое построение матрицы Хаусхолдера i-го этапа преобразований по формулам (1.22)–(1.23) дает матрицу

$$\mathbf{H}_i := \begin{pmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1-2\mu_i^2\tilde{w}_i^2 & -2\mu_i^2\tilde{w}_i a_{i+1,i}^{(i-1)} & \dots & -2\mu_i^2\tilde{w}_i a_{ni}^{(i-1)} \\ 0 & \dots & 0 & -2\mu_i^2\tilde{w}_i a_{i+1,i}^{(i-1)} & 1-2\mu_i^2 a_{i+1,i}^{(i-1)} a_{i+1,i}^{(i-1)} & \dots & -2\mu_i^2 a_{i+1,i}^{(i-1)} a_{ni}^{(i-1)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & -2\mu_i^2\tilde{w}_i a_{ni}^{(i-1)} & -2\mu_i^2 a_{i+1,i}^{(i-1)} a_{ni}^{(i-1)} & \dots & 1-2\mu_i^2 a_{ni}^{(i-1)} a_{ni}^{(i-1)} \end{pmatrix}$$

(ее левый верхний угол — единичная $(i-1)\times(i-1)$-матрица, а $\tilde{w}_i := a_{ii}^{(i-1)} - \beta_i$). Результатом умножения этой матрицы на матрицу $\mathbf{A}_{i-1}$ предыдущего этапа, имеющую вид

$$\mathbf{A}_{i-1} := \begin{pmatrix} \beta_1 & a_{12}^{(1)} & \dots & a_{1,i-1}^{(1)} & a_{1i}^{(1)} & a_{1,i+1}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & \beta_2 & \dots & a_{2,i-1}^{(2)} & a_{2i}^{(2)} & a_{2,i+1}^{(2)} & \dots & a_{2n}^{(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \beta_{i-1} & a_{i-1,i}^{(i-1)} & a_{i-1,i+1}^{(i-1)} & \dots & a_{i-1,n}^{(i-1)} \\ 0 & 0 & \dots & 0 & a_{ii}^{(i-1)} & a_{i,i+1}^{(i-1)} & \dots & a_{in}^{(i-1)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & a_{ni}^{(i-1)} & a_{n,i+1}^{(i-1)} & \dots & a_{nn}^{(i-1)} \end{pmatrix},$$

является матрица

$$\mathbf{A}_i := \mathbf{H}_i \mathbf{A}_{i-1} = \begin{pmatrix} \beta_1 & a_{12}^{(1)} & \dots & a_{1,i-1}^{(1)} & a_{1i}^{(1)} & a_{1,i+1}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & \beta_2 & \dots & a_{2,i-1}^{(2)} & a_{2i}^{(2)} & a_{2,i+1}^{(2)} & \dots & a_{2n}^{(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \beta_{i-1} & a_{i-1,i}^{(i-1)} & a_{i-1,i+1}^{(i-1)} & \dots & a_{i-1,n}^{(i-1)} \\ 0 & 0 & \dots & 0 & \beta_i & a_{i,i+1}^{(i)} & \dots & a_{in}^{(i)} \\ 0 & 0 & \dots & 0 & 0 & a_{i+1,i+1}^{(i)} & \dots & a_{i+1,n}^{(i)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & a_{n,i+1}^{(i)} & & a_{nn}^{(i)} \end{pmatrix}.$$

Ее новые элементы $a_{mj}^{(i)}$ подсчитывают по формуле

$$a_{mj}^{(i)} = a_{mj}^{(i-1)} - 2\mu_i^2 a_{mi}^{(i-1)} \sum_{k=i}^{n} a_{ki}^{(i-1)} a_{kj}^{(i-1)} \qquad (1.24)$$

при $m = i, \dots, n$, $j = i+1, \dots, n$ с учетом того, что здесь для единообразия произведено замещение элемента $a_{ii}^{(i-1)}$ элементом $a_{ii}^{(i-1)} := a_{ii}^{(i-1)} - \beta_i$.

Алгоритм QR-разложения квадратной матрицы

Входные данные: *матрица* $\mathbf{A} := (a_{ij})_{i,j=1}^{n}$.

Выходные данные: *верхняя треугольная матрица* $\mathbf{R} := (r_{ij})_{i,j=1}^{n}$, *ортогональная матрица* $\mathbf{Q} := (q_{ij})_{i,j=1}^{n}$.

1. $\mathbf{R} := \mathbf{0}$, $\mathbf{Q} := \mathbf{E}$, $\mathbf{H} := \mathbf{E}$.
2. Для $i := 1, \dots, n-1$:

   2.1. $s := \sum_{k=i}^{n} a_{ki}^2$.

   2.2. Если $s = 0$, то переход к началу шага 2.

   2.3. Если $a_{ii} \le 0$, то $\beta := \sqrt{s}$, иначе $\beta := -\sqrt{s}$.

2.4. $\gamma := \dfrac{1}{\beta(\beta - a_{ii})}$.

2.5. $a_{ii} := a_{ii} - \beta$.

2.6. $\mathbf{H} := \mathbf{E}$.

2.7. Для $m := i, \ldots, n$:

для $j := i, \ldots, n$:

если $j \neq m$, то $h_{mj} := -\gamma a_{mi} a_{ji}$, иначе $h_{mj} := 1 - \gamma a_{mi}^2$;

$h_{jm} := h_{mj}$.

2.8. Для $m := i, \ldots, n$:

для $j := i+1, \ldots, n$:

$$t_{mj} := a_{mj} - \gamma a_{mi} \sum_{k=i}^{n} a_{ki} a_{kj} .$$

2.9. Для $m := i, \ldots, n$:

для $j := i+1, \ldots, n$:

$a_{mj} := t_{mj}$.

2.10. $a_{ii} := \beta$.

2.11. Для $m := i+1, \ldots, n$:

$a_{mi} := 0$.

2.12. Для $m := 1, \ldots, n$:

для $j := 1, \ldots, n$:

$$t_{mj} := \sum_{k=i}^{n} q_{mk} h_{kj} .$$

2.13. Для $m := 1, \ldots, n$:

для $j := 1, \ldots, n$:

$q_{mj} := t_{mj}$.

3. Для $m := 1, \ldots, n$:

для $j := 1, \ldots, n$:

$r_{mj} := a_{mj}$.

Замечание 1.6. Более простой и прозрачный, но неоптимизированный алгоритм QR-факторизации легко получить, используя операции над матрицами (а не над их элементами) в соответствии с приведенным ранее описанием необходимых для этого действий.

§ 1.5. QR-РАЗЛОЖЕНИЕ НА ОСНОВЕ ПРЕОБРАЗОВАНИЙ ГИВЕНСА

Введем в рассмотрение матрицы вида

$$\mathbf{T}_{ij} := \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ & \ddots & & & & & & & \\ 0 & & c & & 0 & & s & & 0 \\ & & & \ddots & & & & & \\ 0 & & 0 & & 1 & & 0 & & 0 \\ & & & & & \ddots & & & \\ 0 & & -s & & 0 & & c & & 0 \\ & & & & & & & \ddots & \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix} \begin{matrix} \\ \\ i \\ \\ \\ \\ j \\ \\ \end{matrix}, \qquad (1.25)$$

$$\qquad\qquad\qquad\quad i \qquad\qquad\quad j$$

получающиеся из единичной матрицы заменой ее четырех элементов, стоящих на пересечении i-х и j-х строк и столбцов, элементами c и s, расположенными соответствующим образом. Будем всюду далее считать эти числа связанными соотношением

$$c^2 + s^2 = 1 \,. \qquad (1.26)$$

Это позволяет интерпретировать числа c и s в матрице $\mathbf{T}_{ij}$ как косинус и синус некоторого угла α. В таком случае ее 2×2-подматрица

$$\tilde{\mathbf{T}} := \begin{pmatrix} c & s \\ -s & c \end{pmatrix} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}, \qquad (1.27)$$

как известно, задает преобразование, геометрический смысл которого состоит в повороте системы координат в декартовой плоскости на угол α в положительном направлении. Следовательно, и $n\times n$-матрицу (1.25) можно считать матрицей преобразования поворотом в определяемой i-й и j-й строками (столбцами)

гиперплоскости пространства $\mathbb{R}_n$. Отсюда — ее название ***матрица плоских вращений***, или просто ***матрица вращений.*** Иной термин, применяемый для упоминания матрицы вращений, — это ***матрица Гивенса****.

Следует отметить, что матрицами вращений называют не только матрицу вида (1.25), но и другие матрицы подобной структуры, 2×2-подматрицы которых имеют, например, симметричный по отношению к (1.27) вид

$$\tilde{\tilde{\mathbf{T}}} := \begin{pmatrix} c & -s \\ s & c \end{pmatrix};$$

главное, чтобы при этом элементы c и s удовлетворяли условию нормировки (1.26). Легко проверить, что при выполнении этого условия каждая из таких двумерных матриц обладает свойством ортогональности, что, в свою очередь, влечет ортогональность соответствующих $n\times n$-матриц $\mathbf{T}_{ij}$. Таким образом, осуществляемое с помощью матриц вращения ***преобразование вращения*** можно отнести к ***ортогональным преобразованиям.***

Обратим внимание на то, что матрицы вращения определяются двумя параметрами c и s, на которые наложено лишь одно условие. Следовательно, имеется возможность наложить еще какое-то условие, направленное на достижение определенных целей. Обычно такой целью ставят создание нуля на месте какого-нибудь заданного элемента строки или столбца с номерами i и/или j у матрицы – результата применения преобразования вращения (другие строки и столбцы при таком преобразовании не изменяются). Это осуществляется, условно говоря, выбором подходящего угла поворота α.

Зададим последовательность матриц Гивенса:

$$\mathbf{G}_1 := \mathbf{T}_{12}\ ,\quad \mathbf{G}_2 := \mathbf{T}_{13}\ ,\quad \dots\ ,\quad \mathbf{G}_{n-1} := \mathbf{T}_{1n}\ . \tag{1.28}$$

---

* Ги́венс Джеймс Уоллас (1910—1993) — американский математик. Как и Хаусхолдер, в свое время был президентом организации SIAM (общество промышленной и прикладной математики).

С их помощью построим последовательность векторов:

$$\mathbf{x}_1 = \mathbf{G}_1\mathbf{x}\ ,\quad \mathbf{x}_2 = \mathbf{G}_2\mathbf{x}_1\ ,\ \ldots,\ \mathbf{x}_{n-1} = \mathbf{G}_{n-1}\mathbf{x}_{n-2}\ , \tag{1.29}$$

начинающуюся с некоторого произвольно заданного вектора $\mathbf{x} := (x_1; x_2; \ldots; x_n)^T$. При конкретизации матриц вращений $\mathbf{G}_i$ угол поворота α_i будем подбирать так, чтобы при этом преобразовании, называемом ***преобразованием Гивенса***, обращалась в нуль $(i+1)$-я компонента преобразуемого вектора. В итоге $n-1$ таких элементарных преобразований вектор $\mathbf{x}$ с ненулевой первой компонентой должен трансформироваться в вектор, коллинеарный первому орту $\mathbf{e}_1$ пространства $\mathbb{R}_n$ (как и в случае преобразований Хаусхолдера, описанных в предыдущем параграфе).

Итак, имеем:

$$\mathbf{x}_1 := \mathbf{G}_1\mathbf{x} = \begin{pmatrix} c_1 & s_1 & 0 & \cdots & 0 \\ -s_1 & c_1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_n \end{pmatrix} = \begin{pmatrix} c_1x_1 + s_1x_2 \\ -s_1x_1 + c_1x_2 \\ x_3 \\ \cdots \\ x_n \end{pmatrix} =: \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3 \\ \cdots \\ x_n \end{pmatrix}.$$

Наложим условие

$$-s_1x_1 + c_1x_2 = 0\,, \tag{1.30}$$

при котором $x_2^{(1)} = 0$ для любых фиксированных значений x_1, x_2. Если компонента x_2 вектора $\mathbf{x}$ равна нулю, то, очевидно, сразу можно принять $c_1 := 1$, $s_1 := 0$, т.е. этот шаг будет представлять собой тождественное преобразование $\mathbf{G}_1 := \mathbf{E}$. При $x_2 \neq 0$, рассматривая равенство (1.30) совместно с условием нормировки (1.26), находим

$$c_1 = \pm\frac{x_1}{\sqrt{x_1^2 + x_2^2}}\ ,\quad s_1 = \pm\frac{x_2}{\sqrt{x_1^2 + x_2^2}}\,.$$

Это две пары (при соответствии верхних и нижних знаков) значений параметров c_1 и s_1, удовлетворяющих поставленным требо-

ваниям. Ограничимся фиксированием одной пары, отвечающей повороту на острый угол в положительном направлении:

$$c_1 := \frac{x_1}{\sqrt{x_1^2 + x_2^2}}, \qquad s_1 := \frac{x_2}{\sqrt{x_1^2 + x_2^2}}. \tag{1.31}$$

Следующий промежуточный вектор $\mathbf{x}_2$ получаем из вектора $\mathbf{x}_1$ с пересчитанной с помощью вычисленных по формулам (1.31) значений c_1 и s_1 компонентой $x_1^{(1)} := c_1 x_1 + s_1 x_2$. Имеем:

$$\mathbf{x}_2 := \mathbf{G}_2 \mathbf{x}_1 = \begin{pmatrix} c_2 & 0 & s_2 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ -s_2 & 0 & c_2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1^{(1)} \\ 0 \\ x_3 \\ x_4 \\ \cdots \\ x_n \end{pmatrix} = \begin{pmatrix} c_2 x_1^{(1)} + s_2 x_3 \\ 0 \\ -s_2 x_1^{(1)} + c_2 x_3 \\ x_4 \\ \cdots \\ x_n \end{pmatrix} =: \begin{pmatrix} x_1^{(2)} \\ 0 \\ x_3^{(2)} \\ x_4 \\ \cdots \\ x_n \end{pmatrix}.$$

Потребовав, чтобы $x_3^{(2)} = 0$, из системы

$$\begin{cases} -s_2 x_1^{(1)} + c_2 x_3 = 0, \\ s_2^2 + c_2^2 = 1 \end{cases}$$

аналогично предыдущему находим значения

$$c_2 = \frac{x_1^{(1)}}{\sqrt{(x_1^{(1)})^2 + x_3^2}}, \quad s_2 = \frac{x_3}{\sqrt{(x_1^{(1)})^2 + x_3^2}}$$

(или полагаем $c_2 := 1$, $s_2 := 0$, т.е. $\mathbf{G}_2 := \mathbf{E}$, если $x_3 = 0$).

На последнем, $(n-1)$-м шаге такого процесса будет получен вектор

$$\mathbf{x}_{n-1} := \mathbf{G}_{n-1} \mathbf{x}_{n-2} = \left(x_1^{(n-1)}; 0; \ldots; 0 \right)^T,$$

где $$c_{n-1} = \frac{x_1^{(n-2)}}{\sqrt{(x_1^{(n-2)})^2 + x_n^2}}, \qquad s_{n-1} = \frac{x_n}{\sqrt{(x_1^{(n-2)})^2 + x_n^2}},$$

$x_1^{(n-1)} = c_{n-1} x_1^{(n-2)} + s_{n-1} x_n$.

Как отмечалось ранее, произведение ортогональных матриц есть матрица ортогональная. Следовательно, преобразование вектора $\mathbf{x} := (x_1; x_2; \ldots; x_n)^T$ к виду $\mathbf{y} := (y_1; 0; \ldots; 0)^T = \mathbf{x}_{n-1}$ можно представить как $\mathbf{y} = \mathbf{G}\mathbf{x}$, где матрица $\mathbf{G}$ — произведение матриц элементарных вращений:

$$\mathbf{G} := \mathbf{G}_{n-1} \ldots \mathbf{G}_2 \mathbf{G}_1 = \mathbf{T}_{1n} \ldots \mathbf{T}_{13} \mathbf{T}_{12}$$

является ортогональной.

Пример 1.5. Методом Гивенса решим задачу, решенную в примере 1.3 методом Хаусхолдера: преобразуем вектор $\mathbf{x} := \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix}$ в вектор $\mathbf{y} := \begin{pmatrix} y_1 \\ 0 \\ 0 \end{pmatrix}$, коллинеарный орту $\mathbf{e}_1$.

По формулам (1.31) вычисляем:

$$c_1 := \frac{x_1}{\sqrt{x_1^2 + x_2^2}} = \frac{2}{\sqrt{4+4}} = \frac{\sqrt{2}}{2}, \qquad s_1 := \frac{x_2}{\sqrt{x_1^2 + x_2^2}} = \frac{-2}{\sqrt{4+4}} = -\frac{\sqrt{2}}{2};$$

тогда

$$\mathbf{G}_1 = \begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 & 0 \\ \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{y}^{(1)} := \mathbf{G}_1 \mathbf{x} = \begin{pmatrix} 2\sqrt{2} \\ 0 \\ 1 \end{pmatrix}.$$

Далее аналогично находим $c_2 = \frac{2\sqrt{2}}{3}$, $s_2 = \frac{1}{3}$, и, следовательно,

$$\mathbf{G}_2 = \begin{pmatrix} 2\sqrt{2}/3 & 0 & 1/3 \\ 0 & 1 & 0 \\ -1/3 & 0 & 2\sqrt{2}/3 \end{pmatrix}, \qquad \mathbf{y} := \mathbf{y}^{(2)} := \mathbf{G}_2 \mathbf{y}^{(1)} = \mathbf{G}_2 \mathbf{G}_1 \mathbf{x} = \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix}.$$

Теперь рассмотрим применение описанной стратегии ортогональных преобразований вращения к произвольной вещественной матрице $\mathbf{A} := (a_{ij})_{i,\, j=1}^{n}$.

Придавая первому столбцу матрицы $\mathbf{A}$ статус вектора $\mathbf{x}$ в предыдущих рассуждениях и выкладках, определяющих формулы

для подсчета значений параметров c и s матриц вращения, выполняем $n-1$ последовательных пересчетов элементов данной матрицы.

Именно, на первом шаге первого этапа получаем матрицу

$$\mathbf{A}^{(1)} := \mathbf{T}_{12}\mathbf{A} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix},$$

в которой по сравнению с $\mathbf{A}$ изменения претерпели элементы только первых двух строк; очевидно, они должны быть пересчитаны по формулам

$$a_{1j}^{(1)} = c_1 a_{1j} + s_1 a_{2j} \quad (j = 1,\dots,n), \tag{1.32}$$

$$a_{2j}^{(1)} = -s_1 a_{1j} + c_1 a_{2j} \quad (j = 2,\dots,n), \tag{1.33}$$

где $c_1 = \dfrac{a_{11}}{\sqrt{a_{11}^2 + a_{21}^2}}$, $s_1 = \dfrac{a_{21}}{\sqrt{a_{11}^2 + a_{21}^2}}$

$(c_1 := 1,\ s_1 := 0,$ если $a_{21} = 0)$.

На втором шаге этого этапа имеем

$$\mathbf{A}^{(2)} := \mathbf{T}_{13}\mathbf{A}^{(1)} = \mathbf{T}_{13}\mathbf{T}_{12}\mathbf{A} = \begin{pmatrix} a_{11}^{(2)} & a_{12}^{(2)} & a_{13}^{(2)} & \dots & a_{1n}^{(2)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & \dots & a_{3n}^{(1)} \\ a_{41} & a_{42} & a_{43} & \dots & a_{4n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix},$$

где

$$a_{1j}^{(2)} = c_2 a_{1j}^{(1)} + s_2 a_{3j} \quad (j = 1,...,n),$$
$$a_{3j}^{(1)} = -s_2 a_{1j}^{(1)} + c_2 a_{3j} \quad (j = 2,...,n),$$

$$c_2 = \frac{a_{11}^{(1)}}{\sqrt{(a_{11}^{(1)})^2 + a_{31}{}^2}}, \quad s_2 = \frac{a_{31}}{\sqrt{(a_{11}^{(1)})^2 + a_{31}{}^2}},$$

и т.д. Первый этап завершается на $(n-1)$-м шаге построением матрицы

$$\mathbf{A}^{(n-1)} := \mathbf{T}_{1n}\mathbf{A}^{(n-2)} = \mathbf{T}_{1n}...\mathbf{T}_{13}\mathbf{T}_{12}\mathbf{A} = \begin{pmatrix} a_{11}^{(n-1)} & a_{12}^{(n-1)} & \dots & a_{1n}^{(n-1)} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \dots & \dots & \dots & \dots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{pmatrix}.$$

Второй этап состоит в проведении аналогичных преобразований, производимых над финальной матрицей первого этапа $\mathbf{A}_1 := \mathbf{A}^{(n-1)}$. Цель преобразований этого этапа — создание нулей под вторым элементом главной диагонали. Роль элемента, стоящего в позиции (1, 1) и участвовавшего в процессе получения всех нулей в первом столбце, на втором этапе отдается элементу с индексами 2, 2. Следовательно, теперь нужно выстроить последовательность матриц элементарных вращений вида (1.25) так:

$$\mathbf{T}_{23}, \ \mathbf{T}_{24}, ..., \ \mathbf{T}_{2n}, \tag{1.34}$$

а угол поворота (а точнее, параметры c и s) подбирать, опираясь на элементы в позиции (2, 2). При этом первая строка преобразуемой матрицы больше изменяться не будет, поскольку строки и столбцы всех матриц (1.34), имеющие номер 1, являются единичными векторами. Таким образом, результатом второго этапа преобразований вращения, направленных на триангуляризацию данной квадратной матрицы $\mathbf{A}$, будет матрица вида

$$\mathbf{A}_2 := \begin{pmatrix} a_{11}^{(n-1)} & a_{12}^{(n-1)} & a_{13}^{(n-1)} & \dots & a_{1n}^{(n-1)} \\ 0 & a_{22}^{(n-1)} & a_{23}^{(n-1)} & \dots & a_{2n}^{(n-1)} \\ 0 & 0 & a_{33}^{(2)} & \dots & a_{3n}^{(2)} \\ 0 & 0 & a_{43}^{(2)} & \dots & a_{4n}^{(2)} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & a_{n3}^{(2)} & \dots & a_{nn}^{(2)} \end{pmatrix},$$

которую можно рассматривать как результат преобразований

$$\mathbf{A}_2 := \mathbf{A}_1^{(n-2)} = \mathbf{T}_{2n}\dots\mathbf{T}_{24}\mathbf{T}_{23}\mathbf{A}_1 =$$

$$= (\mathbf{T}_{2n}\dots\mathbf{T}_{24}\mathbf{T}_{23})(\mathbf{T}_{1n}\dots\mathbf{T}_{13}\mathbf{T}_{12})\mathbf{A}\,.$$

На следующем этапе с помощью матриц вращения $\mathbf{T}_{34}$, $\mathbf{T}_{35}, \dots, \mathbf{T}_{3n}$ аналогично создаются нули под третьим диагональным элементом, и, наконец, на последнем, $(n-1)$-м этапе матрица $\mathbf{A}$ приводится к треугольному виду. Обозначим эту матрицу через $\mathbf{R}$. Ее выражение через элементы, верхние индексы которых отражают число шагов преобразований (пересчетов), таково:

$$\mathbf{R} := \mathbf{A}_{n-1} := \begin{pmatrix} a_{11}^{(n-1)} & a_{12}^{(n-1)} & \dots & a_{1n}^{(n-1)} \\ 0 & a_{22}^{(n-1)} & \dots & a_{2n}^{(n-1)} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn}^{(n-1)} \end{pmatrix}.$$

Представление матрицы $\mathbf{R}$ через матрицы $\mathbf{T}_{ij}$ показывает порядок, в котором производят пересчеты элементов, если обратить внимание на связь индексов матриц вращений и нижних индексов элементов преобразуемых матриц:

$$\mathbf{R} = (\mathbf{T}_{n-1,n})(\mathbf{T}_{n-2,n}\mathbf{T}_{n-2,n-1})\dots(\mathbf{T}_{2n}\dots\mathbf{T}_{24}\mathbf{T}_{23})(\mathbf{T}_{1n}\dots\mathbf{T}_{13}\mathbf{T}_{12})\mathbf{A} \qquad (1.35)$$

(в скобки заключены сомножители, соответствующие одному этапу описанных преобразований).

Из полученной триангуляризации (1.35) легко понять, что собой представляет **QR**-разложение матрицы **A**: очевидно, в силу свойств ортогональных матриц $\mathbf{T}_{ij}$ можно записать представление

$$\mathbf{A} = \mathbf{QR} ,$$

где $\mathbf{Q} := (\mathbf{T}_{1n}\ldots\mathbf{T}_{13}\mathbf{T}_{12})^T(\mathbf{T}_{2n}\ldots\mathbf{T}_{24}\mathbf{T}_{23})^T\ldots(\mathbf{T}_{n-2,n}\mathbf{T}_{n-2,n-1})^T(\mathbf{T}_{n-1,n})^T$.

Замечание 1.7. Обратим внимание на факт, эксплуатировавшийся при построении матриц Хаусхолдера в § 1.4 (см. (1.14)): ортогональное преобразование не изменяет длины преобразуемого вектора. В данном случае легко непосредственно проверить, что, например, при первом элементарном преобразовании матрицы **A** матрицей плоских вращений $\mathbf{T}_{12}$ евклидова норма каждого столбца останется неизменной. Действительно, при любом $j \in \{2,\ldots,n\}$ в соответствии с формулами (1.32), (1.33) пересчета элементов первых двух строк и с учетом условия (1.26) имеем:

$$\left(a_{1j}^{(1)}\right)^2 + \left(a_{2j}^{(1)}\right)^2 = c_1^2 a_{1j}^2 + 2c_1 s_1 a_{1j} a_{2j} + s_1^2 a_{2j}^2 + s_1^2 a_{1j}^2 - 2c_1 s_1 a_{1j} a_{2j} + c_1^2 a_{2j}^2 =$$

$$= (c_1^2 + s_1^2) a_{1j}^2 + (c_1^2 + s_1^2) a_{2j}^2 = a_{1j}^2 + a_{2j}^2 .$$

При $j = 1$ такое равенство очевидно. Поскольку другие строки матрицы при этом не затрагиваются, справедливо утверждаемое, и такое имеет место при каждом элементарном вращении. Следовательно, в описанном процессе приведения произвольной матрицы **A** к треугольному виду **R** (отображенному формулой (1.35)) не может происходить совокупного роста элементов.

УПРАЖНЕНИЯ

1.1. Матрицу $\mathbf{A} := \begin{pmatrix} 1 & -3 \\ 2 & 5 \end{pmatrix}$ разложите в произведение треугольных матриц.

1.2. Выполните **LU**-разложение матрицы $\mathbf{A} := \begin{pmatrix} 4 & -3 & 2 \\ 8 & -8 & 7 \\ 12 & -5 & 5 \end{pmatrix}$.

1.3. Подсчитайте, сколько может быть различных $\mathbf{U}^T\mathbf{U}$-разложений симметричной положительно определенной матрицы размера $n \times n$.

1.4. Убедитесь, что матрица **A**, фигурирующая в примере 1.2 (§ 1.3), не имеет вещественного $\mathbf{U}^T\mathbf{U}$-разложения.

1.5. Ограничиваясь трехмерным случаем, покажите, что вещественная симметричная матрица $\mathbf{A}$ может быть представлена в виде произведения $\mathbf{U}^T\mathbf{D}\mathbf{U}$ трех вещественных матриц, где $\mathbf{U}$ — некоторая верхняя треугольная матрица, а $\mathbf{D}$ — диагональная матрица с элементами диагонали $+1$ или -1.

1.6. Пользуясь рекуррентными формулами (1.10), выполните $\mathbf{U}^T\mathbf{D}\mathbf{U}$-разложение матрицы $\mathbf{A} := \begin{pmatrix} 1 & 2 & 0 & 0 \\ 2 & 1 & 2 & 0 \\ 0 & 2 & 1 & 2 \\ 0 & 0 & 2 & 1 \end{pmatrix}$.

1.7. Рассмотрите $\mathbf{L}\mathbf{L}^T$- и $\mathbf{L}\mathbf{D}\mathbf{L}^T$-разложения симметричных матриц. Чем они отличаются от описанных в § 1.3 $\mathbf{U}^T\mathbf{U}$- и $\mathbf{U}^T\mathbf{D}\mathbf{U}$-разложений соответственно?

1.8. Докажите, что преобразование Хаусхолдера сохраняет неизменной евклидову длину вектора, непосредственно пользуясь представлением преобразованного вектора по формуле (1.15).

1.9. Сравните преобразованные векторы и итоговые ортогональные матрицы Хаусхолдера и Гивенса в примерах 1.3 (см. § 1.4) и 1.5 (§ 1.5) в свете замечания 1.4.

1.10. Какие изменения следует провести в процессе $\mathbf{QR}$-факторизации квадратной матрицы на основе преобразования Хаусхолдера, чтобы осуществить аналогичное $\mathbf{LQ}$-разложение?

Выполните $\mathbf{LQ}$-разложение матрицы $\mathbf{A} := \begin{pmatrix} 1 & 2 & 2 \\ 3 & 3 & 4 \\ 4 & 4 & 5 \end{pmatrix}$.

1.11. Для матрицы $\mathbf{A} := \begin{pmatrix} 3 & -1 & 4 & -2 \\ 2 & 1 & -5 & 0 \\ 5 & -2 & 3 & 1 \\ 1 & 3 & 2 & 4 \end{pmatrix}$ постройте матрицу $\mathbf{T}$ плоских вращений такую, чтобы матрица $\mathbf{TA}$ имела нуль в позиции $(1, 3)$.

Глава 2

ПРЯМЫЕ МЕТОДЫ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

§ 2.1. МЕТОД ГАУССА (СХЕМА ЕДИНСТВЕННОГО ДЕЛЕНИЯ)

Считается [9], что 75% всех расчетных математических задач приходится на решение систем линейных алгебраических уравнений (СЛАУ). Это неудивительно, так как математические модели тех или иных явлений или процессов либо сразу строятся как линейные алгебраические, либо сводятся к таковым посредством дискретизации и/или линеаризации. Поэтому трудно переоценить роль, которую играет выбор эффективного (в том или ином смысле) способа решения СЛАУ. Современная вычислительная математика располагает большим арсеналом методов, а математическое обеспечение ЭВМ — многими пакетами прикладных программ, позволяющих решать различные возникающие на практике линейные системы. Чтобы ориентироваться среди методов и программ и в нужный момент сделать оптимальный выбор, нужно разбираться в основах построений методов и алгоритмов, учитывающих специфику постановок задач, знать их сильные и слабые стороны и границы применимости. Имеющаяся довольно обширная статистика использования пакетов прикладных компьютерных программ показывает, что на практике чаще всего применяют самые примитивные широко известные численные методы решения тех или иных математически поставленных задач, хотя к ним можно было бы привлечь более эффективные численные методы, содержащиеся в тех же пакетах [58]. Этот факт объясняется только недостаточной подготовкой пользователей прикладных программ в области вычислительной математики.

Все методы решения линейных алгебраических задач (наряду с задачей решения СЛАУ, это и вычисление определителей, и обращение матриц, и задачи на собственные значения) можно разбить на два класса: прямые и итерационные. Как явствует из

n-й степени, а также поиском линейно независимых решений вырожденных СЛАУ. В связи с этим описанный непосредственный подход к решению алгебраической проблемы собственных значений обычно применяют лишь при очень малых размерах матриц $\mathbf{A}$ $(n=2,3)$; уже при $n\geq 4$ на первый план выходят специальные численные методы решения подобных задач. Далее рассмотрены некоторые из этих методов в таком ключе, чтобы можно было понять идеи, лежащие в их основе, и в то же время получить возможность решать поставленные задачи до конца для некоторых классов матриц (более полное и глубокое изложение этой темы см. в монографиях [23, 26, 32, 54, 57, 67, 69] и в учебных пособиях [1, 3, 4, 7, 17, 25, 34, 42]).

Следует заметить, что в недалеком прошлом численные методы решения задач на собственные значения опирались, как правило, на классический подход, т.е. на развертывание «вековых определителей», в частности, в простейшем случае с помощью приведения матрицы $\mathbf{A}$ подходящим преобразованием к так называемой ***сопровождающей матрице***

$$\mathbf{C}:=\begin{pmatrix} c_1 & c_2 & c_3 & \dots & c_{n-1} & c_n \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix},$$

где в первой строке стоят коэффициенты уравнения (4.4), записанного в виде

$$(-1)^n\left(\lambda^n-c_1\lambda^{n-1}-c_2\lambda^{n-2}-\dots-c_{n-1}\lambda-c_n\right)=0.$$

Современные методы решения полной проблемы ориентированы на алгоритмическое построение из матрицы $\mathbf{A}$ такой матрицы, определенные элементы которой являлись бы приближенными значениями собственных чисел $\mathbf{A}$, причем параллельно, по возможности, формировались бы и ее собственные векторы.

Прежде чем приступить к изучению методов нахождения собственных чисел и векторов, вспомним их некоторые простые свойства, требующиеся в дальнейшем.

координатами, а $\mathbf{A} := (a_{ij})_{i,\,j=1}^{n}$ — вещественная $n \times n$-матрица коэффициентов данной системы. Эффективность способов решения системы (2.1) во многом зависит от структуры и свойств матрицы $\mathbf{A}$: размера, обусловленности, симметричности, заполненности (т.е. соотношения между числом ненулевых и нулевых элементов), специфики расположения ненулевых элементов в матрице и др.

Так, размерность системы (т.е. число n) является главным фактором, заставляющим вычислителей отвернуться от весьма привлекательных в теоретическом плане и приемлемых на практике при небольших n (конкретно, равных 2, 3) ***формул Крамера****:

$$x_i = \frac{\det \mathbf{A}_i}{\det \mathbf{A}} \quad (i = 1, 2, \ldots, n),$$

позволяющих находить неизвестные компоненты вектора $\mathbf{x}$ в виде дробей, знаменателем которых является определитель матрицы системы, а числителем — определители матриц $\mathbf{A}_i$, полученных из $\mathbf{A}$ заменой столбца коэффициентов при вычисляемом неизвестном столбцом свободных членов. Если при реализации этих формул определители вычисляются понижением порядка на основе разложения по элементам какой-нибудь строки или столбца матрицы, то на вычисление определителя n-го порядка будет затрачиваться $n!$ операций умножения. Факториальный рост (или какой-то другой очень быстрый рост) числа арифметических операций с увеличением размерности задачи называют «проклятьем размерности». Что это такое, можно представить, зафиксировав, например, $n = 100$. Оценив величину $100! \approx 10^{158}$ и прикинув потенциальные возможности развития вычислительной техники, приходим к выводу о том, что в обозримом будущем системы сотого порядка в принципе не могут быть решены по формулам Крамера [3, 41]. Заметим при этом, что, во-первых, метод Крамера

---

* Кра́мер Габриель (1704–1752) — швейцарский математик. Заложил основы теории определителей.

будет неустойчив, т.е. погрешности округлений будут катастрофически нарастать, во-вторых, размерность $n=100$ для современных задач не так и велика: довольно часто решают системы с сотнями и тысячами неизвестных.

Если осуществлять вычисление обратной матрицы с помощью построения союзной матрицы, т. е. через алгебраические дополнения, то нахождение решения векторно-матричного уравнения (2.1а) по формуле

$$\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}$$

фактически равнозначно применению формул Крамера и также практически непригодно по упомянутым выше причинам для вычислительных целей.

Наиболее известным и популярным способом решения линейных систем вида (2.1) является ***метод Гаусса****. Суть его проста — это последовательное исключение неизвестных. В отличие от курсов аналитической линейной алгебры, нас будут интересовать вычислительные аспекты метода Гаусса, а именно технология получения вектора-решения $\mathbf{x}$ из исходных матрицы $\mathbf{A}$ и вектора $\mathbf{b}$, причем, по возможности, минимизирующая влияние неизбежных ошибок округления. С этой целью, работая с уравнениями системы (2.1), выведем сначала совокупность формул, позволяющих в итоге получить искомые значения неизвестных, а затем на основе этих формул запишем алгоритм решения поставленной задачи.

Будем поэтапно приводить систему (2.1) к треугольному виду, исключая последовательно сначала x_1 из второго, третьего, ... , n-го уравнений, затем x_2 из третьего, четвертого, ... , n-го уравнений преобразованной системы и т.д.

На первом этапе заменим второе, третье, ... , n-е уравнения на уравнения, получающиеся сложением этих уравнений с первым,

---

* Га́усс Карл Фридрих (1777–1855) — знаменитый немецкий математик, оказавший глубокое влияние на развитие алгебры, теории чисел, дифференциальной геометрии, геодезии, астрономии и др.

умножаемым поочередно на величины $-\frac{a_{21}}{a_{11}}$, $-\frac{a_{31}}{a_{11}}$, ..., $-\frac{a_{n1}}{a_{11}}$ соответственно. Результатом этого этапа преобразований будет эквивалентная (2.1) система

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1, \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + \ldots + a_{3n}^{(1)}x_n = b_3^{(1)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n = b_n^{(1)}, \end{cases} \tag{2.2}$$

коэффициенты которой (с верхним индексом (1)) подсчитывают по формулам:

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}}a_{1j}, \quad b_i^{(1)} = b_i - \frac{a_{i1}}{a_{11}}b_1 \quad (i, j = 2, 3, \ldots, n).$$

При этом можно считать, что $a_{11} \neq 0$, так как по предположению система (2.1) однозначно разрешима, значит, все коэффициенты при x_1 не могут одновременно равняться нулю, и на первое место всегда можно поставить уравнение с отличным от нуля первым коэффициентом.

На втором этапе проделываем такие же операции, как и на первом, с подсистемой системы (2.2), получающейся исключением из нее первого уравнения. Эквивалентный (2.2) (а значит, и (2.1)) результат второго этапа будет иметь вид

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1, \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ a_{33}^{(2)}x_3 + \ldots + a_{3n}^{(2)}x_n = b_3^{(2)}, \\ \ldots\ldots\ldots\ldots\ldots \\ a_{n3}^{(2)}x_3 + \ldots + a_{nn}^{(2)}x_n = b_n^{(2)}, \end{cases}$$

где $a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{i2}^{(1)}}{a_{22}^{(1)}} a_{2j}^{(1)}$, $b_i^{(2)} = b_i^{(1)} - \frac{a_{i2}^{(1)}}{a_{22}^{(1)}} b_2^{(1)}$; $i, j = 3, \ldots, n$.

Продолжая этот процесс, на $(n-1)$-м этапе так называемого ***прямого хода*** метода Гаусса данную систему (2.1) приведем к треугольному виду:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{22}^{(1)}x_2 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ \ldots\ldots\ldots \quad \ldots \quad \ldots \\ a_{nn}^{(n-1)}x_n = b_n^{(n-1)}. \end{cases} \tag{2.3}$$

На основе предыдущих рассуждений и формул легко убедиться, что коэффициенты этой системы могут быть получены из коэффициентов исходной системы последовательным пересчетом по формулам

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)}, \qquad b_i^{(k)} = b_i^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} b_k^{(k-1)} , \tag{2.4}$$

где верхний индекс k (номер этапа) должен изменяться от 1 до $n-1$, нижние индексы i и j (в любой очередности) — от $k+1$ до n; по определению полагаем $a_{ij}^{(0)} := a_{ij}$, $b_i^{(0)} := b_i$.

Треугольная, а точнее трапециевидная, структура системы (2.3) позволяет последовательно одно за другим вычислять значения неизвестных, начиная с последнего:

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}} ;$$

$$\ldots\ldots\ldots\ldots\ldots$$

$$x_2 = \frac{b_2^{(1)} - a_{23}^{(1)}x_3 - \ldots - a_{2n}^{(1)}x_n}{a_{22}^{(1)}} ;$$

$$x_1 = \frac{b_1 - a_{12}x_2 - \ldots - a_{1n}x_n}{a_{11}} .$$

Этот процесс последовательного вычисления значений неизвестных называют ***обратным ходом*** метода Гаусса. Очевидно, он определяется одной формулой

$$x_k = \frac{1}{a_{kk}^{(k-1)}}\left(b_k^{(k-1)} - \sum_{j=k+1}^{n} a_{kj}^{(k-1)} x_j\right), \qquad (2.5)$$

где k полагают равным поочередно $n, n-1, \ldots, 2, 1$, и сумма по определению считается равной нулю, если нижний предел суммирования имеет значение больше верхнего.

Итак, решение СЛАУ вида (2.1) методом Гаусса сводится к последовательной реализации вычислений по формулам (2.4) и (2.5). Учитывая цикличность выполняемых при этом операций, а также нецелесообразность хранения промежуточных результатов (пересчитываемых коэффициентов промежуточного этапа), запишем следующий простой **алгоритм** решения линейных систем (2.1) методом Гаусса [51]:

1. Для $k := 1, 2, \ldots, n-1$:
2. для $i := k+1, \ldots, n$:
3. $t_{ik} := a_{ik} / a_{kk}$,
4. $b_i := b_i - t_{ik} b_k$;
5. для $j := k+1, \ldots, n$:
6. $a_{ij} := a_{ij} - t_{ik} a_{kj}$.
7. $x_n := b_n / a_{nn}$;
8. для $k := n-1, \ldots, 2, 1$:
9. $x_k := \frac{1}{a_{kk}}(b_k - \sum_{j=k+1}^{n} a_{kj} x_j)$.

Подав на его вход квадратную матрицу $\left(a_{ij}\right)_{i,j=1}^{n}$ коэффициентов при неизвестных системы (2.1) и вектор $\left(b_i\right)_{i=1}^{n}$ свободных членов и выполнив три вложенных цикла вычислений прямого хода (строки 1–6) и один цикл вычислений обратного хода (стро-

ки 7–9), на выходе алгоритма получим вектор-решение $(x_k)_{k=1}^n$ (его компоненты находятся в обратном порядке). Всё это будет так, если, разумеется, ни один из знаменателей не обращается в нуль и все вычисления проводятся точно.

Описанную реализацию метода Гаусса называют ***схемой единственного деления***.

Так как реальные машинные вычисления производятся не с точными, а с усеченными числами (см. приложение), т.е. неизбежны ошибки округления, то, анализируя, например, формулы (2.4), можно сделать вывод о том, что выполнение алгоритма может прекратиться или привести к неверным результатам, если знаменатели дробей на каком-то этапе окажутся равными нулю или очень маленькими числами. Чтобы уменьшить влияние ошибок округлений и исключить деление на нуль, на каждом этапе прямого хода уравнения системы (точнее, обрабатываемой подсистемы) обычно переставляют так, чтобы деление производилось на наибольший по модулю в данном столбце (обрабатываемом подстолбце) элемент. Числа, на которые производится деление в методе Гаусса, называются ***ведущими***, или ***главными***, ***элементами***. Отсюда название рассматриваемой модификации метода, исключающей деление на нуль и уменьшающей вычислительные погрешности, — ***метод Гаусса с постолбцовым выбором главного элемента*** (или, иначе, с частичным упорядочиванием по столбцам).

Частичное упорядочивание по столбцам требует внесения в алгоритм следующих изменений: между строками 1 и 2 нужно сделать следующую вставку:

⊗ «Найти $m \geq k$, такое, что $|a_{m\,k}| = \max\limits_{i \geq k} \{|a_{i\,k}|\}$;
если $a_{m\,k} = 0$, остановить работу алгоритма («однозначного решения нет»),
иначе поменять местами b_k и b_m, $a_{k\,j}$ и $a_{m\,j}$ при всех $j = k, \ldots, n$».

Более разумным, наверное, является сравнение $|a_{m\,k}|$ не с нулем, а с некоторым малым допуском $\varepsilon > 0$, задаваемым вычис-

лителем в зависимости от различных априорных соображений. Счет останавливается или берется под особый контроль, если окажется $|a_{m\,k}| < \varepsilon$. Заметим, что соответствующая частичному упорядочиванию вставка $\otimes$ в алгоритм позволяет фактически в процессе его выполнения проводить алгоритмическое исследование системы (2.1) на однозначную разрешимость.

Чтобы частичное упорядочивание было более эффективным, перед этим целесообразно произвести ***масштабирование*** (уравновешивание) системы: например, разделить все числа каждой строки на наибольшее число этой строки [51].

Устойчивость алгоритма к погрешностям исходных данных и результатов промежуточных вычислений можно еще усилить, если выполнять деление на каждом этапе на элемент, наибольший по модулю во всей матрице преобразуемой на данном этапе подсистемы. Такая модификация метода Гаусса, называемая ***методом главных элементов***, применяется довольно редко, поскольку сильно усложняет алгоритм, не давая существенного выигрыша по сравнению с применением постолбцового упорядочивания. Усложнение связано как с необходимостью осуществления двумерного поиска главных элементов, так и с необходимостью запоминать номера столбцов, откуда берутся эти элементы (перестановка столбцов означает как бы переобозначение неизвестных, в связи с чем требуется обратная замена).

Как было сказано выше, решения линейных алгебраических систем можно получать с помощью определителей или обратных матриц. Однако нетрудно увидеть, что более эффективно поступать наоборот: вычислять определители и обращать матрицы в рамках решения линейных систем методом Гаусса*.

Действительно, выполняемые в методе Гаусса преобразования прямого хода, приведшие матрицу A системы к треугольному виду (см. (2.3)) , таковы, что они не изменяют определитель матрицы A. Учитывая, что определитель треугольной матрицы равен

---

* Хотя в [57], например, прямо говорится о бесполезности определителей в вычислительной линейной алгебре.

произведению диагональных элементов, имеем:

$$\det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1\,n} \\ 0 & a_{22}^{(1)} & \dots & a_{2\,n}^{(1)} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{n\,n}^{(n-1)} \end{vmatrix} = a_{11} \cdot a_{22}^{(1)} \cdot \dots \cdot a_{n\,n}^{(n-1)} \ .$$

Таким образом, *определитель матрицы равен произведению всех ведущих элементов при ее преобразовании методом Гаусса.*

При желании получить $\det \mathbf{A}$ дополнительно к решению СЛАУ $\mathbf{Ax} = \mathbf{b}$ алгоритм предыдущего пункта должен быть пополнен всего лишь одной следующей строкой:

$$10. \qquad \det \mathbf{A} = \prod_{k=1}^{n} a_{kk} \ .$$

Если метод Гаусса используется только для вычисления определителя, из алгоритма его реализации следует изъять строки 4 и 7–9.

Так как перестановка строк матрицы меняет знак определителя, то при постолбцовом выборе главного элемента, т.е. при включении в алгоритм вставки $\otimes$, нужно в результате учесть число p произведенных перестановок, точнее четность этого числа. Это означает, что при вычислении $\det \mathbf{A}$ с использованием алгоритма Гаусса с частичным упорядочиванием вместо строки 10 должна быть включена строка

$$10'. \qquad \det \mathbf{A} = (-1)^p \prod_{k=1}^{n} a_{kk} \ .$$

Пример 2.1. Пользуясь схемой единственного деления, найти детерминант матрицы $\mathbf{A} := \begin{pmatrix} 2 & -1 & 1 \\ 4 & 3 & 1 \\ 6 & -13 & 6 \end{pmatrix}$.

Для данного случая алгоритм этого метода можно конкретизировать так:

1. Для $k := 1, 2$:
2. для $i := k+1, 3$:
3. $t_{ik} := \dfrac{a_{ik}}{a_{kk}}$;
4. для $j := k+1, 3$:

5. $a_{ij} := a_{ij} - t_{ik}a_{kj}$.

6. $\det \mathbf{A} := a_{11} \cdot a_{22} \cdot a_{33}$.

Воспользуемся им. Фиксируем $k := 1$ и соответственно $i := 2$. Тогда $t_{21} := \dfrac{a_{21}}{a_{11}} = \dfrac{4}{2} = 2$, и при $j = 2, 3$ имеем:

$$a_{22} := a_{22} - t_{21}a_{12} = 3 - 2(-1) = 5,$$
$$a_{23} := a_{23} - t_{21}a_{13} = 1 - 2 \cdot 1 = -1.$$

Переключая i на значение 3, далее подсчитываем значение $t_{31} := \dfrac{a_{31}}{a_{11}} = \dfrac{6}{2} = 3$ и при $j = 2, 3$ получаем

$$a_{32} := a_{32} - t_{31}a_{12} = -13 - 3(-1) = -10,$$
$$a_{33} := a_{33} - t_{31}a_{13} = 6 - 3 \cdot 1 = 3.$$

Теперь полагаем $k := 2$ (т.е. переходим ко второму этапу преобразований). При этом значении k индексы i и j могут принять только одно значение: 3. Следовательно, достаточно подсчитать значения

$$t_{32} := \frac{a_{32}}{a_{22}} = \frac{-10}{5} = -2 \quad \text{и} \quad a_{33} := a_{33} - t_{32}a_{23} = 3 - (-2)(-1) = 1$$

(где вместо элементов a_{22}, a_{23}, a_{32}, a_{33} исходной матрицы подставляем новые, подсчитанные на первом этапе, значения), чтобы получить искомое значение определителя

$$\det \mathbf{A} = 2 \cdot 5 \cdot 1 = 10.$$

Для получения матрицы $\mathbf{A}^{-1}$, обратной по отношению к матрице $\mathbf{A} := (a_{ij})_{i,\,j=1}^{n}$, будем исходить из того, что она является решением матричного уравнения

$$\mathbf{AX} = \mathbf{E}, \tag{2.6}$$

где $\mathbf{E}$ — единичная матрица.

Представляя искомую матрицу $\mathbf{X} := (x_{ij})_{i,\,j=1}^{n}$ как набор (вектор-строку) векторов-столбцов:

$$\mathbf{x}_1 := \begin{pmatrix} x_{11} \\ x_{21} \\ \cdots \\ x_{n1} \end{pmatrix}, \quad \mathbf{x}_2 := \begin{pmatrix} x_{12} \\ x_{22} \\ \cdots \\ x_{n2} \end{pmatrix}, \quad \ldots, \quad \mathbf{x}_n := \begin{pmatrix} x_{1n} \\ x_{2n} \\ \cdots \\ x_{nn} \end{pmatrix},$$

а единичную матрицу $\mathbf{E}$ как набор единичных векторов (ортов):

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \dots \\ 0 \end{pmatrix}, \quad \dots \ , \quad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 1 \end{pmatrix},$$

матричное уравнение (2.6) в соответствии с правилами умножения матриц подменим эквивалентной системой не связанных между собой векторно-матричных уравнений:

$$\mathbf{A}\mathbf{x}_1 = \mathbf{e}_1; \qquad \mathbf{A}\mathbf{x}_2 = \mathbf{e}_2; \qquad \dots \ ; \qquad \mathbf{A}\mathbf{x}_n = \mathbf{e}_n. \tag{2.7}$$

Каждое из уравнений (2.7) имеет вид (2.1а) и может быть решено методом Гаусса. При этом специфичным является то обстоятельство, что все СЛАУ (2.7) имеют одну и ту же матрицу коэффициентов, а это означает, что наиболее трудоемкая часть метода Гаусса — приведение матрицы системы к треугольному виду — общая для всех систем (2.7). Так что, если требуется приспособить рассмотренный выше алгоритм решения СЛАУ методом Гаусса к обращению матриц, целесообразно не просто применить его последовательно n раз к системам (2.7), а слегка подкорректировать: «размножить» строки 4 и 9 так, чтобы в роли вектора $\mathbf{b}$ оказались все единичные векторы $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$. Тогда в результате завершения работы алгоритма будут получаться столбец за столбцом (столбцы «перевернуты»!) элементы обратной матрицы $\mathbf{X} = \mathbf{A}^{-1}$. При этом введение в алгоритм частичного упорядочивания, т.е. постолбцовый выбор главного элемента, не требует запоминаний и обратных замен.

§ 2.2. РЕШЕНИЕ СЛАУ И ОБРАЩЕНИЕ МАТРИЦ НА ОСНОВЕ LU-РАЗЛОЖЕНИЯ

Если матрица $\mathbf{A}$ исходной системы (2.1) разложена в произведение треугольных матриц $\mathbf{L}$ и $\mathbf{U}$, то, значит, вместо уравнения (2.1а) можно записать эквивалентное ему уравнение

$$\mathbf{L}\mathbf{U}\mathbf{x} = \mathbf{b}.$$

Введя вектор вспомогательных переменных $\mathbf{y} := (y_1; y_2; \dots; y_n)^T$, последнее можно переписать в виде системы

$$\begin{cases} \mathbf{Ly} = \mathbf{b}, \\ \mathbf{Ux} = \mathbf{y}. \end{cases}$$

Таким образом, решение данной системы с *квадратной* матрицей коэффициентов свелось к последовательному решению двух систем с *треугольными* матрицами коэффициентов.

Получим сначала формулы для вычисления элементов y_i вспомогательного вектора $\mathbf{y}$. Для этого запишем уравнение $\mathbf{Ly} = \mathbf{b}$ в развернутом виде:

$$\begin{cases} y_1 = b_1, \\ l_{21}y_1 + y_2 = b_2, \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ l_{n\,1}y_1 + l_{n\,2}y_2 + \dots + l_{n,n-1}y_{n-1} + y_n = b_n. \end{cases}$$

Отсюда видно, что все значения y_i могут быть последовательно найдены при $i = 1, 2, \dots, n$ по формуле

$$y_i = b_i - \sum_{k=1}^{i-1} l_{i\,k}y_k. \tag{2.8}$$

Развернем теперь векторно-матричное уравнение $\mathbf{Ux} = \mathbf{y}$:

$$\begin{cases} u_{11}x_1 + u_{12}x_2 + \dots + u_{1\,n}x_n = y_1, \\ \qquad\quad u_{22}x_2 + \dots + u_{2\,n}x_n = y_2, \\ \qquad\quad \dots\dots\dots \quad \dots\dots\dots \\ \qquad\qquad\qquad\quad u_{n\,n}x_n = y_n. \end{cases} \tag{2.9}$$

Из этой системы значения неизвестных x_i находят так же последовательно, как и y_i, но в обратном порядке, что можно отразить одной формулой:

$$x_i = \frac{1}{u_{i\,i}}\left(y_i - \sum_{k=i+1}^{n} u_{i\,k}x_k\right), \quad \text{где } i = n, n-1, \dots, 2, 1. \tag{2.10}$$

Итак, решение СЛАУ посредством LU-факторизации сводится к организации вычислений по четырем формулам: совокупности

формул (1.2), (1.3) для получения матрицы $\mathbf{L}+\mathbf{U}-\mathbf{E}$ (1.1) структурно ненулевых и неединичных элементов матриц $\mathbf{L}$ и $\mathbf{U}$, формулы (2.8) для получения вектора свободных членов треугольной системы (2.9) и формулы (2.10), генерирующей решение исходной системы (2.1). Такой процесс реализации метода Гаусса называют ***компактной схемой Гаусса*** [18, 69], или ***схемой Холецкого**** [25]. Нетрудно усмотреть связь между этой схемой и рассмотренной ранее схемой единственного деления, если воспользоваться показанным в § 1.2 представлением $\mathbf{LU}$-разложения через последовательность простейших матричных преобразований.

Вычисление определителя $\mathbf{LU}$-факторизованной матрицы $\mathbf{A}$ опирается на свойство определителя произведения матриц и сводится к перемножению n чисел:

$$\det \mathbf{A} = \det \mathbf{L} \ \det \mathbf{U} = u_{11} \cdot u_{22} \cdot \ldots \cdot u_{n\,n} .$$

Для обращения матрицы $\mathbf{A}$ с помощью $\mathbf{LU}$-факторизации можно применить прием, который рассмотрен в § 2.1, т.е. n-кратно использовать формулы (2.8) и (2.10) для получения столбцов матрицы $\mathbf{A}^{-1}$. При этом в качестве b_i в (2.8) должны фигурировать только числа 0 или 1: для нахождения первого столбца $\mathbf{A}^{-1}$ полагаем $b_1 := 1$, $b_2 = b_3 = \ldots = b_n := 0$, для второго — $b_2 := 1$, $b_1 = b_3 = \ldots = b_n := 0$, и т.д. Можно, однако, вывести и специальные формулы для выражения элементов обратной матрицы через элементы матриц $\mathbf{L}$ и $\mathbf{U}$. Продемонстрируем это.

Пусть матрицы $\mathbf{A}$ и $\mathbf{U}$ обратимы (матрица $\mathbf{L}$ обратима всегда). Тогда

$$\mathbf{A} = \mathbf{L} \cdot \mathbf{U} \ \Leftrightarrow \ \mathbf{A}^{-1} = \mathbf{U}^{-1} \cdot \mathbf{L}^{-1} .$$

Умножая последнее равенство поочередно на $\mathbf{U}$ слева и на $\mathbf{L}$ справа, имеем

$$\mathbf{U}\mathbf{A}^{-1} = \mathbf{L}^{-1} \quad \text{и} \quad \mathbf{A}^{-1}\mathbf{L} = \mathbf{U}^{-1}. \tag{2.11}$$

---

* Чаще схемой Холецкого называют описываемый в § 2.4 основанный на той же идее способ решения с и м м е т р и ч н ы х линейных систем (метод квадратных корней).

Обозначим искомые элементы матрицы $\mathbf{A}^{-1}$ через x_{ij}. Учитывая, что треугольные матрицы при обращении сохраняют свою структуру, перепишем равенства (2.11) в следующем виде:

$$\begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & u_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ * & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ * & * & \cdots & 1 \end{pmatrix},$$

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix} = \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & * \end{pmatrix}.$$

Звездочкой здесь обозначены некоторые числа, указывающие на структуру матрицы; их знание для дальнейшего не требуется.

Полученные матричные равенства можно рассматривать как систему $2n^2$ уравнений с n^2 неизвестными x_{ij} $(i, j = 1, 2, \ldots, n)$.

Из этих $2n^2$ уравнений ровно n^2 имеют известные правые части (это 0 или 1). Выпишем соответствующую им $n \times n$-матрицу уравнений:

$$u_{11}x_{11} + \ldots + u_{1n}x_{n1} = 1, \quad u_{11}x_{12} + \ldots + u_{1n}x_{n2} = 0, \quad \ldots, \quad u_{11}x_{1n} + \ldots + u_{1n}x_{nn} = 0$$

$$x_{21} + \ldots + x_{2n}l_{n1} = 0, \quad u_{22}x_{22} + \ldots + u_{2n}x_{n2} = 1, \quad \ldots, \quad u_{22}x_{2n} + \ldots + u_{2n}x_{nn} = 0$$

$$\ldots\ldots\ldots\ldots \qquad \ldots\ldots\ldots\ldots \qquad \ldots\ldots\ldots\ldots$$

$$x_{n1} + \ldots + x_{nn}l_{n1} = 0, \qquad x_{n2} + \ldots + x_{nn}l_{n2} = 0, \quad \ldots, \qquad u_{nn}x_{nn} = 1$$

Коротко все эти уравнения могут быть представлены следующими тремя типами связей:

$$\sum_{k=i}^{n} u_{ik}x_{kj} = 0, \quad \text{если } i < j,$$

$$\sum_{k=i}^{n} u_{ik}x_{kj} = 1, \quad \text{если } i = j,$$

и

$$x_{ij} + \sum_{k=j+1}^{n} x_{ik} l_{kj} = 0, \quad \text{если } i > j.$$

Отсюда* можно выразить все элементы x_{ij} искомой обратной матрицы $\mathbf{A}^{-1}$:

$$x_{jj} = \frac{1}{u_{jj}} \left(1 - \sum_{k=j+1}^{n} u_{jk} x_{kj} \right); \tag{2.12}$$

$$x_{ij} = -\frac{1}{u_{ii}} \sum_{k=i+1}^{n} u_{ik} x_{kj} \qquad (i < j); \tag{2.13}$$

$$x_{ij} = -\sum_{k=j+1}^{n} x_{ik} l_{kj} \qquad (i > j). \tag{2.14}$$

Формулы (2.12) – (2.14) позволяют эффективно обращать **LU**-факторизованную матрицу, если соблюдать определенную технологию их использования. А именно, как видно из записанной выше матрицы уравнений, следует сначала из последнего столбца уравнений найти x_{nn}, $x_{n-1,n}$, ..., x_{2n}, x_{1n}, затем из оставшейся части последней строки уравнений найти $x_{n,n-1}, \ldots, x_{n2}, x_{n1}$, потом переключиться на предпоследний столбец и т.д.**

---

* Эти уравнения можно записать еще короче, если использовать ***символ Кронекера*** $\delta_{ij} := \begin{cases} 1, & \text{если } i = j, \\ 0, & \text{если } i \neq j \end{cases}$. Он позволяет первые две формулы совместить: $\sum_{k=i}^{n} u_{ik} x_{kj} = \delta_{ij}$. Это относится и к формулам (2.12), (2.13).

** Можно встретить и иной подход к процедуре обращения матриц на основе **LU**-разложения [29]. Учитывая, что треугольные матрицы обращаются достаточно просто, можно сначала выполнить обращение матриц **L** и **U**, а затем, умножив $\mathbf{U}^{-1}$ на $\mathbf{L}^{-1}$, получить $\mathbf{A}^{-1}$.

Пример 2.2. Применяя LU-разложение матрицы $\mathbf{A}$, выполненное в примере 1.1, найти второй столбец матрицы $\mathbf{A}^{-1}$.

Искомый столбец (обозначим его $\mathbf{x}_2 := (x_{12}; x_{22}; x_{32})^T$) можно считать решением векторно-матричного уравнения

$$\mathbf{A}\,\mathbf{x}_2 = \mathbf{e}_2, \quad \text{где} \quad \mathbf{e}_2 := (0; 1; 0)^T.$$

Знание LU-разложения матрицы $\mathbf{A}$

$$\begin{pmatrix} 2 & -1 & 1 \\ 4 & 3 & 1 \\ 6 & -13 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & -2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & -1 & 1 \\ 0 & 5 & -1 \\ 0 & 0 & 1 \end{pmatrix}$$

позволяет свести это уравнение к двум более простым. Сначала решаем уравнение $\mathbf{L}\mathbf{y} = \mathbf{e}_2$, т.е. систему

$$\begin{cases} y_1 = 0, \\ 2y_1 + y_2 = 1, \\ 3y_1 - 2y_2 + y_3 = 0, \end{cases}$$

получив при этом $\mathbf{y} := (y_1; y_2; y_3)^T = (0; 1; 2)^T$, а затем — аналогичное уравнение $\mathbf{U}\,\mathbf{x}_2 = \mathbf{y}$, также представляющее собой в развернутом виде треугольную систему

$$\begin{cases} 2x_{12} - x_{22} + x_{32} = 0, \\ 5x_{22} - x_{32} = 1, \\ x_{32} = 2, \end{cases}$$

из которой находим $\mathbf{x}_2 := (-0,7; 0,6; 2)^T$. Следовательно, обратная к $\mathbf{A}$ матрица имеет вид

$$\mathbf{A}^{-1} := \begin{pmatrix} * & -0,7 & * \\ * & 0,6 & * \\ * & 2 & * \end{pmatrix}.$$

Ясно, что последние две системы можно было и не выписывать, а выполнить вычисления непосредственно по формулам (2.8), (2.10) (или по формулам (2.12) – (2.14)).

Перемножая элементы $u_{11} := 2$, $u_{22} := 5$, $u_{33} := 1$, легко убедиться, что при этом получается то же значение $\det \mathbf{A}$, которое было получено иначе в примере 2.1.

Решение системы (2.1) с матрицей коэффициентов, факторизованной по формулам (1.5) (т.е. с фиксированием элементов диагонали матрицы $\mathbf{U}$), получают следующим образом:

$$y_i = \frac{1}{l_{i\,i}}\left(b_i - \sum_{k=1}^{i-1} l_{i\,k} y_k\right), \quad i = 1, 2, \ldots, n,$$

$$x_i = y_i - \sum_{k=i+1}^{n} u_{i\,k} x_k, \quad i = n, n-1, \ldots, 1.$$

Детерминант матрицы $\mathbf{A}$ в этом случае равен произведению $l_{11} l_{22} \ldots l_{n\,n}$, а для подсчета элементов обратной матрицы $\mathbf{X} = \mathbf{A}^{-1}$ используют совокупность формул

$$x_{i\,i} = \frac{1}{l_{i\,i}}\left(1 - \sum_{k=i+1}^{n} x_{i\,k} l_{k\,i}\right),$$

$$x_{i\,j} = -\frac{1}{l_{j\,j}} \sum_{k=i+1}^{n} x_{i\,k} l_{k\,j} \qquad (i > j),$$

$$x_{i\,j} = -\sum_{k=i+1}^{n} u_{i\,k} x_{k\,j} \qquad (i < j)$$

с такой организацией вычислений, при которой сначала вычисляют последнюю строку $\left(x_{n\,j}\right)$ при $j = n, n-1, \ldots, 1$, затем последний столбец $\left(x_{i\,n}\right)$ при $i = n-1, \ldots, 1$, потом предпоследнюю строку $\left(x_{n-1,\,j}\right)$ при $j = n-1, \ldots, 1$ и т.д.

В отличие от схемы единственного деления компактная схема Гаусса менее удобна для усовершенствования с целью уменьшения влияния вычислительных погрешностей путем выбора подходящих ведущих элементов. Достоинством же ее можно считать то, что LU-разложение матрицы $\mathbf{A}$ играет роль обратной матрицы, может помещаться в память компьютера на место матрицы $\mathbf{A}$ и использоваться, например, при решении нескольких систем, имеющих одну и ту же матрицу коэффициентов и разные правые части (и вообще для решения разных задач линейной алгебры).

Замечание 2.1. Некоторое усложнение процедуры LU-факторизации позволяет применять ее в более широких условиях. А именно, для любой невырожденной квадратной матрицы $\mathbf{A}$ можно подобрать такие матрицы переста-

новок $\mathbf{P}$ и $\mathbf{Q}$, что будет осуществимо LU-разложение матрицы $\mathbf{PAQ}$ (причем это может быть сделано так, чтобы минимизировались вычислительные погрешности, т.е. реализовывалась стратегия выбора главного элемента). Решение системы $\mathbf{Ax}=\mathbf{b}$ в таком случае находится следующим образом. Полагаем $\mathbf{x}=\mathbf{Qz}$, где $\mathbf{z}$ — вспомогательный вектор. Тогда исходная система принимает вид $\mathbf{AQz}=\mathbf{b}$; умножив последнее равенство слева на матрицу $\mathbf{P}$, приходим к эквивалентной системе $\mathbf{PAQz}=\mathbf{Pb}$, которая, в свою очередь, может быть представлена в виде $\mathbf{LUz}=\mathbf{Pb}$. Далее последовательно решаем треугольные системы $\mathbf{Ly}=\mathbf{Pb}$ и $\mathbf{Uz}=\mathbf{y}$ относительно вспомогательных векторов $\mathbf{y}$ и $\mathbf{z}$ соответственно, после чего вычисляем искомый вектор $\mathbf{x}=\mathbf{Qz}$. Вся сложность реализации такой схемы состоит в конструировании подходящих матриц перестановок $\mathbf{P}$ и $\mathbf{Q}$.

§ 2.3. РЕШЕНИЕ СИММЕТРИЧНЫХ СЛАУ

Пусть матрица $\mathbf{A}=\left(a_{ij}\right)$ системы (2.1) обладает симметрией: $a_{ij}=a_{ji}$. Тогда при наличии ее $\mathbf{U}^T\mathbf{U}$-разложения (см. § 1.3) решение системы $\mathbf{Ax}=\mathbf{b}$ сводится к последовательному решению двух треугольных систем:

$$\mathbf{U}^T\mathbf{y}=\mathbf{b} \quad \text{и} \quad \mathbf{Ux}=\mathbf{y}.$$

Первая из них имеет вид

$$\begin{cases} u_{11}y_1 = b_1, \\ u_{12}y_1 + u_{22}y_2 = b_2, \\ \dots\dots\dots\dots\dots\dots \\ u_{1\,n}y_1 + u_{2\,n}y_2 + \dots + u_{n\,n}y_n = b_n, \end{cases}$$

откуда последовательно получаем значения вспомогательных неизвестных $y_1, y_2, \dots, y_n$ по единой формуле

$$y_i = \frac{1}{u_{i\,i}}\left(b_i - \sum_{k=1}^{i-1} u_{k\,i}y_k\right), \tag{2.15}$$

полагая в ней $i := 1, 2, \dots, n$. Из второй системы

$$\begin{cases} u_{11}x_1 + u_{12}x_2 + \ldots + u_{1\,n}x_n = y_1, \\ \qquad u_{22}x_2 + \ldots + u_{2\,n}x_n = y_2, \\ \qquad\qquad \cdots\cdots\cdots \\ \qquad\qquad\qquad u_{n\,n}x_n = y_n \end{cases}$$

находим искомые значения x_i в обратном порядке, т.е. последовательной подстановкой значений $i := n,\ n-1,\ \ldots,\ 1$ в формулу

$$x_i = \frac{1}{u_{i\,i}}\left(y_i - \sum_{k=i+1}^{n} u_{i\,k}x_k\right). \qquad (2.16)$$

Решение симметричных СЛАУ посредством совокупности формул (1.6) – (1.7), (2.15) – (2.16) называют ***методом квадратных корней*** или ***схемой Холецкого***. В случае систем с положительно определенными матрицами можно рассчитывать на хорошие результаты применения такого метода (особенно если в процессе решения делать проверку на немалость величин $|u_{i\,i}|$, чтобы избежать большого роста погрешностей). В других случаях нет, например, гарантии, что в процессе разложения не появятся чисто мнимые числа, что, кстати, может не отразиться на результатах, если в алгоритме реализации метода квадратных корней предусмотреть возможность появления мнимых чисел [69].

Аналогично решаются симметричные системы (2.1) при помощи $\mathbf{U}^T\mathbf{D}\mathbf{U}$-разложения, процесс выполнения которого показан в § 1.3 (см. итоговые формулы (1.9)). В этом случае эквивалентным уравнению $\mathbf{A}\mathbf{x} = \mathbf{b}$ является уравнение

$$\mathbf{U}^T\mathbf{D}\mathbf{U}\mathbf{x} = \mathbf{b}\ .$$

Введя вектор вспомогательных переменных $\mathbf{y} = (y_1; y_2; \ldots; y_n)^T$, последнее можно переписать в виде системы

$$\begin{cases} \mathbf{U}^T\mathbf{D}\mathbf{y} = \mathbf{b}, \\ \mathbf{U}\mathbf{x} = \mathbf{y}\ . \end{cases} \qquad (2.17)$$

Первое из уравнений этой системы в развернутом виде суть

$$\begin{cases} d_{11}y_1 = b_1, \\ u_{12}d_{11}y_1 + d_{22}y_2 = b_2, \\ \cdots\cdots\cdots\cdots\cdots \\ u_{1\,n}d_{11}y_1 + u_{2\,n}d_{22}y_2 + \ldots + d_{nn}y_n = b_n . \end{cases}$$

Отсюда, придавая i значения 1, 2, ... , n , последовательно получаем значения вспомогательных неизвестных y_1, y_2, ..., y_n:

$$y_i = \frac{1}{d_{ii}}\left(b_i - \sum_{k=1}^{i-1} u_{ki}d_{kk}y_k\right) . \qquad (2.18)$$

Развернув второе уравнение векторно-матричной системы (2.17), получаем систему

$$\begin{cases} x_1 + u_{12}x_2 + \ldots + u_{1\,n}x_n = y_1 , \\ \qquad x_2 + \ldots + u_{2\,n}x_n = y_2 , \\ \qquad\qquad \cdots\cdots\cdots \\ \qquad\qquad x_n = y_n . \end{cases}$$

Значения $x_n, x_{n-1},\ldots ,x_1$ находят из нее по формуле

$$x_i = y_i - \sum_{k=i+1}^{n} u_{ik}x_k \ , \quad \text{где } i := n,\ n-1,\ \ldots,\ 1 . \qquad (2.19)$$

Таким образом, решение симметричной системы вида (2.1), основанное на $\mathbf{U}^T\mathbf{DU}$-разложении, означает: получение соответствующей факторизации с помощью связки формул (1.9) и затем вычисление решения последовательным применением формул (2.18) и (2.19) .

§ 2.4. МЕТОД ПРОГОНКИ

Часто возникает необходимость в решении линейных алгебраических систем, матрицы которых, являясь слабо заполненными, т.е. содержащими немного ненулевых элементов, имеют вполне конкретную четкую структуру. Среди них выделим системы с такими матрицами, в которых ненулевые элементы располагаются на главной диагонали и на нескольких примыкающих к ней диагоналях. Для решения этих систем с ленточными матрицами

коэффициентов метод Гаусса можно модифицировать так, чтобы его применение стало более эффективным.

Рассмотрим наиболее простой случай ***ленточных систем***, к которым сводится решение задач сплайн-интерполяции функций, дискретизации краевых задач для дифференциальных уравнений методами конечных разностей, конечных элементов и др. А именно, будем искать решение такой системы, каждое уравнение которой связывает в векторе неизвестных $\mathbf{x} := (x_1; x_2; \ldots; x_n)^T$ три «соседних» компоненты; подобная связь позволяет записать систему одним уравнением:

$$b_i x_{i-1} + c_i x_i + d_i x_{i+1} = r_i\,, \qquad (2.20)$$

где $i = 1, 2, \ldots, n$; $b_1 := 0$, $d_n := 0$. Уравнения вида (2.20) называют ***трехточечными разностными уравнениями второго порядка****. Матрица системы (2.20) является ***трехдиагональной***, что хорошо видно из следующего эквивалентного (2.20) векторно-матричного представления:

$$\begin{pmatrix} c_1 & d_1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ b_2 & c_2 & d_2 & 0 & \ldots & 0 & 0 & 0 \\ 0 & b_3 & c_3 & d_3 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & b_{n-1} & c_{n-1} & d_{n-1} \\ 0 & 0 & 0 & 0 & \ldots & 0 & b_n & c_n \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \ldots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ \ldots \\ r_{n-1} \\ r_n \end{pmatrix}.$$

Ставя цель избавиться от ненулевых элементов в поддиагональной части матрицы системы, предположим, что существуют такие наборы чисел δ_i и λ_i $(i = 1, 2, \ldots, n)$, при которых имеет

---

* Чаще вместо принятой здесь записи системы (2.20), предполагающей как бы наличие фиктивных неизвестных x_0 и x_{n+1} с нулевыми коэффициентами, считают в (2.20) текущий индекс i изменяющимся от 2 до $n-1$, выделяя первое и последнее уравнения системы соответственно:

$$c_1 x_1 + d_1 x_2 = r_1 \quad \text{и} \quad b_n x_{n-1} + c_n x_n = r_n$$

— в отдельные строки (их называют ***краевыми условиями*** разностного уравнения (2.20)).

место равенство

$$x_i = \delta_i x_{i+1} + \lambda_i, \tag{2.21}$$

означающее, что трехточечное уравнение второго порядка (2.20) преобразуется в двухточечное уравнение первого порядка (2.21).

Уменьшим в связи (2.21) индекс на единицу и полученное при этом выражение $x_{i-1} = \delta_{i-1} x_i + \lambda_{i-1}$ подставим в данное уравнение (2.20):

$$b_i \delta_{i-1} x_i + b_i \lambda_{i-1} + c_i x_i + d_i x_{i+1} = r_i,$$

откуда получаем

$$x_i = -\frac{d_i}{c_i + b_i \delta_{i-1}} x_{i+1} + \frac{r_i - b_i \lambda_{i-1}}{c_i + b_i \delta_{i-1}}. \tag{2.22}$$

Равенство (2.22) имеет вид равенства (2.21) и будет точно с ним совпадать, иначе говоря, представление (2.21) будет иметь место, если при всех $i = 1, 2, \ldots, n$ выполняются рекуррентные соотношения

$$\delta_i = -\frac{d_i}{c_i + b_i \delta_{i-1}}, \quad \lambda_i = \frac{r_i - b_i \lambda_{i-1}}{c_i + b_i \delta_{i-1}}. \tag{2.23}$$

Легко видеть, что в силу условия $b_1 := 0$ процесс вычисления δ_i, λ_i может быть начат со значений

$$\delta_1 := -\frac{d_1}{c_1}, \quad \lambda_1 := \frac{r_1}{c_1}$$

и продолжен далее по формулам (2.23) последовательно при $i := 2, 3, \ldots, n$, причем при $i = n$, в силу $d_n := 0$, получим $\delta_n = 0$. Следовательно, полагая в (2.21) $i := n$, будем иметь

$$x_n = \lambda_n = \frac{r_n - b_n \lambda_{n-1}}{c_n + b_n \delta_{n-1}},$$

где λ_{n-1}, δ_{n-1} — уже известные с предыдущего шага числа. Далее по формулам (2.21) последовательно находим значения x_{n-1}, x_{n-2}, ..., x_1 при $i := n-1, n-2, \ldots, 1$ соответственно.

Таким образом, решение уравнений вида (2.20) выведенным выше ***методом прогонки**** сводится к вычислениям по трем простым формулам. Сначала, полагая $i := 1, 2, \ldots, n$, делают цикл вычислений так называемых ***прогоночных коэффициентов*** δ_i, λ_i по формулам (2.23) (***прямая прогонка***), а затем по формуле (2.21) при $i = n, n-1, \ldots, 1$ вычисляют значения неизвестных x_i (***обратная прогонка***).

Для успешного применения метода прогонки нужно, чтобы в процессе вычислений не возникало ситуаций с делением на нуль, а при больших размерностях систем не было быстрого роста погрешностей округлений.

Будем называть прогонку ***корректной****, если знаменатели прогоночных коэффициентов* (2.23) *не обращаются в нуль, и* ***устойчивой****, если* $|\delta_i| < 1$ *при всех* $i \in \{1, 2, \ldots, n\}$.

Приведем простые достаточные условия корректности и устойчивости прогонки, которые во многих приложениях метода автоматически выполняются.

Теорема 2.1. *Пусть коэффициенты* b_i *и* d_i *уравнения* (2.20) *при* $i = 2, 3, \ldots, n-1$ *отличны от нуля и пусть*

$$|c_i| > |b_i| + |d_i| \quad \forall\ i \in \{1, 2, \ldots, n\}\ . \qquad (2.24)$$

Тогда прогонка (2.23), (2.21) *корректна и устойчива* (*т.е.* $c_i + b_i\delta_{i-1} \neq 0$, $|\delta_i| < 1$).

Доказательство. Воспользуемся методом математической индукции для установления обоих нужных неравенств одновременно.

При $i := 1$, в силу (2.24), имеем:

$$|c_1| > |d_1| \geq 0\,,$$

т.е. неравенство нулю знаменателя первой пары прогоночных коэффициентов, а также

---

* Термин, характерный в основном для отечественной литературы по вычислительной математике, введен в 50-х гг. XX в. (см., например, [1, 22, 44]).

$$|\delta_1| = |-d/c_1| < 1 .$$

Предположим, что знаменатель $(i-1)$-х прогоночных коэффициентов не равен нулю и что $|\delta_{i-1}| < 1$. Тогда, используя свойства модулей, условия теоремы и индукционные предположения, получаем

$$|c_i + b_i\delta_{i-1}| \geq |c_i| - |b_i\delta_{i-1}| > |b_i| + |d_i| - |b_i| \cdot |\delta_{i-1}| =$$
$$= |d_i| + |b_i|\left(1 - |\delta_{i-1}|\right) > |d_i| > 0 ,$$

а с учетом этого

$$|\delta_i| = \left| -\frac{d_i}{c_i + b_i\delta_{i-1}} \right| = \frac{|d_i|}{|c_i + b_i\delta_{i-1}|} < \frac{|d_i|}{|d_i|} = 1 .$$

Следовательно, $c_i + b_i\delta_{i-1} \neq 0$ и $|\delta_i| < 1$ при всех $i \in \{1, 2, \ldots, n\}$, т.е. имеет место утверждаемая в данных условиях корректность и устойчивость прогонки. Теорема доказана.

Имеются более слабые условия корректности и устойчивости прогонки, чем требуемое в теореме 2.1 условие строгого диагонального преобладания в матрице **A** (соответствующие утверждения см., например, в книгах [19, 44, 61, 62]). В частности, *для корректности и устойчивости прогонки, определяемой формулами* (2.21), (2.23), *достаточно, чтобы хотя бы одно из неравенств* (2.24) *было строгим* [62].

Пусть **A** — матрица коэффициентов данной системы (2.20), удовлетворяющих условиям теоремы 2.2, и пусть

$$\delta_1 := -\frac{d_1}{c_1} , \quad \delta_i := -\frac{d_i}{c_i + b_i\delta_{i-1}} \quad (i = 2, 3, \ldots, n-1) , \quad \delta_n := 0$$

— прогоночные коэффициенты, определяемые первой из формул (2.23), а

$$\Delta_i := c_i + b_i\delta_{i-1} \quad (i = 2, 3, \ldots, n)$$

— знаменатели этих коэффициентов (отличные от нуля согласно утверждению теоремы 2.1).

Непосредственной проверкой легко убедиться, что имеет место представление $\mathbf{A} = \mathbf{LU}$, в котором матрицы $\mathbf{L}$ и $\mathbf{U}$ имеют следующий вид:

$$\mathbf{L} := \begin{pmatrix} c_1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ b_2 & \Delta_2 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & b_3 & \Delta_3 & 0 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & b_{n-1} & \Delta_{n-1} & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & b_n & \Delta_n \end{pmatrix},$$

$$\mathbf{U} := \begin{pmatrix} 1 & -\delta_1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -\delta_2 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 1 & -\delta_3 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & -\delta_{n-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix}.$$

В силу утверждения теоремы 2.1, оно единственно. При этом выведенный ранее ***метод правой прогонки***, очевидно, соответствует описанной в § 1.2 **LU**-факторизации с фиксированием единичной диагонали у правой треугольной матрицы $\mathbf{U}$.

Учет ленточной структуры факторизуемой матрицы $\mathbf{A}$ позволяет отказаться от приведенных для произвольной матрицы формул (1.5) в пользу менее затратного процесса преобразований. Именно, будем считать, что факторизуется трехдиагональная матрица $\mathbf{A} := (a_{ij})_{i,j=1}^{n}$, соответствующая матрице системы (2.20), т.е. для ее элементов имеют место соотношения

$$a_{ii} := c_i \qquad \forall\ i \in \{1, 2, \dots, n\},$$

$$a_{i+1,i} := b_{i+1}, \quad a_{i,i+1} := d_i \qquad \forall\ i \in \{1, 2, \dots, n-1\},$$

$$a_{ij} := 0 \quad \text{при } |i-j| > 1.$$

Тогда при сформулированных в теореме 2.1 условиях такую мат-

рицу можно представить в виде произведения двух ***бидиагональных матриц*** следующего вида:

$$\mathbf{L}=\begin{pmatrix} l_{11} & 0 & 0 & \cdots & 0 & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 & 0 \\ 0 & l_{32} & l_{33} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & l_{n,n-1} & l_{nn} \end{pmatrix} \text{ и } \mathbf{U}=\begin{pmatrix} 1 & u_{12} & 0 & \cdots & 0 & 0 \\ 0 & 1 & u_{23} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & u_{n-1,n} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Осуществить такое представление можно, применяя следующий простейший **алгоритм LU-разложения трехдиагональной матрицы:**

1. $\delta_0 := 0;\quad b_1 := 0;\quad d_n := 0$.
2. Для $i := 1, 2, \ldots, n$:
3. $\quad l_{ii} := c_i + b_i\delta_{i-1}$,
4. $\quad \delta_i = -\dfrac{d_i}{l_{ii}}$.
5. Для $i := 1, 2, \ldots, n-1$:
6. $\quad l_{i+1,i} := b_{i+1}$,
7. $\quad u_{i,i+1} := -\delta_i$.

При необходимости попутно может быть вычислен определитель:

$$\det \mathbf{A} = \prod_{i=1}^{n} l_{ii} = c_1 \prod_{i=2}^{n} \Delta_i . \qquad (2.25)$$

Пример 2.3. Методом прогонки решить систему

$$\begin{cases} 2x_1 + x_2 = -10, \\ 2x_1 + 9x_2 + 2x_3 = -26, \\ 4x_2 + 17x_3 - 4x_4 = -16, \\ 4x_3 + 15x_4 - 8x_5 = -2, \\ 2x_4 + 3x_5 = 16 \end{cases}$$

и вычислить определитель матрицы ее коэффициентов.

Данная система состоит из уравнений вида (2.20), где b_i, c_i, d_i, r_i суть элементы векторов

$$\mathbf{b} := \begin{pmatrix} 0 \\ 2 \\ 4 \\ 4 \\ 2 \end{pmatrix}, \quad \mathbf{c} := \begin{pmatrix} 2 \\ 9 \\ 17 \\ 15 \\ 3 \end{pmatrix}, \quad \mathbf{d} := \begin{pmatrix} 1 \\ 2 \\ -4 \\ -8 \\ 0 \end{pmatrix} \quad \text{и} \quad \mathbf{r} := \begin{pmatrix} -10 \\ -26 \\ -16 \\ -2 \\ 16 \end{pmatrix}$$

соответственно. Налицо диагональное преобладание в матрице системы, означающее, что ни при каком $i \in \{1, 2, \ldots, 5\}$ величина

$$\Delta_i := c_i + b_i \delta_{i-1}$$

не обратится в нуль (корректность) и что при вычислении прогоночных коэффициентов δ_i, λ_i по формулам

$$\delta_i = -\frac{d_i}{\Delta_i}, \quad \lambda_i = \frac{r_i - b_i \lambda_{i-1}}{\Delta_i} \qquad (i = 1, 2, \ldots, 5)$$

абсолютные величины всех δ_i окажутся меньшими единицы (устойчивость).

Полагая последовательно $i := 1, 2, \ldots, 5$, имеем (прямая прогонка):

$$\Delta_1 := c_1 = 2, \qquad \delta_1 := -\frac{d_1}{\Delta_1} = -\frac{1}{2}, \qquad \lambda_1 := \frac{r_1}{\Delta_1} = \frac{-10}{2} = -5;$$

$$\Delta_2 := c_2 + b_2\delta_1 = 8, \qquad \delta_2 := -\frac{d_2}{\Delta_2} = -\frac{1}{4}, \quad \lambda_2 := \frac{r_2 - b_2\lambda_1}{\Delta_2} = -2;$$

$$\Delta_3 := c_3 + b_3\delta_2 = 16, \qquad \delta_3 := -\frac{d_3}{\Delta_3} = \frac{1}{4}, \qquad \lambda_3 := \frac{r_3 - b_3\lambda_2}{\Delta_3} = -\frac{1}{2};$$

$$\Delta_4 := c_4 + b_4\delta_3 = 16, \qquad \delta_4 := -\frac{d_4}{\Delta_4} = \frac{1}{2}, \qquad \lambda_4 := \frac{r_4 - b_4\lambda_3}{\Delta_4} = 0;$$

$$\Delta_5 := c_5 + b_5\delta_4 = 4, \qquad \delta_5 := -\frac{d_5}{\Delta_5} = 0, \qquad \lambda_5 := \frac{r_5 - b_5\lambda_4}{\Delta_5} = 4.$$

Обратная прогонка по формуле (2.21) при $i := 5, 4, \ldots, 1$ дает искомые значения неизвестных:

$$x_5 := \lambda_5 = 4,$$

$$x_4 := \delta_4 x_5 + \lambda_4 = \frac{1}{2} \cdot 4 + 0 = 2,$$

$$x_3 := \delta_3 x_4 + \lambda_3 = \frac{1}{4} \cdot 2 + (-\frac{1}{2}) = 0,$$

$$x_2 := \delta_2 x_3 + \lambda_2 = -\frac{1}{4} \cdot 0 + (-2) = -2,$$

$$x_1 := \delta_1 x_2 + \lambda_1 = -\frac{1}{2} \cdot (-2) + (-5) = -4.$$

Для вычисления определителя матрицы **A** коэффициентов данной системы, согласно формуле (2.25), достаточно перемножить пять чисел:

$$\det \mathbf{A} := \prod_{i=1}^{5} \Delta_i = 2 \cdot 8 \cdot 16 \cdot 16 \cdot 4 = 16384 .$$

В заключение этого параграфа заметим, что применяется ряд других, отличных от рассмотренной нами правой прогонки, методов подобного типа, решающих как поставленную здесь задачу (2.20) для систем с трехдиагональными матрицами (левая прогонка, встречная прогонка, немонотонная, циклическая, ортогональная прогонки и т.д.), так и для более сложных систем с матрицами ленточной структуры или блочно-матричной структуры (например, матричная прогонка, пятидиагональная прогонка). Коротко охарактеризуем некоторые из них.

Левая прогонка. Отличается от правой тем, что при выводе этого метода решения систем (2.20) вместо равенства (2.21) предполагается наличие двухточечной связи между «соседними» неизвестными x_i и x_{i+1} вида

$$x_{i+1} = \tilde{\delta}_i x_i + \tilde{\lambda}_i .$$

Из этого предположения вытекают формулы для прогоночных коэффициентов $\tilde{\delta}_i$ и $\tilde{\lambda}_i$, аналогичные формулам (2.23). Подсчет значений $\tilde{\delta}_i$, $\tilde{\lambda}_i$ по новым формулам должен производиться при изменении i от n до 1, а сами значения неизвестных теперь подсчитываются в прямом направлении изменения i (от 1 до n). Очевидно, левую прогонку можно связать с **LU**-разложением трехдиагональной матрицы данной системы при фиксировании единичной диагонали у матрицы **L** .

Встречная прогонка представляет собой комбинацию правой и левой прогонок. Ее целесообразно применять в случаях, когда при значительной размерности n требуется находить не все значения неизвестных, а какое-то одно, например, x_m, где $1 < m < n$ (или несколько компонент вектора неизвестных, группируемых около x_m). Суть в том, чтобы вычислить значения δ_i, λ_i при i от 1 до m по формулам правой прогонки и значения $\tilde{\delta}_i$,

$\tilde{\lambda}_i$ при i от n до m по формулам левой прогонки, а затем из равенств

$$x_m = \delta_m x_{m+1} + \lambda_m, \quad x_{m+1} = \tilde{\delta}_m x_m + \tilde{\lambda}_m,$$

исключив x_{m+1}, найти значение x_m. Зная его, теперь можно последовательно вычислять как $x_{m+1}, x_{m+2}, \ldots$, так и $x_{m-1}, x_{m-2}, \ldots$.

Немонотонная прогонка — модификация правой прогонки, направленная на расширение границ применимости и повышение вычислительной устойчивости метода прогонки решения систем вида (2.20). В противовес обычной прогонке, которой можно поставить в соответствие рассмотренную выше (§ 2.1) схему единственного деления, реализующую метод Гаусса с монотонным исключением неизвестных в заранее заданном порядке, немонотонная прогонка строится как метод Гаусса с построчным выбором главного (ведущего) элемента для данной специфичной системы с трехдиагональной матрицей.

Ортогональная прогонка имеет ту же цель расширения возможностей устойчивого решения систем уравнений (2.20), но построение этого метода существенно отличается от предыдущей методики и опирается, явно или неявно, на идеи ортогональных преобразований (вращения). Сложность этого метода вряд ли оправдывает его достоинства.

Циклическая прогонка применяется для решения систем, определяемых трехточечным уравнением типа (2.20) в следующей специфической постановке:

считаем, что в (2.20) $i := 0, \pm 1, \pm 2, \ldots$ и что при этом имеют место равенства

$$b_i = b_{i+N}, \quad c_i = c_{i+N}, \quad d_i = d_{i+N}, \quad r_i = r_{i+N}.$$

Заданная периодичность коэффициентов влечет за собой периодичность решения с тем же периодом N: $x_i = x_{i+N}$. Последнее означает, что достаточно найти только N расположенных подряд значений $x_0, x_1, \ldots, x_{N-1}$ компонент бесконечномерного вектора решения. С этой целью из данной бесконечной системы выделяют конечную подсистему

$$\begin{cases} b_0 x_{N-1} + c_0 x_0 + d_0 x_1 = r_0\,, \\ b_i x_{i-1} + c_i x_i + d_i x_{i+1} = r_i \quad (i = 1, 2, \ldots, N-1), \\ x_N = x_0\,. \end{cases}$$

Поскольку у этой системы нарушена трехдиагональная структура, ее решение находят в виде

$$x_i = u_i + v_i x_0 \quad (i = 0, 1, \ldots, N)\,,$$

где вспомогательные неизвестные u_i и v_i должны быть соответственно решениями задач

$$\begin{cases} b_i u_{i-1} + c_i u_i + d_i u_{i+1} = r_i \quad (i = 1, 2, \ldots, N-1)\,, \\ u_0 := 0\,, \qquad u_N := 0 \end{cases}$$

и

$$\begin{cases} b_i v_{i-1} + c_i v_i + d_i v_{i+1} = 0 \quad (i = 1, 2, \ldots, N-1)\,, \\ v_0 := 1\,, \qquad v_N := 1\,. \end{cases}$$

Две последние системы, очевидно, можно решить обычной прогонкой, а требуемое для нахождения произвольного значения x_i значение x_{i+1} можно выразить из первого уравнения предыдущей системы, подставляя в него

$$x_{N-1} := u_{N-1} + v_{N-1} x_0 \quad \text{и} \quad x_1 := u_1 + v_1 x_0\,.$$

Другие методы прогонки предназначены для решения задач, постановки которых более существенно отличаются от (2.20), чем предыдущие. Так, ***пятидиагональную прогонку*** выводят и обосновывают для решения СЛАУ с пятидиагональными матрицами. Различные варианты ***матричной прогонки*** строят для случаев, когда системы уравнений имеют вид, аналогичный (2.20), но неизвестные $x_i := \mathbf{x}_i$ являются векторами, а коэффициенты — соответственно матрицами. При этом имеется определенное разнообразие как в постановках разностных задач с векторными неизвестными, так и в способах построения матричных прогонок.

Вывод формул и исследование различных вариантов метода прогонки можно найти, например, в [33, 62].

§ 2.5. МЕТОДЫ ОТРАЖЕНИЙ И ВРАЩЕНИЙ

Возвращаясь к линейной системе вида (2.1) с матрицей $\mathbf{A}$, не обладающей какими-либо структурными особенностями, предложим два метода ее решения, основанных на ортогональной триангуляризации, описанной в § 1.4, 1.5.

Сначала рассмотрим метод, опирающийся на преобразования Хаусхолдера и носящий название ***метод отражений***.

Возможны две ситуации.

1. Предположим, что уже известно $\mathbf{QR}$-разложение матрицы $\mathbf{A}$:

$$\mathbf{A} = \mathbf{QR}, \tag{2.26}$$

где $\mathbf{Q} := (q_{ij})_{i,\,j=1}^{n}$ — матрица ортогональная, а $\mathbf{R} := (r_{ij})_{i,\,j=1}^{n}$ — правая треугольная матрица. Тогда после подстановки (2.26) в (2.1а) имеем эквивалентную систему

$$\mathbf{QRx} = \mathbf{b},$$

которая, в свою очередь, ввиду ортогональности матрицы $\mathbf{Q}$ равносильна системе

$$\mathbf{Rx} = \mathbf{Q}^T\mathbf{b}. \tag{2.27}$$

Введя вспомогательный вектор $\mathbf{y} := \mathbf{Q}^T\mathbf{b}$ с координатами

$$y_i := \sum_{k=1}^{n} q_{ki} b_k, \quad i = 1, 2, \ldots, n,$$

систему (2.27) представляем следующим образом:

$$\begin{cases} r_{11}x_1 + r_{12}x_2 + \ldots + r_{1n}x_n = y_1, \\ \qquad r_{22}x_2 + \ldots + r_{2n}x_n = y_2, \\ \qquad\qquad \cdots\cdots\cdots\cdots \\ \qquad\qquad\qquad r_{nn}x_n = y_n. \end{cases} \tag{2.28}$$

Отсюда видно, что для получения искомого решения $\mathbf{x}$ достаточно только выполнить обратный ход метода Гаусса. Если диагональ матрицы $\mathbf{R}$ не содержит нулей, т.е. если

$$r_{ii} \neq 0 \quad \forall\, i \in \{1, 2, \ldots, n\},$$

то неизвестные системы (2.28), а значит, и исходной системы (2.1)

можно последовательно найти по формулам

$$x_n = \frac{y_n}{r_{nn}}, \quad x_i = \frac{y_i - \sum_{k=i+1}^{n} r_{ik} x_k}{r_{ii}}, \quad \text{где } i = n-1, \ldots, 2, 1. \qquad (2.29)$$

Описанная ситуация может иметь место, когда, например, нужно решить несколько СЛАУ с одной и той же матрицей коэффициентов, но с разными правыми частями, или если решаются разные линейные задачи для одной и той же матрицы, которую в таком случае целесообразно хранить в факторизованном виде.

2. Будем теперь исходить из того, что готового **QR**-разложения матрицы $\mathbf{A}$ нет и, вообще говоря, оно в явном виде не требуется, а нужно получить решение системы (2.1), используя преобразования отражения.

Промежуточной целью здесь опять является приведение данной СЛАУ к виду (2.28), служащему стартовым для обратного хода метода Гаусса. Это означает, что одинаковыми преобразованиями, сохраняющими эквивалентность систем, матрицу $\mathbf{A}$ нужно привести к верхней треугольной матрице $\mathbf{R}$, а вектор $\mathbf{b}$ — к вектору $\mathbf{y} = \mathbf{Q}^T\mathbf{b}$, где $\mathbf{R}$ и $\mathbf{Q}$ отвечают представлению (2.26). В других терминах расширенную матрицу $\mathbf{B} := (\mathbf{A}|\mathbf{b})$ системы (2.1) ортогональными преобразованиями нужно перевести в расширенную матрицу $\mathbf{C} := (\mathbf{R}|\mathbf{y})$ системы (2.28). Ясно, что это можно сделать, применяя последовательно к столбцам матрицы $\mathbf{B}$, например, преобразования Хаусхолдера по схеме **QR**-факторизации из § 1.4.

Действительно, так как в этом случае

$$\mathbf{R} = \mathbf{H}_{n-1} \ldots \mathbf{H}_2 \mathbf{H}_1 \mathbf{A}, \quad \mathbf{Q}^T = \mathbf{H}_{n-1} \ldots \mathbf{H}_2 \mathbf{H}_1$$

и, значит, $\mathbf{Q}^T\mathbf{B} = \mathbf{C}$, то отсюда имеем следующее «технологичное» представление $n \times (n+1)$-матрицы $\mathbf{C} := (\mathbf{R}|\mathbf{y})$:

$$\mathbf{C} = \mathbf{H}_{n-1}\left(\ldots\left(\mathbf{H}_2\left(\mathbf{H}_1\mathbf{B}\right)\right)\ldots\right),$$

где $\mathbf{H}_i$ $(i = 1, 2, \ldots, n-1)$ — матрица Хаусхолдера i-го этапа, определяемая формулами (1.22) – (1.23).

Пример 2.4. Решим методом отражений систему

$$\begin{cases} x_1 - 2x_2 + x_3 = 1, \\ 2x_1 \quad - 3x_3 = 8, \\ 2x_1 - x_2 - x_3 = 5. \end{cases} \tag{2.30}$$

Для этого нужно выполнить два этапа преобразований Хаусхолдера над расширенной матрицей

$$\mathbf{B} := (\mathbf{A}|\mathbf{b}) = \left(\begin{array}{ccc|c} 1 & -2 & 1 & 1 \\ 2 & 0 & -3 & 8 \\ 2 & -1 & -1 & 5 \end{array}\right), \tag{2.31}$$

а затем сделать обратный ход по формулам (2.29).

На первом этапе имеем:

$$\beta_1 := \operatorname{sgn}_+(-a_{11})\sqrt{a_{11}^2 + a_{21}^2 + a_{31}^2} = -\sqrt{1 + 2^2 + 2^2} = -3,$$

$$\mu_1 := \frac{1}{\sqrt{2\beta_1^2 - 2\beta_1 a_{11}}} = \frac{1}{\sqrt{2(-3)^2 - 2(-3)1}} = \frac{1}{2\sqrt{6}},$$

$$2\mathbf{w}\mathbf{w}^T = \frac{1}{3}\begin{pmatrix} 4 & 2 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{pmatrix}, \quad \mathbf{H}_1 = \frac{1}{3}\begin{pmatrix} -1 & -2 & -2 \\ -2 & 2 & -1 \\ -2 & -1 & 2 \end{pmatrix}, \quad \mathbf{H}_1\mathbf{B} = \frac{1}{3}\left(\begin{array}{ccc|c} -9 & 4 & 7 & -27 \\ 0 & 5 & -7 & 9 \\ 0 & 2 & -1 & 0 \end{array}\right).$$

Результаты вычислений на втором этапе (округленные до третьего знака после запятой) следующие:

$$\beta_2 := \operatorname{sgn}_+\left(-a_{22}^{(1)}\right)\sqrt{\left(a_{22}^{(1)}\right)^2 + \left(a_{32}^{(1)}\right)^2} = -1,795, \quad \mu_2 := \frac{1}{\sqrt{2\beta_2^2 - 2\beta_2 a_{22}^{(1)}}} = 0,284,$$

$$\mathbf{H}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0,932 & -0,371 \\ 0 & -0,371 & 0,928 \end{pmatrix}, \quad \mathbf{C} := \mathbf{H}_2(\mathbf{H}_1\mathbf{B}) = \left(\begin{array}{ccc|c} -3 & 1,333 & 2,333 & -9 \\ 0 & -1,795 & 2,290 & -2,785 \\ 0 & 0 & 0,557 & -1,114 \end{array}\right).$$

Далее при помощи обратного хода метода Гаусса по формулам (2.29) получаем решение системы:

$$x_3 = \frac{-1,114}{0,557} = -2,000,$$

$$x_2 = \frac{-2,785 - 2,290 \cdot (-2)}{-1,795} = -1,000,$$

$$x_1 = \frac{-9 - 1,333(-1) - 2,333(-2)}{-3} = 1,000.$$

Если иметь в виду формирование элементов новой расширенной матрицы системы на месте массива $n\times(n+1)$ элементов расширенной матрицы $(\mathbf{A}|\mathbf{b})$ исходной системы (что позволяет отказаться от верхних индексов — номеров этапов), то в соответствии с формулами (1.23), (1.24) пересчет можно осуществлять следующим образом.

Фиксируя индекс $i := 1, 2, \ldots, n-1$, вычисляем $s := \sum_{k=i}^{n} a_{ki}^2$. Если $s \neq 0$, тогда находим значения

$$\beta_i := \begin{cases} -\sqrt{s} & \text{при } a_{ii} > 0, \\ \sqrt{s} & \text{при } a_{ii} < 0 \end{cases} \quad \text{и} \quad \gamma_i := \frac{1}{\beta_i^2 - \beta_i a_{ii}}.$$

Заменив a_{ii} на $a_{ii} - \beta_i$, далее, полагая $m := i, \ldots, n$, а $j := i+1, \ldots, n$, подсчитываем

$$a_{mj} := a_{mj} - \gamma_i a_{mi} \sum_{k=i}^{n} a_{ki} a_{kj},$$

$$b_m := b_m - \gamma_i a_{mi} \sum_{k=i}^{n} a_{ki} b_k.$$

В таком случае в системе (2.28) можно принять $y_i := b_i$, а

$$r_{ij} := \begin{cases} \beta_i & \text{при } j = i, \\ a_{ij} & \text{при } j > i, \\ (0 & \text{при } j < i), \end{cases}$$

и получить решение исходной системы (2.1) по формулам (2.29).

Условия применимости рассмотренного метода решения СЛАУ оформим в виде следующего утверждения.

Теорема 2.2. *Если решение линейной системы* $\mathbf{Ax} = \mathbf{b}$ *с вещественной квадратной матрицей* $\mathbf{A}$ *существует и единственно, то оно может быть найдено методом отражений.*

Справедливость этой теоремы вытекает непосредственно из установленного ранее факта разложимости произвольной квадратной матрицы посредством преобразований отражения в произведение ортогональной $\mathbf{Q}$ и треугольной $\mathbf{R}$ матриц, однозначности арифметических операций, конечное число которых, как установлено, приводит к решению, и следующей леммы, показывающей, что при вычислениях по формулам (2.29) не может встретиться деление на нуль.

Лемма 2.1. *Если матрица* $\mathbf{A}$ *не вырождена, то в результате ее* **QR**-*разложения на диагонали треугольной матрицы* $\mathbf{R}$ *не может быть нулей.*

Доказательство. Так как транспонирование матрицы не изменяет ее определителя [18, 54], то из равенства $\mathbf{E} = \mathbf{Q}\mathbf{Q}^T$ имеем:

$$1 = \det \mathbf{E} = \det \mathbf{Q} \cdot \det \mathbf{Q}^T = \left(\det \mathbf{Q}\right)^2 \quad \Rightarrow \quad \det \mathbf{Q} = \pm 1.$$

Но тогда из представления (2.26) следует

$$\det \mathbf{A} = \det \mathbf{Q} \cdot \det \mathbf{R} = \pm \det \mathbf{R} = \pm r_{11} r_{22} \ldots r_{nn},$$

так что неравенство нулю определителя матрицы $\mathbf{A}$ возможно лишь в случае, когда ни один из диагональных элементов матрицы $\mathbf{R}$ не равен нулю.

Лемма доказана.

Более основательные сведения о методе отражений можно почерпнуть, например, в книгах [21, 68]).

Если для матрицы $\mathbf{A}$ системы (2.1) факторизация (2.26) осуществляется плоскими вращениями Гивенса (§ 1.5), то базирующийся на нем прямой метод решения СЛАУ называется ***методом вращений***. При наличии готового гивенсова **QR**-разложения получение решения $\mathbf{x}$ системы (2.1) в этом случае ничем не отличается от описанной выше ситуации 1 для метода отражений. Посмотрим, что собой представляет метод вращений в ситуации 2. Очевидно, здесь для приведения исходной системы к виду (2.28), чтобы в дальнейшем воспользоваться формулами обратного хода (2.29), нужно над вектором $\mathbf{b}$ выполнить всю последовательность элементарных преобразований вращения, превративших матрицу

$\mathbf{A}$ в треугольную матрицу $\mathbf{R}$ (см. формулу (1.35)). Как и при построении метода отражений в этой ситуации, целесообразно сразу работать с расширенной матрицей $\mathbf{B} := (\mathbf{A}|\mathbf{b})$ системы (2.1), трансформируя ее в трапециевидную форму (соответствующую системе (2.28)) посредством матриц Гивенса

$$\mathbf{T}_{12}, \mathbf{T}_{13}, \ldots, \mathbf{T}_{1n} \; ; \quad \mathbf{T}_{23}, \mathbf{T}_{24}, \ldots, \mathbf{T}_{2n} \; ; \ldots ; \; \mathbf{T}_{n-2,n}, \mathbf{T}_{n-2,n-1} \; ; \; \mathbf{T}_{n-1,n}$$

в указанной последовательности.

Пример 2.5. Рассмотрим решение трехмерной системы уравнений (2.30) предыдущего примера методом вращений. Представленную в (2.31) расширенную матрицу $\mathbf{B} := (\mathbf{A}|\mathbf{b})$ этой системы теперь приведем к трапециевидной форме плоскими вращениями Гивенса. Априори, их должно быть три.

На первом шаге имеем (см. формулы (1.31) и/или (1.32) – (1.33)):

$$c_1 := \frac{a_{11}}{\sqrt{{a_{11}}^2 + {a_{21}}^2}} = \frac{1}{\sqrt{1^2 + 2^2}} = \frac{1}{\sqrt{5}} \; , \qquad s_1 := \frac{a_{21}}{\sqrt{{a_{11}}^2 + {a_{21}}^2}} = \frac{2}{\sqrt{1^2 + 2^2}} = \frac{2}{\sqrt{5}} \; ,$$

$$\mathbf{T}_{12}\mathbf{B} = \begin{pmatrix} 1/\sqrt{5} & 2/\sqrt{5} & 0 \\ -2/\sqrt{5} & 1/\sqrt{5} & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \left(\begin{array}{ccc|c} 1 & -2 & 1 & 1 \\ 2 & 0 & -3 & 8 \\ 2 & -1 & -1 & 5 \end{array}\right) \approx \left(\begin{array}{ccc|c} 2{,}236 & -0{,}894 & -2{,}236 & 7{,}603 \\ 0 & 1{,}789 & -2{,}236 & 2{,}683 \\ 2 & -1 & -1 & 5 \end{array}\right).$$

Далее аналогично получаем:

$$\mathbf{T}_{13}(\mathbf{T}_{12}\mathbf{B}) \approx \begin{pmatrix} 0{,}745 & 0 & 0{,}667 \\ 0 & 1 & 0 \\ -0{,}667 & 0 & 0{,}745 \end{pmatrix} \cdot (\mathbf{T}_{12}\mathbf{B}) \approx \left(\begin{array}{ccc|c} 3{,}000 & -1{,}333 & -2{,}333 & 9{,}000 \\ 0 & 1{,}789 & -2{,}236 & 2{,}683 \\ 0 & -0{,}149 & 0{,}745 & -1{,}342 \end{array}\right);$$

$$\mathbf{T}_{23}(\mathbf{T}_{13}\mathbf{T}_{12}\mathbf{B}) \approx \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0{,}997 & -0{,}083 \\ 0 & 0{,}083 & 0{,}997 \end{pmatrix} \cdot (\mathbf{T}_{13}(\mathbf{T}_{12}\mathbf{B})) \approx \left(\begin{array}{ccc|c} 3{,}000 & -1{,}333 & -2{,}333 & 9{,}000 \\ 0 & 1{,}795 & -2{,}290 & 2{,}785 \\ 0 & 0 & 0{,}557 & -1{,}114 \end{array}\right).$$

Сопоставляя полученной в итоге трапециевидной матрице систему

$$\begin{cases} 3{,}000x_1 - 1{,}333x_2 - 2{,}333x_3 = \; 9{,}000, \\ \qquad 1{,}795x_2 \quad - 2{,}290x_3 = \; 2{,}785, \\ \qquad\qquad\qquad\quad 0{,}557x_3 = -1{,}114 \; , \end{cases}$$

эквивалентную данной системе (2.30), отсюда последовательно находим $x_3 :\approx -2{,}000$, $x_2 :\approx -1{,}000$, $x_1 :\approx 1{,}000$.

(Заметим, что вычисления здесь проводились с некоторым запасом десятичных знаков, и округления до трех знаков после запятой проведены уже при записи результатов.)

УПРАЖНЕНИЯ

2.1. Используя схему единственного деления,

а) подсчитайте определитель $\begin{vmatrix} 1 & 1 & 1 \\ 1 & 0 & -1 \\ 1 & 2 & 1 \end{vmatrix}$;

б) решите систему $\begin{cases} x_1 + x_2 + x_3 = 6, \\ x_1 - x_3 = -2, \\ x_1 + 2x_2 + x_3 = 8; \end{cases}$

в) найдите матрицу, обратную матрице $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & -1 \\ 1 & 2 & 1 \end{pmatrix}$.

2.2. Дана система

$$\begin{cases} 0{,}1x_1 + 2x_2 - 10x_3 = 0{,}6, \\ 0{,}3x_1 + 6{,}01x_2 - 25x_3 = 1{,}852, \\ 0{,}4x_1 + 8{,}06x_2 + 10{,}001x_3 = 2{,}91201. \end{cases}$$

А) Решите систему методом Гаусса, пошагово выполняя предписания алгоритма из § 2.1.

Б) Перемножением ведущих элементов метода Гаусса найдите детерминант матрицы коэффициентов данной системы.

В) Выполните задания А), Б), имитируя работу модельного компьютера, в котором под запись мантиссы числа в режиме с плавающей запятой выделено три десятичных разряда (см. приложение).

Проделайте то же, проводя частичное упорядочивание по столбцам.

Сравните результаты В) с результатами А) и Б).

Г) Каковы невязки приближенных решений данной системы, полученных в задании В)? Произведите итерационное уточнение этих решений в той же вычислительной среде.

2.3. Используя LU-разложение, полученное в упр.1.2, найдите решение системы $\mathbf{Ax} = \mathbf{b}$, где $\mathbf{b} := (0\,; -12\,;\ 4)^T$, а также найдите матрицу $\mathbf{A}^{-1}$ двумя способами: 1) решая подсистемы $\mathbf{Ax}_i = \mathbf{e}_i$, где $\mathbf{x}_i$ и $\mathbf{e}_i$ — столбцы соответственно искомой и единичной матриц, 2) реализуя вычисления по формулам (2.12) – (2.13).

2.4. Дана система

$$\begin{cases} 16x_1 - 8x_2 - 4x_3 = -8, \\ -8x_1 + 13x_2 - 4x_3 - 3x_4 = 7, \\ -4x_1 - 4x_2 + 9x_3 = 6, \\ \quad - 3x_2 + 3x_4 = -3. \end{cases}$$

А) Решите систему методом квадратных корней.

Б) С помощью полученного в А) $\mathbf{U}^T\mathbf{U}$-разложения Холецкого найдите детерминант матрицы коэффициентов данной системы и обратную ей матрицу.

2.5. Выведите формулы для вычисления элементов матрицы $\mathbf{A}^{-1}$, обратной к симметричной матрице $\mathbf{A}$, на основе $\mathbf{U}^T\mathbf{U}$-разложения (подобные формулам (2.12) – (2.13)). Используйте их для обращения матрицы $\begin{pmatrix} 4 & 1 & -2 \\ 1 & 8 & 3 \\ -2 & 3 & 10 \end{pmatrix}$.

2.6. Примените $\mathbf{U}^T\mathbf{DU}$-разложение для решения системы

$$\begin{cases} x_1 + 2x_2 + 3x_3 = -3, \\ 2x_1 + x_2 + 4x_3 = -5, \\ 3x_1 + 4x_2 + x_3 = 5. \end{cases}$$

2.7. Запишите расчетные формулы для решения системы

$$\begin{cases} 3x_1 - x_2 = 1, \\ x_1 + 4x_2 - 2x_3 = 2, \\ 2x_2 + 6x_3 - 3x_4 = 3, \\ \dots\dots\dots\dots\dots\dots \\ 98x_{98} + 198x_{99} - 99x_{100} = 99, \\ 99x_{99} + 200x_{100} = 100 \end{cases}$$

методом прогонки. Будет ли прогонка устойчивой?

2.8. Установите, при каких $n \in \mathbb{N}$ можно гарантировать корректность и устойчивость метода прогонки для решения системы (2.20), где:

$$b_i := 1 + i \quad \text{при } i = 2,\ 3,\ \dots,\ n;$$
$$c_i := 15 + i \quad \text{при } i = 1,\ 2,\ \dots,\ n;$$
$$d_i := -i \quad \text{при } i = 1,\ 2,\ \dots,\ n-1.$$

Сравните два подхода:

а) применяя достаточное условие корректности и устойчивости;

б) по определению.

2.9. Выведите формулы левой прогонки для решения системы (2.20) (т.е. такие, при которых неизвестные x_i вычислялись бы в порядке возрастания индексов).

2.10. Решите систему упражнения 2.8 при значениях $n := 7$, $r_7 := 202$, $r_i := i^2 + 14i - 1$ (где $i := 1, \ldots, 6$) по формулам правой и левой прогонок.

2.11. Пусть в (2.20) $b_{i+1} := d_i$ при всех $i \in \{1, 2, \ldots, n-1\}$. Запишите для этого случая расчетные формулы метода квадратных корней.

2.12. Выведите формулы для получения решения системы

$$\begin{cases} \alpha_1 x_1 + \beta_1 x_2 = b_1, \\ \beta_{i-1} x_{i-1} + \alpha_i x_i + \beta_i x_{i+1} = b_i \quad (i = 2, \ldots, n-1), \\ \beta_{n-1} x_{n-1} + \alpha_n x_n = b_n \end{cases}$$

на основе $\mathbf{U}^T\mathbf{D}\mathbf{U}$-разложения трехдиагональных матриц.

2.13. Решите систему из упр. 2.2 методом вращений:

а) используя всю разрядную сетку калькулятора или компьютера;

б) работая, как и в упр. 2.2В), с тремя значащими цифрами.

Проанализируйте результаты, сравнивая их с результатами упражнений 2.2А) и 2.2В).

2.14. Методом отражений решите систему $\begin{cases} x - 2y + z = 0, \\ 2x - y - 2z = 6, \\ 2x - y - z = 9. \end{cases}$

2.15. Запишите алгоритм обращения матрицы методом отражений. Пользуясь им, найдите матрицу, обратную для матрицы

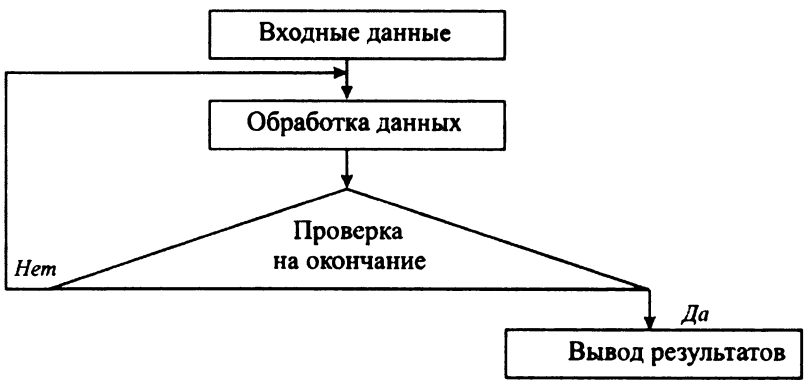
$$\mathbf{A} := \begin{pmatrix} 2 & 6 & 5 \\ 5 & 3 & -2 \\ 7 & 4 & 3 \end{pmatrix}.$$

Глава 3

ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ СЛАУ

§ 3.1. НЕКОТОРЫЕ ОБЩИЕ СВЕДЕНИЯ ОБ ИТЕРАЦИОННЫХ ПРОЦЕССАХ

Слово ***итерация*** в переводе с латинского языка означает *повторение*. ***Итерационный метод*** решения уравнений или других задач — это построение последовательности приближений к искомому значению на основе многократного выполнения одинаковых, как правило, достаточно простых действий. Примитивно итерационный процесс можно отобразить следующей блок-схемой:



Обработку данных здесь осуществляют с помощью одних и тех же операций, и вместе с проверкой на окончание она составляет полный ***итерационный шаг***.

Сразу заметим, что не всякая циклическая обработка данных может быть отнесена к итерационным методам: есть методы, итерационные по форме, но не являющиеся итерационными по сути, поскольку в них на разных шагах однотипно решаются разные подзадачи исходной задачи, а не происходит последовательное уточнение решения. Примером таких методов может служить,

например, любой разностный метод решения начальных задач для дифференциальных уравнений. Формальные признаки того, какой метод — итерационный или шаговый — «зашит» в показанной выше простейшей блок-схеме, должны быть отражены в проверке на окончание. Для итерационных методов здесь характерно наличие таких проверок, как проверка того, что на текущем итерационном шаге достигнута заданная точность (в том или ином смысле по тому или иному критерию), что не нарушается процесс сближения членов строящейся итерационной последовательности (как следствия взаимодействия ошибок округления), что не исчерпан установленный предел допустимого числа итераций (подсчитываемый на основе каких-либо теоретических оценок или задаваемый просто из каких-то рациональных соображений).

Пусть некий итерационный метод (иначе, ***итерационный процесс***) порождает в пространстве $\mathbb{R}_n$ последовательность элементов $(\mathbf{x}^{(k)})$, рассматриваемую как потенциально пригодную для того, чтобы считать ее последовательностью приближений к искомому элементу $\mathbf{x}^* \in \mathbb{R}_n$. Если имеет место сходимость $\mathbf{x}^{(k)} \to \mathbf{x}^*$ при $k \to \infty$, то в таком случае говорят, что соответствующий ***итерационный метод сходится.*** Факт сходимости итерационного метода решения некоторой задачи означает, что, остановив на k-м шаге процесс вычисления приближений, можно положить $\mathbf{x}^* \approx \mathbf{x}^{(k)}$ с абсолютной погрешностью $\| \mathbf{x}^* - \mathbf{x}^{(k)} \|$. Задание получить итерационным методом решение $\mathbf{x}^*$ с точностью ε обычно расценивается как проведение итераций до выполнения неравенства $\| \mathbf{x}^* - \mathbf{x}^{(k)} \| \le \varepsilon$ (если при этом речь не идет об относительной точности получения решения). Ясно, что для одной и той же задачи могут быть предложены разные итерационные методы, и одним из главных критериев выбора лучшего из них является то, насколько быстро убывают погрешности $\| \mathbf{x}^* - \mathbf{x}^{(k)} \|$ с ростом k (разумеется, при этом нужно учитывать и «цену» одной итерации). Характер убывания этих величин определяет ***скорость*** и ***порядок сходимости*** итерационного метода.

Определение 3.1. *Сходимость последовательности* $(\mathbf{x}^{(k)})$ *к* $\mathbf{x}^*$ *называется* ***линейной*** *(соответственно итерационный процесс называется* ***линейно сходящимся****), если существует такая постоянная* $C \in (0, 1)$ *и такой номер* k_0, *что*

$$\| \mathbf{x}^* - \mathbf{x}^{(k+1)} \| \leq C \| \mathbf{x}^* - \mathbf{x}^{(k)} \| \quad \forall k \geq k_0, \tag{3.1}$$

и ***сверхлинейной****, если существует такая положительная последовательность* $(C_k)_{k=0}^{\infty}$, *что* $C_k \underset{k\to\infty}{\to} 0$ *и*

$$\| \mathbf{x}^* - \mathbf{x}^{(k+1)} \| \leq C_k \| \mathbf{x}^* - \mathbf{x}^{(k)} \| \quad \forall k \in \mathbb{N}_0. \tag{3.2}$$

Определение 3.2. *Говорят, что последовательность* $(\mathbf{x}^{(k)})$ ***сходится*** *к* $\mathbf{x}^*$ ***по меньшей мере с p-м порядком*** *(соответственно* ***итерационный процесс имеет по меньшей мере p-й порядок****), если найдутся такие константы* $C > 0$ *и* $p \geq 1$, *что*

$$\| \mathbf{x}^* - \mathbf{x}^{(k+1)} \| \leq C \| \mathbf{x}^* - \mathbf{x}^{(k)} \|^p \tag{3.3}$$

при всех $k \in \mathbb{N}_0$, *начиная с некоторого* $k := k_0$.

Фиксируя в определении 3.2 значение $p := 1$, видим, что линейно сходящийся процесс можно называть ***процессом первого порядка***; значению $p := 2$ в (3.3) соответствует ***квадратично сходящийся процесс***, $p := 3$ означает ***кубическую сходимость***.

К линейной сходимости применяют также термин ***сходимость со скоростью геометрической прогрессии***. Объяснение ему можно найти в том, что определяющее линейную сходимость неравенство (3.1) между абсолютными погрешностями $(k+1)$-го и k-го приближений к предельной точке $\mathbf{x}^*$ означает существование последовательности положительных чисел ε_k, мажорирующих эти погрешности и связанных соотношением $\varepsilon_{k+1} = C\varepsilon_k$, т.е. являющихся членами геометрической прогрессии со знаменателем $C := \varepsilon_{k+1}/\varepsilon_k$ (= const). Отсюда следует также естественность в определении 3.1 условия $C < 1$, чтобы последо-

вательность погрешностей была убывающей, иначе и речи не может быть о сходимости (в определении 3.2 для предельного случая $p := 1$ также следует ограничить C единицей; при $p > 1$ в этом, вообще говоря, нет необходимости; проанализируйте почему).

Если требуемой в неравенстве (3.1) константы C не удается найти, но установлено неравенство (3.2) с $C_k \xrightarrow[k\to\infty]{} C \in (0,1)$, то в этом случае говорят об ***асимптотически линейной сходимости***. Аналогично можно определить ***асимптотически p-й порядок***.

Среди нескольких способов охарактеризовать скорость сходимости итерационных последовательностей наиболее четко оформились два способа: 1) опирающийся на ***q-сходимость*** (от англ. *quotient* — частное) и 2) опирающийся на ***r-сходимость*** (от англ. *root* — корень). Происхождение этих терминов можно связать соответственно с признаками Даламбера (через отношение) и Коши (через арифметический корень), применяемыми в одномерном случае для установления абсолютной сходимости числового ряда

$$x_0 + (x_1 - x_0) + (x_2 - x_1) + \ldots + (x_k - x_{k-1}) + \ldots, \qquad (3.4)$$

что равносильно установлению сходимости последовательности (x_k), так как сходимость ряда (3.4) означает существование предела его частичных сумм

$$S_1 = x_0, \quad S_2 = x_1, \quad S_3 = x_2, \quad \ldots, \quad S_{k+1} = x_k, \quad \ldots .$$

Четкие определения и различия между двумя упомянутыми типами сходимостей можно найти в [52]. Даже грубое представление о порядке сходимости того или иного метода дает возможность сравнить его с другими методами. Более точно для этого подходит знание порядка q-сходимости (что и определено выше); порядок r-сходимости говорит лишь о наличии такой последовательности положительных чисел, которая, сходясь к нулю с этим порядком, мажорирует последовательность величин $\| \mathbf{x}^* - \mathbf{x}^{(k)} \|$. Не вникая в тонкости, можно сказать, что обычно, изучая итерационный метод, устанавливают факт сходимости итерационной последовательности $(\mathbf{x}^{(k)})$ к искомому элементу $\mathbf{x}^*$ и получают апостериор-

ные и априорные оценки погрешности. О порядке же метода (в том или ином смысле, не всегда уточняя, в каком) судят или на основе неравенства типа (3.3), или по априорной оценке погрешности вида

$$\| \mathbf{x}^* - \mathbf{x}^{(k)} \| \leq C\nu^{p^k}, \tag{3.5}$$

где $C > 0$ и $\nu \in (0, 1)$ — некоторые константы, а $p \geq 1$ — порядок метода, или по неравенству вида

$$\| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \| \leq C \| \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} \|^p, \tag{3.6}$$

показывающему скорость сближения членов итерационной последовательности и являющемуся ключевым для установления сходимости и получения оценок погрешностей. Чаще всего, разные способы приводят к одному и тому же значению p, хотя это и не гарантировано.

Имеются и другие представления о том, что понимать под термином *скорость сходимости*, как правило, в контексте построения итерационных методов решения той или иной задачи. В частности, в следующем параграфе вводится важная характеристика линейно сходящихся методов — ***средняя скорость сходимости***.

Коснемся еще одного аспекта понятия сходимости итерационного метода. В приведенных выше определениях отождествлялись сходимость итерационного процесса и сходимость итерационной последовательности, порождаемой данным процессом; при этом негласно считалось, что последовательность $(\mathbf{x}^{(k)})$ уже как бы фиксирована заданием начальной точки $\mathbf{x}^{(0)}$. Варьирование $\mathbf{x}^{(0)}$ в границах некоторой области порождает множество итерационных последовательностей. Итерационные методы, дающие в пределе решение данной задачи при любом начальном приближении $\mathbf{x}^{(0)}$, называют ***глобально сходящимися***. Если же сходимость итерационной последовательности $(\mathbf{x}^{(k)})$ к искомому элементу $\mathbf{x}^*$ имеет место лишь при задании $\mathbf{x}^{(0)}$ из некоторой, вообще говоря, достаточно малой окрестности $\mathbf{x}^*$, то соответствующий итерационный метод называют ***локально сходящимся***.

§ 3.2. МЕТОД ПРОСТЫХ ИТЕРАЦИЙ

Система линейных алгебраических уравнений, в векторно-матричной записи имеющая стандартный вид

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{3.7}$$

где $\mathbf{A} := (a_{ij})_{i,j=1}^{n}$ — $n \times n$-матрица, а $\mathbf{x} := (x_1; x_2; \ldots; x_n)^T$ и $\mathbf{b} := (b_1; b_2; \ldots; b_n)^T$ — n-мерные векторы-столбцы, тем или иным способом (таких способов существует бесконечное множество; некоторые из них будут рассмотрены далее) может быть преобразована к эквивалентной ей системе вида

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{c}, \tag{3.8}$$

где $\mathbf{x}$ — тот же вектор неизвестных, а $\mathbf{B}$ и $\mathbf{c}$ — некоторые новые матрица и вектор соответственно. Систему вида (3.8) можно трактовать как задачу о неподвижной точке линейного отображения $\mathbf{B}$ в пространстве $\mathbb{R}_n$ и определить последовательность приближений $\mathbf{x}^{(k)}$ к неподвижной точке $\mathbf{x}^*$ рекуррентным равенством

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{c}, \quad k = 0, 1, 2, \ldots . \tag{3.9}$$

Итерационный процесс (3.9), начинающийся с некоторого вектора $\mathbf{x}^{(0)} := \left(x_1^{(0)}; x_2^{(0)}; \ldots; x_n^{(0)}\right)^T$, будем далее называть ***методом простых итераций*** (сокращенно **МПИ**).

Изучим комплекс вопросов о сходимости этого процесса. А именно:

1. Какие нужно предъявить требования к $\mathbf{B}$, $\mathbf{c}$ и $\mathbf{x}^{(0)}$, чтобы последовательность $(\mathbf{x}^{(k)})$ при $k \to \infty$ имела пределом $\mathbf{x}^*$ — неподвижную точку задачи (3.8) (и значит, решение эквивалентной (3.8) исходной системы (3.7))?

2. С какой скоростью сходится этот процесс, т.е. каков закон убывания абсолютных погрешностей получаемых по формуле (3.9) приближений?

3. Сколько нужно сделать итераций по формуле (3.9), чтобы при заданном начальном приближении $\mathbf{x}^{(0)}$ найти решение задачи (3.8) с заданной точностью?

Ответы на подобные вопросы теории итерационных методов в $\mathbb{R}_n$ часто опираются на следующие два утверждения о сходимости матричной геометрической прогрессии, являющиеся частными случаями соответствующих утверждений о сходимости степенных рядов в пространствах линейных операторов и называемых иногда соответственно ***леммой Неймана*** (см., например, [52]) и ***теоремой Банаха*** ([35] и др.). Во втором из них, а также всюду далее под нормой матрицы следует понимать мультипликативную норму, такую что $\|\mathbf{E}\| = 1$ ($\mathbf{E}$ — единичная матрица).

Лемма 3.1. *Условие, что все собственные числа матрицы* $\mathbf{B}$ *по модулю меньше единицы**, *является необходимым и достаточным для того, чтобы*:

1) $\mathbf{B}^k \to \mathbf{0}$ *при* $k \to \infty$ $(k \in \mathbb{N})$;

2) *матрица* $\mathbf{E} - \mathbf{B}$ *имела обратную матрицу и было справедливым представление*

$$(\mathbf{E}-\mathbf{B})^{-1} = \mathbf{E} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^k + \ldots .$$

Лемма 3.2. *Если* $\|\mathbf{B}\| \le q < 1$, *то матрица* $\mathbf{E} - \mathbf{B}$ *имеет обратную матрицу* $(\mathbf{E}-\mathbf{B})^{-1} = \sum\limits_{k=0}^{\infty} \mathbf{B}^k$ *и при этом справедливо неравенство*

$$\left\|(\mathbf{E}-\mathbf{B})^{-1}\right\| \le \frac{1}{1-q}.$$

Доказательство леммы 3.2. Рассмотрим матричный ряд

$$\mathbf{E} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^k + \ldots . \tag{3.10}$$

В силу условия леммы и вытекающего из мультипликативного

---

* В иной формулировке: «... *спектральный радиус матрицы* $\mathbf{B}$ *меньше единицы*...» или, короче, $\rho(\mathbf{B}) < 1$.

свойства нормы неравенства $\|\mathbf{B}^k\| \le \|\mathbf{B}\|^k$ этот ряд можно промажорировать сходящимся числовым рядом:

$$\|\mathbf{E}\| + \|\mathbf{B}\| + \|\mathbf{B}^2\| + \ldots + \|\mathbf{B}^k\| + \ldots \le 1 + q + q^2 + \ldots + q^k + \ldots = \frac{1}{1-q}.$$

Следовательно, ряд (3.10) сходится, т.е. существует матрица

$$\mathbf{V} := \mathbf{E} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^k + \ldots$$

такая, что $\|\mathbf{V}\| \le \frac{1}{1-q}$. Так как в силу перестановочности произведения степеней матриц имеет место равенство

$$(\mathbf{E}-\mathbf{B})\mathbf{V} = (\mathbf{E}-\mathbf{B})\left(\mathbf{E}+\mathbf{B}+\mathbf{B}^2+\ldots+\mathbf{B}^k+\ldots\right) =$$

$$= \mathbf{E}+\mathbf{B}+\mathbf{B}^2+\ldots+\mathbf{B}^k+\ldots-\mathbf{B}-\mathbf{B}^2-\mathbf{B}^3-\ldots-\mathbf{B}^{k+1}-\ldots = \mathbf{E}$$

и аналогичное равенство $\mathbf{V}(\mathbf{E}-\mathbf{B}) = \mathbf{E}$, то $\mathbf{V} = (\mathbf{E}-\mathbf{B})^{-1}$. Лемма доказана.

Доказательство леммы 3.1 более сложно. Его можно найти во многих учебных пособиях по вычислительной математике и по функциональному анализу (см., например, [7, 25, 35, 42, 52]).

Теорема 3.1. *Необходимым и достаточным условием сходимости метода простых итераций* (3.9) *при любом начальном векторе* $\mathbf{x}^{(0)}$ *к решению* $\mathbf{x}^*$ *системы* (3.8) *является требование, чтобы все собственные числа* $\lambda_{\mathbf{B}}$ *матрицы* $\mathbf{B}$ *были по модулю меньше* 1.

Доказательство. *Достаточность.* Пусть $\max|\lambda_{\mathbf{B}}| < 1$, тогда по лемме 3.1 общий член $\mathbf{B}^k$ ряда (3.10) стремится к нуль-матрице и существует матрица $(\mathbf{E}-\mathbf{B})^{-1}$, являющаяся пределом частичных сумм $\left(\mathbf{E}+\mathbf{B}+\mathbf{B}^2+\ldots+\mathbf{B}^k\right)$ при $k \to \infty$. Применяя рекурсию в равенстве (3.9), определяющем МПИ, получим:

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{c} = \mathbf{B}^2\mathbf{x}^{(k-1)} + (\mathbf{B}+\mathbf{E})\mathbf{c} = \ldots = \\ = \mathbf{B}^{k+1}\mathbf{x}^{(0)} + \left(\mathbf{E}+\mathbf{B}+\mathbf{B}^2+\ldots+\mathbf{B}^k\right)\mathbf{c}. \tag{3.11}$$

В силу сказанного выше предел последнего выражения существует при любом фиксированном $\mathbf{x}^{(0)}$ и равен $(\mathbf{E}-\mathbf{B})^{-1}\mathbf{c}$. Следовательно, итерационный процесс (3.9) сходится, т.е. существует

$$\mathbf{x}^* := \lim_{k\to\infty} \mathbf{x}^{(k)} = (\mathbf{E}-\mathbf{B})^{-1}\mathbf{c}.$$

Подставляя $\mathbf{x}^*$ в уравнение (3.8), преобразованное к виду $(\mathbf{E}-\mathbf{B})\mathbf{x} = \mathbf{c}$, имеем равенство

$$(\mathbf{E}-\mathbf{B})(\mathbf{E}-\mathbf{B})^{-1}\mathbf{c} = \mathbf{c},$$

означающее, что вектор $\mathbf{x}^* = \lim\limits_{k\to\infty} \mathbf{x}^{(k)}$ удовлетворяет системе (3.8). (Заметим, что этот вектор $\mathbf{x}^*$ — единственное решение уравнения (3.8). Действительно, допустив, что наряду с $\mathbf{x}^*$ таким, что $\mathbf{x}^* = \mathbf{B}\mathbf{x}^* + \mathbf{c}$, имеется вектор $\mathbf{x}^{**}$, удовлетворяющий такому же равенству $\mathbf{x}^{**} = \mathbf{B}\mathbf{x}^{**} + \mathbf{c}$, получаем $\mathbf{x}^* - \mathbf{x}^{**} = \mathbf{B}(\mathbf{x}^* - \mathbf{x}^{**})$. Последнее означает, что число $\lambda := 1$ по определению является собственным значением матрицы $\mathbf{B}$, что противоречит условию).

Необходимость. Как видно из представления общего члена итерационной последовательности $(\mathbf{x}^{(k)})$ в форме (3.11), существование $\lim\limits_{k\to\infty} \mathbf{x}^{(k+1)}$ при любых векторах $\mathbf{x}^{(0)}$ и $\mathbf{c}$ (в том числе и нулевых, что гарантирует существование предела каждого слагаемого в правой части (3.11)) влечет сходимость матриц $\mathbf{B}^{k+1}$ к нуль-матрице и сходимость ряда $\sum\limits_{k=0}^{\infty} \mathbf{B}^k$ к матрице $(\mathbf{E}-\mathbf{B})^{-1}$. Согласно лемме 3.1, это равносильно выполнению условия $|\lambda_{\mathbf{B}}| < 1$ для каждого собственного числа матрицы $\mathbf{B}$. Теорема доказана.

Теорема 3.2. *Пусть* $\|\mathbf{B}\| \le q < 1$. *Тогда при любом начальном векторе* $\mathbf{x}^{(0)}$ *МПИ* (3.9) *сходится к единственному решению* $\mathbf{x}^*$ *задачи* (3.8) *и при всех* $k \in \mathbb{N}$ *справедливы оценки погрешности*:

1) $\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \dfrac{q}{1-q}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\|$ (***апостериорная***);

2) $\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \dfrac{q^k}{1-q}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|$ (***априорная***).*

(Одно и то же обозначение $\|\cdot\|$ здесь используется для матричных и векторных норм, согласованных между собой, т.е. таких, что $\|\mathbf{Ax}\| \le \|\mathbf{A}\| \cdot \|\mathbf{x}\|$.)

Доказательство. Вычитая из равенства (3.9) равенство $\mathbf{x}^{(k)} = \mathbf{Bx}^{(k-1)} + \mathbf{c}$, имеем $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} = \mathbf{B}\left(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right)$. Переходя в последнем к нормам, получаем неравенство

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le q\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\|, \tag{3.12}$$

из которого видно, в силу условия $q < 1$, что элементы итерационной последовательности $(\mathbf{x}^{(k)})$ сближаются с ростом номера k. С помощью (3.12) оценим разность между $(k+m)$-м и k-м членами этой последовательности при некотором $m \in \mathbb{N}$:

$$\left\|\mathbf{x}^{(k+m)} - \mathbf{x}^{(k)}\right\| = \left\|\mathbf{x}^{(k+m)} - \mathbf{x}^{(k+m-1)} + \mathbf{x}^{(k+m-1)} - \mathbf{x}^{(k+m-2)} + \right.$$

$$\left. + \mathbf{x}^{(k+m-2)} - \ldots - \mathbf{x}^{(k+1)} + \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le$$

---

* Латинские слова *a priori* и *a posteriori* означают соответственно «до опыта» и «из опыта», т.е. априорной оценкой можно воспользоваться до начала счета, а апостериорной — лишь после проведения k-й итерации.

$$\leq \left\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\right\|+\left\|\mathbf{x}^{(k+2)}-\mathbf{x}^{(k+1)}\right\|+\ldots+\left\|\mathbf{x}^{(k+m)}-\mathbf{x}^{(k+m-1)}\right\| \leq$$

$$\leq q\left\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\right\|+q^2\left\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\right\|+\ldots+q^m\left\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\right\|=$$

$$=\frac{q\left(1-q^m\right)}{1-q}\left\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\right\| \leq \frac{q^k}{1-q}\left(1-q^m\right)\left\|\mathbf{x}^{(1)}-\mathbf{x}^{(0)}\right\|.$$

Рассматривая итоговое неравенство при $k \to \infty$ и фиксированном m, видим, что $(\mathbf{x}^{(k)})$ является фундаментальной последовательностью и, в силу полноты пространства $\mathbb{R}_n$, имеет предел. Обозначим его $\mathbf{x}^*$. Переходя к пределу в равенстве (3.9), получаем равенство $\mathbf{x}^* = \mathbf{B}\mathbf{x}^* + \mathbf{c}$, означающее, что $\mathbf{x}^* = \lim\limits_{k\to\infty} \mathbf{x}^{(k)}$ — решение уравнения (3.8). При этом $\mathbf{x}^*$ — единственное решение (3.8), так как, предположив существование другого решения $\mathbf{x}^{**} \neq \mathbf{x}^*$ и нормируя равенство $\mathbf{x}^* - \mathbf{x}^{**} = \mathbf{B}\left(\mathbf{x}^* - \mathbf{x}^{**}\right)$, приходим к противоречащему условию теоремы неравенству $\|\mathbf{B}\| \geq 1$.

Справедливость утверждаемых в теореме оценок погрешности видна из неравенств

$$\left\|\mathbf{x}^{(k+m)}-\mathbf{x}^{(k)}\right\| \leq \frac{q\left(1-q^m\right)}{1-q}\left\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\right\| \leq \frac{q^k}{1-q}\left(1-q^m\right)\left\|\mathbf{x}^{(1)}-\mathbf{x}^{(0)}\right\|,$$

если в них теперь зафиксировать k и перейти к пределу при $m \to \infty$. Теорема доказана.

Замечание 3.1. Последние неравенства говорят еще о том, что априорная оценка, как правило, грубее апостериорной.

Замечание 3.2. Теорема 3.2 могла быть доказана на основе леммы 3.2 и теоремы 3.1. В частности, сходимость последовательности $\left(\mathbf{x}^{(k)}\right)$ к решению $\mathbf{x}^*$ системы (3.8) сразу следует из теоремы 3.1, в силу соотношений $|\lambda_{\mathbf{B}}| \leq \|\mathbf{B}\| < 1$.

Из леммы 3.2 также легко вывести другую априорную оценку погрешности k-го приближения: вычитая из равенства

$$\mathbf{x}^* = (\mathbf{E}-\mathbf{B})^{-1}\mathbf{c} = \left(\mathbf{E}+\mathbf{B}+\ldots+\mathbf{B}^k+\ldots\right)\mathbf{c}$$

равенство

$$\mathbf{x}^{(k)} = \mathbf{B}^k\mathbf{x}^{(0)} + \left(\mathbf{E}+\mathbf{B}+\ldots+\mathbf{B}^{k-1}\right)\mathbf{c}$$

(см. (3.11)) и далее переходя к нормам, имеем:

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| = \left\|\left(\mathbf{B}^k + \mathbf{B}^{k+1} + \ldots\right)\mathbf{c} - \mathbf{B}^k\mathbf{x}^{(0)}\right\| \le \left\|\mathbf{B}^k\right\| \cdot \left\|(\mathbf{E}-\mathbf{B})^{-1}\mathbf{c} - \mathbf{x}^{(0)}\right\|,$$

т.е.

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le q^k\left(\left\|\mathbf{x}^{(0)}\right\| + \frac{\|\mathbf{c}\|}{1-q}\right). \qquad (3.13)$$

Замечание 3.3. Априорная оценка позволяет подсчитывать заранее число итераций k, достаточное для получения решения $\mathbf{x}^*$ с заданной точностью ε (в смысле допустимого уровня абсолютных погрешностей) при выбранном начальном векторе $\mathbf{x}^{(0)}$. Для этого нужно найти наименьшее целое решение неравенства

$$\frac{q^k}{1-q}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\| \le \varepsilon$$

относительно переменной k (или неравенства $q^k\left(\left\|\mathbf{x}^{(0)}\right\| + \frac{\|\mathbf{c}\|}{1-q}\right) \le \varepsilon$ в соответствии с результатом (3.13) предыдущего замечания). Апостериорной же оценкой удобно пользоваться непосредственно в процессе вычислений и останавливать этот процесс, как только выполнится неравенство

$$\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \frac{1-q}{q}\varepsilon. \qquad (3.14)$$

Отметим, что упрощенное по сравнению с (3.14) неравенство $\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \varepsilon$ будет гарантией выполнения неравенства $\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \varepsilon$ только в том случае, когда $q \le \frac{1}{2}$.

Замечание 3.4. По поводу выбора начального приближения.

Как установлено выше, сходимость МПИ (3.9) при условии $|\lambda_{\mathbf{B}}| < 1$ является глобальной, т.е. гарантируется при *любом* начальном векторе $\mathbf{x}^{(0)}$. Очевидно, итераций потребуется тем меньше, чем ближе $\mathbf{x}^{(0)}$ к $\mathbf{x}^*$. Если нет никакой дополнительной информации о решении задачи (3.8) (например, может

быть известным решение близкой задачи или грубое решение данной задачи), за $\mathbf{x}^{(0)}$ обычно принимают вектор $\mathbf{c}$ свободных членов системы (3.8). Мотивация этого может быть такой: матрица $\mathbf{B}$ «мала», значит вектор $\mathbf{Bx}$ «мал», следовательно, и вектор $\mathbf{x}^*$ не должен сильно отличаться от вектора $\mathbf{c}$. При выборе $\mathbf{x}^{(0)} := \mathbf{c}$ фигурирующая в теореме 3.2 априорная оценка принимает вид

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{\|\mathbf{c}\|}{1-q} \cdot q^{k+1} \quad \forall\, k \in \mathbb{N}. \tag{3.15}$$

Оценка (3.13) в этом случае приводит к несколько худшему, чем (3.15), результату: $\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \|\mathbf{c}\| \cdot \frac{2-q}{1-q} q^k$.

Определение 3.3. *Средней скоростью сходимости метода простых итераций* (3.9) *называется вещественное число* $r(\mathbf{B}) := -\lg \rho(\mathbf{B})$, *где* $\rho(\mathbf{B}) := \max\left|\lambda_{\mathbf{B}}\right|$ — *спектральный радиус матрицы* $\mathbf{B}$.

Смысл введенной этим определением характеристики линейно сходящегося итерационного метода следующий.

Вычитая определяющее метод равенство (3.9) из равенства $\mathbf{x}^* = \mathbf{Bx}^* + \mathbf{c}$, имеем связь между ошибками k-го и $(k+1)$-го приближений

$$\mathbf{x}^* - \mathbf{x}^{(k+1)} = \mathbf{B}\left(\mathbf{x}^* - \mathbf{x}^{(k)}\right).$$

Отсюда, переходя нормированием к связи между абсолютными погрешностями

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\right\| \le \|\mathbf{B}\| \cdot \left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\|, \tag{3.16}$$

видим, что в соответствии с определением 3.1 МПИ (3.9) — линейно сходящийся (при $\|\mathbf{B}\| < 1$) метод. Логарифмируя (3.16), получаем неравенство

$$\lg\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| - \lg\left\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\right\| \ge -\lg\|\mathbf{B}\|. \tag{3.17}$$

Предположим, что $\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \approx 10^{-m}$, а $\left\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\right\| \approx 10^{-l}$.

Тогда $\lg\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \approx -m$, $\lg\left\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\right\| \approx -l$, т.е. левую часть неравенства (3.17) можно интерпретировать как число $l-m$ прибавления верных десятичных знаков за одну итерацию МПИ (3.9). Согласно (3.17) это число не меньше, чем $-\lg\|\mathbf{B}\|$. Более точно его характеризует величина $r(\mathbf{B})$ $(\geq -\lg\|\mathbf{B}\|$, в силу известного свойства $\rho(\mathbf{B}) \leq \|\mathbf{B}\|$ для любых норм); при этом неравенство типа (3.17) с $r(\mathbf{B})$ вместо $\|\mathbf{B}\|$ рассматривается уже в специальных нормах, строящихся в соответствии со свойством

$$\forall \varepsilon > 0 \quad \exists \|\cdot\|_\varepsilon : \ \|\mathbf{B}\|_\varepsilon \leq \rho(\mathbf{B}) + \varepsilon .$$

Ясно, что при сравнении методов вида (3.9) в условиях теоремы 3.2 можно обходиться и без введенного выше показателя скорости сходимости: чем меньше норма *матрицы итерирования* $\mathbf{B}$, тем быстрее сходится метод. В условиях же теоремы 3.1, вкупе с леммой 3.1, утверждающей, что лишь с какого-то момента начнет выполняться неравенство $\|\mathbf{B}^k\| < 1$, обеспечивающее сходимость, оценивание качества метода по этому асимптотическому в данном случае показателю становится весьма актуальным.

Замечание 3.5. В соответствующей литературе можно встретить как аналогичное определению 3.3 определение скорости сходимости через натуральный логарифм (вместо десятичного) [29], так и несколько отличающиеся от введенного термины (*асимптотическая скорость сходимости* [29] или просто *скорость сходимости* [26]).

Пример 3.1. Для системы

$$\begin{cases} 1{,}1x_1 - 0{,}2x_2 + 0{,}1x_3 = 1{,}6, \\ 0{,}1x_1 - 1{,}2x_2 - 0{,}2x_3 = 2{,}3, \\ 0{,}2x_1 - 0{,}1x_2 + 1{,}1x_3 = 1{,}5 \end{cases}$$

записать какой-нибудь сходящийся процесс простых итераций. За сколько шагов этого процесса, начатого с нуль-вектора, можно гарантированно достичь точности $\varepsilon = 0{,}001$ по норме-максимум? Найти третье приближение, оценить его абсолютную погрешность и сравнить ее с истинной погрешностью, зная точное решение системы $\mathbf{x}^* := (1; -2; 1)^T$.

Учитывая очевидную близость матрицы данной системы к единичной матрице, вычленим единицы из диагональных элементов, в результате чего система преобразуется к виду

$$\begin{cases} x_1 = -0,1x_1 + 0,2x_2 - 0,1x_3 + 1,6, \\ x_2 = \ 0,1x_1 - 0,2x_2 - 0,2x_3 - 2,3, \\ x_3 = -0,2x_1 + 0,1x_2 - 0,1x_3 + 1,5. \end{cases}$$

Эта система равносильна исходной и имеет форму уравнения (3.8), в котором можно считать

$$\mathbf{B} := \begin{pmatrix} -0,1 & 0,2 & -0,1 \\ 0,1 & -0,2 & -0,2 \\ -0,2 & 0,1 & -0,1 \end{pmatrix}, \quad \mathbf{c} := \begin{pmatrix} 1,6 \\ -2,3 \\ 1,5 \end{pmatrix}.$$

Так как $\|\mathbf{B}\|_\infty = 0,5 < 1$, можно воспользоваться теоремой 3.2, полагая в ней $q := 0,5$. Согласно этой теореме метод простых итераций

$$\begin{cases} x_1^{(k+1)} = -0,1x_1^{(k)} + 0,2x_2^{(k)} - 0,1x_3^{(k)} + 1,6, \\ x_2^{(k+1)} = \ 0,1x_1^{(k)} - 0,2x_2^{(k)} - 0,2x_3^{(k)} - 2,3, \\ x_3^{(k+1)} = -0,2x_1^{(k)} + 0,1x_2^{(k)} - 0,1x_3^{(k)} + 1,5 \end{cases}$$

(где $k = 0, 1, 2, \ldots$) определяет сходящуюся к решению $\mathbf{x}^*$ последовательность векторов $\mathbf{x}^{(k)} := \left(x_1^{(k)}; x_2^{(k)}; x_3^{(k)}\right)^T$, априорная оценка погрешностей которых есть

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\|_\infty \le \frac{0,5^k}{1-0,5}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|_\infty.$$

При заданном векторе $\mathbf{x}^{(0)} := \mathbf{0}$ первым приближением $\mathbf{x}^{(1)}$, очевидно, служит вектор $\mathbf{c}$ свободных членов и, значит, $\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|_\infty = \|\mathbf{c}\|_\infty = 2,3$. Следовательно, требуемое число итерационных шагов, *достаточное* для достижения точности 0,001, может быть найдено как первое из последовательности чисел $k \in \mathbf{N}$, удовлетворяющих неравенству $0,5^{k-1} \cdot 2,3 \le 0,001$. Таковым является значение $k = 13$.

Вычислим приближения $\mathbf{x}^{(2)}$ и $\mathbf{x}^{(3)}$:

$$\begin{cases} x_1^{(2)} = -0,1 \cdot 1,6 + 0,2 \cdot (-2,3) - 0,1 \cdot 1,5 + 1,6 = 0,83, \\ x_2^{(2)} = 0,1 \cdot 1,6 - 0,2 \cdot (-2,3) - 0,2 \cdot 1,5 - 2,3 = -1,98, \\ x_3^{(2)} = -0,2 \cdot 1,6 + 0,1 \cdot (-2,3) - 0,1 \cdot 1,5 + 1,5 = 0,8; \end{cases}$$

$$\begin{cases} x_1^{(3)} = -0{,}1 \cdot 0{,}83 + 0{,}2 \cdot (-1{,}98) - 0{,}1 \cdot 0{,}8 + 1{,}6 = 1{,}041, \\ x_2^{(3)} = 0{,}1 \cdot 0{,}83 - 0{,}2 \cdot (-1{,}98) - 0{,}2 \cdot 0{,}8 - 2{,}3 = -1{,}981, \\ x_3^{(3)} = -0{,}2 \cdot 0{,}83 + 0{,}1 \cdot (-1{,}98) - 0{,}1 \cdot 0{,}8 + 1{,}5 = 1{,}056. \end{cases}$$

Априорная оценка погрешности третьего приближения дает

$$\left\| \mathbf{x}^* - \mathbf{x}^{(3)} \right\|_\infty \le \frac{0{,}5^3}{1 - 0{,}5} \left\| \mathbf{x}^{(1)} - \mathbf{x}^{(0)} \right\|_\infty = 0{,}25 \cdot 2{,}3 = 0{,}575,$$

в то время как истинная ошибка составляет величину

$$\left\| \mathbf{x}^* - \mathbf{x}^{(3)} \right\|_\infty = \left\| \begin{pmatrix} -0{,}041 \\ -0{,}019 \\ -0{,}044 \end{pmatrix} \right\|_\infty = 0{,}044,$$

что на порядок лучше прогнозируемой ошибки. Это говорит о том, что найденное выше априорное число итерационных шагов заведомо больше необходимого (оценка есть оценка!). Если воспользоваться апостериорной оценкой погрешности, то для того же приближения $\mathbf{x}^{(3)}$ получим

$$\left\| \mathbf{x}^* - \mathbf{x}^{(3)} \right\|_\infty \le \frac{0{,}5}{1 - 0{,}5} \left\| \mathbf{x}^{(3)} - \mathbf{x}^{(2)} \right\|_\infty = \left\| \begin{pmatrix} 0{,}211 \\ -0{,}001 \\ 0{,}256 \end{pmatrix} \right\|_\infty = 0{,}256.$$

Имеем несколько лучший результат, как и следовало ожидать в соответствии с замечанием 3.1. Нетрудно понять, что сравнительная с априорными точность апостериорных оценок заметно увеличивается с увеличением номера итерации.

Средняя скорость сходимости построенного здесь МПИ характеризуется величиной $\eta(\mathbf{B}) \approx -\lg 0{,}290 \approx 0{,}538$.

§ 3.3. МЕТОДЫ ЯКОБИ, ЗЕЙДЕЛЯ И ПВР (SOR)

Вернемся к рассмотрению СЛАУ в виде (3.7). После выяснения условия, которому должна удовлетворять матрица коэффициентов приведенной системы (3.8) для сходимости МПИ (3.9), следует осуществить приведение системы (3.7) к виду (3.8) так, чтобы это условие выполнялось. Рассмотрим один из способов такого приведения, достаточно эффективный в определенных случаях.

Представим матрицу $\mathbf{A}$ системы (3.7) в виде

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{R},$$

где $\mathbf{D}$ — диагональная, а $\mathbf{L}$ и $\mathbf{R}$ — соответственно левая и правая

строго треугольные (т.е. с нулевой диагональю) матрицы. Тогда система (3.7) может быть записана в виде

$$\mathbf{L}\mathbf{x}+\mathbf{D}\mathbf{x}+\mathbf{R}\mathbf{x}=\mathbf{b}, \tag{3.18}$$

и если на диагонали исходной матрицы нет нулей, то эквивалентной (3.7) задачей вида (3.8) будет

$$\mathbf{x}=-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})\mathbf{x}+\mathbf{D}^{-1}\mathbf{b}, \tag{3.19}$$

т.е. в равенствах (3.8) и (3.9) следует положить

$$\mathbf{B}:=-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R}), \quad \mathbf{c}:=\mathbf{D}^{-1}\mathbf{b}.$$

Основанный на таком приведении системы (3.7) к виду (3.8) метод простых итераций (3.9) называют ***методом Якоби****. В векторно-матричных обозначениях он определяется формулой

$$\mathbf{x}^{(k+1)}=-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})\mathbf{x}^{(k)}+\mathbf{D}^{-1}\mathbf{b}, \quad k=0,1,2,\ldots. \tag{3.20}$$

Чтобы записать метод Якоби (3.20) решения системы (3.7) в развернутом виде, достаточно заметить, что обратной матрицей к матрице $\mathbf{D}=\left(a_{ii}\right)_{i=1}^{n}$ служит диагональная матрица $\mathbf{D}^{-1}$ с элементами диагонали $d_{ii}=1/a_{ii}$. Поэтому представление (3.19) системы (3.7), записанной в виде (3.18), равнозначно выражению «диагональных неизвестных» через остальные:

$$\begin{cases} x_1=-\left(a_{12}x_2+a_{13}x_3+\ldots+a_{1n}x_n-b_1\right)/a_{11}, \\ x_2=-\left(a_{21}x_1+a_{23}x_3+\ldots+a_{2n}x_n-b_2\right)/a_{22}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n=-\left(a_{n1}x_1+a_{n2}x_2+\ldots+a_{n,n-1}x_{n-1}-b_n\right)/a_{nn}. \end{cases} \tag{3.21}$$

Теперь для записи итерационного процесса (3.20) осталось в равенствах системы (3.21) только «навесить» индексы, соответствующие номерам приближений (т.е. $k=0,1,2,\ldots$):

---

* Яко́би Карл Густав Якоб (1804–1851) — немецкий математик.

$$\begin{cases} x_1^{(k+1)} = -\left(a_{12}x_2^{(k)} + a_{13}x_3^{(k)} + \ldots + a_{1n}x_n^{(k)} - b_1\right)\Big/a_{11}, \\ x_2^{(k+1)} = -\left(a_{21}x_1^{(k)} + a_{23}x_3^{(k)} + \ldots + a_{2n}x_n^{(k)} - b_2\right)\Big/a_{22}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(k+1)} = -\left(a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} + \ldots + a_{n,n-1}x_{n-1}^{(k)} - b_n\right)\Big/a_{nn}. \end{cases} \tag{3.20a}$$

Установим простой достаточный признак сходимости метода Якоби к решению системы (3.7).

Теорема 3.3. *В случае диагонального преобладания в матрице* **A** *системы* (3.7) *метод Якоби* (3.20) *сходится.*

Доказательство. ***Диагональное преобладание*** в матрице **A** означает, что

$$\left|a_{ii}\right| > \sum_{\substack{j=1 \\ (j\neq i)}}^{n} \left|a_{ij}\right| \quad \forall\, i \in \{1, \ldots, n\}.$$

Следовательно, в *любой* строке ***матрицы итерирования*** (иначе, ***матрицы перехода***)

$$\mathbf{B} := \begin{pmatrix} 0 & -\dfrac{a_{12}}{a_{11}} & -\dfrac{a_{13}}{a_{11}} & \ldots & -\dfrac{a_{1,n-1}}{a_{11}} & -\dfrac{a_{1n}}{a_{11}} \\ -\dfrac{a_{21}}{a_{22}} & 0 & -\dfrac{a_{23}}{a_{22}} & \ldots & -\dfrac{a_{2,n-1}}{a_{22}} & -\dfrac{a_{2n}}{a_{22}} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ -\dfrac{a_{n1}}{a_{nn}} & -\dfrac{a_{n2}}{a_{nn}} & -\dfrac{a_{n3}}{a_{nn}} & \ldots & -\dfrac{a_{n,n-1}}{a_{nn}} & 0 \end{pmatrix} \tag{3.22}$$

метода Якоби сумма модулей элементов меньше единицы. Значит, по крайней мере, одна из норм матрицы **B** (именно, $\|\mathbf{B}\|_\infty$ согласованная с векторной нормой-максимум) меньше единицы. Таким образом, существуют нормы, в которых к методу Якоби, рассматриваемому как метод простых итераций, применима теорема 3.2,

т.е. метод Якоби сходится. Так как сходимость по одной норме в пространстве $\mathbb{R}_n$ означает сходимость по любой другой, тем самым теорема 3.3 доказана.

Замечание 3.6. Обратим внимание на то, что к методу Якоби при условии диагонального преобладания в матрице $\mathbf{A}$ относится полностью заключение теоремы 3.2, а также предыдущие замечания; нужно лишь учесть в них, что матрица $\mathbf{B}$ определяется с помощью (3.22), а вектор $\mathbf{c}$ — равенством

$$\mathbf{c} := \left(\frac{b_1}{a_{11}}; \frac{b_2}{a_{22}}; \ldots; \frac{b_n}{a_{nn}} \right)^T .$$

При этом матрица $\mathbf{D}$ в представлении системы (3.18) заведомо обратима.

Пример 3.2. Решим поставленную в примере 3.1 задачу методом Якоби.

В соответствии с формулами (3.20а) МПИ в форме метода Якоби можно записать так:

$$\begin{cases} x_1^{(k+1)} = \frac{1}{11}\left(16 + 2x_2^{(k)} - x_3^{(k)}\right), \\ x_2^{(k+1)} = \frac{1}{12}\left(-23 + x_1^{(k)} - 2x_3^{(k)}\right), \\ x_3^{(k+1)} = \frac{1}{11}\left(15 - 2x_1^{(k)} + x_2^{(k)}\right), \qquad k = 0, 1, 2, \ldots . \end{cases}$$

Матрица перехода $\mathbf{B}$ в этом процессе имеет следующие характеристики:

$$\|\mathbf{B}\|_\infty = \left\| \begin{pmatrix} 0 & 2/11 & -1/11 \\ 1/12 & 0 & -1/6 \\ -2/11 & 1/11 & 0 \end{pmatrix} \right\|_\infty = \frac{3}{11} \approx 0{,}273 \ , \ \rho(\mathbf{B}) \approx 0{,}201 .$$

Видим, что сходимость последовательности векторов $\mathbf{x}^{(k)} := \left(x_1^{(k)}; x_2^{(k)}; x_3^{(k)}\right)^T$ к точному решению $\mathbf{x}^* := (1; -2; 1)^T$ гарантирована (со средней скоростью сходимости $\eta(\mathbf{B}) \approx -\lg 0{,}201 := \approx 0{,}697$).

Начиная процесс итераций, как и в примере 3.1, с вектора $\mathbf{x}^{(0)} := \mathbf{0}$, последовательно получаем:

$$\mathbf{x}^{(1)} = \begin{pmatrix} 16/11 \\ -23/12 \\ 15/11 \end{pmatrix} \approx \begin{pmatrix} 1{,}455 \\ -1{,}917 \\ 1{,}364 \end{pmatrix}, \quad \mathbf{x}^{(2)} \approx \begin{pmatrix} 0{,}982 \\ -2{,}023 \\ 0{,}925 \end{pmatrix}, \quad \mathbf{x}^{(3)} \approx \begin{pmatrix} 1{,}003 \\ -1{,}989 \\ 1{,}001 \end{pmatrix} .$$

Поскольку в данном случае $\left\| \mathbf{x}^{(1)} - \mathbf{x}^{(0)} \right\|_\infty = \left\| \mathbf{x}^{(1)} \right\|_\infty = \frac{23}{12}$, оценку числа

шагов, достаточного для достижения точности 0,001 по норме-максимум, находим из неравенства $\frac{(3/11)^k}{1-3/11}\cdot\frac{23}{12}\le 0{,}001$. Наименьшим натуральным значением k, удовлетворяющим этому неравенству, является число $k=6$ (существенно меньшее аналогичного показателя итерационного процесса, организованного в предыдущем примере).

Апостериорная оценка абсолютной погрешности третьего приближения суть $\left\|\mathbf{x}^*-\mathbf{x}^{(3)}\right\|_\infty \le \frac{3/11}{1-3/11}\cdot\left\|\mathbf{x}^{(3)}-\mathbf{x}^{(2)}\right\|_\infty = \frac{3}{8}\cdot 0{,}076 = 0{,}0285$ (при фактической его погрешности $\left\|\mathbf{x}^*-\mathbf{x}^{(3)}\right\|_\infty = \left\|(-0{,}003;\ -0{,}011;\ -0{,}001)^T\right\|_\infty = 0{,}011$).

Следствием теоремы 3.1, устанавливающим необходимые и достаточные условия сходимости метода Якоби, является следующая теорема.

Теорема 3.4. *Метод Якоби* (3.20) *сходится к решению системы* (3.7) *в том и только в том случае, когда все корни уравнения*

$$\begin{vmatrix} a_{11}\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}\lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn}\lambda \end{vmatrix} = 0$$

по модулю меньше единицы.

Действительно, чтобы все собственные числа матрицы $\mathbf{B}:=-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})$ были по модулю меньше единицы, как этого требует теорема 3.1 для данного случая, нужно, чтобы меньше единицы были модули всех корней характеристического уравнения

$$\det\left(-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})-\lambda\mathbf{E}\right)=0.$$

Последнее же эквивалентно уравнению

$$\det(\mathbf{L}+\mathbf{R}+\lambda\mathbf{D})=0,$$

которое в записи через элементы исходной матрицы $\mathbf{A}$ и фигурирует в формулировке теоремы.

Под ***методом Зейделя***[*] обычно понимают такое видоизменение метода простых итераций (3.9) решения СЛАУ, приведенных к виду (3.8), при котором для подсчета i-й компоненты $(k+1)$-го приближения к искомому вектору $\mathbf{x}^*$ используют уже найденные на этом, т.е. $(k+1)$-м шаге, новые значения первых $i-1$ компонент. Это означает, что если система (3.7) тем или иным способом сведена к системе (3.8) с матрицей коэффициентов $\mathbf{B} := (b_{ij})_{i,j=1}^{n}$ и вектором свободных членов $\mathbf{c} := (c_i)_{i=1}^{n}$, то приближения к ее решению по методу Зейделя определяются системой равенств

$$\begin{cases} x_1^{(k+1)} = b_{11}x_1^{(k)} + b_{12}x_2^{(k)} + \ldots + b_{1n}x_n^{(k)} + c_1, \\ x_2^{(k+1)} = b_{21}x_1^{(k+1)} + b_{22}x_2^{(k)} + \ldots + b_{2n}x_n^{(k)} + c_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(k+1)} = b_{n1}x_1^{(k+1)} + b_{n2}x_2^{(k+1)} + \ldots + b_{nn}x_n^{(k)} + c_n, \end{cases} \tag{3.23}$$

где $k = 0, 1, 2, \ldots$, а $x_i^{(0)}$ — компоненты вектора $\mathbf{x}^{(0)}$, задаваемого в начале итерационного процесса.

Применительно к компьютерным расчетам один полный, т.е. векторный, итерационный шаг метода Зейделя может интерпретироваться как реализация формулы (3.8), где под знаком равенства следует понимать знак присваивания, а под $\mathbf{x}$ — один и тот же линейный массив из n элементов, на нулевом шаге заполненный компонентами заданного начального вектора $\mathbf{x}^{(0)}$. В таком случае на обычной однопроцессорной вычислительной машине элементы массива $\mathbf{x}$ будут постепенно замещаться новыми элементами.

Аналогичный взгляд на МПИ (3.9) показывает, что для компьютерной реализации одного его шага нужно целиком сохранять n-элементный массив $\mathbf{x}$, подставляемый в правую часть, до тех пор,

---

[*] Зе́йдель Филипп Людвиг (1821–1896) — немецкий астроном и математик.

шагов, достаточного для достижения точности 0,001 по норме-максимум, находим из неравенства $\frac{(3/11)^k}{1-3/11}\cdot\frac{23}{12}\le 0{,}001$. Наименьшим натуральным значением k, удовлетворяющим этому неравенству, является число $k=6$ (существенно меньшее аналогичного показателя итерационного процесса, организованного в предыдущем примере).

Апостериорная оценка абсолютной погрешности третьего приближения суть $\left\|\mathbf{x}^*-\mathbf{x}^{(3)}\right\|_\infty \le \frac{3/11}{1-3/11}\cdot\left\|\mathbf{x}^{(3)}-\mathbf{x}^{(2)}\right\|_\infty = \frac{3}{8}\cdot 0{,}076 = 0{,}0285$ (при фактической его погрешности $\left\|\mathbf{x}^*-\mathbf{x}^{(3)}\right\|_\infty = \left\|(-0{,}003;\ -0{,}011;\ -0{,}001)^T\right\|_\infty = 0{,}011$).

Следствием теоремы 3.1, устанавливающим необходимые и достаточные условия сходимости метода Якоби, является следующая теорема.

Теорема 3.4. *Метод Якоби* (3.20) *сходится к решению системы* (3.7) *в том и только в том случае, когда все корни уравнения*

$$\begin{vmatrix} a_{11}\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}\lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn}\lambda \end{vmatrix} = 0$$

по модулю меньше единицы.

Действительно, чтобы все собственные числа матрицы $\mathbf{B}:=-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})$ были по модулю меньше единицы, как этого требует теорема 3.1 для данного случая, нужно, чтобы меньше единицы были модули всех корней характеристического уравнения

$$\det\left(-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})-\lambda\mathbf{E}\right)=0.$$

Последнее же эквивалентно уравнению

$$\det(\mathbf{L}+\mathbf{R}+\lambda\mathbf{D})=0,$$

которое в записи через элементы исходной матрицы $\mathbf{A}$ и фигурирует в формулировке теоремы.

Под ***методом Зейделя***[*] обычно понимают такое видоизменение метода простых итераций (3.9) решения СЛАУ, приведенных к виду (3.8), при котором для подсчета i-й компоненты $(k+1)$-го приближения к искомому вектору $\mathbf{x}^*$ используют уже найденные на этом, т.е. $(k+1)$-м шаге, новые значения первых $i-1$ компонент. Это означает, что если система (3.7) тем или иным способом сведена к системе (3.8) с матрицей коэффициентов $\mathbf{B} := (b_{ij})_{i,j=1}^{n}$ и вектором свободных членов $\mathbf{c} := (c_i)_{i=1}^{n}$, то приближения к ее решению по методу Зейделя определяются системой равенств

$$\begin{cases} x_1^{(k+1)} = b_{11}x_1^{(k)} + b_{12}x_2^{(k)} + \ldots + b_{1n}x_n^{(k)} + c_1, \\ x_2^{(k+1)} = b_{21}x_1^{(k+1)} + b_{22}x_2^{(k)} + \ldots + b_{2n}x_n^{(k)} + c_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(k+1)} = b_{n1}x_1^{(k+1)} + b_{n2}x_2^{(k+1)} + \ldots + b_{nn}x_n^{(k)} + c_n, \end{cases} \tag{3.23}$$

где $k = 0, 1, 2, \ldots$, а $x_i^{(0)}$ — компоненты вектора $\mathbf{x}^{(0)}$, задаваемого в начале итерационного процесса.

Применительно к компьютерным расчетам один полный, т.е. векторный, итерационный шаг метода Зейделя может интерпретироваться как реализация формулы (3.8), где под знаком равенства следует понимать знак присваивания, а под $\mathbf{x}$ — один и тот же линейный массив из n элементов, на нулевом шаге заполненный компонентами заданного начального вектора $\mathbf{x}^{(0)}$. В таком случае на обычной однопроцессорной вычислительной машине элементы массива $\mathbf{x}$ будут постепенно замещаться новыми элементами.

Аналогичный взгляд на МПИ (3.9) показывает, что для компьютерной реализации одного его шага нужно целиком сохранять n-элементный массив $\mathbf{x}$, подставляемый в правую часть, до тех пор,

---

[*] Зейдель Филипп Людвиг (1821–1896) — немецкий астроном и математик.

пока не сформируется полностью другой n-элементный массив — результат данного итерационного шага. В связи с такой интерпретацией метод Зейделя называют ***методом последовательных смещений***, а метод простых итераций — ***методом одновременных смещений***.

Остановимся подробнее на случае, когда приведение системы (3.7) к виду (3.8) основано на ее представлении в виде (3.18), т.е. когда метод Зейделя есть модификация метода Якоби. Запись соответствующих расчетных формул здесь сводится к верхней индексации системы (3.21) по типу (3.23):

$$\begin{cases} x_1^{(k+1)} = -\left(a_{12}x_2^{(k)} + \ a_{13}x_3^{(k)} + \ldots + \ a_{1n}x_n^{(k)} - b_1\right)\Big/a_{11}, \\ x_2^{(k+1)} = -\left(a_{21}x_1^{(k+1)} + a_{23}x_3^{(k)} + \ldots + a_{2n}x_n^{(k)} - b_2\right)\Big/a_{22}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(k+1)} = -\left(a_{n1}x_1^{(k+1)} + a_{n2}x_2^{(k+1)} + \ldots + a_{n,\,n-1}x_{n-1}^{(k+1)} - b_n\right)\Big/a_{nn}, \end{cases} \tag{3.24}$$

где при заданном $\mathbf{x}^{(0)} := \left(x_1^{(0)}, \ldots, x_n^{(0)}\right)^T$ полагают $k = 0, 1, 2, \ldots$.

Для анализа сходимости метода Зейделя (3.24) обратимся к его векторно-матричной форме. Легко видеть, что если неявный вид метода Якоби, вытекающий из представления (3.18) системы (3.7), есть

$$\mathbf{L}\mathbf{x}^{(k)} + \mathbf{D}\mathbf{x}^{(k+1)} + \mathbf{R}\mathbf{x}^{(k)} = \mathbf{b} \quad \text{(сравните с (3.20))},$$

то равнозначный (3.24) неявный вид метода Зейделя в векторно-матричных обозначениях таков:

$$\mathbf{L}\mathbf{x}^{(k+1)} + \mathbf{D}\mathbf{x}^{(k+1)} + \mathbf{R}\mathbf{x}^{(k)} = \mathbf{b}.$$

Следовательно, тот же вектор $\mathbf{x}^{(k+1)}$, который фигурирует в левой части совокупности равенств (3.24), может быть получен по формуле

$$\mathbf{x}^{(k+1)} = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}\mathbf{x}^{(k)} + (\mathbf{L}+\mathbf{D})^{-1}\mathbf{b}. \tag{3.25}$$

Последнее выражение определяет не что иное, как МПИ (3.9) для системы вида (3.8), где

$$\mathbf{B} := -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}, \quad \mathbf{c} := (\mathbf{L}+\mathbf{D})^{-1}\mathbf{b},$$

т.е. результат применения одного шага метода Зейделя (3.24), полученного на основе $(\mathbf{L}+\mathbf{D}+\mathbf{R})$-разложения матрицы $\mathbf{A}$, можно расценивать как шаг МПИ для эквивалентной (3.7) задачи о неподвижной точке

$$\mathbf{x} = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}\mathbf{x} + (\mathbf{L}+\mathbf{D})^{-1}\mathbf{b} \tag{3.26}$$

(разумеется, если треугольная матрица $\mathbf{L}+\mathbf{D}$ обратима). Эта связь между методом Зейделя и методом простых итераций позволяет легко переформулировать некоторые утверждения о сходимости МПИ применительно к методу Зейделя (3.24).

Теорема 3.5. *Для сходимости метода Зейделя* (3.24) *необходимо и достаточно, чтобы все корни уравнения*

$$\begin{vmatrix} a_{11}\lambda & a_{12} & \dots & a_{1n} \\ a_{21}\lambda & a_{22}\lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1}\lambda & a_{n2}\lambda & \dots & a_{nn}\lambda \end{vmatrix} = 0 \tag{3.27}$$

были по модулю меньше единицы.

Доказательство. Применяя теорему 3.1 к МПИ (3.25), составляем характеристическое уравнение, определяющее собственные числа λ матрицы $\mathbf{B} := -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}$:

$$\det\left(-(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R} - \lambda\mathbf{E}\right) = 0.$$

Это уравнение равносильно уравнению

$$\det\left((\mathbf{L}+\mathbf{D})\lambda + \mathbf{R}\right) = 0,$$

которое с учетом представления $\mathbf{A} = \mathbf{L}+\mathbf{D}+\mathbf{R}$ совпадает с (3.27).

Замечание 3.7. Уравнение (3.27), а также метод (3.24), являющийся частным случаем более общей формы метода Зейделя (3.23), называют иногда соответственно *уравнением* и *методом Некрасова* [69]. Метод (3.23) называют еще и *методом Гаусса–Зейделя* [50–52], хотя, если верить автору книги [57], «этот метод не был известен Зейделю и презирался Гауссом как бесполезный».

Прямым следствием теоремы 3.2 для метода Зейделя (3.24) является следующая теорема.

Теорема 3.6. *Пусть* $\left\|(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}\right\| \le t < 1$. *Тогда при любом начальном векторе* $\mathbf{x}^{(0)}$ *метод Зейделя* (3.24) *сходится к решению* $\mathbf{x}^*$ *системы* (3.7) *и справедливы оценки погрешности*

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{t}{1-t}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \frac{t^k}{1-t}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|. \qquad (3.28)$$

Ясно, что для непосредственного использования оценок (3.28) нужно предварительно выполнить обращение треугольной матрицы $\mathbf{L}+\mathbf{D}$ и перемножить матрицы $(\mathbf{L}+\mathbf{D})^{-1}$ и $\mathbf{R}$. В таком случае частично теряется смысл в поэлементной реализации метода Зейделя (3.24); вместо этого можно проводить итерации по векторно-матричной формуле (3.25) до тех пор, пока не выполнится условие $\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \frac{1-t}{t}\varepsilon$, где $\varepsilon > 0$ — требуемая точность. В частности, такой подход может быть рекомендован при решении СЛАУ методом Зейделя на компьютерах с векторной обработкой информации.

Более подходящие для использования оценки погрешности метода Зейделя (3.24) можно получить, разлагая матрицу $\mathbf{B} := -\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})$ (см. (3.22)) в сумму двух строго треугольных матриц, т.е. полагая

$$\mathbf{B} = \mathbf{B}_L + \mathbf{B}_R,$$

где

$$\mathbf{B}_L := \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ -\dfrac{a_{21}}{a_{22}} & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ -\dfrac{a_{n1}}{a_{nn}} & -\dfrac{a_{n2}}{a_{nn}} & \dots & -\dfrac{a_{n,n-1}}{a_{nn}} & 0 \end{pmatrix},$$

$$\mathbf{B}_R := \begin{pmatrix} 0 & -\dfrac{a_{12}}{a_{11}} & \dots & -\dfrac{a_{1,n-1}}{a_{11}} & -\dfrac{a_{1n}}{a_{11}} \\ 0 & 0 & \dots & -\dfrac{a_{2,n-1}}{a_{22}} & -\dfrac{a_{2n}}{a_{22}} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

С ними эквивалентное (3.7) уравнение (3.8) приобретает вид

$$\mathbf{x} = \mathbf{B}_L\mathbf{x} + \mathbf{B}_R\mathbf{x} + \mathbf{c},$$

т.е. неподвижная точка $\mathbf{x}^*$ удовлетворяет равенству

$$\mathbf{x}^* = \mathbf{B}_L\mathbf{x}^* + \mathbf{B}_R\mathbf{x}^* + \mathbf{c},$$

а метод Зейделя (3.24) определяется неявной формулой

$$\mathbf{x}^{(k+1)} = \mathbf{B}_L\mathbf{x}^{(k+1)} + \mathbf{B}_R\mathbf{x}^{(k)} + \mathbf{c}.$$

Из двух последних равенств получаем следующее:

$$\mathbf{x}^* - \mathbf{x}^{(k+1)} = \mathbf{B}_L(\mathbf{x}^* - \mathbf{x}^{(k+1)}) + \mathbf{B}_R(\mathbf{x}^* - \mathbf{x}^{(k)}).$$

Этот результат приводим к виду

$$\mathbf{x}^* - \mathbf{x}^{(k)} = \mathbf{B}_L(\mathbf{x}^* - \mathbf{x}^{(k)}) + \mathbf{B}_R(\mathbf{x}^* - \mathbf{x}^{(k-1)}), \qquad (3.29)$$

который можно расценивать как точную связь между погрешностями k-го и $(k-1)$-го приближений в методе Зейделя (3.24). Отсюда, переходя к нормам, легко вывести априорную оценку погрешности, что можно оформить в виде сформулированного далее утверждения [1].

Теорема 3.7. *Пусть* $\dfrac{\|\mathbf{B}_R\|}{1-\|\mathbf{B}_L\|} \le p < 1$. *Тогда метод Зейделя* (3.24) *определяет сходящуюся последовательность* $(x^{(k)})$ *при любом начальном векторе* $\mathbf{x}^{(0)}$ *и имеет место оценка*

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le p^k \left\|\mathbf{x}^* - \mathbf{x}^{(0)}\right\| \quad \forall\, k \in \mathbb{N}.$$

Как и у предыдущей, у этой теоремы имеются свои недостатки, затрудняющие ее применение: нужно знать меру близости начального приближения $\mathbf{x}^{(0)}$ к решению $\mathbf{x}^*$. Ценность ее скорее в том, что в ней фигурирует легко вычисляемый коэффициент $\dfrac{\|\mathbf{B}_R\|}{1-\|\mathbf{B}_L\|}$ связи ошибок результатов двух соседних итерационных шагов, характеризующий быстроту сходимости метода Зейделя (3.24). При организации практических вычислений по формулам (3.24) целесообразнее ориентироваться на следующий результат.

Теорема 3.8. *Пусть* $\|\mathbf{B}\| < 1$ (*где* $\mathbf{B}$ — *матрица* (3.22)). *Тогда для определяемой методом Зейделя* (3.24) *последовательности приближений справедлива апостериорная оценка погрешности*

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{\|\mathbf{B}_R\|}{1-\|\mathbf{B}\|}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \quad \forall\, k \in \mathbb{N}.$$

Для доказательства этого утверждения подставим $\mathbf{B}_L = \mathbf{B} - \mathbf{B}_R$ в равенство (3.29). Имеем

$$(\mathbf{E}-\mathbf{B})\left(\mathbf{x}^* - \mathbf{x}^{(k)}\right) = -\mathbf{B}_R(\mathbf{x}^* - \mathbf{x}^{(k)}) + \mathbf{B}_R(\mathbf{x}^* - \mathbf{x}^{(k-1)}),$$

что в условиях теоремы (с учетом леммы 3.2) равносильно равенству

$$\mathbf{x}^* - \mathbf{x}^{(k)} = (\mathbf{E}-\mathbf{B})^{-1}\mathbf{B}_R(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}).$$

Отсюда, переходя к нормам, получаем нужную оценку.

Из теоремы 3.8 вытекает следующая, более удобная на практике, формулировка.

Следствие 3.1. *Пусть* k_ε *— первое в ряду натуральных чисел* k, *с которым при заданном* $\varepsilon > 0$ *для генерируемой процессом Зейделя* (3.24) *последовательности векторов* $\mathbf{x}^{(k)} := \left(x_1^{(k)};\ x_2^{(k)};\ldots;\ x_n^{(k)}\right)^T$ *в некоторых согласованных нормах выполняется неравенство*

$$\left\| \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} \right\| \le \frac{1 - \|\mathbf{B}\|}{\|\mathbf{B}_R\|} \cdot \varepsilon \ .$$

Тогда за решение $\mathbf{x}^*$ *системы* (3.7) *может быть принят вектор* $\mathbf{x}^{\left(k_\varepsilon\right)}$ *и абсолютная погрешность при этом не будет превышать* ε (*в выбранной норме*).

Условия сходимости методов Зейделя и простых итераций, вообще говоря, различаются. Но некоторые достаточные условия можно применять к обоим методам одновременно.

Теорема 3.9. *Если в матрице* $\mathbf{A}$ *системы* (3.7) *имеет место диагональное преобладание, то метод Зейделя* (3.24) *сходится, причем быстрее, чем метод Якоби* (3.20а).

Доказательство. Вычитая тождественное (3.24) равенство (3.25) из равенства (3.26), рассматриваемого как верное равенство при подстановке в него решения $\mathbf{x}^*$, получаем

$$\mathbf{x}^* - \mathbf{x}^{(k+1)} = -(\mathbf{L} + \mathbf{D})^{-1}\mathbf{R}\left(\mathbf{x}^* - \mathbf{x}^{(k)}\right).$$

Введем вектор ошибок $\Delta^{(k)} := \mathbf{x}^* - \mathbf{x}^{(k)}$ с компонентами $\delta_i^{(k)} := x_i^* - x_i^{(k)}$. Тогда это равенство через элементы матрицы $\mathbf{A}$ исходной системы (3.7) можно записать так (см. соответствие между (3.25) и (3.24)):

$$\delta_i^{(k+1)} = -\frac{1}{a_{ii}} \sum_{j=1}^{i-1} a_{ij}\delta_j^{(k+1)} - \frac{1}{a_{ii}} \sum_{j=i+1}^{n} a_{ij}\delta_j^{(k)} \ ,$$

где $i = 1, 2, \ldots, n$; $k = 0, 1, 2, \ldots$. Переходя к модулям, отсюда имеем:

$$\left|\delta_i^{(k+1)}\right| \le \frac{1}{|a_{ii}|}\sum_{j=1}^{i-1}|a_{ij}|\cdot\left|\delta_j^{(k+1)}\right| + \frac{1}{|a_{ii}|}\sum_{j=i+1}^{n}|a_{ij}|\cdot\left|\delta_j^{(k)}\right| \le$$

$$\le \left(\max_j\left|\delta_j^{(k+1)}\right|\right)\cdot\frac{1}{|a_{ii}|}\sum_{j=1}^{i-1}|a_{ij}| + \left(\max_j\left|\delta_j^{(k)}\right|\right)\cdot\frac{1}{|a_{ii}|}\sum_{j=i+1}^{n}|a_{ij}|.$$

Обозначим $\alpha_i := \frac{1}{|a_{ii}|}\sum_{j=1}^{i-1}|a_{ij}|$, $\beta_i := \frac{1}{|a_{ii}|}\sum_{j=i+1}^{n}|a_{ij}|$, $\left\|\Delta^{(k)}\right\|_\infty := \max_i\left|\delta_i^{(k)}\right|$ (где $\left\|\Delta^{(k)}\right\|_\infty$ может трактоваться как абсолютная погрешность k-го приближения по методу Зейделя*). В этих обозначениях последнее неравенство приобретает вид

$$\left|\delta_i^{(k+1)}\right| \le \alpha_i\left\|\Delta^{(k+1)}\right\|_\infty + \beta_i\left\|\Delta^{(k)}\right\|_\infty. \qquad (3.30)$$

Пусть $m \in \{1, 2, \ldots, n\}$ — значение индекса i, при котором реализуется равенство $\left|\delta_m^{(k+1)}\right| = \max_i\left|\delta_i^{(k+1)}\right| = \left\|\Delta^{(k+1)}\right\|_\infty$. Тогда из (3.30) следует

$$\left\|\Delta^{(k+1)}\right\|_\infty \le \alpha_m\left\|\Delta^{(k+1)}\right\|_\infty + \beta_m\left\|\Delta^{(k)}\right\|_\infty,$$

т.е. при этом фиксированном $i := m$ выполняется неравенство

$$\left\|\Delta^{(k+1)}\right\|_\infty \le \frac{\beta_m}{1-\alpha_m}\left\|\Delta^{(k)}\right\|_\infty. \qquad (3.31)$$

Так как в условиях диагонального преобладания справедливо неравенство $\alpha_i + \beta_i < 1$, а это неравенство, в свою очередь, влечет неравенство $\frac{\beta_i}{1-\alpha_i} \le \alpha_i + \beta_i$ (проверьте!), причем равенство в последнем случае имеет место лишь при $i = 1$, то абсолютная погрешность приближений по методу Зейделя (3.24) согласно (3.31)

---

* Индекс ∞ у знака нормы использован согласно обозначению соответствующего частного случая l_p-***нормы*** (см. приложение).

убывает со скоростью геометрической прогрессии, знаменатель которой, вообще говоря, меньше, чем для соответствующего этому случаю метода Якоби (3.20а). Такое мажорирование последовательности величин $\left\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\right\|$ позволяет сделать заключение о справедливости доказываемой теоремы.

Замечание 3.8. В соответствии с последней теоремой можно утверждать, что *в методе Зейделя* (3.24) *в указанных условиях допустимо использование оценок погрешности метода Якоби.* Естественно, они во многих случаях бывают грубее.

Пример 3.3. Возьмем за основу расчетные формулы МПИ, записанные в примере 3.1 для заданной там системы, и в соответствии с (3.23) преобразуем их в расчетные формулы метода Зейделя:

$$\begin{cases} x_1^{(k+1)} = -0,1x_1^{(k)} + 0,2x_2^{(k)} - 0,1x_3^{(k)} + 1,6, \\ x_2^{(k+1)} = 0,1x_1^{(k+1)} - 0,2x_2^{(k)} - 0,2x_3^{(k)} - 2,3, \\ x_3^{(k+1)} = -0,2x_1^{(k+1)} + 0,1x_2^{(k+1)} - 0,1x_3^{(k)} + 1,5. \end{cases}$$

Начиная процесс вычислений с того же начального приближения $\mathbf{x}^{(0)} := \mathbf{0}$, далее при $k := 0, 1, 2$ последовательно получаем:

$$\begin{cases} x_1^{(1)} = 1,6, \\ x_2^{(1)} = 0,1 \cdot 1,6 - 2,3 = -2,14, \\ x_3^{(1)} = -0,2 \cdot 1,6 + 0,1 \cdot (-2,14) + 1,5 = 0,966; \end{cases}$$

$$\begin{cases} x_1^{(2)} = -0,1 \cdot 1,6 + 0,2 \cdot (-2,14) - 0,1 \cdot 0,966 + 1,6 :\approx 0,915, \\ x_2^{(2)} = 0,1 \cdot 0,915 - 0,2 \cdot (-2,14) - 0,2 \cdot 0,966 - 2,3 :\approx -1,974, \\ x_3^{(2)} = -0,2 \cdot 0,915 + 0,1 \cdot (-1,974) - 0,1 \cdot 0,966 + 1,5 :\approx 1,023; \end{cases}$$

$$\begin{cases} x_1^{(3)} = -0,1 \cdot 0,915 + 0,2 \cdot (-1,974) - 0,1 \cdot 1,023 + 1,6 :\approx 1,011, \\ x_2^{(3)} = 0,1 \cdot 1,011 - 0,2 \cdot (-1,974) - 0,2 \cdot 1,023 - 2,3 :\approx -2,009, \\ x_3^{(3)} = -0,2 \cdot 1,011 + 0,1 \cdot (-2,009) - 0,1 \cdot 1,023 + 1,5 :\approx 0,995. \end{cases}$$

Вектор ошибок третьего приближения $\mathbf{x}^{(3)}$ по методу Зейделя к точному решению $\mathbf{x}^* := (1; -2; 1)^T$ данной в примере 3.1 линейной системы есть

$(-0{,}011;\ 0{,}009;\ 0{,}005)^T$. Его норма-максимум составляет величину 0,011, что в 4 раза меньше, чем при применении в тех же условиях соответствующего метода простых итераций. Поскольку показанное улучшение не требует дополнительных вычислений, налицо эффективность подобной модификации МПИ.

Если применительно к данной задаче метод Зейделя рассматривать как модификацию метода Якоби, то соответствующие расчетные формулы, согласно (3.24), таковы:

$$\begin{cases} x_1^{(k+1)} = \dfrac{1}{11}\left(16 + 2x_2^{(k)} - x_3^{(k)}\right), \\ x_2^{(k+1)} = \dfrac{1}{12}\left(-23 + x_1^{(k+1)} - 2x_3^{(k)}\right), \\ x_3^{(k+1)} = \dfrac{1}{11}\left(15 - 2x_1^{(k+1)} + x_2^{(k+1)}\right), \end{cases}$$

где $k = 0, 1, 2, \ldots$; $x_1^{(0)} := 0$, $x_2^{(0)} := 0$, $x_3^{(0)} := 0$. Счет по этим формулам (с округлением до третьего знака) дает:

$$1)\, k := 0, \begin{cases} x_1^{(1)} = 1{,}455, \\ x_2^{(1)} = -1{,}795, \\ x_3^{(1)} = 0{,}936; \end{cases} \quad 2)\, k := 1, \begin{cases} x_1^{(2)} = 1{,}043, \\ x_2^{(2)} = -1{,}986, \\ x_3^{(2)} = 0{,}993; \end{cases} \quad 3)\, k := 2, \begin{cases} x_1^{(3)} = 1{,}003, \\ x_2^{(3)} = -1{,}999, \\ x_3^{(3)} = 1{,}000. \end{cases}$$

Полученный результат третьей итерации более точен, чем в предыдущем варианте метода Зейделя (здесь абсолютная погрешность $\left\|\mathbf{x}^* - \mathbf{x}^{(3)}\right\|_\infty := 0{,}003$), а его апостериорная оценка через $\left\|\mathbf{x}^{(3)} - \mathbf{x}^{(2)}\right\|_\infty := 0{,}040$ составляет величину 0,015 как при использовании для этих целей оценки МПИ-Якоби, так и оценки теоремы 3.8 (осмыслите, почему).

Остановимся еще на одном важном для приложений классе систем вида (3.7), для которых имеет место сходимость метода Зейделя (3.24).

Определение 3.4 [25]. *Система* $\mathbf{Ax} = \mathbf{b}$ *называется* ***нормальной***, *если матрица* $\mathbf{A}$ — *симметричная положительно определенная.*

Теорема 3.10. *Если система* (3.7) — *нормальная, то метод Зейделя* (3.24) *сходится.*

Доказательство этой теоремы заключается в проверке того, что положительная определенность матрицы $\mathbf{A}=\mathbf{L}+\mathbf{D}+\mathbf{L}^T$ влечет выполнимость условия теоремы 3.5 (т.е. собственные числа матрицы $-(\mathbf{L}+\mathbf{D})^{-1}\mathbf{L}^T$ по модулю меньше единицы). Это доказательство можно найти, например, в [7, 25].

Любая линейная система $\mathbf{Ax}=\mathbf{b}$ легко может быть симметризована умножением на матрицу $\mathbf{A}^T$. Переход от системы $\mathbf{Ax}=\mathbf{b}$ к системе $\mathbf{A}^T\mathbf{Ax}=\mathbf{A}^T\mathbf{b}$ (или к $\mathbf{A}^*\mathbf{Ax}=\mathbf{A}^*\mathbf{b}$ в более общем случае, когда $a_{ij}\in\mathbb{C}$) называют ***симметризацией Гаусса***.

Справедлива следующая теорема.

Теорема 3.11 [25]. *Пусть* $\det\mathbf{A}\neq 0$. *Тогда система* $\mathbf{A}^T\mathbf{Ax}=\mathbf{A}^T\mathbf{b}$ — *нормальная.*

Таким образом, если, например, известно, что система (3.7) однозначно разрешима, но в ее матрице коэффициентов нет диагонального преобладания, метод Зейделя типа (3.24) можно применять к системе $\mathbf{A}^T\mathbf{Ax}=\mathbf{A}^T\mathbf{b}$. Правда, здесь возникают трудности со своевременным окончанием процесса итерирования, обеспечивающим заданную точность приближенного решения, так как приведенные ранее оценки погрешности (см. теорему 3.6 и замечание 3.8) в этом случае часто не работают. Да и сходимость может оказаться весьма медленной. И, кроме того, такая симметризация может ухудшить обусловленность системы (это еще будет обсуждаться в дальнейшем).

Наряду с рассмотренными, применяют и другие способы приведения систем (3.7) к виду (3.8) для их решения методами простых итераций и Зейделя. Достаточно общий подход к этой процедуре заключается в том, что эквивалентное (3.7) уравнение $\mathbf{0}=\mathbf{b}-\mathbf{Ax}$ умножают на некоторую неособенную матрицу $\mathbf{H}$ (матричный параметр) и к обеим частям прибавляют вектор $\mathbf{x}$. Полученное уравнение

$$\mathbf{x}=\mathbf{x}+\mathbf{H}(\mathbf{b}-\mathbf{Ax}),$$

переписанное в виде

$$\mathbf{x} = (\mathbf{E} - \mathbf{HA})\mathbf{x} + \mathbf{Hb},$$

имеет структуру (3.8). Проблема теперь заключается в подборе матрицы $\mathbf{H}$, такой, чтобы матрица $\mathbf{B} := \mathbf{E} - \mathbf{HA}$ обладала нужными свойствами для сходимости применяемых методов; для некоторых классов матриц $\mathbf{A}$ имеются определенные рекомендации [25, 42]. Заметим, что матрица $\mathbf{H}$ может быть как постоянной (в этой ситуации говорят о ***стационарном*** итерационном процессе), так и изменяющейся от шага к шагу. В последнем случае данное уравнение $\mathbf{Ax} = \mathbf{b}$ подменяется последовательностью эквивалентных ему уравнений $\mathbf{x} = \mathbf{B}_k\mathbf{x} + \mathbf{c}_k$, и соответствующий итерационный процесс называется ***нестационарным***.

Если применение оценок погрешностей в методах простых итераций и Зейделя невозможно из-за отсутствия констант $q < 1$ или $t < 1$, ограничивающих сверху какие-либо нормы матрицы итерирования соответствующего метода (см. теоремы 3.2 и 3.6), эти методы неэффективны и, более того, как будет далее показано, малонадежны ввиду медленной сходимости. Рассмотрим одно обобщение метода Зейделя, позволяющее иногда в несколько раз ускорить сходимость итерационной последовательности.

Пусть $z_i^{(k)}$ — обозначение i-й компоненты k-го приближения к решению системы (3.7) по методу Зейделя, а обозначение $x_i^{(k)}$ будем использовать для i-й компоненты k-го приближения, получаемого новым методом. Этот метод определим равенством

$$x_i^{(k+1)} = x_i^{(k)} + \omega\left(z_i^{(k+1)} - x_i^{(k)}\right), \qquad (3.32)$$

где $i = 1, 2, \ldots, n$; $k = 0, 1, 2, \ldots$; $x_i^{(0)}$ — задаваемые начальные значения; ω — числовой параметр, который называют ***параметром релаксации***. Очевидно, при $\omega = 1$ метод (3.32), называемый ***методом релаксации*** (*ослабления*), совпадает с методом Зейделя.

(Метод Зейделя в качестве представителя семейства релаксационных методов иногда называют ***методом полной релаксации.***)

Конкретизируем метод релаксации для случая, когда исходная система (3.7) представляется в виде (3.19) и, следовательно, метод Зейделя имеет вид (3.24).

Пользуясь (3.24) и введенными здесь обозначениями, запишем дополнительное к (3.32) равенство для выражения компонент векторов $\mathbf{z}^{(k)} := \left(z_i^{(k)}\right)_{i=1}^{n}$ через компоненты векторов $\mathbf{x}^{(k)} := \left(x_i^{(k)}\right)_{i=1}^{n}$:

$$z_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right). \qquad (3.33)$$

Таким образом, метод релаксации можно понимать как поочередное применение формул (3.33) и (3.32) при каждом $k = 0, 1, 2, \ldots$. Действительно, задав начальные значения $x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}$ и параметр ω, при $k := 0$, полагая $i := 1, 2, \ldots, n$, вычислим

$$z_1^{(1)}, x_1^{(1)};\quad z_2^{(1)}, x_2^{(1)};\ \ldots;\ z_n^{(1)}, x_n^{(1)};$$

при $k := 1$, также полагая $i := 1, 2, \ldots, n$, находим

$$z_1^{(2)}, x_1^{(2)};\quad z_2^{(2)}, x_2^{(2)};\ \ldots;\ z_n^{(2)}, x_n^{(2)}$$

и т.д. Но можно избавиться от вспомогательной последовательности $\left(\mathbf{z}^{(k)}\right)$, подставив (3.33) в (3.32). Для $i = 1, 2, \ldots, n$ будем иметь:

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right). \qquad (3.34)$$

От формулы (3.34), объединяющей формулы (3.33) и (3.32) и пригодной для проведения покоординатных вычислений, мало отличающихся от вычислений по методу Зейделя, легко перейти к векторно-матричной записи процесса релаксации. С этой целью перепишем (3.34) в виде

$$a_{ii}x_i^{(k+1)} + \omega\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} = (1-\omega)a_{ii}x_i^{(k)} - \omega\sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + \omega b_i$$

и далее, учитывая аддитивное представление матрицы $\mathbf{A}=\mathbf{L}+\mathbf{D}+\mathbf{R}$, получаем векторно-матричный итерационный процесс в неявной форме

$$(\mathbf{D}+\omega\mathbf{L})\mathbf{x}^{(k+1)}=(1-\omega)\mathbf{D}\mathbf{x}^{(k)}-\omega\mathbf{R}\mathbf{x}^{(k)}+\omega\mathbf{b}.$$

Умножив последнее равенство слева на матрицу $(\mathbf{D}+\omega\mathbf{L})^{-1}$, приходим к эквивалентному (3.34) методу простых итераций

$$\mathbf{x}^{(k+1)}=(\mathbf{D}+\omega\mathbf{L})^{-1}\big((1-\omega)\mathbf{D}-\omega\mathbf{R}\big)\mathbf{x}^{(k)}+\omega(\mathbf{D}+\omega\mathbf{L})^{-1}\mathbf{b} \quad (3.35)$$

(подстановка сюда значения $\omega:=1$ превращает (3.35) в МПИ (3.25), эквивалентный методу Зейделя (3.24)).

Представление метода релаксации (3.34) в виде (3.35) позволяет сделать для него некоторые утверждения о сходимости, привлекая для этого соответствующие теоремы сходимости МПИ. Например, можно применить теоремы 3.1 и 3.2, полагая в них $\mathbf{B}:=(\mathbf{D}+\omega\mathbf{L})^{-1}\big((1-\omega)\mathbf{D}-\omega\mathbf{R}\big)$; правда, получаемые при таком подходе утверждения вряд ли будут вызывать большой интерес. Более глубокие результаты на этом пути получают, изучая спектральные свойства таких матриц $\mathbf{B}$. В частности, установлено, что *для сходимости процесса* (3.34) *необходимо, чтобы* $\omega\in(0,\ 2)$. Для некоторых классов СЛАУ (3.7) это требование к параметру релаксации является и достаточным. Справедлива следующая теорема, обобщающая теорему 3.8.

Теорема 3.12 (*Островского–Рейча* [50, 51, 62]). *Для нормальной системы* $\mathbf{A}\mathbf{x}=\mathbf{b}$ *метод релаксации* (3.34) *сходится при любом* $\mathbf{x}^{(0)}$ *и любом* $\omega\in(0,\ 2)$.

Поскольку итерационный процесс (3.34) содержит параметр, естественно распорядиться им так, чтобы сходимость последовательности $\left(\mathbf{x}^{(k)}\right)$ была наиболее быстрой. Очевидно, это достигается в том случае, когда спектральный радиус матрицы $\mathbf{B}:=(\mathbf{D}+\omega\mathbf{L})^{-1}\big((1-\omega)\mathbf{D}-\omega\mathbf{R}\big)$ будет минимальным. В общем случае задача нахождения оптимального значения $\omega:=\omega_0$ не ре-

шена, и в практических расчетах применяют метод проб и ошибок. Однако для отдельных важных классов задач такие значения удается выразить через собственные числа матрицы $\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})$ (т.е. через корни уравнения, фигурирующего в теореме 3.4) и даже оценить ускорение, достигаемое введением в процесс Зейделя оптимального параметра релаксации. Существенно отметить, что это оптимальное значение $\omega_0 \in (1, 2)$. При значениях $\omega \in (1, 2)$ метод (3.34) называют ***методом последовательной верхней релаксации*** (сокращенно ***ПВР-*** или ***SOR-методом****). Ввиду низкой эффективности метода (3.34) при $\omega \in (0, 1)$, называемого в этом случае ***методом нижней релаксации***, название «метод ПВР» в последнее время относят ко всему семейству методов (3.34), т.е. для любых $\omega \in (0, 2)$. При этом случай $\omega \in (1, 2)$ называют ***сверхрелаксацией***.

Покажем возможный выигрыш при использовании метода ПВР на простейшем примере.

Пример 3.4. Для системы

$$\begin{cases} 2x_1 + x_2 = 1, \\ x_1 + 2x_2 = -1 \end{cases}$$

с симметричной положительно определенной матрицей $\mathbf{A} := \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ и очевидным решением $\mathbf{x}^* := (1; -1)^T$ выполним по три итерационных шага, начиная с $\mathbf{x}^{(0)} := \mathbf{0}$, методами Якоби, Зейделя и ПВР соответственно по формулам

$$\begin{cases} x_1^{(k+1)} = -0{,}5x_2^{(k)} + 0{,}5, \\ x_2^{(k+1)} = -0{,}5x_1^{(k)} - 0{,}5, \end{cases} \qquad \begin{cases} x_1^{(k+1)} = -0{,}5x_2^{(k)} + 0{,}5, \\ x_2^{(k+1)} = -0{,}5x_1^{(k+1)} - 0{,}5 \end{cases}$$

и

$$\begin{cases} x_1^{(k+1)} = (1-\omega)x_1^{(k)} + \frac{\omega}{2}(1 - x_2^{(k)}), \\ x_2^{(k+1)} = (1-\omega)x_2^{(k)} - \frac{\omega}{2}(1 + x_1^{(k+1)}) \end{cases} \quad \text{при } \omega := 1{,}1.$$

---

* От англ. *Successive Over Relaxation.*

Сравнительные результаты третьего шага представлены следующей таблицей.

Таблица 3.1

	Метод Якоби	Метод Зейделя	Метод ПВР (с $\omega := 1{,}1$)
$x_1^{(3)}$	0,875	$\approx 0{,}969$	$\approx 1{,}0008$
$x_2^{(3)}$	–0,875	$\approx -0{,}984$	$\approx -1{,}0009$
$\left\|x^* - x^{(3)}\right\|_\infty$	0,125	$\approx 0{,}031$	$< 0{,}001$

Значение параметра релаксации ω здесь взято близким к оптимальному, которое для матриц, «*упорядоченных согласованно со свойством A*» [51], находят по формуле

$$\omega_{\text{опт}} := \frac{2}{1+\sqrt{1-\rho^2(\mathbf{B})}},$$

где $\rho(\mathbf{B})$ — спектральный радиус матрицы $\mathbf{B} := \mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})$ (в данном случае $\mathbf{B} := \begin{pmatrix} 0 & 0{,}5 \\ 0{,}5 & 0 \end{pmatrix}$, $\rho(\mathbf{B}) := 0{,}5$, $\omega_{\text{опт}} := \approx 1{,}0718$).

§ 3.4. О ДРУГИХ ПОДХОДАХ К ПОСТРОЕНИЮ ИТЕРАЦИОННЫХ МЕТОДОВ

В основе построения и изучения или, по крайней мере, понимания многих итерационных методов лежит связь между системами алгебраических уравнений и методами дискретизации дифференциальных уравнений, их порождающими.

В простейшем абстрактном, но далеко не самом общем случае легко установить такую связь между СЛАУ (3.7) и абстрактным дифференциальным уравнением

$$\frac{d\mathbf{y}}{dt} + \mathbf{A}\mathbf{y}(t) = \mathbf{b} \tag{3.36}$$

с начальным условием $\mathbf{y}(0) := \mathbf{x}^{(0)}$, где t — абстрактная скаляр-

ная переменная, изменяющаяся на промежутке $[0, +\infty)$, а матрица $\mathbf{A}$ и вектор $\mathbf{b}$ те же, что и в уравнении (3.7).

Пусть постоянный вектор $\mathbf{x}$ и переменный вектор $\mathbf{y} := \mathbf{y}(t)$ — решения задач (3.7) и (3.36) соответственно. Введем вектор

$$\mathbf{z}(t) := \mathbf{x} - \mathbf{y}(t).$$

Учитывая равенство $\frac{d\mathbf{z}}{dt} = -\frac{d\mathbf{y}}{dt}$, из совместного рассмотрения (3.7) и (3.36) выясняем, что $\mathbf{z}(t)$ удовлетворяет однородному дифференциальному уравнению

$$\frac{d\mathbf{z}}{dt} = -\mathbf{A}\mathbf{z}(t)$$

с начальным условием $\mathbf{z}(0) = \mathbf{x} - \mathbf{x}^{(0)}$. Решением этой начальной задачи служит вектор

$$\mathbf{z}(t) = e^{-\mathbf{A}t} \cdot \mathbf{z}(0),$$

и если спектр $\mathbf{A}$ лежит в правой полуплоскости (в частности, если, например, матрица $\mathbf{A}$ положительно определена), то $\mathbf{z}(t) \xrightarrow[t \to \infty]{} \mathbf{0}$ при любых $\mathbf{z}(0)$. Таким образом, решение $\mathbf{x}$ системы (3.7) (стационарной задачи) может быть получено как предел при $t \to \infty$ решения $\mathbf{y}(t)$ задачи Коши (3.36) (эволюционной задачи) с произвольным начальным вектором $\mathbf{x}^{(0)}$.

Методы приближенного решения стационарных задач, основанные на нахождении решений нестационарных задач, асимптотически эквивалентных данным задачам для достаточно больших значений искусственной скалярной переменной, называются ***методами установления***.

Замечание 3.9. Иногда к дифференциальным уравнениям переходят не от исходной стационарной задачи, а от какого-то конкретного итерационного метода ее решения. Получающуюся при этом асимптотически эквивалентную дифференциальную задачу называют ***непрерывным аналогом*** соответствующего итерационного метода.

Будем далее считать параметром скалярную величину τ_k, которую применительно к задаче (3.36) можно интерпретировать как шаг (вообще говоря, переменный), с которым на полуоси $[0, +\infty)$ фиксируются точки

$$t_0 (:= 0),\ t_1,\ t_2,\ \ldots,$$

т.е. $t_{k+1} = t_k + \tau_k$, где $k = 0, 1, 2, \ldots$.

При «замораживании» $t := t_k$ уравнение (3.36) принимает вид

$$\left.\frac{d\mathbf{y}}{dt}\right|_{t=t_k} = -\mathbf{A}\mathbf{y}(t_k) + \mathbf{b}. \tag{3.37}$$

Для производной в его левой части при малых τ_k на основе определения можно записать приближенное равенство

$$\left.\frac{d\mathbf{y}}{dt}\right|_{t=t_k} = \lim_{\tau_k \to 0} \frac{\mathbf{y}(t_k + \tau_k) - \mathbf{y}(t_k)}{\tau_k} \approx \frac{\mathbf{y}(t_{k+1}) - \mathbf{y}(t_k)}{\tau_k}.$$

Теперь ясно, что, полагая $\mathbf{x}^{(k)} := \mathbf{y}(t_k)$ (заметим, что $\mathbf{y}(t_0) = \mathbf{y}(0) = \mathbf{x}^{(0)}$), равенство (3.37) можно приближенно заменить равенством

$$\frac{\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}}{\tau_k} = -\mathbf{A}\mathbf{x}^{(k)} + \mathbf{b}. \tag{3.38}$$

Последняя формула при $k = 0, 1, 2, \ldots$ задает ***явный*** итерационный процесс. Его называют ***двухслойным* итерационным методом*** [62] или ***методом Ричардсона*** [61].

Семейство двухслойных итерационных методов примет более общий вид, если ввести в равенство (3.38) невырожденный

---

* Смысл термина «двухслойный» становится понятным при изучении численных процессов решения уравнений математической физики. Обращаясь же к численному интегрированию систем ОДУ, обнаруживаем, что (3.38) есть не что иное, как явный метод Эйлера (с переменным шагом) для задачи (3.36).

матричный параметр $\mathbf{B}_k$:

$$\mathbf{B}_k \frac{\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}}{\tau_k} = -\mathbf{A}\mathbf{x}^{(k)} + \mathbf{b}. \tag{3.39}$$

Различные конкретные итерационные процессы решения СЛАУ (3.7) (в том числе и все рассмотренные выше) получаются из (3.39) фиксированием матриц $\mathbf{B}_k$ и скаляров τ_k. При этом, если $\mathbf{B}_k$ и τ_k не зависят от k, т.е. одни и те же на каждой итерации, то (3.39) определяет ***стационарный метод***, в противном случае — ***нестационарный***. В общем случае, за исключением $\mathbf{B}_k \equiv \mathbf{E}$, (3.39) — ***неявный метод***.

Выбор параметров $\mathbf{B}_k$, τ_k в (3.39) осуществляют, добиваясь удовлетворения каких-либо отдельных или совокупности нескольких, возможно в чем-то противоречивых требований, таких, как простота, хорошая структура и легкая обращаемость матриц $\mathbf{B}_k$, и в то же время как можно более быстрая сходимость последовательности $(\mathbf{x}^{(k)})$ к решению $\mathbf{x}^*$ системы (3.7). Разумеется, оптимальность или, скорее, квазиоптимальность некоторых методов рассматриваемого семейства удается установить лишь при очень жестких ограничениях на решаемую систему (3.7).

Так, например, доказано [61, 62], что если система (3.7) — нормальная с известными границами $\lambda_{\min} > 0$, $\lambda_{\max} > 0$ спектра ее матрицы коэффициентов, то при заранее зафиксированном (максимальном в реализуемом процессе) числе итераций K метод (3.38) будет обеспечивать наименьшую погрешность, другими словами, минимизировать величину $\|\mathbf{x}^* - \mathbf{x}^{(K)}\|$, в том случае, когда параметры τ_k вычисляются по формуле

$$\tau_k = \frac{2}{(\lambda_{\max} + \lambda_{\min}) + (\lambda_{\max} - \lambda_{\min}) t_{k+1}}, \tag{3.40}$$

где $k = 0, 1, \ldots, K-1$, а $t_k := \cos \dfrac{(2k-1)\pi}{2K}$ — корни полинома

Чебышева K-й степени. Совокупность формул (3.38), (3.40) называют ***явным итерационным методом с чебышевским набором параметров***. Имеется обобщение приведенного утверждения и на неявный случай.

Дальнейшее формальное развитие методы установления получают как сугубо неявные методы вида (3.39) с матрицами $\mathbf{B}_k$, представляемыми в виде произведения простых, легко обращаемых (например, ленточных) матриц, в связи с чем такие методы называются ***методами расщепления***. Из методов расщепления наиболее известными являются ***методы переменных направлений**** и ***попеременно-треугольный метод***. Неформальное изучение этих методов более целесообразно по месту их применения — при численном решении многомерных задач математической физики.

Рассматриваются также ***трехслойные итерационные методы*** (в частности, с чебышевскими параметрами), связывающие уже не два, а три соседних приближения: $\mathbf{x}^{(k+1)}$, $\mathbf{x}^{(k)}$ и $\mathbf{x}^{(k-1)}$. В отличие от предыдущих, такие методы являются ***двухшаговыми***.

Другой большой класс методов итерационного решения СЛАУ (3.7) — это так называемые ***методы вариационного типа***. К ним относятся методы минимальных невязок, минимальных поправок, минимальных итераций, наискорейшего спуска, сопряженных градиентов и т.п. Хорошего понимания и обоснования таких методов можно достигнуть лишь с привлечением теории оптимизации, ибо решение линейной алгебраической системы здесь подменяется решением эквивалентной экстремальной задачи.

А именно: пусть $\mathbf{Ax} = \mathbf{b}$ — нормальная n-мерная система, т.е. $\mathbf{A}$ — положительно определенная симметричная матрица, и пусть $(\cdot,\cdot)$ — скалярное произведение в пространстве $\mathbb{R}_n$.

---

* В зарубежной литературе для таких методов используется аббревиатура ADI — *Alternating Direction Implicite* [18, 50].

Образуем квадратичный функционал

$$\Phi(\mathbf{x}) := (\mathbf{Ax}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x}) + c, \tag{3.41}$$

где $c \in \mathbb{R}_1$ — произвольная постоянная. Задача решения нормальной системы (3.7) и задача минимизации функционала (3.41) эквивалентны ([59] и др.). Действительно, нормальная система имеет решение и притом единственное; обозначим его $\mathbf{x}^*$. Тогда при любом векторе $\mathbf{x}$, представленном в виде $\mathbf{x} = \mathbf{x}^* + \Delta$, справедливо

$$\Phi(\mathbf{x}) = \Phi(\mathbf{x}^* + \Delta) = (\mathbf{A}(\mathbf{x}^* + \Delta),\ \mathbf{x}^* + \Delta) - 2(\mathbf{b},\ \mathbf{x}^* + \Delta) + c =$$

$$= (\mathbf{Ax}^*,\ \mathbf{x}^*) + (\mathbf{A}\Delta,\ \mathbf{x}^*) + (\mathbf{Ax}^*,\ \Delta) + (\mathbf{A}\Delta,\ \Delta) - 2(\mathbf{b},\ \mathbf{x}^*) - 2(\mathbf{b},\ \Delta) + c =$$

$$= \Phi(\mathbf{x}^*) + (\mathbf{A}\Delta,\ \mathbf{x}^*) + (\mathbf{Ax}^*,\ \Delta) - 2(\mathbf{Ax}^*,\ \Delta) + (\mathbf{A}\Delta,\ \Delta) =$$

$$= \Phi(\mathbf{x}^*) + (\mathbf{A}\Delta,\ \Delta) \geq \Phi(\mathbf{x}^*),$$

в силу самосопряженности и положительности $\mathbf{A}$; значит,

$$\Phi(\mathbf{x}^*) = \min_{\mathbf{x} \in \mathbf{R}_n} \Phi(\mathbf{x}).$$

Следовательно, для нахождения решения нормальной системы (3.7) можно применять различные методы численной минимизации функционала $\Phi(\mathbf{x})$ (в конечном итоге функции n переменных $x_1, x_2, \ldots, x_n$).

Одним из наиболее популярных и хорошо разработанных методов подобного типа является ***метод сопряженных градиентов***. Приведем без вывода алгоритм, быть может, недостаточно подробный, но вполне определенный, чтобы с его помощью можно было решать нормальные СЛАУ (3.7) таким способом [51]. Фигурирующим в нем переменным можно придать следующий оптимизационный смысл:

$\mathbf{x}^{(k)}$ — k-е приближение к искомому решению $\mathbf{x}^*$;

$\xi^{(k)}$ — невязка k-го приближения, играющая роль антиградиента функции $\Phi(\mathbf{x})$;

$\mathbf{p}^{(k)}$ — направление минимизации функции $\Phi(\mathbf{x})$ в точке $\mathbf{x}^{(k)}$;

α_k — коэффициент, обеспечивающий минимум $\Phi(\mathbf{x})$ в направлении вектора $\mathbf{p}^{(k)}$;

$-\beta_k$ — коэффициент при $\mathbf{p}^{(k)}$ в формуле для вычисления направления $\mathbf{p}^{(k+1)}$, обеспечивающий $\mathbf{A}$-сопряженность векторов $\mathbf{p}^{(k)}$ и $\mathbf{p}^{(k+1)}$ (т.е. равенство $\left(\mathbf{A}\mathbf{p}^{(k)}, \mathbf{p}^{(k+1)}\right) = 0$);

$\mathbf{q}^{(k)}$ — вспомогательный вектор.

Алгоритм МСГ

Шаг 1.1. Задать $\mathbf{x}^{(0)}$ (начальный вектор) и число $\varepsilon > 0$ (допустимый уровень абсолютных погрешностей).

Шаг 1.2. Вычислить вектор $\xi^{(0)} := \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$ (невязка начального приближения).

Шаг 1.3. Положить $\mathbf{p}^{(0)} := \xi^{(0)}$,
$k := 0$ (номер итерации).

Шаг 2.1. Вычислить вектор $\mathbf{q}^{(k)} := \mathbf{A}\mathbf{p}^{(k)}$.

Шаг 2.2. Вычислить скаляр $\alpha_k := \left(\xi^{(k)}, \mathbf{p}^{(k)}\right) \Big/ \left(\mathbf{q}^{(k)}, \mathbf{p}^{(k)}\right)$ (шаговый множитель).

Шаг 2.3. Вычислить вектор $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}$ (очередное приближение).

Шаг 2.4. Вычислить вектор $\xi^{(k+1)} = \xi^{(k)} - \alpha_k \mathbf{q}^{(k)}$ (невязка $(k+1)$-го приближения*).

Шаг 2.5. Проверить выполнение неравенства $\left\|\xi^{(k+1)}\right\|_2 \leq \varepsilon$; если оно выполняется, остановить работу алгоритма и вывести результаты.

Шаг 3.1. Вычислить скаляр $\beta_k := \left(\xi^{(k+1)}, \mathbf{q}^{(k)}\right) \Big/ \left(\mathbf{p}^{(k)}, \mathbf{q}^{(k)}\right)$.

Шаг 3.2. Вычислить вектор $\mathbf{p}^{(k+1)} = \xi^{(k+1)} - \beta_k \mathbf{p}^{(k)}$ (новое направление минимизации).

Шаг 3.3. Положить $k := k+1$ и вернуться к шагу 2.1.

Интересно определить место, которое занимает этот метод в общей классификации методов решения линейных алгебраических систем. Дело в том, что метод сопряженных градиентов, являясь по форме итерационным, фактически должен быть отнесен к прямым методам, ибо доказано, что с его помощью минимум квадратичной функции (3.41) от n переменных, иначе, решение n-мерной линейной системы (3.7), достигается ровно за n шагов при любом начальном векторе $\mathbf{x}^{(0)}$. Применяют же метод сопряженных градиентов именно как итерационный метод (что видно и из приведенного алгоритма), имея в виду два обстоятельства. Во-первых, реальный вычислительный процесс может быть довольно далек от идеального и вследствие неизбежных ошибок округления на n-м шаге может быть не достигнута нужная точность. Во-вторых, если размерность n решаемой задачи велика, то число

---

* Полагая $\xi^{(k)} := \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}$, видим, что (в силу 2.3 и 2.1)

$$\xi^{(k+1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k+1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)} - \mathbf{A}\alpha_k \mathbf{p}^{(k)} = \xi^{(k)} - \alpha_k \mathbf{A}\mathbf{p}^{(k)} = \xi^{(k)} - \alpha_k \mathbf{q}^{(k)}.$$

Использование такого выражения невязки $\xi^{(k+1)}$ позволяет обходиться без вычисления вектора $\mathbf{A}\mathbf{x}^{(k+1)}$. Однако нужно понимать, что подобная экономия в арифметических операциях может отразиться на вычислительной устойчивости метода.

шагов, достаточное для получения решения системы с нужной точностью (т.е. выход по критерию 2.4), может оказаться значительно меньшим этой (n) теоретической величины.

Покажем сначала возможности метода сопряженных градиентов на очень простом примере, где точное решение найдется раньше, чем будет выполнен полный цикл предложенного алгоритма.

Пример 3.5. Для решения системы $\begin{cases} 2x_1 + x_2 = 1, \\ x_1 + 2x = -1 \end{cases}$ с положительно определенной матрицей $\mathbf{A} := \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ и правой частью $\mathbf{b} := \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ методом сопряженных градиентов возьмем начальное приближение $\mathbf{x}^{(0)} := \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. Его невязка $\xi^{(0)} := \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(0)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. В соответствии с алгоритмом МСГ далее имеем:

$$\mathbf{p}^{(0)} := \xi^{(0)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \mathbf{q}^{(0)} := \mathbf{A}\,\mathbf{p}^{(0)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

и, следовательно,

$$\alpha_0 := \left(\xi^{(0)},\, \mathbf{p}^{(0)}\right) \Big/ \left(\mathbf{q}^{(0)},\, \mathbf{p}^{(0)}\right) = 1.$$

Таким образом, находим первое приближение

$$\mathbf{x}^{(1)} := \mathbf{x}^{(0)} + \alpha_0 \mathbf{p}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + 1 \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

с невязкой

$$\xi^{(1)} := \xi^{(0)} - \alpha_0 \mathbf{q}^{(0)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} - \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

говорящей о том, что $\mathbf{x}^* = \mathbf{x}^{(1)} := \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

Теперь рассмотрим пример более типичного поведения метода сопряженных градиентов, применяя его к трехмерной СЛАУ с симметричной положительно определенной матрицей коэффициентов.

Пример 3.6. Дана система $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, где $\mathbf{A} := \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & -1 & 2 \end{pmatrix}$, $\mathbf{b} := \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}$.

Продемонстрируем все расчеты, выполняя шаг за шагом предписания алгоритма сопряженных градиентов, игнорируя лишь шаг 2.5, поскольку задавать ϵ здесь нет смысла ввиду малой размерности системы.

Приняв за начальное приближение $\mathbf{x}^{(0)}$ нуль-вектор, далее последовательно вычисляем:

$$\xi^{(0)} := \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(0)} = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}, \quad \mathbf{p}^{(0)} := \xi^{(0)} = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}, \quad \mathbf{q}^{(0)} := \mathbf{A}\,\mathbf{p}^{(0)} = \begin{pmatrix} 0 \\ 4 \\ -5 \end{pmatrix},$$

$$\alpha_0 := \frac{\left(\xi^{(0)},\ \mathbf{p}^{(0)}\right)}{\left(\mathbf{q}^{(0)},\ \mathbf{p}^{(0)}\right)} = \frac{9}{18} = 0{,}5, \quad \mathbf{x}^{(1)} := \mathbf{x}^{(0)} + \alpha_0 \mathbf{p}^{(0)} = \begin{pmatrix} 0{,}5 \\ 1 \\ -1 \end{pmatrix},$$

$$\xi^{(1)} := \xi^{(0)} - \alpha_0 \mathbf{q}^{(0)} = \begin{pmatrix} 1 \\ 0 \\ 0{,}5 \end{pmatrix}, \quad \beta_0 := \frac{\left(\xi^{(1)},\ \mathbf{q}^{(0)}\right)}{\left(\mathbf{q}^{(0)},\ \mathbf{p}^{(0)}\right)} = -\frac{5}{36},$$

$$\mathbf{p}^{(1)} := \xi^{(1)} - \beta_0\,\mathbf{p}^{(0)} = \frac{1}{36} \cdot \begin{pmatrix} 41 \\ 10 \\ 8 \end{pmatrix}, \quad \mathbf{q}^{(1)} := \mathbf{A}\,\mathbf{p}^{(1)} = \frac{1}{36} \cdot \begin{pmatrix} 90 \\ 2 \\ 47 \end{pmatrix},$$

$$\alpha_1 := \frac{\left(\xi^{(1)},\ \mathbf{p}^{(1)}\right)}{\left(\mathbf{q}^{(1)},\ \mathbf{p}^{(1)}\right)} = \frac{90}{227} \approx 0{,}396476, \quad \mathbf{x}^{(2)} := \mathbf{x}^{(1)} + \alpha_1 \mathbf{p}^{(1)} = \begin{pmatrix} 0{,}951542 \\ 1{,}110132 \\ -0{,}911894 \end{pmatrix},$$

$$\xi^{(2)} := \xi^{(1)} + \alpha_1 \mathbf{q}^{(1)} \approx \begin{pmatrix} 0{,}008810 \\ -0{,}022026 \\ -0{,}017621 \end{pmatrix}, \quad \beta_1 := \frac{\left(\xi^{(2)},\ \mathbf{p}^{(1)}\right)}{\left(\mathbf{q}^{(1)},\ \mathbf{p}^{(1)}\right)} \approx -0{,}000699,$$

$$\mathbf{p}^{(2)} := \xi^{(2)} - \beta_1 \mathbf{p}^{(1)} \approx \begin{pmatrix} 0{,}009606 \\ -0{,}021832 \\ -0{,}017466 \end{pmatrix}, \quad \mathbf{q}^{(2)} := \mathbf{A}\,\mathbf{p}^{(2)} \approx \begin{pmatrix} 0{,}001746 \\ -0{,}004366 \\ -0{,}003494 \end{pmatrix},$$

$$\alpha_2 := \frac{\left(\xi^{(2)},\ \mathbf{p}^{(2)}\right)}{\left(\mathbf{q}^{(2)},\ \mathbf{p}^{(2)}\right)} \approx 5{,}044386, \quad \mathbf{x}^{(3)} := \mathbf{x}^{(2)} + \alpha_2 \mathbf{p}^{(2)} = \begin{pmatrix} 0{,}999998 \\ 1{,}000003 \\ -0{,}999999 \end{pmatrix}.$$

Отличие вектора $\mathbf{x}^{(3)}$ от истинного решения $\mathbf{x}^* = (1;\ 1;\ -1)^T$ обусловлено лишь ошибками округления, а о том, что $\mathbf{x}^{(3)}$ действительно можно считать искомым решением с определенной точностью (не зная $\mathbf{x}^*$), следует судить, выполнив еще один неполный шаг, а именно подсчитав невязку $\xi^{(3)}$.

Простейший вариант ***метода минимальных невязок*** определяется совокупностью формул:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \tau_k \xi^{(k)}, \quad \xi^{(k)} = \mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}, \quad \tau_k = \frac{\left(\mathbf{A}\xi^{(k)},\ \xi^{(k)}\right)}{\left(\mathbf{A}\xi^{(k)},\ \mathbf{A}\xi^{(k)}\right)}.$$

Его можно рассматривать как явный двухслойный итерационный процесс (3.38), в котором параметр τ_k на каждом итерационном шаге $k=0,1,2,\ldots$ выбирается таким, чтобы минимизировалась евклидова норма невязки $\xi^{(k+1)}$ получаемого приближения $\mathbf{x}^{(k+1)}$.

Действительно, вычтем из вектора $\xi^{(k+1)}=\mathbf{A}\mathbf{x}^{(k+1)}-\mathbf{b}$ вектор $\xi^{(k)}$. Имеем

$$\xi^{(k+1)}-\xi^{(k)}=\mathbf{A}\mathbf{x}^{(k+1)}-\mathbf{A}\mathbf{x}^{(k)}=\mathbf{A}\mathbf{x}^{(k)}-\mathbf{A}\tau_k\xi^{(k)}-\mathbf{A}\mathbf{x}^{(k)},$$

т.е.

$$\xi^{(k+1)}=\xi^{(k)}-\tau_k\mathbf{A}\xi^{(k)}.$$

Возводя последнее равенство в квадрат (в смысле скалярного умножения векторов), получаем

$$\left(\xi^{(k+1)},\xi^{(k+1)}\right)=\left(\xi^{(k)},\xi^{(k)}\right)-2\tau_k\left(\mathbf{A}\xi^{(k)},\xi^{(k)}\right)+\tau_k^2\left(\mathbf{A}\xi^{(k)},\mathbf{A}\xi^{(k)}\right),$$

или, что то же,

$$\left\|\xi^{(k+1)}\right\|_2^2=\left\|\xi^{(k)}\right\|_2^2-2\tau_k\left(\mathbf{A}\xi^{(k)},\ \xi^{(k)}\right)+\tau_k^2\cdot\left\|\mathbf{A}\xi^{(k)}\right\|_2^2.$$

Легко видеть, что минимум этой положительной квадратичной функции (значит, и величины $\|\xi^{(k+1)}\|$) достигается именно при указанном в записи метода значении τ_k.

В случае нормальной системы для метода минимальных невязок можно получить ту же оценку скорости сходимости, что и для метода простой итерации:

$$\mathbf{x}^{(k+1)}=\left(\mathbf{E}-\tau\mathbf{A}\right)\mathbf{x}^{(k)}+\tau\mathbf{b}$$

при оптимальном значении параметра $\tau_0:=\dfrac{2}{\lambda_{\min}+\lambda_{\max}}$ (в предположении, что известны границы $\lambda_{\min}$ и $\lambda_{\max}$ спектра матрицы $\mathbf{A}$) [61].

Рассмотренные здесь методы далеко не исчерпывают все многообразие итерационных способов решения СЛАУ. В частности, нами совсем не затрагивалась проблема решения больших разреженных систем, где на первый план выходят блочные методы, максимально сохраняющие исходную разреженность матриц (см., например, [9, 27, 50, 75]).

§ 3.5. ИТЕРАЦИОННОЕ ОБРАЩЕНИЕ МАТРИЦ

Согласно утверждениям лемм 3.1, 3.2 (см. § 3.2), если матрица $\mathbf{B} := \mathbf{E} - \mathbf{A}$ мала (в смысле ее нормы или спектрального радиуса), то обратная к $\mathbf{A}$ матрица

$$\mathbf{A}^{-1} = (\mathbf{E} - \mathbf{B})^{-1} = \mathbf{E} + \mathbf{B} + \mathbf{B}^2 + \dots,$$

в принципе, может быть найдена сколь угодно точно приближенным суммированием данного матричного ряда. Однако такой непосредственный подход к вычислению имеет два очевидных недостатка: во-первых, реально его можно применить лишь для обращения матриц, близких к единичной, во-вторых, сходимость последовательностей частичных сумм этого ряда будет медленной даже при достаточно малых нормах матриц $\mathbf{B}$. Поэтому, пользуясь отмеченным фактом лишь как теоретической основой, построим итерационный процесс, определяющий существенно более быстро сходящуюся последовательность приближений к обратной для $\mathbf{A}$ матрице $\mathbf{A}^{-1}$. Будем далее обозначать эти приближения, получаемые на k-м шаге, через $\mathbf{U}_k$, а их ***невязки*** $\mathbf{E} - \mathbf{A}\mathbf{U}_k$ — через $\mathbf{\Psi}_k$.

Лемма 3.3. *Если для матрицы* $\mathbf{A}$ *найдется такая обратимая матрица* $\mathbf{U}_0$, *что модули всех собственных чисел матрицы* $\mathbf{\Psi}_0 := \mathbf{E} - \mathbf{A}\mathbf{U}_0$ *меньше единицы, то матрица* $\mathbf{A}$ *обратима и для обратной матрицы справедливо представление*

$$\mathbf{A}^{-1} = \mathbf{U}_0 (\mathbf{E} - \mathbf{\Psi}_0)^{-1} = \mathbf{U}_0 \left(\mathbf{E} + \mathbf{\Psi}_0 + \mathbf{\Psi}_0^2 + \dots \right). \qquad (3.42)$$

Доказательство. Из равенства

$$\mathbf{A}\mathbf{U}_0 = \mathbf{E} - \mathbf{\Psi}_0, \tag{3.43}$$

в силу обратимости матриц $\mathbf{U}_0$ и $\mathbf{E}-\mathbf{\Psi}_0$ (последнее по лемме 3.1), имеем

$$\mathbf{A} = (\mathbf{E}-\mathbf{\Psi}_0)\mathbf{U}_0^{-1} = \left(\left((\mathbf{E}-\mathbf{\Psi}_0)^{-1}\right)^{-1}\mathbf{U}_0^{-1}\right) = \left(\mathbf{U}_0(\mathbf{E}-\mathbf{\Psi}_0)^{-1}\right)^{-1},$$

т.е. матрица $\mathbf{A}$ обратима и справедливо представление

$$\mathbf{A}^{-1} = \mathbf{U}_0(\mathbf{E}-\mathbf{\Psi}_0)^{-1}.$$

Доказательство завершается разложением матрицы $(\mathbf{E}-\mathbf{\Psi}_0)^{-1}$ в матричный ряд (лемма 3.1).

Очевидным следствием лемм 3.2 и 3.3 является следующая лемма.

Лемма 3.4. *Пусть матрица* $\mathbf{U}_0$ *обратима и* $\|\mathbf{\Psi}_0\| < 1$. *Тогда*:

1) *существует матрица* $\mathbf{A}^{-1}$;

2) *справедливо представление* $\mathbf{A}^{-1}$ *по формуле* (3.42);

3) *имеет место оценка* $\|\mathbf{A}^{-1}\| \le \dfrac{\|\mathbf{U}_0\|}{1-\|\mathbf{\Psi}_0\|}$.

Для построения итерационного процесса зафиксируем в разложении (3.42) $m+1$ первых слагаемых и будем считать первым приближением к $\mathbf{A}^{-1}$ матрицу

$$\mathbf{U}_1 = \mathbf{U}_0\left(\mathbf{E}+\mathbf{\Psi}_0+\ldots+\mathbf{\Psi}_0^m\right).$$

Найдем выражение невязки $\mathbf{\Psi}_1$ этого приближения через невязку $\mathbf{\Psi}_0$ предыдущего (в данном случае начального) приближения $\mathbf{U}_0$:

$$\begin{aligned}\mathbf{\Psi}_1 &:= \mathbf{E}-\mathbf{A}\mathbf{U}_1 = \mathbf{E}-\mathbf{A}\mathbf{U}_0\left(\mathbf{E}+\mathbf{\Psi}_0+\ldots\mathbf{\Psi}_0^m\right) = \\ &= \mathbf{E}-(\mathbf{E}-\mathbf{\Psi}_0)\left(\mathbf{E}+\mathbf{\Psi}_0+\ldots+\mathbf{\Psi}_0^m\right) = \mathbf{E}-\left(\mathbf{E}-\mathbf{\Psi}_0^{m+1}\right) = \mathbf{\Psi}_0^{m+1}.\end{aligned} \tag{3.44}$$

Благодаря полученной связи между невязками можно утверждать, что если выполняются условия лемм 3.3 или 3.4 по отношению к матрицам $\mathbf{U}_0$, $\mathbf{\Psi}_0$, то для матриц $\mathbf{U}_1$, $\mathbf{\Psi}_1$ они тем более будут выполнены. Следовательно, к матрицам $\mathbf{U}_1, \mathbf{\Psi}_1$ можно применить все рассуждения, проведенные выше для $\mathbf{U}_0$, $\mathbf{\Psi}_0$. Таким образом, приходим к итерационному процессу

$$\begin{cases} \mathbf{\Psi}_k = \mathbf{E} - \mathbf{A}\mathbf{U}_k, \\ \mathbf{U}_{k+1} = \mathbf{U}_k\left(\mathbf{E} + \mathbf{\Psi}_k + \ldots + \mathbf{\Psi}_k^m\right), \end{cases} \tag{3.45}$$

где $k = 0, 1, 2, \ldots$ — номер итерации; $\mathbf{U}_0$ — задаваемая начальная матрица, близкая к $\mathbf{A}^{-1}$ в указанном выше смысле, а $m \in \mathbb{N}$ — параметр метода.

Изучим сходимость этого процесса.

Теорема 3.13. *Пусть квадратные матрицы* $\mathbf{A}$ *и* $\mathbf{U}_0$ *таковы, что матрица* $\mathbf{U}_0$ *обратима и* $\|\mathbf{\Psi}_0\| < 1$. *Тогда существует обратная к* $\mathbf{A}$ *матрица* $\mathbf{A}^{-1}$ *и к ней сходится последовательность матриц* $\mathbf{U}_k$, *определяемая итерационным процессом* (3.45). *При этом имеет место точное равенство*

$$\mathbf{A}^{-1} - \mathbf{U}_k = \left(\mathbf{A}^{-1} - \mathbf{U}_0\right)\mathbf{\Psi}_0^{(m+1)^k - 1} \tag{3.46}$$

и справедливы оценки погрешности:

$$1)\ \left\|\mathbf{A}^{-1} - \mathbf{U}_k\right\| \le \frac{\|\mathbf{U}_k\mathbf{\Psi}_k\|}{1 - \|\mathbf{\Psi}_k\|};$$

$$2)\ \left\|\mathbf{A}^{-1} - \mathbf{U}_k\right\| \le \frac{\|\mathbf{U}_0\|}{1 - \|\mathbf{\Psi}_0\|} \cdot \|\mathbf{\Psi}_0\|^{(m+1)^k}.$$

Доказательство. Существование $\mathbf{A}^{-1}$ следует из леммы 3.4. Упомянутая повторяемость рассуждений и выкладок, проведенных на первом итерационном шаге, позволяет считать

очевидными равенства типа (3.42), (3.44) для k-й итерации:

$$\mathbf{A}^{-1} = \mathbf{U}_k(\mathbf{E}-\mathbf{\Psi}_k)^{-1} = \mathbf{U}_k\left(\mathbf{E}+\mathbf{\Psi}_k+\mathbf{\Psi}_k^2+\ldots\right), \tag{3.47}$$

$$\mathbf{\Psi}_k = \mathbf{E}-\mathbf{A}\mathbf{U}_k = \mathbf{\Psi}_{k-1}^{m+1} = \mathbf{\Psi}_{k-2}^{(m+1)^2} = \ldots = \mathbf{\Psi}_0^{(m+1)^k}. \tag{3.48}$$

Из (3.42) имеем

$$\begin{aligned}\mathbf{A}^{-1}-\mathbf{U}_0 &= \mathbf{U}_0\left(\mathbf{E}+\mathbf{\Psi}_0+\mathbf{\Psi}_0^2+\ldots\right)-\mathbf{U}_0 = \\ &= \mathbf{U}_0\left(\mathbf{E}+\mathbf{\Psi}_0+\mathbf{\Psi}_0^2+\ldots\right)\mathbf{\Psi}_0 = \mathbf{A}^{-1}\mathbf{\Psi}_0,\end{aligned} \tag{3.49}$$

а из (3.47) аналогично (с учетом (3.48)) получаем

$$\mathbf{A}^{-1}-\mathbf{U}_k = \mathbf{A}^{-1}\mathbf{\Psi}_k = \mathbf{A}^{-1}\mathbf{\Psi}_0^{(m+1)^k}. \tag{3.50}$$

Заменяя здесь в правой части $\mathbf{A}^{-1}\mathbf{\Psi}_0$ на $\mathbf{A}^{-1}-\mathbf{U}_0$ (см. (3.49)), получаем утверждаемое в теореме равенство (3.46). Переходя в нем к нормам, в соответствии с условием заключаем, что

$$\left\|\mathbf{A}^{-1}-\mathbf{U}_k\right\| \le \left\|\mathbf{A}^{-1}-\mathbf{U}_0\right\|\cdot\left\|\mathbf{\Psi}_0\right\|^{(m+1)^k-1} \xrightarrow[k\to\infty]{} 0,$$

т.е. имеет место сходимость последовательности $\left(\mathbf{U}_k\right)_{k=1}^{\infty}$ к матрице $\mathbf{A}^{-1}$ по норме, а значит, и поэлементная сходимость.

Для доказательства первой оценки (апостериорной) вычтем $\mathbf{U}_k$ из равенства (3.47):

$$\mathbf{A}^{-1}-\mathbf{U}_k = \mathbf{U}_k\left(\mathbf{E}+\mathbf{\Psi}_k+\mathbf{\Psi}_k^2+\ldots\right)-\mathbf{U}_k = \mathbf{U}_k\mathbf{\Psi}_k\left(\mathbf{E}-\mathbf{\Psi}_k\right)^{-1}.$$

Отсюда по лемме 3.2 с учетом (3.48) получаем требуемую оценку 1.

Вторая оценка (априорная) может быть найдена в результате загрубления первой. Но можно вывести ее непосредственно из равенства (3.50), подставив в его правую часть вместо $\mathbf{A}^{-1}$ выражение $\mathbf{U}_0\left(\mathbf{E}-\mathbf{\Psi}_0\right)^{-1}$ (см. (3.42)):

$$\mathbf{A}^{-1}-\mathbf{U}_k = \mathbf{U}_0\left(\mathbf{E}-\mathbf{\Psi}_0\right)^{-1}\mathbf{\Psi}_0^{(m+1)^k}.$$

Переход к нормам в последнем равенстве и привлечение леммы 3.2 завершает доказательство теоремы.

Равенства (3.45) определяют фактически не один, а целое семейство итерационных методов обращения. Фиксированием параметра $m = 1, 2, \ldots$ можно получать конкретные процессы $(m+1)$-го порядка скорости сходимости. Этот порядок может быть сколь угодно большим, однако обычно ограничиваются процессами второго ($m = 1$) и третьего ($m = 2$) порядков. Приоритет процесса второго порядка связан с его простотой и более ранним появлением: первая публикация об этом методе относится к 1933 г. и принадлежит Г. Шульцу, в связи с чем и все семейство (3.45) естественно называть ***методом Шульца****. Метод третьего порядка, как показано М. Альтманом, близок к оптимальному по вычислительным затратам в семействе методов (3.45).

Замечание 3.10. Как сам быстросходящийся итерационный процесс (3.45), так и представленные теоремой 3.13 результаты можно без каких-либо особых дополнительных условий отнести к более общей задаче обращения линейных ограниченных операторов в полных нормированных пространствах.

Процесс (3.45) построения приближений к обратной матрице легко видоизменить подобно тому, как это было сделано с методом простых итераций решения СЛАУ, когда для более оперативного учета получаемой на текущей итерации информации перешли от него к методу Зейделя (см. § 3.3). Например, зейделева модификация метода Шульца второго порядка может быть определена равенствами

$$\begin{cases} \Psi_k = \mathbf{E} - \mathbf{A}\mathbf{U}_k, \\ \mathbf{U}_{k+1} = \mathbf{U}_k + \mathbf{U}_k \underline{\Psi_k} + \mathbf{U}_{k+1} \overline{\Psi_k}, \end{cases} \tag{3.51}$$

где $k = 0, 1, 2, \ldots$; $\Psi_k = \underline{\Psi_k} + \overline{\Psi_k}$, а $\underline{\Psi_k}$ и $\overline{\Psi_k}$ — соответственно нижняя треугольная и строго верхняя треугольная матрицы. При реализации этой модификации — назовем ее ***методом Шульца–Зейделя*** — нужно либо расписывать формулы (3.51) поэле-

---

* В разных литературных источниках можно встретить и другие названия этого метода: ***Хотеллинга***, ***Бодевига*** (Бодвига), а также ***Hobo***.

ментно (чтобы не работать с заведомо нулевыми элементами), либо формировать матрицу $\mathbf{U}_{k+1}$ постепенным замещением старых элементов новыми, осуществляя на k-й итерации цикл присвоений

$$\mathbf{U} := \mathbf{U} + \mathbf{U}\boldsymbol{\Psi},$$

где до начала цикла в правой части в двумерном массиве $\mathbf{U}$ должна содержаться матрица $\mathbf{U}_k$, а в двумерном массиве $\boldsymbol{\Psi}$ — матрица $\boldsymbol{\Psi}_k$ (заполнение массивов новыми элементами производят по строкам).

Если для метода Шульца второго порядка

$$\boldsymbol{\Psi}_k = \mathbf{E} - \mathbf{A}\mathbf{U}_k, \quad \mathbf{U}_{k+1} = \mathbf{U}_k + \mathbf{U}_k\boldsymbol{\Psi}_k, \tag{3.52}$$

служащего отправной точкой для процесса (3.51), невязки соседних приближений связаны равенством $\boldsymbol{\Psi}_k = \boldsymbol{\Psi}_{k-1}^2$ (см. (3.48)), то для метода Шульца–Зейделя аналогичная связь, как нетрудно убедиться, имеет вид

$$\boldsymbol{\Psi}_k = \boldsymbol{\Psi}_{k-1}\underline{\boldsymbol{\Psi}_{k-1}}\left(\mathbf{E} - \overline{\boldsymbol{\Psi}_{k-1}}\right)^{-1}.$$

На ее основе выводится неравенство

$$\left\|\mathbf{A}^{-1} - \mathbf{U}_{k+1}\right\| \le \frac{\left\|\mathbf{U}_k\boldsymbol{\Psi}_k\right\|}{1-\left\|\boldsymbol{\Psi}_k\right\|} \cdot \frac{\left\|\underline{\boldsymbol{\Psi}_k}\right\|}{1-\left\|\overline{\boldsymbol{\Psi}_k}^2\right\|}, \tag{3.53}$$

которое можно назвать ***субапостериорной оценкой погрешности***, поскольку в ее правой части фигурирует не последнее найденное приближение, а предпоследнее.

Из оценки (3.53) при $k := 0$ видно, что если строго нижняя треугольная матрица $\underline{\boldsymbol{\Psi}_0}$ окажется нулевой, то обращение матрицы $\mathbf{A} := (a_{ij})_{i,j=1}^n$ процессом (3.51) может быть выполнено точно за один шаг. Такая ситуация гарантирована, например, в случае, когда матрица $\mathbf{A}$ — верхняя треугольная с ненулевыми элементами диагонали и за начальную принимают матрицу $\mathbf{U}_0 := \operatorname{diag}\left(1/a_{ii}\right)$

(убедитесь, что процессу Шульца (3.52) для этого потребуется $n-1$ итераций; аналогичная (3.53) субапостериорная оценка для него есть $\left\|\mathbf{A}^{-1}-\mathbf{U}_{k+1}\right\| \le \frac{\left\|\mathbf{U}_k\boldsymbol{\Psi}_k\right\|}{1-\left\|\boldsymbol{\Psi}_k\right\|}\cdot\frac{\left\|\boldsymbol{\Psi}_k\right\|}{1-\left\|\boldsymbol{\Psi}_k{}^2\right\|}$).

Рассмотренный тривиальный случай с обращением треугольной матрицы методом Шульца–Зейделя (который, кстати, можно принять за некоторую новую схему реализации метода Гаусса:

$$\mathbf{U}:=\mathbf{D}^{-1},\quad \boldsymbol{\Psi}:=-\mathbf{RU},\quad \mathbf{U}:=\mathbf{U}+\mathbf{U}\boldsymbol{\Psi},\quad \mathbf{A}^{-1}:=\mathbf{U},$$

если в разложении $\mathbf{A}=\mathbf{L}+\mathbf{D}+\mathbf{R}$ строго нижняя треугольная матрица $\mathbf{L}$ — нулевая) позволяет рассчитывать на преимущества модификации (3.51) над методом Шульца (3.52) при обращении асимметричных матриц. При этом, если в обращаемой матрице доминирует нижний треугольник, несложно перейти к «зеркальному» по отношению к (3.51) процессу, изменив порядок перебора индексов при перемножении матриц.

Потребность метода Шульца и его модификаций в хорошем начальном приближении $\mathbf{U}_0$, обеспечивающем малость начальной невязки $\boldsymbol{\Psi}_0$, ограничивает сферу применения таких методов и не позволяет считать их универсальными способами обращения матриц. Представляются возможными следующие варианты с выбором $\mathbf{U}_0$:

А) Хорошее приближение к $\mathbf{A}^{-1}$ известно из постановки задачи приближенного обращения.

Б) Найдено грубое приближение к $\mathbf{A}^{-1}$ каким-либо иным методом и требуется его уточнение.

В) Матрица $\mathbf{A}$ имеет диагональное преобладание. Тогда за $\mathbf{U}_0$ можно принять матрицу $\operatorname{diag}\left(1/a_{ii}\right)$.

Г) Матрица $\mathbf{A}$ — симметричная положительно определенная с верхней границей спектра β. Тогда можно принять $\mathbf{U}_0:=\alpha\mathbf{E}$, где $\alpha\in\left(0,2/\beta\right)$.

Д) Матрица $\mathbf{A}$ — произвольная невырожденная и γ — верхняя граница спектра матрицы $\mathbf{AA}^T$. В этом случае можно положить $\mathbf{U}_0 := \alpha \mathbf{A}^T$ с $\alpha \in (0, 2/\gamma)$.

В последних двух вариантах гарантируется малость спектрального радиуса матрицы $\mathbf{\Psi}_0$ (т.е. выполнимость условия леммы 3.3) [7], что влечет, в принципе, сходимость метода Шульца; но при этом может оказаться $\|\mathbf{\Psi}_0\| > 1$ и, как следствие, «неработающие» оценки и затянутый переходный процесс к проявлению характеризуемой этими оценками быстрой сходимости. Без знания величин β и γ здесь можно обойтись, положив $\alpha := 1/\|\mathbf{A}\|$ в варианте Г и $\alpha := 1/\|\mathbf{AA}^T\|$ в варианте Д, что допустимо в силу известного неравенства между спектральным радиусом и нормой.

УПРАЖНЕНИЯ

3.1. Запишите расчетные формулы для решения методом простых итераций следующих систем: а) $\begin{cases} 2x - y = 5, \\ x + 7y = -5 \end{cases}$; б) $\begin{cases} x + 7y = 19, \\ 8x - 2y = 13 \end{cases}$. Найдите их решения с точностью до 0,01 по норме-максимум.

3.2. Для системы

$$\begin{cases} 10x_1 + x_2 + 2x_3 = 1, \\ 2x_1 + 10x_2 + 3x_3 = 2, \\ 3x_1 + 4x_2 + 10x_3 = 3 \end{cases}$$

запишите совокупность формул метода простых итераций, позволяющих получить ее решение с заданной точностью $\varepsilon := 0{,}001$. За сколько шагов можно гарантированно получить решение с такой точностью ?

3.3. Через сколько шагов сойдется начатый с нуль-вектора итерационный процесс

$$\begin{cases} x_{k+1} = 0{,}2x_k - 0{,}2z_k + 0{,}7, \\ y_{k+1} = 0{,}12x_k - 0{,}3y_k + 0{,}18z_k + 1{,}5, \\ z_{k+1} = 0{,}1x_k + 0{,}2y_k - 0{,}2z_k - 0{,}8 \end{cases} \quad (k = 0,\ 1,\ 2,\ \ldots)$$

с точностью до 0,001 по норме-максимум?

3.4. Запишите расчетные формулы для решения системы $\begin{cases} 6x+2y=-8, \\ 5x-y=-12 \end{cases}$ методом Зейделя.

3.5. Подготовьте расчетные формулы для решения системы

$$\begin{cases} 2x-7y+z=15, \\ 3x-y+8z=12, \\ 10x-2y+3z=18 \end{cases}$$

с точностью $\varepsilon := 10^{-5}$ методами Якоби и Зейделя.

3.6. Проверьте, выполняются ли необходимые условия сходимости методов Якоби и Зейделя, примененных к системе

$$\begin{cases} x_1 + x_2 = 2, \\ x_1 + 2x_2 + x_3 = 4, \\ x_2 + 2x_3 = 3. \end{cases}$$

3.7. Сделайте по пять итераций методов Якоби и Зейделя для системы

$$\begin{cases} 10x_1 + x_2 - 2x_3 = 10, \\ x_1 - 5x_2 + x_3 = 10, \\ 3x_1 - x_2 + 10x_3 = -5. \end{cases}$$

Сколько верных знаков можно гарантировать в приближенных решениях, полученных тем и другим способами?

3.8. Докажите, что при любом начальном векторе $\left(x^{(0)};\ y^{(0)};\ z^{(0)}\right)^T$ последовательности векторов $\left(x_1^{(k)};\ y_1^{(k)};\ z_1^{(k)}\right)^T$ и $\left(x_2^{(k)};\ y_2^{(k)};\ z_2^{(k)}\right)^T$, определяемые при $k = 0,\ 1,\ 2,\ \ldots$ равенствами

$$\begin{cases} x_1^{(k+1)} = 0{,}1x_1^{(k)} + 0{,}2y_1^{(k)} - 3, \\ y_1^{(k+1)} = 0{,}2x_1^{(k)} - 0{,}1y_1^{(k)} + 0{,}1z_1^{(k)} + 2, \\ z_1^{(k+1)} = -0{,}3x_1^{(k)} + 0{,}2z_1^{(k)} - 1 \end{cases}$$

и

$$\begin{cases} x_2^{(k+1)} = \left(2y_2^{(k)} - 30\right)/9, \\ y_2^{(k+1)} = \left(2x_2^{(k)} + z_2^{(k)} + 20\right)/11, \\ z_2^{(k+1)} = -\left(3x_2^{(k)} + 10\right)/8, \end{cases}$$

сходятся, причем к одному и тому же предельному вектору $\left(x^*;\ y^*;\ z^*\right)^T$.

Запишите линейную систему (в стандартном виде), решением которой служит этот предельный вектор. За сколько шагов итераций по данным формулам можно получить предельный вектор с точностью $\varepsilon := 10^{-6}$ (по норме-максимум), если начать счет с нулевого вектора?

3.9. Для линейной системы

$$\begin{cases} 10x_1 + 2x_2 + 3x_3 + 4x_4 = 1, \\ 2x_1 + 5x_2 + x_3 = -6, \\ 3x_1 + x_2 + 10x_3 - x_4 = -7, \\ 4x_1 - x_3 + 10x_4 = -6 \end{cases}$$

запишите метод Зейделя и обоснуйте его сходимость. Каковы расчетные формулы метода ПВР в этом случае?

3.10. Убедитесь, что к системе

$$\begin{cases} x_1 + 2x_2 = 3, \\ 2x_1 + 2x_2 + x_3 = 5, \\ x_2 + 2x_3 = 3 \end{cases}$$

неприменим напрямую метод Якоби и что ее матрица не является положительно определенной. Выполните симметризацию Гаусса и убедитесь, что новая система нормальная. Запишите для нее процессы Зейделя и релаксации. Опробуйте последний при значениях параметра релаксации ω, больших и меньших единицы (например, полагая $\omega := 1{,}2$ и $\omega := 0{,}8$). Сравните результаты применения методов верхней, нижней и полной релаксации.

3.11. Покажите, что метод Якоби решения системы $\mathbf{Ax} = \mathbf{b}$ можно считать стационарным методом простых итераций вида $\mathbf{x}^{(k+1)} = (\mathbf{E} - \mathbf{HA})\mathbf{x}^{(k)} + \mathbf{Hb}$ с матрицей $\mathbf{H} := \mathbf{D}^{-1}$.

3.12. Дана система

$$\begin{cases} 7x_1 + 5x_2 + x_3 = 2{,}2, \\ 5x_1 + 8x_2 + 2x_3 = 2{,}4, \\ x_1 + 2x_2 + 4x_3 = 1{,}6. \end{cases}$$

Найдите четвертое приближение к ее решению по методу минимальных невязок, начиная итерационный процесс с нулевого вектора. За сколько итераций по методу Якоби достигается примерно такая же величина евклидовой нормы невязки? Сравните вычислительные затраты, требующиеся для реализации одного шага каждого из этих методов.

3.13. Методом сопряженных градиентов решите систему

$$\begin{cases} 3x_1 + 2x_2 + x_3 = 3, \\ 2x_1 + 2x_2 - x_3 = -3, \\ x_1 - x_2 + 2x_3 = 6. \end{cases}$$

3.14. Запишите процесс Шульца для обращения матрицы $\mathbf{A} := \begin{pmatrix} 1 & -1 \\ -2 & 5 \end{pmatrix}$, начинающийся с $\mathbf{U}_0 := \begin{pmatrix} 1{,}6 & 0{,}3 \\ 0{,}6 & 0{,}3 \end{pmatrix}$. Будет ли он сходящимся?

3.15. Даны матрицы $\mathbf{A} := \begin{pmatrix} 1 & -2 & 3 \\ -1 & 1 & 2 \\ 2 & -1 & -1 \end{pmatrix}$ и $\mathbf{U}_0 := \begin{pmatrix} -0{,}1 & 0{,}6 & 0{,}9 \\ -0{,}4 & 0{,}9 & 0{,}6 \\ 0{,}1 & 0{,}4 & 0{,}1 \end{pmatrix}$.

А) Подсчитав невязку $\mathbf{\Psi}_0 := \mathbf{E} - \mathbf{A}\mathbf{U}_0$, убедитесь в существовании матрицы $\mathbf{A}^{-1}$ и оцените какую-либо ее норму.

Б) Сделайте по два приближения к $\mathbf{A}^{-1}$ методом Шульца второго и третьего порядков и оцените близость полученных приближений к $\mathbf{A}^{-1}$.

В) Сравните о ц е н к и погрешностей приближений с истинными ошибками, найдя $\mathbf{A}^{-1}$ каким-нибудь прямым методом.

3.16. На матрице $\mathbf{A} := \begin{pmatrix} 10 & 3 & -1 \\ 0{,}5 & 5 & -2 \\ 0{,}1 & 1 & 10 \end{pmatrix}$ сравните поведение методов Шульца второго порядка (формула (3.45) при $m := 1$) и Шульца–Зейделя (3.51), выполнив по две итерации с начальной матрицей $\mathbf{U}_0$, взятой с учетом диагонального преобладания в данной матрице.

Глава 4

ЗАДАЧИ НА СОБСТВЕННЫЕ ЗНАЧЕНИЯ

§ 4.1. СОБСТВЕННЫЕ ПАРЫ МАТРИЦЫ И ИХ ПРОСТЕЙШИЕ СВОЙСТВА

Одна из задач, очевидным образом приводящая к так называемой алгебраической проблеме собственных значений, следующая.

Пусть $\mathbf{A}$ — вещественная $n\times n$-матрица, $\mathbf{y}=\mathbf{y}(t)$ — n-мерная векторная функция скалярного аргумента t, и пусть необходимо найти нетривиальные решения системы дифференциальных уравнений

$$\frac{d\mathbf{y}}{dt}=\mathbf{A}\mathbf{y} \tag{4.1}$$

в виде $\mathbf{y}=e^{\lambda t}\mathbf{x}$, где $\mathbf{x}\in\mathbb{C}_n$, $\lambda\in\mathbb{C}$. Подставляя $\mathbf{y}$ и $\frac{d\mathbf{y}}{dt}$ в уравнение (4.1), получаем

$$\lambda e^{\lambda t}\mathbf{x}=\mathbf{A}e^{\lambda t}\mathbf{x},$$

т.е. система (4.1) действительно будет иметь решения заданного вида в том и только том случае, если будут найдены такие пары чисел λ и ненулевых векторов $\mathbf{x}$, что

$$\mathbf{A}\mathbf{x}=\lambda\mathbf{x}. \tag{4.2}$$

Имеется ряд других примеров из областей, лежащих за пределами линейной алгебры, в которых также приходят к необходимости решать подобные (4.2) алгебраические задачи, называемые ***задачами на собственные значения*** (см., например, [38]). При этом различают ***полную*** (***алгебраическую***, или, иначе, ***матричную***) ***проблему собственных значений***, предполагающую нахождение всех ***собственных пар*** $\{\lambda,\mathbf{x}\}$ матрицы $\mathbf{A}$, и ***частичные проблемы собственных значений***, состоящие, как правило, в нахождении одного или нескольких ***собственных чисел*** λ и, возможно, соответствующих им

собственных векторов* $\mathbf{x}$. Чаще всего в последнем случае речь идет о нахождении наибольшего и наименьшего по модулю собственных чисел; знание таких характеристик матрицы позволяет, например, делать заключения о сходимости тех или иных итерационных методов, оптимизировать параметры итерационных методов, учитывать влияние на результаты решения алгебраических задач погрешностей исходных данных и вычислительных погрешностей (потребность в таких числах неоднократно возникала в гл. 3). Встречаются и несколько иные постановки частичных проблем [3, 4, 19, 38].

Трактуя $\mathbf{A}$ в равенстве (4.2) как матрицу линейного преобразования в пространстве $\mathbb{R}_n$, задачу на собственные значения можно сформулировать так: для каких чисел λ и ненулевых векторов $\mathbf{x}$ линейное преобразование вектора с помощью матрицы $\mathbf{A}$ не изменяет направления этого вектора в $\mathbb{R}_n$, т.е. сводится к «растяжению» этого вектора в λ раз? Такая задача, очевидно, эквивалентна задаче исследования однородной СЛАУ с параметром: при каких λ система

$$(\mathbf{A}-\lambda\mathbf{E})\mathbf{x}=\mathbf{0} \tag{4.3}$$

имеет нетривиальные решения? Найти эти решения.

Теоретически эта задача легко решаема: нужно найти корни так называемого ***характеристического***, или, иначе, ***«векового»***, ***уравнения***

$$\det(\mathbf{A}-\lambda\mathbf{E})=0 \tag{4.4}$$

и, подставляя их поочередно в (4.3), получать из соответствующих вырожденных систем собственные векторы.

Практическая реализация такого, в сущности, простого подхода сопряжена с рядом трудностей, возрастающих с ростом размерности решаемой задачи. Трудности эти, в частности, обусловлены необходимостью развертывать ***«вековой определитель»*** $\det(\mathbf{A}-\lambda\mathbf{E})$ и вычислять корни получающегося здесь многочлена

---

* Выражения «собственное число» и «собственное значение» в данном контексте — синонимы. В более общем случае, когда $\mathbf{A}$ в (4.2) — некоторый оператор, ***собственные элементы*** $\mathbf{x}$ могут иметь другую природу.

n-й степени, а также поиском линейно независимых решений вырожденных СЛАУ. В связи с этим описанный непосредственный подход к решению алгебраической проблемы собственных значений обычно применяют лишь при очень малых размерах матриц $\mathbf{A}$ $(n=2,3)$; уже при $n\geq 4$ на первый план выходят специальные численные методы решения подобных задач. Далее рассмотрены некоторые из этих методов в таком ключе, чтобы можно было понять идеи, лежащие в их основе, и в то же время получить возможность решать поставленные задачи до конца для некоторых классов матриц (более полное и глубокое изложение этой темы см. в монографиях [23, 26, 32, 54, 57, 67, 69] и в учебных пособиях [1, 3, 4, 7, 17, 25, 34, 42]).

Следует заметить, что в недалеком прошлом численные методы решения задач на собственные значения опирались, как правило, на классический подход, т.е. на развертывание «вековых определителей», в частности, в простейшем случае с помощью приведения матрицы $\mathbf{A}$ подходящим преобразованием к так называемой ***сопровождающей матрице***

$$\mathbf{C}:=\begin{pmatrix} c_1 & c_2 & c_3 & \dots & c_{n-1} & c_n \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix},$$

где в первой строке стоят коэффициенты уравнения (4.4), записанного в виде

$$(-1)^n\left(\lambda^n - c_1\lambda^{n-1} - c_2\lambda^{n-2} - \dots - c_{n-1}\lambda - c_n\right)=0.$$

Современные методы решения полной проблемы ориентированы на алгоритмическое построение из матрицы $\mathbf{A}$ такой матрицы, определенные элементы которой являлись бы приближенными значениями собственных чисел $\mathbf{A}$, причем параллельно, по возможности, формировались бы и ее собственные векторы.

Прежде чем приступить к изучению методов нахождения собственных чисел и векторов, вспомним их некоторые простые свойства, требующиеся в дальнейшем.

Свойство 4.1. *Если* $\{\lambda, \mathbf{x}\}$ *— собственная пара матрицы* $\mathbf{A}$, *а* $\alpha (\neq 0)$ *— некоторое число, то* $\{\lambda, \alpha\mathbf{x}\}$ *также является собственной парой для* $\mathbf{A}$.

Действительно, умножив верное для данных λ и $\mathbf{x}$ равенство (4.2) на число α, получаем верное равенство

$$\mathbf{A}(\alpha\mathbf{x}) = \lambda(\alpha\mathbf{x}).$$

Оно означает, что каждому собственному числу λ соответствует бесчисленное множество собственных векторов, различающихся лишь скалярным множителем. Такие векторы задают одно и то же направление в n-мерном пространстве; в соответствие этому направлению можно поставить нормированный вектор или орт (вообще говоря, одному собственному числу может соответствовать и несколько линейно независимых собственных векторов).

Свойство 4.2. *Пусть* $\{\mu, \mathbf{x}\}$ *— собственная пара матрицы* $\mathbf{A} - p\mathbf{E}$ *при некотором* $p \in \mathbb{R}$. *Тогда* $\{\lambda := \mu + p, \mathbf{x}\}$ *— собственная пара матрицы* $\mathbf{A}$.

Чтобы убедиться в этом, заметим, что по условию равенство

$$(\mathbf{A} - p\mathbf{E})\mathbf{x} = \mu\mathbf{x} \tag{4.5}$$

при данных μ и $\mathbf{x}$ — верное. Рассмотрим равенство $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ при $\lambda := \mu + p$:

$$\mathbf{A}\mathbf{x} = (\mu + p)\mathbf{x}.$$

Оно равносильно (4.5), и значит, справедливо; с другой стороны, говорит о том, что $\{\lambda, \mathbf{x}\}$ — собственная пара матрицы $\mathbf{A}$.

Как видим, *прибавление к данной матрице* $\mathbf{A}$ *скалярной матрицы* $p\mathbf{E}$ *не изменяет ее собственных векторов и смещает спектр исходной матрицы на число* p (влево при $p > 0$).

Свойство 4.3. *Если* $\{\lambda, \mathbf{x}\}$ *— собственная пара обратимой матрицы* $\mathbf{A}$, *то* $\{1/\lambda, \mathbf{x}\}$ *— собственная пара матрицы* $\mathbf{A}^{-1}$.

Справедливость этого свойства почти очевидна: умножив верное для данных λ и $\mathbf{x}$ равенство $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ слева на матрицу $\frac{1}{\lambda}\mathbf{A}^{-1}$, получаем равенство $\frac{1}{\lambda}\mathbf{x} = \mathbf{A}^{-1}\mathbf{x}$, что и означает утверждаемое.

Свойство 4.4. *Собственными числами диагональных и треугольных матриц являются их диагональные элементы.*

Этот факт легко усматривается из очевидного представления характеристических уравнений (4.4) для таких матриц в виде

$$\prod_{i=1}^{n}(\lambda - a_{ii}) = 0.$$

Последнее равенство свидетельствует о том, что

диагональные и треугольные вещественные матрицы имеют только вещественные собственные значения (ровно n с учетом возможной их кратности). *Вещественность всех собственных чисел (**спектра**) присуща и очень важному в приложениях классу симметричных матриц* [1, 18, 64].

Определение 4.1. ***Отношением Рэлея**** *для $n \times n$-матрицы $\mathbf{A}$ называется функционал $\rho(\mathbf{x}) := \dfrac{(\mathbf{A}\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}$, определенный на множестве ненулевых n-мерных векторов $\mathbf{x}$.*

Свойство 4.5. *Пусть $\mathbf{x}^*$ — собственный вектор матрицы $\mathbf{A}$, тогда $\rho(\mathbf{x}^*)$ — ее собственное число.*

Для доказательства этого утверждения обозначим через λ^* собственное число матрицы $\mathbf{A}$, соответствующее вектору $\mathbf{x}^*$. Подставляя $\mathbf{A}\mathbf{x}^* = \lambda^*\mathbf{x}^*$ в вытекающее из определения 4.1 равенство

$$\left(\mathbf{A}\mathbf{x}^*,\ \mathbf{x}^*\right) = \rho\left(\mathbf{x}^*\right)\left(\mathbf{x}^*,\ \mathbf{x}^*\right),$$

---

* Лорд Рэле́й (до получения титула лорда — Стретт Джон Уильям (1842–1919)) — английский физик, один из основоположников теории колебаний.

имеем

$$\lambda^*\left(\mathbf{x}^*,\ \mathbf{x}^*\right)=\rho\left(\mathbf{x}^*\right)\left(\mathbf{x}^*,\ \mathbf{x}^*\right),$$

откуда после деления на $\left(\mathbf{x}^*,\ \mathbf{x}^*\right)\neq 0$ получаем утверждаемое: $\lambda^*=\rho(\mathbf{x}^*)$.

Отношение Рэлея обладает рядом других ценных свойств. Например, если матрица $\mathbf{A}$ — симметричная положительно определенная со спектром $\lambda_1>\lambda_2>\ \ldots\ >\lambda_n$, то $\max\ \rho(\mathbf{x})=\lambda_1$, $\min\ \rho(\mathbf{x})=\lambda_n$, $\rho(\mathbf{x})\in[\lambda_n,\lambda_1]$ при любых n-мерных ненулевых векторах $\mathbf{x}$ и, кроме того, равенство $\operatorname{grad}\rho(\mathbf{x})=\mathbf{0}$ справедливо тогда и только тогда, когда $\mathbf{x}$ — собственный вектор матрицы $\mathbf{A}$ (см. [1, 54]). Эти свойства служат основой для некоторых способов локализации собственных значений и построения градиентных методов их вычисления.

В дальнейшем (§ 4.2, 4.3) будет полезно следующее экстремальное свойство отношения Рэлея.

Свойство 4.6. *Минимум евклидовой нормы вектора* $\xi(\lambda):=\mathbf{A}\mathbf{x}-\lambda\mathbf{x}$ *для любого фиксированного ненулевого вектора* $\mathbf{x}$ *достигается при* $\lambda=\rho(\mathbf{x})$.

Смысл этого факта в следующем: *если некоторый вектор* $\mathbf{x}$ *считать приближением к собственному вектору матрицы* $\mathbf{A}$ (а значит, вектор ξ — его ***невязкой***), *то отношение Рэлея* $\rho(\mathbf{x})$ *будет наилучшим приближением к соответствующему этому вектору собственному числу в смысле евклидовой метрики.*

Доказательство свойства 4.6 для симметричных $\mathbf{A}$ и вещественных $\mathbf{x}$ весьма просто. Действительно, рассмотрим квадрат евклидовой нормы невязки

$$\|\xi(\lambda)\|_2^2=(\mathbf{A}\mathbf{x}-\lambda\mathbf{x},\ \mathbf{A}\mathbf{x}-\lambda\mathbf{x})=$$

$$=(\mathbf{A}\mathbf{x},\ \mathbf{A}\mathbf{x})-2\lambda(\mathbf{A}\mathbf{x},\ \mathbf{x})+\lambda^2(\mathbf{x},\ \mathbf{x})=q(\lambda)(\mathbf{x},\ \mathbf{x}),$$

где $q(\lambda):=\lambda^2-2\lambda\rho(\mathbf{x})+\dfrac{(\mathbf{A}\mathbf{x},\ \mathbf{A}\mathbf{x})}{(\mathbf{x},\ \mathbf{x})}$. Очевидно, квадратный

трехчлен $q(\lambda)$ всегда имеет минимум при $\lambda := \rho(\mathbf{x})$, а поскольку $(\mathbf{x}, \mathbf{x}) > 0$, это значение λ доставляет минимум величине $\|\xi(\lambda)\|_2^2$ и, следовательно, величине $\|\xi(\lambda)\|_2$.

Следующие два свойства связаны с мультипликативной операцией над матрицами, называемой ***преобразованием подобия.***

Определение 4.2. *Матрицы* $\mathbf{A}$ *и* $\mathbf{B}$ ***называются подобными,*** *если они связаны соотношением* $\mathbf{B} = \mathbf{C}^{-1}\mathbf{AC}$, *где* $\mathbf{C}$ *— произвольная невырожденная матрица.*

Свойство 4.7. *Пусть* $\{\lambda, \mathbf{x}\}$ *— собственная пара матрицы* $\mathbf{B} = \mathbf{C}^{-1}\mathbf{AC}$. *Тогда* $\{\lambda, \mathbf{Cx}\}$ *— собственная пара матрицы* $\mathbf{A}$.

Чтобы убедиться в справедливости этого свойства, достаточно подставить выражение $\mathbf{B} = \mathbf{C}^{-1}\mathbf{AC}$ в верное для пары $\{\lambda, \mathbf{x}\}$ равенство $\mathbf{Bx} = \lambda\mathbf{x}$: имеем $\mathbf{C}^{-1}\mathbf{ACx} = \lambda\mathbf{x}$, откуда после умножения слева на матрицу $\mathbf{C}$ получаем равенство $\mathbf{ACx} = \lambda\mathbf{Cx}$, означающее истинность утверждения.

Как видим, *преобразование подобия сохраняет неизменным спектр любой матрицы.*

Определение 4.3. *Матрица* $\mathbf{A}$ *размера* $n \times n$ ***называется матрицей простой структуры,*** *если она имеет ровно* n *линейно независимых собственных векторов.*

Свойство 4.8. *Пусть* $\mathbf{A}$ *—* $n \times n$*-матрица простой структуры, а матрицы* $\Lambda := \operatorname{diag}(\lambda_i)$ *и* $\mathbf{X} := (\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_n)$ *образованы из ее собственных чисел и собственных векторов соответственно. Тогда справедливо равенство* $\Lambda = \mathbf{X}^{-1}\mathbf{AX}$.

Действительно, то, что $\{\lambda_i, \mathbf{x}_i\}$ являются собственными парами матрицы $\mathbf{A}$, означает, что

$$\mathbf{Ax}_i = \lambda_i \mathbf{x}_i \quad \forall i \in \{1, 2, \ldots, n\}.$$

Нетрудно убедиться, что эти n векторных равенств могут быть записаны в виде одного матричного равенства

$$\mathbf{AX} = \mathbf{X\Lambda}.$$

В силу простой структуры матрицы $\mathbf{A}$, все ее собственные векторы, т.е. столбцы матрицы $\mathbf{X}$, линейно независимы, поэтому матрица $\mathbf{X}$ обратима. Умножив последнее равенство слева на матрицу $\mathbf{X}^{-1}$, получим нужное представление $\mathbf{\Lambda} = \mathbf{X}^{-1}\mathbf{AX}$.

Так как для диагональной матрицы $\mathbf{\Lambda}$, образованной из собственных чисел, собственными векторами могут служить единичные векторы исходного базиса (действительно, $\mathbf{\Lambda e}_i = \lambda_i \mathbf{e}_i$, $\forall i \in \{1, 2, \ldots, n\}$), то, применяя к последнему случаю свойство 4.7 с $\mathbf{C} := \mathbf{X}$ и с $\mathbf{x} := \mathbf{e}_i$ (т.е. с $\mathbf{Cx} = \mathbf{Xe}_i = \mathbf{x}_i$), приходим к другой формулировке свойства 4.8:

Если $\{\lambda_i, \mathbf{e}_i\}$ *является собственной парой матрицы* $\mathbf{\Lambda} = \operatorname{diag}(\lambda_i) = \mathbf{X}^{-1}\mathbf{AX}$, *то* $\{\lambda_i, \mathbf{x}_i\}$ *есть собственная пара матрицы* $\mathbf{A}$ (обозначения те же, что и выше).

Родственным преобразованию подобия является матричное ***преобразование конгруэнтности.***

Определение 4.4. *Матрицы* $\mathbf{A}$ *и* $\mathbf{B}$ ***называются конгруэнтными,*** *если существует такая невырожденная матрица* $\mathbf{P}$, *что* $\mathbf{B} = \mathbf{P}^T\mathbf{AP}$.

Примером конгруэнтного преобразования может служить рассмотренное в § 1.3 $\mathbf{U}^T\mathbf{DU}$-разложение симметричных матриц.

Очевидно, если в определяющем конгруэнтность $\mathbf{A}$ и $\mathbf{B}$ равенстве матрица $\mathbf{P}$ окажется ортогональной, то из $\mathbf{P}^T = \mathbf{P}^{-1}$ в соответствии с определением 4.2 следует подобие матриц $\mathbf{A}$ и $\mathbf{B}$.

Преобразование конгруэнтности играет важную роль при построении метода бисекции нахождения собственных чисел симметричных вещественных матриц (§ 4.5). Отметим два свойства такого преобразования.

Свойство 4.9. *Симметричную матрицу конгруэнтное преобразование переводит в симметричную.*

Действительно, если $\mathbf{A}^T = \mathbf{A}$, то

$$\mathbf{B}^T = (\mathbf{P}^T\mathbf{A}\mathbf{P})^T = \mathbf{P}^T\mathbf{A}^T(\mathbf{P}^T)^T = \mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{B}.$$

Второе свойство связано с так называемым ***законом инерции***, характеризующим один замечательный инвариант конгруэнции симметричных матриц (и соответствующих им квадратичных форм). Имеется много редакций этого закона [18, 20, 26, 46, 54 и др.]. Одна из формулировок связана с понятием инерции матрицы.

Определение 4.5. ***Инерцией*** *симметричной матрицы* $\mathbf{A}$ *называется тройка целых чисел* (m_-, m_0, m_+), *где* m_-, m_0 *и* m_+ — *количество отрицательных, нулевых и положительных собственных чисел матрицы* $\mathbf{A}$ *соответственно.*

Свойство 4.10 (***теорема Сильвестра****). *Симметричные матрицы* $\mathbf{A}$ *и* $\mathbf{B}$, *связанные соотношением конгруэнтности, имеют одну и ту же инерцию.*

(Доказательство см., например, в [26].)

Сформулированное свойство означает, что если симметричную матрицу конгруэнтными преобразованиями привели, например, к диагональному виду, то по знакам элементов диагонали преобразованной матрицы можно точно определить, сколько положительных, отрицательных и нулевых собственных чисел имеется в спектре исходной матрицы.

Заметим при этом, что если матрица $\mathbf{A}$ имеет размер $n \times n$, то

$$m_- + m_0 + m_+ = n,$$
$$m_- + m_+ = \operatorname{rank} \mathbf{A}$$

(разность $m_+ - m_-$ ***индексов положительности*** и ***отрицательности*** называют ***сигнатурой*** матрицы $\mathbf{A}$).

---

* Сильве́стр Джеймс Джозеф (1814–1897) — английский математик. Известен своими трудами по алгебре, теории чисел, теории вероятностей, механике и математической физике. Основатель первого американского математического журнала.

§ 4.2. СТЕПЕННОЙ МЕТОД

Рассмотрим простейший метод решения частичных проблем собственных значений, который вряд ли может быть отнесен к широко применяемым методам решения таких задач, но который много значит для понимания и построения других, более эффективных методов.

Пусть о вещественной $n \times n$-матрице $\mathbf{A}$ известно, что она является матрицей простой структуры. Тогда у нее есть ровно n линейно независимых собственных векторов, имеющих в исходном базисе пространства $\mathbb{R}_n$ представление

$$\mathbf{x}_1 := \begin{pmatrix} x_{11} \\ x_{12} \\ \dots \\ x_{1n} \end{pmatrix}, \quad \mathbf{x}_2 := \begin{pmatrix} x_{21} \\ x_{22} \\ \dots \\ x_{2n} \end{pmatrix}, \quad \dots, \quad \mathbf{x}_n := \begin{pmatrix} x_{n1} \\ x_{n2} \\ \dots \\ x_{nn} \end{pmatrix} \tag{4.6}$$

и могущих также служить базисом n-мерного векторного пространства. Будем считать, что нумерация этих векторов отвечает упорядочению соответствующих им собственных чисел по убыванию модулей (где первое из неравенств — строгое):

$$|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|. \tag{4.7}$$

Ставим **задачу приближенного вычисления наибольшего по модулю собственного числа** λ_1 (вещественного, в силу предположения о строгом доминировании его модуля) **и соответствующего ему собственного вектора** $\mathbf{x}_1$ данной матрицы $\mathbf{A}$.

Возьмем произвольный ненулевой вектор $\mathbf{y}^{(0)}$ и запишем его разложение по базису из собственных векторов $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$:

$$\mathbf{y}^{(0)} := c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \dots + c_n\mathbf{x}_n. \tag{4.8}$$

При этом без ограничения общности можно считать, что $c_1 \neq 0$, так как в противном (маловероятном) случае можно взять другой начальный вектор $\mathbf{y}^{(0)}$.

Выполним первую итерацию вектора $\mathbf{y}^{(0)}$ умножением равенства (4.8) слева на матрицу $\mathbf{A}$:

$$\mathbf{y}^{(1)} = \mathbf{A}\mathbf{y}^{(0)} = c_1\mathbf{A}\mathbf{x}_1 + c_2\mathbf{A}\mathbf{x}_2 + \ldots + c_n\mathbf{A}\mathbf{x}_n .$$

Так как $\{\lambda_i, \mathbf{x}_i\}$ при всех значениях $i \in \{1, 2, \ldots, n\}$ по предположению являются собственными парами матрицы $\mathbf{A}$, то, в силу (4.2), последнее можно переписать в виде

$$\mathbf{y}^{(1)} = c_1\lambda_1\mathbf{x}_1 + c_2\lambda_2\mathbf{x}_2 + \ldots + c_n\lambda_n\mathbf{x}_n .$$

Для второй итерации по тому же принципу получаем

$$\mathbf{y}^{(2)} = \mathbf{A}\mathbf{y}^{(1)} = \mathbf{A}^2\mathbf{y}^{(0)} =$$
$$= c_1\lambda_1\mathbf{A}\mathbf{x}_1 + c_2\lambda_2\mathbf{A}\mathbf{x}_2 + \ldots + c_n\lambda_n\mathbf{A}\mathbf{x}_n =$$
$$= c_1\lambda_1^2\mathbf{x}_1 + c_2\lambda_2^2\mathbf{x}_2 + \ldots + c_n\lambda_n^2\mathbf{x}_n .$$

Очевидно, k-я итерация вектора $\mathbf{y}^{(0)}$ с помощью матрицы $\mathbf{A}$ дает вектор

$$\mathbf{y}^{(k)} = \mathbf{A}\mathbf{y}^{(k-1)} = \mathbf{A}^k\mathbf{y}^{(0)} = c_1\lambda_1^k\mathbf{x}_1 + c_2\lambda_2^k\mathbf{x}_2 + \ldots + c_n\lambda_n^k\mathbf{x}_n \qquad (4.9)$$

или, с учетом представления $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ в исходном базисе (см. (4.6)),

$$\mathbf{y}^{(k)} := \begin{pmatrix} y_1^{(k)} \\ y_2^{(k)} \\ \cdots \\ y_n^{(k)} \end{pmatrix} = c_1\lambda_1^k \begin{pmatrix} x_{11} \\ x_{12} \\ \cdots \\ x_{1n} \end{pmatrix} + c_2\lambda_2^k \begin{pmatrix} x_{21} \\ x_{22} \\ \cdots \\ x_{2n} \end{pmatrix} + \ldots + c_n\lambda_n^k \begin{pmatrix} x_{n1} \\ x_{n2} \\ \cdots \\ x_{nn} \end{pmatrix}.$$

Беря отношения компонент ***итерированного вектора*** $\mathbf{y}^{(k)}$ к соответствующим компонентам предыдущего вектора $\mathbf{y}^{(k-1)}$, будем иметь:

$$\frac{y_i^{(k)}}{y_i^{(k-1)}} = \frac{c_1\lambda_1^k x_{1i} + c_2\lambda_2^k x_{2i} + \ldots + c_n\lambda_n^k x_{ni}}{c_1\lambda_1^{k-1}x_{1i} + c_2\lambda_2^{k-1}x_{2i} + \ldots + c_n\lambda_n^{k-1}x_{ni}} =$$

$$= \lambda_1 \frac{1 + \frac{c_2}{c_1} \cdot \frac{x_{2i}}{x_{1i}} \left(\frac{\lambda_2}{\lambda_1}\right)^k + \ldots + \frac{c_n}{c_1} \cdot \frac{x_{ni}}{x_{1i}} \left(\frac{\lambda_n}{\lambda_1}\right)^k}{1 + \frac{c_2}{c_1} \cdot \frac{x_{2i}}{x_{1i}} \left(\frac{\lambda_2}{\lambda_1}\right)^{k-1} + \ldots + \frac{c_n}{c_1} \cdot \frac{x_{ni}}{x_{1i}} \left(\frac{\lambda_n}{\lambda_1}\right)^{k-1}}. \qquad (4.10)$$

Предел дроби в последнем равенстве при сделанных допущениях равен 1 в процессе $k \to \infty$, и, значит, $y_i^{(k)} \Big/ y_i^{(k-1)} \xrightarrow[k\to\infty]{} \lambda_1$ для каждого $i \in \{1, 2, \ldots, n\}$, при котором $x_{1i} \neq 0$ (заметим, что числа $x_{11}, x_{12}, \ldots, x_{1n}$ не могут быть одновременно нулями, так как $\mathbf{x}_1$ — базисный вектор и поэтому не может быть нулевым).

Представляя вектор $\mathbf{y}^{(k)}$ на основе (4.9) в виде

$$\mathbf{y}^{(k)} = c_1 \lambda_1^k \left[\mathbf{x}_1 + \frac{c_2}{c_1} \left(\frac{\lambda_2}{\lambda_1}\right)^k \mathbf{x}_2 + \ldots + \frac{c_n}{c_1} \left(\frac{\lambda_n}{\lambda_1}\right)^k \mathbf{x}_n \right], \qquad (4.11)$$

можно сделать вывод, что при тех же исходных допущениях, в силу $\left|\lambda_i / \lambda_1\right|^k \xrightarrow[k\to\infty]{(i\neq 1)} 0$, в фигурирующей в скобках выражения (4.11) линейной комбинации векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ с ростом k начнет доминировать первое слагаемое. Это означает, что вектор $\mathbf{y}^{(k)}$ от итерации к итерации будет давать все более хорошие приближения к собственному вектору $\mathbf{x}_1$ по направлению, т.е. с точностью до скалярного множителя $c_1 \lambda_1^k$ (см. свойство 4.1).

Таким образом, как показывают приведенные рассуждения, метод нахождения «старшей» собственной пары матрицы простой структуры, называемый ***степенным методом****, в своей основе

---

* Этимология данного термина совершенно ясна. Можно встретить и другие названия: ***счет на установление*** [34], ***итерационный метод фон Мизеса*** [40]. Иногда применяют латинскую аббревиатуру **PM** (от англ. *Power Method*) [54].

весьма примитивен и состоит в следующем: берут произвольный вектор $\mathbf{y}^{(0)}\ (\neq \mathbf{0})$, простыми итерациями $\mathbf{y}^{(k)} = \mathbf{A}\mathbf{y}^{(k-1)}$ строят последовательность векторов $\mathbf{y}^{(k)}$ и параллельно рассматривают последовательности отношений соответствующих компонент векторов k-й и $(k-1)$-й итераций (отношения с чрезвычайно малыми по модулю знаменателями следует игнорировать). Как только будут установлены несколько первых цифр во всех этих отношениях (что выясняется проверкой выполнения приближенных равенств $\frac{y_i^{(k)}}{y_i^{(k-1)}} \approx \frac{y_i^{(k-1)}}{y_i^{(k-2)}}$), так можно допустить, что найдено наибольшее по модулю собственное число с точностью, определяемой последним установившимся в отношениях знаком, и соответствующий ему собственный вектор, за который принимают последний итерированный вектор $\mathbf{y}^{(k)}$.

Для практической реализации такая схема нахождения старшей собственной пары малопригодна по многим причинам и требует определенной доводки. Рассмотрим некоторые из этих причин и соответственно пути модификации вышеописанного простейшего алгоритма.

Анализируя выражение $\mathbf{y}^{(k)}$ в форме (4.11), видим, что при достаточно большом числе итераций k за счет множителя λ_1^k в процессе счета может произойти либо превышение допустимого множества используемых компьютерных чисел, если $|\lambda_1| > 1$, либо пропадание значащих цифр итерированных векторов, если $|\lambda_1| < 1$. Устранить это явление можно достаточно легко, введя в итерационный процесс нормировку итерированных векторов (т.е. приведение к единичной длине по той или иной метрике) на каждой итерации или через некоторое фиксированное число итерационных шагов.

Так, пошаговая нормировка векторов порождает следующий

РМ-*алгоритм*.

Шаг 1. Ввести $n \times n$-матрицу $\mathbf{A}$, задать n-мерный вектор $\mathbf{y}^{(0)}$, вычислить $\|\mathbf{y}^{(0)}\|$ и вектор $\mathbf{x}^{(0)} := \mathbf{y}^{(0)} / \|\mathbf{y}^{(0)}\|$; положить $k := 1$.

Шаг 2. Вычислить вектор $\mathbf{y}^{(k)} = \mathbf{A}\mathbf{x}^{(k-1)}$.

Шаг 3. Вычислить $\|\mathbf{y}^{(k)}\|$ и $\mathbf{x}^{(k)} := \mathbf{y}^{(k)} / \|\mathbf{y}^{(k)}\|$.

Шаг 4. Вычислить отношения $\lambda_i^{(k)} = y_i^{(k)} / x_i^{(k-1)}$ (соответствующих координат векторов $\mathbf{y}^{(k)}$ и $\mathbf{x}^{(k-1)}$) при значениях $i \in \{1, 2, \ldots, n\}$ таких, что $|x_i^{(k-1)}| > \delta$, где $\delta > 0$ — некоторое задаваемое малое число (допуск).

Шаг 5. Подвергнуть числа $\lambda_i^{(k)}$ тесту на сходимость. Если обнаруживается совпадение требуемого числа знаков в значениях $\lambda_i^{(k)}$ и $\lambda_i^{(k-1)}$ ($\lambda^{(0)}$ можно задавать произвольно), то работу алгоритма прекратить и за старшее собственное число λ_1 принять усредненное (по i) значение $\lambda_i^{(k)}$, а за нормированный старший собственный вектор $\mathbf{x}_1$ — вектор $\mathbf{x}^{(k)}$. В противном случае — вернуться к шагу 2 с $k := k+1$.

Слабым местом данного алгоритма, очевидно, является последний шаг, т.е. решение проблемы своевременного останова работы алгоритма. Этот шаг описан из рациональных соображений и не может гарантировать во всех случаях (даже при сделанных допущениях) получения собственной пары $\{\lambda_1, \mathbf{x}_1\}$ с наперед заданной точностью, поскольку при разработке метода не было получено никаких оценок погрешности.

Относительно характера сходимости степенного метода можно утверждать (см. формулы (4.10) и (4.11)), что в указанных условиях итерационный процесс является линейным, т.е. сходится со скоростью геометрической прогрессии, знаменатель которой

определяется в основном величиной $|\lambda_2/\lambda_1|$. Это означает, что сходимость будет тем лучше и, как следствие, критерий останова в шаге 5 тем надежнее, чем сильнее доминирует в спектре матрицы $\mathbf{A}$ собственное число λ_1. Подмеченный факт вкупе со свойством 4.2 позволяет существенно ускорить нахождение наибольшего по модулю собственного числа матрицы $\mathbf{A}$ путем удачного смещения ее спектра, чему могут способствовать какие-либо априорные сведения об исходной задаче*.

То же свойство 4.2 собственных пар позволяет применять степенной метод непосредственно **для нахождения наименьшего по модулю собственного числа** λ_n знакоопределенной матрицы $\mathbf{A}$ в случае, когда наибольшее λ_1 уже найдено. Для этого достаточно найти наибольшее по модулю собственное число Λ матрицы $\mathbf{A}-\lambda_1\mathbf{E}$; соответствующий ему собственный вектор этой матрицы и число $\lambda_n := \Lambda + \lambda_1$ будут образовывать искомую собственную пару.

Действительно, вычитая из верного для собственной пары $\{\lambda_n, \mathbf{x}_n\}$ равенства $\mathbf{A}\mathbf{x}_n = \lambda_n\mathbf{x}_n$ тождество $\lambda_1\mathbf{x}_n = \lambda_1\mathbf{x}_n$, получаем верное равенство

$$\left(\mathbf{A}-\lambda_1\mathbf{E}\right)\mathbf{x}_n = \left(\lambda_n-\lambda_1\right)\mathbf{x}_n,$$

означающее, что $\Lambda := \lambda_n - \lambda_1$ и $\mathbf{x}_n$ служат собственной парой матрицы $\mathbf{A}-\lambda_1\mathbf{E}$. Так как для знакоопределенной матрицы справед-

---

* См. по этому поводу [67, 69]. В книге [67] приведен пример, наглядно показывающий эффективность подходящего сдвига: если некая матрица $\mathbf{A}$ шестого порядка имеет собственные числа $\lambda_i = 21-i$, то непосредственное применение степенного метода к вычислению λ_1 порождает итерационный процесс, сходящийся со скоростью порядка $(19/20)^k$, в то время как сдвиг на величину $p=17$, оставляющий соответствующее λ_1 число $\mu_1 := \lambda_1 - p = 3$ старшим в спектре матрицы $\mathbf{A}-p\mathbf{E}$, позволяет найти его степенным методом, сходящимся уже со скоростью порядка $(2/3)^k$.

ливо неравенство $|\lambda_n - \lambda_1| \geq |\lambda_i - \lambda_1|$ при любом $i \in \{1, 2, \ldots, n\}$, то, следовательно, Λ — наибольшее по модулю собственное число матрицы $\mathbf{A} - \lambda_1 \mathbf{E}$ и может быть найдено степенным методом.

Знание старшего собственного числа λ_1 матрицы $\mathbf{A}$ простой структуры, получаемого в процессе прямых итераций по формулам (4.9), (4.10), в предположении, что

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \ldots \geq |\lambda_n|,$$

позволяет без больших дополнительных затрат найти **приближенное значение второго по модулю собственного числа** λ_2. Это можно сделать по формуле

$$\lambda_2 :\approx \frac{y_i^{(k+1)} - \lambda_1 y_i^{(k)}}{y_i^{(k)} - \lambda_1 y_i^{(k-1)}}, \qquad (4.12)$$

вычисляя фигурирующие в правой части отношения для достаточно больших k и для всех $i \in \{1, 2, \ldots, n\}$, при которых абсолютная величина знаменателя не меньше некоторого порогового значения, и затем усредняя результат. Понятно, что при этом неизбежна потеря точности.

Для обоснования* приближенного равенства (4.12) подставим в его правую часть выражения компонент $(k+1)$-го, k-го и $(k-1)$-го итерированных векторов в соответствии с их представлением (4.9) в исходном базисе. После взаимного уничтожения по паре первых членов в числителе и знаменателе будем иметь:

$$\frac{y_i^{(k+1)} - \lambda_1 y_i^{(k)}}{y_i^{(k)} - \lambda_1 y_i^{(k-1)}} = \frac{c_2\lambda_2^{k+1}x_{2i} - c_2\lambda_1\lambda_2^{k}x_{2i} + \ldots + c_n\lambda_n^{k+1}x_{ni} - c_n\lambda_1\lambda_n^{k}x_{ni}}{c_2\lambda_2^{k}x_{2i} - c_2\lambda_1\lambda_2^{k-1}x_{2i} + \ldots + c_n\lambda_n^{k}x_{ni} - c_n\lambda_1\lambda_n^{k-1}x_{ni}} =$$

---

* Другое обоснование см., например, в [25]. Там же показано, что за соответствующий λ_2 собственный вектор $\mathbf{x}_2$ можно принять нормированный вектор $\mathbf{y}^{(k+1)} - \lambda_1 \mathbf{y}^{(k)}$.

$$= \frac{c_2\lambda_2^{k+1}x_{2i}\left(1-\frac{\lambda_1}{\lambda_2}+\sum_{j=3}^{n}\frac{\lambda_j^{k+1}-\lambda_1\lambda_j^k}{\lambda_2^{k+1}}\cdot\frac{c_j}{c_2}\cdot\frac{x_{ji}}{x_{2i}}\right)}{c_2\lambda_2^k x_{2i}\left(1-\frac{\lambda_1}{\lambda_2}+\sum_{j=3}^{n}\frac{\lambda_j^k-\lambda_1\lambda_j^{k-1}}{\lambda_2^k}\cdot\frac{c_j}{c_2}\cdot\frac{x_{ji}}{x_{2i}}\right)} \xrightarrow[k\to\infty]{} \lambda_2 ,$$

так как $\frac{\lambda_j^{k+1}-\lambda_1\lambda_j^k}{\lambda_2^{k+1}} := \left(\frac{\lambda_j}{\lambda_2}\right)^{k+1} - \frac{\lambda_1}{\lambda_2}\left(\frac{\lambda_j}{\lambda_2}\right)^k \xrightarrow[k\to\infty]{} 0$ при всех значениях $j = 3, \ldots, n$.

Вернемся к вопросу о недостатках степенного метода нахождения наибольшего по модулю собственного числа и путях их устранения. При этом далее будем рассматривать в основном класс **симметричных положительно определенных матриц**. Известно, что такие матрицы имеют положительный вещественный спектр $\lambda_1, \lambda_2, \ldots, \lambda_n$, ортонормированный базис из собственных векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ и, естественно, являются матрицами простой структуры.

Обсудим шаг 4 предложенного выше РМ-алгоритма.

Вычисление на каждом итерационном шаге отношений в с е х пар соответствующих компонент векторов $\mathbf{x}$ и $\mathbf{y} := \mathbf{Ax}$, да еще с определенными проверками, при больших значениях n требует значительных вычислительных затрат, хотя и дает о старшем собственном числе λ_1 дополнительную информацию: как утверждается в [40], значение λ_1 заключено между наименьшим и наибольшим из этих отношений, т.е. имеются двусторонние оценки величины λ_1 на каждой итерации.

В целях упрощения соответствующей шагу 4 РМ-алгоритма процедуры проведем следующие рассуждения.

Пусть $\mathbb{R}_n$ — евклидово пространство, $\mathbf{A}$ — симметричная положительно определенная матрица и пусть последовательность итерированных векторов $\mathbf{y}^{(k)}$ строится, как и ранее, по формулам (4.9).

Найдем выражения скалярных произведений $\mathbf{y}^{(k)}$ на $\mathbf{y}^{(k)}$ и $\mathbf{y}^{(k)}$ на $\mathbf{y}^{(k-1)}$ через собственные числа. Выполняя умножение правых частей (4.9) по правилам умножения многочленов и учитывая ортонормированность собственных векторов, т.е. условие $\left(\mathbf{x}_i, \mathbf{x}_j\right) = \delta_{ij}$ при $i, j \in \{1, 2, \ldots, n\}$, имеем:

$$\left(\mathbf{y}^{(k)},\ \mathbf{y}^{(k)}\right) = c_1^2\lambda_1^{2k} + c_2^2\lambda_2^{2k} + \ldots + c_n^2\lambda_n^{2k},$$

$$\left(\mathbf{y}^{(k)},\ \mathbf{y}^{(k-1)}\right) = c_1^2\lambda_1^{2k-1} + c_2^2\lambda_2^{2k-1} + \ldots + c_n^2\lambda_n^{2k-1}.$$

Отношение этих чисел

$$\frac{\left(\mathbf{y}^{(k)},\ \mathbf{y}^{(k)}\right)}{\left(\mathbf{y}^{(k)},\ \mathbf{y}^{(k-1)}\right)} = \lambda_1 \frac{1 + \left(\frac{c_2}{c_1}\right)^2\left(\frac{\lambda_2}{\lambda_1}\right)^{2k} + \ldots + \left(\frac{c_n}{c_1}\right)^2\left(\frac{\lambda_n}{\lambda_1}\right)^{2k}}{1 + \left(\frac{c_2}{c_1}\right)^2\left(\frac{\lambda_2}{\lambda_1}\right)^{2k-1} + \ldots + \left(\frac{c_n}{c_1}\right)^2\left(\frac{\lambda_n}{\lambda_1}\right)^{2k-1}} \qquad (4.13)$$

в оговоренных выше условиях при $k \to \infty$ имеет пределом наибольшее собственное число λ_1, причем скорость сходимости к пределу будет больше, чем в степенном методе, опирающемся на отношения (4.10) ($O\left(\left|\lambda_2/\lambda_1\right|^{2k}\right)$ против $O\left(\left|\lambda_2/\lambda_1\right|^{k}\right)$).

Базирующаяся на таком подходе модификация степенного метода называется ***методом скалярных произведений***. Его реализацией может служить, например, следующий

SP-*алгоритм.**

Шаг 1. Ввести: данную симметричную $n \times n$-матрицу $\mathbf{A}$, произвольный n-мерный начальный вектор $\mathbf{y}^{(0)}$ $(\neq \mathbf{0})$, малое число $\varepsilon > 0$ (определяющее допустимую абсолютную погрешность искомого собственного числа λ_1), число $\lambda^{(0)}$ для начального

* SP от англ. *Scalar Product*.

сравнения (например, $\lambda^{(0)} := 0$). Положить $k := 1$ (включить счетчик итераций).

Шаг 2. Вычислить скаляры $s^{(0)} := \left(\mathbf{y}^{(0)}, \mathbf{y}^{(0)}\right)$, $\|\mathbf{y}^{(0)}\|_2 := \sqrt{s^{(0)}}$ и вектор $\mathbf{x}^{(0)} := \mathbf{y}^{(0)} / \|\mathbf{y}^{(0)}\|_2$.

Шаг 3. Вычислить: $\mathbf{y}^{(k)} := \mathbf{A}\mathbf{x}^{(k-1)}$ (итерация нормированного вектора).

Шаг 4. Вычислить: $s^{(k)} := \left(\mathbf{y}^{(k)}, \mathbf{y}^{(k)}\right)$ и $t^{(k)} := \left(\mathbf{y}^{(k)}, \mathbf{x}^{(k-1)}\right)$ (скалярные произведения), $\|\mathbf{y}^{(k)}\|_2 := \sqrt{s^{(k)}}$, $\mathbf{x}^{(k)} := \mathbf{y}^{(k)} / \|\mathbf{y}^{(k)}\|_2$ (приближение к нормированному собственному вектору), $\lambda^{(k)} := s^{(k)} / t^{(k)}$ (приближение к собственному числу λ_1).

Шаг 5. Если $|\lambda^{(k)} - \lambda^{(k-1)}| > \varepsilon$, положить $k := k+1$ и вернуться к шагу 3, иначе завершить вычисления, считая $\lambda_1 :\approx \lambda^{(k)}$, $\mathbf{x}_1 :\approx \mathbf{x}^{(k)}$.

Замечание 4.1. Данный алгоритм позволяет более быстро (т.е. за меньшее число итераций), чем РМ-алгоритм, найти с нужной точностью наибольшее собственное число симметричной матрицы, но при этом точность приближенного равенства $\mathbf{x}_1 :\approx \mathbf{x}^{(k)}$ для соответствующего собственного вектора может оказаться недостаточной (объясните почему?).

Замечание 4.2. Очевидно, в методе скалярных произведений вместо отношения (4.13), стремящегося к λ_1 при $k \to \infty$, можно с тем же успехом взять отношение $\left(\mathbf{y}^{(k+1)}, \mathbf{y}^{(k)}\right)$ к $\left(\mathbf{y}^{(k)}, \mathbf{y}^{(k)}\right)$, имеющее тот же предел. Последнее же есть не что иное, как отношение Рэлея: $\dfrac{\left(\mathbf{y}^{(k+1)}, \mathbf{y}^{(k)}\right)}{\left(\mathbf{y}^{(k)}, \mathbf{y}^{(k)}\right)} = \dfrac{\left(\mathbf{A}\mathbf{y}^{(k)}, \mathbf{y}^{(k)}\right)}{\left(\mathbf{y}^{(k)}, \mathbf{y}^{(k)}\right)} = \rho\left(\mathbf{y}^{(k)}\right)$; отсюда другое название метода скалярных произведений — ***метод частных Рэлея***. В соответствии со свойствами 4.5, 4.6 предыдущего параграфа можно сказать, что этим методом на каждом итерационном шаге находят наилучшее для вычисленного вектора $\mathbf{y}^{(k)}$ приближение к собственному числу λ_1 в смысле евклидовой нормы невязки.

Пример 4.1. На матрице $\mathbf{A} := \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ покажем *процесс* построения приближений к старшему собственному числу λ_1 (и соответствующему ему собственному вектору $\mathbf{x}_1$) методом скалярных произведений.

Приняв за начальный вектор $\mathbf{y}^{(0)}$ первый орт $\mathbf{e}_1 = (1; 0)^T$, далее следуем SP-алгоритму. Имеем:

$$s^{(0)} := \left(\mathbf{y}^{(0)}, \mathbf{y}^{(0)}\right) = 1, \quad \|\mathbf{y}^{(0)}\|_2 := \sqrt{s^{(0)}} = 1, \quad \mathbf{x}^{(0)} := \frac{\mathbf{y}^{(0)}}{\|\mathbf{y}^{(0)}\|_2} = \begin{pmatrix} 1 \\ 0 \end{pmatrix};$$

1-я итерация:

$$\mathbf{y}^{(1)} := \mathbf{A}\,\mathbf{x}^{(0)} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \quad s^{(1)} := \left(\mathbf{y}^{(1)}, \mathbf{y}^{(1)}\right) = 5,$$

$$t^{(1)} := \left(\mathbf{y}^{(1)}, \mathbf{x}^{(0)}\right) = 2, \quad \left\|\mathbf{y}^{(1)}\right\|_2 := \sqrt{s^{(1)}} = \sqrt{5},$$

$$\mathbf{x}^{(1)} := \frac{\mathbf{y}^{(1)}}{\|\mathbf{y}^{(1)}\|_2} = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 \\ -1 \end{pmatrix} \approx \begin{pmatrix} 0{,}894 \\ -0{,}447 \end{pmatrix}, \quad \lambda^{(1)} := \frac{s^{(1)}}{t^{(1)}} = \frac{5}{2} = 2{,}5;$$

2-я итерация:

$$\mathbf{y}^{(2)} := \mathbf{A}\,\mathbf{x}^{(1)} = \frac{1}{\sqrt{5}}\begin{pmatrix} 5 \\ -4 \end{pmatrix}, \quad s^{(2)} := \left(\mathbf{y}^{(2)}, \mathbf{y}^{(2)}\right) = \frac{41}{5},$$

$$t^{(2)} := \left(\mathbf{y}^{(2)}, \mathbf{x}^{(1)}\right) = \frac{14}{5}, \quad \left\|\mathbf{y}^{(2)}\right\|_2 := \sqrt{s^{(2)}} = \sqrt{\frac{41}{5}},$$

$$\mathbf{x}^{(2)} := \frac{\mathbf{y}^{(2)}}{\|\mathbf{y}^{(2)}\|_2} = \frac{1}{\sqrt{41}}\begin{pmatrix} 5 \\ -4 \end{pmatrix} \approx \begin{pmatrix} 0{,}781 \\ -0{,}625 \end{pmatrix}, \quad \lambda^{(2)} := \frac{s^{(2)}}{t^{(2)}} = \frac{41}{14} \approx 2{,}929;$$

3-я итерация:

$$\mathbf{y}^{(3)} := \mathbf{A}\,\mathbf{x}^{(2)} = \frac{1}{\sqrt{41}}\begin{pmatrix} 14 \\ -13 \end{pmatrix}, \quad s^{(3)} := \left(\mathbf{y}^{(3)}, \mathbf{y}^{(3)}\right) = \frac{365}{41},$$

$$t^{(3)} := \left(\mathbf{y}^{(3)}, \mathbf{x}^{(2)}\right) = \frac{122}{41}, \quad \left\|\mathbf{y}^{(3)}\right\|_2 := \sqrt{s^{(3)}} = \sqrt{\frac{365}{41}},$$

$$\mathbf{x}^{(3)} := \frac{\mathbf{y}^{(3)}}{\|\mathbf{y}^{(3)}\|_2} = \frac{1}{\sqrt{365}}\begin{pmatrix} 14 \\ -13 \end{pmatrix} \approx \begin{pmatrix} 0{,}733 \\ -0{,}680 \end{pmatrix}, \quad \lambda^{(3)} := \frac{s^{(3)}}{t^{(3)}} = \frac{365}{122} \approx 2{,}992.$$

По значениям величин $\left|\lambda^{(2)}-\lambda^{(1)}\right| :\approx 0,429$, $\left|\lambda^{(3)}-\lambda^{(2)}\right| :\approx 0,063$ можно судить о сближении с каждой итерацией членов последовательности $\lambda^{(1)}$, $\lambda^{(2)}$, $\lambda^{(3)}$, являющихся приближениями к наибольшему по модулю собственному числу λ_1. Последнюю из этих величин можно считать нестрогой оценкой абсолютной погрешности равенства $\lambda_1 \approx \lambda^{(3)}$ (на самом деле знание $\lambda_1 := 3$ показывает его более высокую точность: $\left|\lambda_1-\lambda^{(3)}\right| \approx 0,008 < 0,01$). Для собственного вектора $\mathbf{x}_1$ приближенное равенство $\mathbf{x}_1 \approx \mathbf{x}^{(3)}$ можно оценить величиной

$$\left\|\mathbf{x}^{(3)}-\mathbf{x}^{(2)}\right\| :\approx \left\|\begin{pmatrix}-0,048\\-0,055\end{pmatrix}\right\|_2 \approx 0,073.$$

Наличие ортонормированного базиса из собственных векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ матрицы $\mathbf{A}$ позволяет применять степенной метод (метод скалярных произведений) для последовательного вычисления собственных пар $\left\{\lambda_i, \mathbf{x}_i\right\}$, следующих за известной старшей парой $\left\{\lambda_1, \mathbf{x}_1\right\}$ более совершенными, чем определяемый формулой (4.12), способами. Рассмотрим один из них.

Пусть первая собственная пара $\left\{\lambda_1, \mathbf{x}_1\right\}$ уже найдена, причем вектор $\mathbf{x}_1$ таков, что $\|\mathbf{x}_1\| := \sqrt{(\mathbf{x}_1, \mathbf{x}_1)} = 1$. Возьмем произвольный ненулевой вектор $\mathbf{z}^{(0)}$ и образуем вектор

$$\mathbf{y}^{(0)} := \mathbf{z}^{(0)} - \left(\mathbf{z}^{(0)}, \mathbf{x}_1\right)\mathbf{x}_1. \qquad (4.14)$$

Так как

$$\left(\mathbf{y}^{(0)}, \mathbf{x}_1\right) = \left(\mathbf{z}^{(0)}, \mathbf{x}_1\right) - \left(\mathbf{z}^{(0)}, \mathbf{x}_1\right)(\mathbf{x}_1, \mathbf{x}_1) = 0,$$

то вектор $\mathbf{y}^{(0)}$ ортогонален $\mathbf{x}_1$, т.е. его проекция на первый базисный вектор системы векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ равна нулю. Значит, разложение (4.8) вектора $\mathbf{y}^{(0)}$ по этому базису имеет вид

$$\mathbf{y}^{(0)} = c_2\mathbf{x}_2 + c_3\mathbf{x}_3 + \ldots + c_n\mathbf{x}_n,$$

и соответственно степенные итерации этого вектора типа (4.9) порождают векторы

$$\mathbf{y}^{(k)} = c_2\lambda_2^k\mathbf{x}_2 + c_3\lambda_3^k\mathbf{x}_3 + \ldots + c_n\lambda_n^k\mathbf{x}_n . \qquad (4.15)$$

Легко видеть (ср. с (4.13)), что если имеет место неравенство $|\lambda_2| > |\lambda_i|$ при всех значениях индекса $i \in \{3, \ldots, n\}$, то $\left(\mathbf{y}^{(k)}, \mathbf{y}^{(k)}\right) \Big/ \left(\mathbf{y}^{(k)}, \mathbf{y}^{(k-1)}\right) \xrightarrow[k\to\infty]{} \lambda_2$ со скоростью $O\left((\lambda_3/\lambda_2)^{2k}\right)$ и $\mathbf{x}^{(k)} = \mathbf{y}^{(k)} \Big/ \| \mathbf{y}^{(k)} \| \xrightarrow[k\to\infty]{} \mathbf{x}_2$ со скоростью $O\left((\lambda_3/\lambda_2)^k\right)$.

Следующая собственная пара $\{\lambda_3, \mathbf{x}_3\}$ может быть найдена приближенно тем же методом, если за начальный вектор последовательности $\left(\mathbf{y}^{(k)}\right)$ принять вектор

$$\mathbf{y}^{(0)} := \mathbf{z}^{(0)} - \left(\mathbf{z}^{(0)}, \mathbf{x}_1\right)\mathbf{x}_1 - \left(\mathbf{z}^{(0)}, \mathbf{x}_2\right)\mathbf{x}_2 ,$$

ортогональный одновременно $\mathbf{x}_1$ и $\mathbf{x}_2$ при любом $\mathbf{z}^{(0)}$, и т.д.

Известны и другие способы последовательного нахождения собственных пар, опирающиеся на непосредственное применение степенного метода. При этом имеются возможности понижения размерности решаемой задачи при нахождении каждой последующей собственной пары.

Замечание 4.3. В реальных расчетах, в силу неизбежных ошибок округлений, в представлении (4.15) итерированного вектора $\mathbf{y}^{(k)}$ при вычислении второй собственной пары появится малое, но растущее с увеличением номера k слагаемое, соответствующее проекции $\mathbf{y}^{(k)}$ на первый собственный вектор $\mathbf{x}_1$. Поэтому реальный алгоритм должен предусматривать возврат к началу процесса итерирования, т.е. проведение операции ортогонализации по формуле (4.14) с $\mathbf{z}^{(0)} := \mathbf{y}^{(m)}$ через некоторое число итераций $k := m$.

Замечание 4.4. Не всегда известно, выполняются ли оговоренные выше условия, при которых изучался степенной метод. В таких ситуациях при его применении нужно принимать особые меры предосторожности. Целесообразно, например, контролировать, сближаются ли члены последовательности $(\lambda^{(k)})$

посредством проверки неравенств

$$\left|\lambda^{(k+1)}-\lambda^{(k)}\right|<\left|\lambda^{(k)}-\lambda^{(k-1)}\right|$$

подобно тому, как это делалось в примере 4.1 (такая проверка называется ***приемом Гарвика*** [1, 34]), а также осуществлять запуск итерационного процесса с разных начальных векторов. В случае кратности находимого собственного числа последнее просто необходимо для вычисления степенным методом разных соответствующих ему собственных векторов.

Замечание 4.5. Как отмечалось ранее, степенной метод сходится линейно, точнее, имеет лишь асимптотическую скорость сходимости геометрической прогрессии. При слабом доминировании модуля вычисляемого собственного числа эта сходимость может оказаться чрезвычайно медленной. Ускорения итерационного процесса можно достигнуть за счет быстрого накопления степеней матриц по схеме

$$\mathbf{A}\cdot\mathbf{A}=\mathbf{A}^2,\quad \mathbf{A}^2\cdot\mathbf{A}^2=\mathbf{A}^4$$

и т.д., что позволяет проводить не последовательное, пошаговое, а скачкообразное построение последовательности $(\mathbf{y}^{(k)})$ с помощью равенств вида

$$\mathbf{y}^{(k)}=\mathbf{A}^{k-m}\mathbf{y}^{(m)}$$

при фиксированных значениях $m\in\{0,\ 1,\ \ldots,\ k-1\}$ (таких, при которых $k-m$ является некоторой целой степенью двойки) и нормированием сразу после очередного скачка. Здесь, правда, нужно особенно внимательно относиться к риску выхода за границы диапазона компьютерных чисел в процессе счета.

Если не требуется находить собственный вектор $\mathbf{x}_1$, то более быстро вычислять максимальное по модулю собственное число λ_1 можно на основе соотношения [25]

$$\lambda_1^k+\lambda_2^k+\ldots+\lambda_n^k=\operatorname{Sp}\mathbf{A}^k\quad\forall k\in\mathbf{N}.$$

Вычислив $\mathbf{A}^k$ по закону удвоения степеней, а затем $\mathbf{A}^{k+1}=\mathbf{A}^k\cdot\mathbf{A}$, находим отношение ***следов*** (сумм диагональных элементов) этих матриц:

$$\frac{\operatorname{Sp}\mathbf{A}^{k+1}}{\operatorname{Sp}\mathbf{A}^k}=\frac{\lambda_1^{k+1}\left(1+\left(\frac{\lambda_2}{\lambda_1}\right)^{k+1}+\ldots+\left(\frac{\lambda_n}{\lambda_1}\right)^{k+1}\right)}{\lambda_1^k\left(1+\left(\frac{\lambda_2}{\lambda_1}\right)^{k}+\ldots+\left(\frac{\lambda_n}{\lambda_1}\right)^{k}\right)}\xrightarrow[k\to\infty]{}\lambda_1.$$

Замечание 4.6. Более популярный способ ускорения сходимости степенного метода — это применение Δ^2-*процесса Эйткена**. Считается, что если $\lambda^{(k-1)}$, $\lambda^{(k)}$, $\lambda^{(k+1)}$ являются тремя последовательными приближениями к собственному числу, полученными степенным методом, то число

$$\tilde{\lambda} := \lambda^{(k-1)} - \frac{\left(\lambda^{(k)} - \lambda^{(k-1)}\right)^2}{\lambda^{(k+1)} - 2\lambda^{(k)} + \lambda^{(k-1)}}$$

ближе к пределу строящейся последовательности, чем каждое из них. Этот факт может быть использован в реальных алгоритмах либо через несколько итерационных шагов (например, через два на третий), либо на завершающем этапе вычислений. Для искомого собственного вектора такое ускорение может проводиться п о к о о р д и н а т н о .

§ 4.3. МЕТОД ОБРАТНЫХ ИТЕРАЦИЙ И RQI-АЛГОРИТМ

В предыдущем параграфе было показано, что при определенных условиях наименьшее по модулю собственное число λ_n может быть найдено степенным методом, когда уже известно наибольшее число λ_1. Если же проблема состоит в нахождении лишь младшей собственной пары матрицы $\mathbf{A}$, то можно обойтись и без вычисления λ_1, применяя степенной метод к матрице $\mathbf{A}^{-1}$.

В самом деле если данная матрица $\mathbf{A}$ имеет собственные пары

$$\{\lambda_1, \mathbf{x}_1\},\ \{\lambda_2, \mathbf{x}_2\},\ \ldots,\ \{\lambda_{n-1}, \mathbf{x}_{n-1}\},\ \{\lambda_n, \mathbf{x}_n\},$$

то по свойству 4.3 собственными парами матрицы $\mathbf{A}^{-1}$ будут

$$\left\{\frac{1}{\lambda_1}, \mathbf{x}_1\right\},\ \left\{\frac{1}{\lambda_2}, \mathbf{x}_2\right\},\ \ldots,\ \left\{\frac{1}{\lambda_{n-1}}, \mathbf{x}_{n-1}\right\},\ \left\{\frac{1}{\lambda_n}, \mathbf{x}_n\right\}.$$

При этом упорядочиванию спектра $\mathbf{A}$ вида

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_{n-1}| > |\lambda_n|$$

---

* Читается «дельта-два-процесс». Посвященную ему работу A.Aitken опубликовал в 1931 году.

соответствует цепочка неравенств

$$\left|\frac{1}{\lambda_n}\right| > \left|\frac{1}{\lambda_{n-1}}\right| \geq \ldots \geq \left|\frac{1}{\lambda_2}\right| \geq \left|\frac{1}{\lambda_1}\right|$$

для собственных чисел

$$\gamma_1 := \frac{1}{\lambda_n},\ \gamma_2 := \frac{1}{\lambda_{n-1}},\ \ldots\ ,\ \gamma_{n-1} := \frac{1}{\lambda_2},\ \gamma_n := \frac{1}{\lambda_1}$$

матрицы $\mathbf{A}^{-1}$. Это значит, что наименьшим по модулю собственным числом данной матрицы $\mathbf{A}$ является величина, обратная наибольшему по модулю собственному числу матрицы $\mathbf{A}^{-1}$. Последнее же может быть получено прямыми итерациями произвольного начального вектора $\mathbf{y}^{(0)}$ посредством матрицы $\mathbf{A}^{-1}$ по аналогичной (4.9) формуле

$$\mathbf{y}^{(k)} = \mathbf{A}^{-1}\mathbf{y}^{(k-1)}, \quad k = 1, 2, \ldots . \tag{4.16}$$

При достаточно больших $k \in \mathbb{N}$ последовательность отношений одноименных координат векторов $\mathbf{y}^{(k)}$ и $\mathbf{y}^{(k-1)}$ должна давать приближенное значение $\frac{1}{\lambda_n}$, а вектор $\mathbf{y}^{(k)}$ (желательно его нормирование) можно принять за собственный вектор $\mathbf{x}_n$.

Вместо прямых итераций (4.16), требующих предварительного обращения исходной матрицы $\mathbf{A}$, обычно предпочитают строить ту же последовательность векторов $\mathbf{y}^{(k)}$, решая при $k = 1, 2, 3, \ldots$ линейные системы

$$\mathbf{A}\mathbf{y}^{(k)} = \mathbf{y}^{(k-1)}. \tag{4.17}$$

Так как все эти системы имеют одну и ту же матрицу коэффициентов, то самая трудоемкая часть метода Гаусса для их решения — $\mathbf{LU}$-факторизация матрицы $\mathbf{A}$ — может быть выполнена лишь один раз.

Построение последовательности векторов, приближающих собственный вектор $\mathbf{x}_n$ по неявной формуле (4.17), называют

обратными итерациями, а процесс решения частичных проблем собственных значений на этой основе — ***методом обратных итераций*** (иначе, ***обратным степенным методом*** [9]).

Применение обратных итераций к нахождению младшей собственной пары матрицы $\mathbf{A}$ не требует написания специального алгоритма, достаточно лишь заменить один шаг в алгоритмах предыдущего параграфа. А именно, наполнение шага 2 в PM-алгоритме для матриц простой структуры и шага 3 в SP-алгоритме для симметричных матриц должно быть следующим:

решить уравнение $\mathbf{A}\mathbf{y}^{(k)} = \mathbf{y}^{(k-1)}$.

Соответствующий такой замене алгоритм называют ***INVIT-алгоритмом*** [54] (от англ. *Inverse Iteration*).

Метод обратных итераций, а точнее, ***обратные итерации со сдвигами*** часто применяют в тех случаях, когда нужно с большой точностью найти собственный вектор, отвечающий какому-либо собственному числу из спектра заданной матрицы при условии, что известно приближенное значение этого числа. Здесь, очевидно, прямое решение однородной системы (4.3) заведомо неприменимо, так как подстановка в нее значения λ, хоть сколько-нибудь отличного от собственного, сделает систему однозначно разрешимой, т.е. допускающей только тривиальное решение. Рассмотрим суть обратных итераций со сдвигами.

Пусть для собственного числа λ_j матрицы простой структуры $\mathbf{A}$ известно его приближение σ такое, что

$$|\lambda_j - \sigma| < |\lambda_i - \sigma| \quad \forall\, i \neq j, \tag{4.18}$$

т.е. число σ ближе к собственному числу λ_j, чем к какому-либо другому собственному числу матрицы $\mathbf{A}$.

Начиная с вектора $\mathbf{x}^{(0)}$ такого, что $\|\mathbf{x}^{(0)}\| = 1$, образуем последовательность нормированных векторов $(\mathbf{x}^{(k)})$ по формулам

$$(\mathbf{A} - \sigma\mathbf{E})\mathbf{y}^{(k)} = \mathbf{x}^{(k-1)}; \tag{4.19}$$

$$\mathbf{x}^{(k)} = \mathbf{y}^{(k)} / \|\mathbf{y}^{(k)}\|, \quad k = 1, 2, \ldots . \tag{4.20}$$

Изучим поведение этой последовательности, для чего запишем разложение векторов $\mathbf{y}^{(k)}$ и $\mathbf{x}^{(k-1)}$ по базису из собственных векторов $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ с некоторыми коэффициентами $c_i^{(k)}$ и $b_i^{(k-1)}$ соответственно:

$$\mathbf{y}^{(k)} = c_1^{(k)}\mathbf{x}_1 + c_2^{(k)}\mathbf{x}_2 + \ldots + c_n^{(k)}\mathbf{x}_n, \tag{4.21}$$

$$\mathbf{x}^{(k-1)} = b_1^{(k-1)}\mathbf{x}_1 + b_2^{(k-1)}\mathbf{x}_2 + \ldots + b_n^{(k-1)}\mathbf{x}_n.$$

Подставляя это в (4.19) и учитывая, что по определению

$$\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i \quad \forall i \in \{1, 2, \ldots, n\},$$

имеем

$$(\lambda_1 - \sigma)c_1^{(k)}\mathbf{x}_1 + (\lambda_2 - \sigma)c_2^{(k)}\mathbf{x}_2 + \ldots + (\lambda_n - \sigma)c_n^{(k)}\mathbf{x}_n = \\ = b_1^{(k-1)}\mathbf{x}_1 + b_2^{(k-1)}\mathbf{x}_2 + \ldots + b_n^{(k-1)}\mathbf{x}_n,$$

откуда, в силу единственности разложения вектора по базису, следует

$$(\lambda_i - \sigma)c_i^{(k)} = b_i^{(k-1)} \quad \forall i \in \{1, 2, \ldots, n\}.$$

Анализируя получающиеся отсюда выражения

$$c_i^{(k)} = \frac{b_i^{(k-1)}}{\lambda_i - \sigma} \tag{4.22}$$

коэффициентов разложения вектора $\mathbf{y}^{(k)}$ по базису из собственных векторов, видим, что вследствие малости модуля знаменателя $\lambda_j - \sigma$ по сравнению с другими знаменателями $\lambda_i - \sigma$ (см. (4.18)) можно рассчитывать на преимущественное возрастание коэффициентов $c_j^{(k)}$ именно при собственном векторе $\mathbf{x}_j$ с ростом k. Значит, чем сильнее неравенство в (4.18), тем сильнее (быстрее) будет доминировать составляющая собственного вектора $\mathbf{x}_j$ в представлении (4.21) вектора $\mathbf{y}^{(k)}$, а значит, и вектора $\mathbf{x}^{(k)}$,

получаемого из $\mathbf{y}^{(k)}$ нормированием (4.20). Последнее же говорит о том, что, каков бы ни был начальный вектор $\mathbf{x}^{(0)}$ $(\neq \mathbf{0})^*$, быстрое доминирование $c_j^{(k)}$ среди остальных коэффициентов $c_i^{(k)}$ происходит еще и за счет числителей дробей (4.22).

Следует заметить, что обратные итерации со сдвигами (4.19), (4.20) позволяют не только найти собственный вектор $\mathbf{x}_j$, но и служат основой для уточнения приближенного равенства $\lambda_j \approx \sigma$.

Действительно, формулы (4.19), (4.20) определяют не что иное, как метод обратных итераций для нахождения наименьшего по модулю собственного числа матрицы $\mathbf{A}-\sigma\mathbf{E}$, и, если σ существенно ближе к λ_j, чем к любому другому собственному числу λ_i матрицы $\mathbf{A}$, то уточняющие λ_j значения, согласно свойству 4.2, можно получать при $k = 1, 2, \ldots$ по формуле

$$\lambda_j^{(k)} = \sigma + \left\langle \frac{x_i^{(k-1)}}{y_i^{(k)}} \right\rangle, \tag{4.23}$$

где $x_i^{(k-1)}$ и $y_i^{(k)}$ — координаты векторов $\mathbf{x}^{(k-1)}$ и $\mathbf{y}^{(k)}$ соответственно, а $\langle \cdot \rangle$ — знак усреднения по всем тем значениям $i \in \{1, 2, \ldots, n\}$, при которых $y_i^{(k)} \neq 0$ [34].

Как показывает практика вычислений, сходимость процесса обратных итераций со сдвигами характеризуется более высокой скоростью по сравнению с обычным степенным методом. Но существенно более быстрая сходимость может быть получена введением переменных сдвигов, определяемых какой-нибудь последовательностью чисел $\sigma_0, \sigma_1, \sigma_2, \ldots$, сходящейся к находимому собственному числу. Не вызывает сомнений целесообразность

---

* Лишь бы при его выборе не попасть на ортогональный $\mathbf{x}_j$ вектор; часто берут вектор $\mathbf{x}^{(0)}$ с равными координатами.

использования в роли таких чисел приближений $\lambda_j^{(k)}$ к собственному числу λ_j, получаемых по формуле (4.23).

Таким образом, ***обратные итерации с переменными сдвигами*** можно определить совокупностью равенств:

$$\left(\mathbf{A}-\lambda_j^{(k-1)}\mathbf{E}\right)\mathbf{y}^{(k)}=\mathbf{x}^{(k-1)},$$

$$\mathbf{x}^{(k)}=\mathbf{y}^{(k)}\Big/\left\|\mathbf{y}^{(k)}\right\|, \qquad (4.24)$$

$$\lambda_j^{(k)}=\lambda_j^{(k-1)}+\left\langle\frac{x_i^{(k-1)}}{y_i^{(k)}}\right\rangle,$$

где $k=1, 2, \ldots,$ а число $\lambda_j^{(0)}$ $\left(\approx\lambda_j\right)$ и вектор $\mathbf{x}^{(0)}$ (с нормой, равной единице) задаются.

Скорость сходимости процесса (4.24) — квадратичная [18, 34], в то время как в случае постоянных сдвигов лишь линейная, хотя и с малыми, как правило, знаменателями геометрической прогрессии. Зачастую бывает достаточно сделать 2–3 итерации по формулам (4.24), чтобы получить заданную собственную пару с реально возможной точностью. Нужно только видеть разницу в цене реализации обратных итераций с постоянными и переменными сдвигами: в первом случае при каждом k решают системы линейных алгебраических уравнений с одной и той же матрицей коэффициентов (как это было и при обратных итерациях (4.17) без сдвигов), во втором случае на разных шагах приходится решать совершенно различные системы.

В методе (4.24), так же как и в предыдущем, неясно, как подбирать начальный сдвиг $\sigma:=\lambda_j^{(0)}$, за исключением случаев, когда решают частичную проблему заведомо в такой постановке, при которой требуется найти собственное число, ближайшее к заданному значению, и соответствующий ему собственный вектор.

Более определенной в этом смысле, к тому же более быстро

сходящейся является следующая модификация метода (4.24) — ***обратные итерации с отношениями Рэлея***, применяемые для решения симметричных задач на собственные значения. Основу этой модификации составляет

RQI-*алгоритм**:

Шаг 0. Задать вектор $\mathbf{x}^{(0)}$ такой, что $\|\mathbf{x}^{(0)}\|=1$.

Шаг 1. Для $k=1, 2, \ldots$:

1.1. Вычислить $\rho_{k-1} := \left(\mathbf{A}\mathbf{x}^{(k-1)}, \mathbf{x}^{(k-1)}\right) \Big/ \left(\mathbf{x}^{(k-1)}, \mathbf{x}^{(k-1)}\right)$.

1.2. Найти $\mathbf{y}^{(k)}$ из уравнения $\left(\mathbf{A}-\rho_{k-1}\mathbf{E}\right)\mathbf{y}^{(k)} = \mathbf{x}^{(k-1)}$.

1.3. Нормировать $\mathbf{y}^{(k)}$, т.е. положить $\mathbf{x}^{(k)} := \mathbf{y}^{(k)} \big/ \|\mathbf{y}^{(k)}\|$.

1.4. Проверить ρ_{k-1}, $\mathbf{x}^{(k)}$ на сходимость. Перейти на шаг 1.1 или остановиться.

После «штатного» останова работы RQI-алгоритма** при некотором значении $k := k_0$ в качестве собственной для данной матрицы $\mathbf{A}$ объявляется пара $\left\{\rho_{k_0-1}, \mathbf{x}^{(k_0)}\right\}$ (или $\left\{\rho_{k_0}, \mathbf{x}^{(k_0)}\right\}$, если выполнить еще один дополнительный шаг 1.1).

Сдвиги на отношения Рэлея при наличии ортонормированного базиса из собственных векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ обеспечивают асимптотически кубическую скорость сходимости ***последовательности Рэлея*** $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ к некоторому из векторов этого базиса [18, 26, 54]. К какому именно собственному вектору — зависит от выбора начального вектора этой последова-

---

* RQI — *Rayleigh Quotient Iteration* (англ.). Можно встретить и другое название: ***RQ-итерация*** [26].

** Наряду с типичной проверкой на малость нормы невязки $\mathbf{A}\mathbf{x}^{(k)} - \rho_k\mathbf{x}^{(k)}$, рекомендуют и такой вариант останова: $\|\mathbf{y}^{(k)}\| > C$, где $C > 0$ — задаваемая большая константа [54].

тельности; беря различные линейно независимые векторы $\mathbf{x}^{(0)}$, можно получать разные собственные пары данной симметричной матрицы $\mathbf{A}$. При этом, правда, без дополнительных условий (типа (4.18) применительно к ρ_0 в роли σ) нельзя гарантировать, что найденное как предел последовательности $\rho_0, \rho_1, \rho_2, \ldots$ собственное число будет ближайшим к числу ρ_0. Для организации предсказуемого поведения итераций с отношениями Рэлея требуется изучение достаточно тонких свойств таких отношений [64].

Чтобы если не обосновать, то хотя бы осмыслить RQI-алгоритм, нужно вспомнить свойство 4.6, согласно которому при выбранном векторе $\mathbf{x}^{(0)}$ вычисленное на первом шаге при $k=1$ отношение Рэлея $\rho_0 := \left(\mathbf{A}\mathbf{x}^{(0)}, \mathbf{x}^{(0)}\right) \Big/ \left(\mathbf{x}^{(0)}, \mathbf{x}^{(0)}\right)$ можно считать некоторым приближением к собственному числу, связанному с заданным в $\mathbb{R}_n$ направлением $\mathbf{x}^{(0)}$. С этим начальным приближением к какому-то собственному числу λ_j далее выполняются обратные итерации с переменными сдвигами, как и в (4.24), только приближения к λ_j находят не через отношения координат векторов $\mathbf{y}^{(k)}$ и $\mathbf{x}^{(k-1)}$, а через отношения Рэлея (как в методе скалярных произведений, см. замечание 4.2 в § 4.2), причем, поскольку здесь ρ_k приближает собственное число *данной* матрицы $\mathbf{A}$, а не «сдвинутой», нет необходимости корректировать получаемое значение на величину смещения спектра, что непременно требовалось делать в соответствии с формулой (4.23) и с последней из формул (4.24).

Замечание 4.7. RQI-алгоритм допускает использование любых векторных норм. Более естественно здесь применение евклидовой нормы; в таком случае, в силу $\|\mathbf{x}^{(k-1)}\|_2 := \sqrt{\left(\mathbf{x}^{(k-1)}, \mathbf{x}^{(k-1)}\right)} = 1$, вычисление ρ_{k-1} можно проводить по формуле

$$\rho_{k-1} = \left(\mathbf{A}\mathbf{x}^{(k-1)}, \mathbf{x}^{(k-1)}\right).$$

Замечание 4.8. Ясно, что применение переменных сдвигов в методе обратных итераций сильно ухудшает от шага к шагу обусловленность матриц решаемых там СЛАУ (они быстро приближаются к вырожденным). Однако, как показали проведенные Уилкинсоном* исследования [67], это не сказывается на достижимой точности получаемых таким методом результатов. Более того, Парлетт [54] аргументированно утверждает полезность плохой обусловленности (редкий случай!) матриц линейных систем в методах обратных итераций с хорошими сдвигами; объяснением этому парадоксальному явлению служит сосредоточение ошибок округлений именно в направлении искомого собственного вектора, что только ускоряет доминирование нужной составляющей.

Пример 4.2. К симметричной матрице $\mathbf{A} := \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ из примера 4.1 применим RQI-алгоритм.

Примем за начальный вектор $\mathbf{x}^{(0)} := \frac{1}{\sqrt{5}}\begin{pmatrix} 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 0{,}6 \\ 0{,}8 \end{pmatrix}$ с равной единице евклидовой нормой и посмотрим, как поведет себя процесс обратных итераций с отношениями Рэлея (которые будем подсчитывать по упрощенной формуле, учитывая замечание 4.7). Следуя алгоритму, последовательно получаем:

$$\rho_0 := \left(\mathbf{A}\mathbf{x}^{(0)},\ \mathbf{x}^{(0)}\right) = \left(\begin{pmatrix} 0{,}4 \\ 1 \end{pmatrix},\ \begin{pmatrix} 0{,}6 \\ 0{,}8 \end{pmatrix}\right) = 1{,}04\,;$$

$$\left(\mathbf{A} - \rho_0\mathbf{E}\right)\mathbf{y}^{(1)} = \mathbf{x}^{(0)} \quad \Leftrightarrow \quad \begin{cases} 0{,}96\, y_1^{(1)} - \ y_2^{(1)} = 0{,}6, \\ -y_1^{(1)} + 0{,}96\, y_2^{(1)} = 0{,}8 \end{cases} \quad \Leftrightarrow \quad \mathbf{y}^{(1)} = \begin{pmatrix} -17{,}551020 \\ -17{,}448980 \end{pmatrix},$$

$$\| \mathbf{y}^{(1)} \|_2 \approx 24{,}748843\,, \quad \mathbf{x}^{(1)} = \frac{\mathbf{y}^{(1)}}{\| \mathbf{y}^{(1)} \|_2} \approx \begin{pmatrix} -0{,}709165 \\ -0{,}705043 \end{pmatrix};$$

далее при $k := 2$ аналогично имеем:

$$\rho_1 := \left(\mathbf{A}\mathbf{x}^{(1)},\ \mathbf{x}^{(1)}\right) \approx \left(\begin{pmatrix} -0{,}713288 \\ -0{,}700921 \end{pmatrix},\ \begin{pmatrix} -0{,}709165 \\ -0{,}705043 \end{pmatrix}\right) \approx 1{,}000018\,;$$

---

* Уи́лкинсон Джеймс Харди (1919–1986) — английский математик, сподвижник Тьюринга в деле конструирования и тестирования первых британских компьютеров, заложивший основы способа исследования методов на вычислительную устойчивость, называемого *обратным анализом ошибок* [36] (см. об этом в § 8.2).

$$(\mathbf{A}-\rho_1\mathbf{E})\mathbf{y}^{(2)}=\mathbf{x}^{(1)} \quad \Leftrightarrow \quad \mathbf{y}^{(2)}\approx\begin{pmatrix}39283,201\\39283,203\end{pmatrix},$$

$$\|\mathbf{y}^{(2)}\|_2\approx 55554,84\,, \quad \mathbf{x}^{(2)}:=\frac{\mathbf{y}^{(2)}}{\|\mathbf{y}^{(2)}\|_2}\approx\begin{pmatrix}0,70710674\\0,70710678\end{pmatrix}.$$

Вектор $\mathbf{x}^{(2)}$ и скаляр $\rho_2:=\left(\mathbf{A}\mathbf{x}^{(2)},\ \mathbf{x}^{(2)}\right)\approx 0,99999994$ с достаточно высокой точностью представляют младшую собственную пару $\{\lambda_2,\ \mathbf{x}_2\}$ данной матрицы $\mathbf{A}$.

Кроме предсказанной высокой скорости сходимости метода, обратим внимание на быстрый рост здесь величин $\|\mathbf{y}^{(k)}\|$, что, как ранее подмечено, можно положить в основу критерия окончания итерационного процесса. Использование одного естественного критерия $|\rho_k-\rho_{k-1}|<\varepsilon$ в некоторых случаях может подвести по причине зацикливания. Чтобы убедиться в этом, достаточно провести в условиях данного примера простейшие, буквально устные, вычисления, начинаемые с вектора $\mathbf{x}^{(0)}=(0;\ 1)^T$. Подчеркнем, что выбор начального вектора в данном методе играет весьма существенную роль.

§ 4.4. МЕТОД ВРАЩЕНИЙ ЯКОБИ РЕШЕНИЯ СИММЕТРИЧНОЙ ПОЛНОЙ ПРОБЛЕМЫ СОБСТВЕННЫХ ЗНАЧЕНИЙ

Все дальнейшие способы приближенного вычисления собственных чисел вещественных квадратных матриц будут так или иначе базироваться на построении последовательностей матриц, подобных данной, имеющих пределом матрицу такого вида, из которого непосредственно можно извлечь искомые значения.

Пусть $\mathbf{A}$ — с и м м е т р и ч н а я вещественная матрица. Пользуясь известным фактом о наличии у таких матриц полной ортонормированной системы собственных векторов, т.е. тем, что матрица $\mathbf{X}$ из таких собственных векторов в этом случае является ортогональной ($\mathbf{X}^{-1}=\mathbf{X}^T$), запишем как следствие свойства 4.8 (см. § 4.1) равенство

$$\mathbf{\Lambda}=\mathbf{X}^T\mathbf{A}\mathbf{X}. \tag{4.25}$$

Значит, для всякой симметричной матрицы $\mathbf{A}$ найдется диагональная матрица $\mathbf{\Lambda}$, ей ортогонально подобная. Вопрос теперь состоит в том, как реализовать хотя бы приближенно равенство

(4.25), которое позволило бы найти сразу все собственные числа матрицы $\mathbf{A}$ (элементы диагонали матрицы $\mathbf{\Lambda}$) и все соответствующие им нормированные собственные векторы (столбцы матрицы $\mathbf{X}$)? Один из возможных ответов на этот вопрос состоит в применении к $\mathbf{A}$ последовательности однотипных преобразований, сохраняющих спектр и приводящих в пределе данную матрицу к диагональному виду.

Для этих целей будем использовать ортогональные преобразования с помощью матриц плоских вращений $\mathbf{T}_{ij}$, описание которых дано в § 1.5.

Выполним последовательно умножение матриц: $\mathbf{A}$ на $\mathbf{T}_{ij}$, затем $\mathbf{T}_{ij}^{T}$ на $\mathbf{AT}_{ij}$. Учитывая специфику структуры матрицы $\mathbf{T}_{ij}$, в силу которой такое преобразование затронет только i-е и j-е столбцы и строки матрицы $\mathbf{A}$, имеем:

$$\mathbf{AT}_{ij} = \begin{pmatrix} a_{11} & \cdots & a_{1i} & \cdots & a_{1j} & \cdots & a_{1n} \\ \cdots & & \cdots & & \cdots & & \cdots \\ a_{1i} & \cdots & a_{ii} & \cdots & a_{ij} & \cdots & a_{in} \\ \cdots & & \cdots & & \cdots & & \cdots \\ a_{1j} & \cdots & a_{ij} & \cdots & a_{jj} & \cdots & a_{jn} \\ \cdots & & \cdots & & \cdots & & \cdots \\ a_{1n} & \cdots & a_{in} & \cdots & a_{jn} & \cdots & a_{nn} \end{pmatrix} \cdot \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \cdots & & \cdots & & \cdots & & \cdots \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ \cdots & & \cdots & & \cdots & & \cdots \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ \cdots & & \cdots & & \cdots & & \cdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} a_{11} & \cdots & ca_{1i} - sa_{1j} & \cdots & sa_{1i} + ca_{1j} & \cdots & a_{1n} \\ \cdots & & \cdots & & \cdots & & \cdots \\ a_{1i} & \cdots & ca_{ii} - sa_{ij} & \cdots & sa_{ii} + ca_{ij} & \cdots & a_{in} \\ \cdots & & \cdots & & \cdots & & \cdots \\ a_{1j} & \cdots & ca_{ij} - sa_{jj} & \cdots & sa_{ij} + ca_{jj} & \cdots & a_{jn} \\ \cdots & & \cdots & & \cdots & & \cdots \\ a_{1n} & \cdots & ca_{in} - sa_{jn} & \cdots & sa_{in} + ca_{jn} & \cdots & a_{nn} \end{pmatrix};$$

$$\mathbf{B} := \begin{pmatrix} b_{11} \cdots b_{1i} \cdots b_{1j} \cdots b_{1n} \\ \cdots \quad \cdots \quad \cdots \quad \cdots \\ b_{1i} \cdots b_{ii} \cdots b_{ij} \cdots b_{in} \\ \cdots \quad \cdots \quad \cdots \quad \cdots \\ b_{1j} \cdots b_{ij} \cdots b_{jj} \cdots b_{jn} \\ \cdots \quad \cdots \quad \cdots \quad \cdots \\ b_{1n} \cdots b_{in} \cdots b_{jn} \cdots b_{nn} \end{pmatrix} := \mathbf{T}_{ij}^{T}\mathbf{AT}_{ij} = \begin{pmatrix} 1 \cdots 0 \cdots 0 \cdots 0 \\ \cdots \quad \cdots \quad \cdots \quad \cdots \\ 0 \cdots c \cdots -s \cdots 0 \\ \cdots \quad \cdots \quad \cdots \quad \cdots \\ 0 \cdots s \cdots c \cdots 0 \\ \cdots \quad \cdots \quad \cdots \quad \cdots \\ 0 \cdots 0 \cdots 0 \cdots 1 \end{pmatrix} \cdot \mathbf{AT}_{ij} =$$

$$= \begin{pmatrix} a_{11} & \cdots & ca_{1i} - sa_{1j} & \cdots & sa_{1i} + ca_{1j} & \cdots & a_{1n} \\ \cdots & & \cdots & & \cdots & & \cdots \\ ca_{1i} - sa_{1j} \cdots & & c^2 a_{ii} + s^2 a_{jj} - 2csa_{ij} & & \cdots cs(a_{ii} - a_{jj}) + (c^2 - s^2)a_{ij} \cdots & & ca_{in} - sa_{jn} \\ \cdots & & \cdots & & \cdots & & \cdots \\ sa_{1i} + ca_{1j} \cdots & & cs(a_{ii} - a_{jj}) + (c^2 - s^2)a_{ij} \cdots & & s^2 a_{ii} + c^2 a_{jj} + 2csa_{ij} & & \cdots sa_{in} + ca_{jn} \\ \cdots & & \cdots & & \cdots & & \cdots \\ a_{1n} & \cdots & ca_{in} - sa_{jn} & \cdots & sa_{in} + ca_{jn} & \cdots & a_{nn} \end{pmatrix}.$$

Помня о том, что числа c и s в матрице $\mathbf{T}_{ij}$ должны быть связаны соотношением

$$c^2 + s^2 = 1, \qquad (4.26)$$

т.е. могут интерпретироваться как косинус и синус угла поворота α в плоскости, определяемой индексами i и j, подбираем этот угол так, чтобы на месте элемента a_{ij} (его называют ***ключевым элементом***) в преобразованной матрице появился нуль (в связи с этим a_{ij} называют еще ***обреченным элементом*** [54]). Ввиду очевидной симметричности полученной выше матрицы **B** нуль при этом одновременно окажется и в позиции (j,i). Таким образом,

$$b_{ij} \equiv b_{ji} = 0 \quad \Leftrightarrow \quad cs(a_{ii} - a_{jj}) + (c^2 - s^2)a_{ij} = 0 \ .$$

Последнее равенство представляем в виде

$$\frac{cs}{c^2 - s^2} = \frac{a_{ij}}{a_{jj} - a_{ii}},$$

что, в свою очередь, полагая $c := \cos\alpha$, $s := \sin\alpha$ и применяя формулы $cs = \dfrac{\sin 2\alpha}{2}$, $c^2 - s^2 = \cos 2\alpha$, записываем как

$$\operatorname{tg} 2\alpha = \frac{2a_{ij}}{a_{jj} - a_{ii}} \text{ (для определенности считают } \alpha \in \left(-\frac{\pi}{4}, \frac{\pi}{4}\right]).$$

Ясно, что нет необходимости находить непосредственно угол α, поскольку нужные для выполнения преобразований числа c и s можно получить через значение $\operatorname{tg} 2\alpha$ по формулам тригонометрии. При этом сразу отметим, что наибольшие требования к точ-

ности в описываемом методе предъявляются именно на стадии вычисления c и s, так как здесь возможны наибольшие потери точности, а искажение значений c и s нарушает ортогональность матриц $\mathbf{T}$, что ведет к неустранимым погрешностям (метод вращений, итерационный по форме, не является итерационным по сути: ему не присуща самоисправляемость методов последовательных приближений).

Матрица $\mathbf{T}_{ij}$ ортогональна при любых $i, j \in \{1, 2, \ldots, n\}$ и, значит, матрица

$$\mathbf{B} := \mathbf{T}_{ij}^T \mathbf{A} \mathbf{T}_{ij} \tag{4.27}$$

подобна $\mathbf{A}$, т.е. имеет тот же набор собственных чисел, что и матрица $\mathbf{A}$.

Классический итерационный метод вращений решения полной симметричной проблемы собственных значений, предложенный Якоби (1846), состоит в построении последовательности матриц

$$\mathbf{B}_0 (:= \mathbf{A}),\ \mathbf{B}_1,\ \mathbf{B}_2,\ \ldots,\ \mathbf{B}_k,\ \ldots$$

с помощью преобразований типа (4.27):

$$\mathbf{B}_k = \mathbf{T}_{ij}^T \mathbf{B}_{k-1} \mathbf{T}_{ij}\,. \tag{4.28}$$

Условие, накладываемое на эти преобразования подобия такое, что *на k-м шаге должен аннулироваться максимальный по модулю элемент матрицы $\mathbf{B}_{k-1}$ предыдущего шага* (а значит, и симметричный ему элемент). Данная стратегия определяет способ фиксирования ключевого элемента и тем самым пары индексов i и j, задающих позиции $(i, i), (j, j), (i, j), (j, i)$ «существенных» элементов в матрице вращения $\mathbf{T}_{ij}$, и угол поворота α, конкретизирующий значения этих элементов, т.е. $c = \cos\alpha$ и $\pm s = \pm \sin\alpha$. Каждый шаг таких преобразований требует пересчета только двух строк (или двух столбцов, что неважно, в силу симметрии) матрицы предыдущего шага. Хотя, к сожалению, нельзя рассчитывать, что таким путем за конечное число шагов можно точно найти диагональную матрицу $\mathbf{\Lambda}$, ибо полученные на некотором этапе

преобразований нулевые элементы на следующем этапе станут, вообще говоря, ненулевыми, но нужное предельное поведение

$$\mathbf{B}_k \xrightarrow[k\to\infty]{} \mathbf{\Lambda},$$

как будет показано ниже, есть.

Определив идею и выполнив подготовительную работу по выводу формул, опишем теперь **один шаг метода вращений Якоби.**

Для того чтобы не перегружать формулы лишними индексами, будем считать, что преобразуется матрица $\mathbf{A} := (a_{ml})_{m,\,l=1}^{n}$ в матрицу $\mathbf{B} := (b_{ml})_{m,\,l=1}^{n}$ согласно равенству (4.27), хотя на самом деле на k-м шаге должно применяться преобразование (4.28) к матрице $\mathbf{B}_{k-1} := (b_{ml}^{(k-1)})$ с результатом $\mathbf{B}_k := (b_{ml}^{(k)})$.

Итак, пусть a_{ij} — ключевой элемент преобразуемой матрицы $\mathbf{A}$. Матрица $\mathbf{B}$, подобная $\mathbf{A}$, формируется следующим образом (вычисление значений c и s здесь организовано так, чтобы минимизировались их погрешности):

1. Вычисляют $p := 2a_{ij}, \quad q := a_{jj} - a_{ii}, \qquad d := \sqrt{p^2 + q^2}$.
2. Если $q \neq 0$, то $r := |q|/(2d), \qquad c := \sqrt{0.5 + r}$, $s := \sqrt{0.5 - r} \cdot \operatorname{sgn}_+(pq)$ (если $|p| << |q|$, то лучше $s := |p| \cdot \operatorname{sgn}_+(pq)/(2cd)$);
   если же $q = 0$, то $c = s := \sqrt{2}/2$.
3. Вычисляют новые диагональные элементы:
   $$b_{ii} := c^2 a_{ii} + s^2 a_{jj} - 2csa_{ij}, \quad b_{jj} := s^2 a_{ii} + c^2 a_{jj} + 2csa_{ij}.$$
4. Полагают $b_{ij} = b_{ji} := 0$ (или для контроля вычисляют
   $$b_{ij} = b_{ji} := \left(c^2 - s^2\right)a_{ij} + cs\left(a_{ii} - a_{jj}\right)).$$
5. При $m = 1, 2, \ldots, n$ таких, что $m \neq i$, $m \neq j$, вычисляют изменяющиеся внедиагональные элементы:
   $$\begin{aligned} b_{im} = b_{mi} &:= ca_{mi} - sa_{mj}, \\ b_{jm} = b_{mj} &:= sa_{mi} + ca_{mj}. \end{aligned} \tag{4.29}$$

6. Для всех остальных пар индексов m, l принимают $b_{ml} := a_{ml}$.

Конечно, в реальных вычислениях, если это считать основой алгоритма, не все записанное здесь следует выполнять. В частности, не нужно делать последних переприсвоений.

Убедимся теперь, что, *если в качестве ключевого, или, в иной терминологии, обреченного, элемента на каждом шаге преобразований подобия по указанным формулам брать максимальный по модулю элемент преобразуемой матрицы, то в пределе получится диагональная матрица.*

Для доказательства этого, т.е. доказательства сходимости последовательности $(\mathbf{B}_k)$ к Λ, используем ***норму Фробениуса****:

$$\|\mathbf{A}\|_F := \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}\,.$$

Проследим за поведением норм матриц (точнее, квадратов этих норм), получающихся из матриц $\mathbf{B}_k$ заменой диагональных элементов нулями. Такие матрицы, определяемые внедиагональными элементами матриц $\mathbf{B}_k$, будем обозначать $\mathbf{V}_{\mathbf{B}_k}$. При этом опять с целью упрощения записей будем пока рассматривать переход от $\mathbf{A}$ к $\mathbf{B}$, совершаемый в соответствии с формулой (4.27).

Найдем выражение суммы квадратов внедиагональных элементов матрицы

$$\mathbf{B} := \mathbf{T}_{ij}^T \mathbf{A} \mathbf{T}_{ij} = \begin{pmatrix} a_{11} & \dots & b_{1i} & \dots & b_{1j} & \dots & a_{1n} \\ \dots & & \dots & & \dots & & \dots \\ b_{i1} & \dots & b_{ii} & \dots & 0 & \dots & b_{in} \\ \dots & & \dots & & \dots & & \dots \\ b_{j1} & \dots & 0 & \dots & b_{jj} & \dots & b_{jn} \\ \dots & & \dots & & \dots & & \dots \\ a_{n1} & \dots & b_{ni} & \dots & b_{nj} & \dots & a_{nn} \end{pmatrix},$$

принадлежащих изменяющимся по сравнению с $\mathbf{A}$ строке i и столбцу j, через элементы матрицы $\mathbf{A}$ (здесь, в отличие от предыдущих записей, в элементах не отражена явным образом

---

* Фробениус Фердинанд Георг (1849–1917) — немецкий математик.

симметричность, имеющая место на самом деле). С учетом (4.29), (4.26) и условия аннулирования элементов $b_{ij} = b_{ji}$ (что позволяет формально внести их в рассматриваемую сумму) имеем:

$$\sum_{m\neq i} b_{mi}^2 + \sum_{m\neq j} b_{mj}^2 = \sum_{m\neq i,\, j} \left(c^2 a_{mi}^2 - 2csa_{mi}a_{mj} + s^2 a_{mj}^2\right) +$$

$$+ b_{ij}^2 + \sum_{m\neq j,\, i} \left(s^2 a_{mi}^2 + 2csa_{mi}a_{mj} + c^2 a_{mj}^2\right) + b_{ji}^2 =$$

$$= \sum_{m\neq i,\, j} \left[\left(c^2 + s^2\right)a_{mi}^2 + \left(c^2 + s^2\right)a_{mj}^2\right] = \sum_{m\neq i,\, j} \left(a_{mi}^2 + a_{mj}^2\right).$$

Аналогичные суммы квадратов j-й строки и i-го столбца, в силу симметрии, дадут такое же выражение. Это означает, что если полученное равенство удвоить и дополнить левую и правую части суммой квадратов всех остальных внедиагональных элементов матрицы $\mathbf{A}$ (служащих соответствующими элементами и матрицы $\mathbf{B}$), то в левой части будет стоять сумма квадратов всех внедиагональных элементов матрицы $\mathbf{B}$, а в правой части — сумма квадратов всех внедиагональных элементов матрицы $\mathbf{A}$, кроме a_{ji}^2 и a_{ij}^2. Следовательно, справедливо равенство

$$\|\mathbf{V_B}\|_F^2 = \|\mathbf{V_A}\|_F^2 - 2a_{ij}^2, \tag{4.30}$$

говорящее об *убывании сумм квадратов внедиагональных элементов в рассматриваемом процессе преобразований подобия.*

Пусть $|a_{ij}| = \max\limits_{m\neq l}\{|a_{ml}|\}$. Тогда можно считать, что значение $a_{ij}^2 = \max\limits_{m\neq l}\left\{a_{ml}^2\right\}$ не меньше, чем среднее значение множества из $n^2 - n$ квадратов всех внедиагональных элементов $n\times n$-матрицы $\mathbf{A}$, т.е. не превосходит величины

$$\frac{1}{n(n-1)}\sum_{m\neq l} a_{ml}^2 = \frac{1}{n(n-1)}\|\mathbf{V_A}\|_F^2.$$

Подставляя полученную оценку $a_{ij}^2 \geq \frac{1}{n(n-1)}\|\mathbf{V_A}\|_F^2$ в равенство (4.30), приходим к неравенству

$$\|\mathbf{V_B}\|_F^2 \leq \left(1-\frac{2}{n(n-1)}\right)\|\mathbf{V_A}\|_F^2 .$$

На основании этого для последовательности матриц $\mathbf{B}_k$, представляемых в виде

$$\mathbf{B}_k = \operatorname{diag}\left(b_{ii}^{(k)}\right) + \mathbf{V}_{\mathbf{B}_k},$$

можно записать:

$$\|\mathbf{V}_{\mathbf{B}_k}\|_F^2 \leq \left(1-\frac{2}{n(n-1)}\right)\|\mathbf{V}_{\mathbf{B}_{k-1}}\|_F^2 \leq \ldots \leq$$

$$\leq \left(1-\frac{2}{n(n-1)}\right)^k \|\mathbf{V_A}\|_F^2 \xrightarrow[k\to\infty]{} 0 .$$

Значит, при указанном способе выбора ключевого элемента последовательность подобных матриц $\mathbf{B}_k$ гарантированно сходится к диагональной матрице $\mathbf{\Lambda}$ из собственных значений, по крайней мере, со скоростью геометрической прогрессии.

Описанный выше классический вариант метода вращений Якоби, как показывают более тонкие оценки, на самом деле имеет асимптотически квадратичную скорость сходимости [17, 43, 54, 67]. Однако при больших размерностях n его реализация наталкивается на существенные потери машинных ресурсов, связанные с поиском наибольшего по модулю ключевого элемента. Поэтому обычно применяют другие версии метода вращений Якоби: ***циклический метод Якоби*** и ***циклический с барьерами*** (иначе, *преградами* [67]).

В циклическом методе вращений в качестве ключевого элемента поочередно используются все внедиагональные элементы верхнего или нижнего треугольника преобразуемой матрицы (перебираемые, например, построчно). Сходимость итерационного

процесса при этом, естественно, замедляется, но остается асимптотически квадратичной.

Стратегия выбора ключевого элемента в циклическом методе Якоби с барьерами такова: устраивают циклический перебор всех над- или поддиагональных элементов матрицы $\mathbf{A}$ (точнее, $\mathbf{B}_{k-1}$) для их использования в роли обреченного, но при этом пропускают элементы, абсолютные величины которых меньше некоторого заданного положительного числа — барьера. Этот барьер может быть изменяемым от итерации к итерации (уменьшающимся по какому-либо осмысленному принципу).

Следует отметить успешную работу таких итерационных процессов и в случаях, когда исходная матрица имеет кратные и, что хуже, близкие собственные числа.

В завершение остается вспомнить, что в соответствии со свойствами 4.7, 4.8 и проведенными рассуждениями за собственные векторы $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ матрицы $\mathbf{A}$ (имеющие единичную евклидову норму) могут быть приближенно приняты столбцы результирующей матрицы, получающейся справа от $\mathbf{A}$ в следующей цепочке преобразований подобия:

$$\mathbf{B}_1 = \mathbf{T}_{i_0 j_0}^T \mathbf{A} \mathbf{T}_{i_0 j_0},$$

$$\mathbf{B}_2 = \mathbf{T}_{i_1 j_1}^T \mathbf{B}_1 \mathbf{T}_{i_1 j_1} = \mathbf{T}_{i_1 j_1}^T \mathbf{T}_{i_0 j_0}^T \mathbf{A} \mathbf{T}_{i_0 j_0} \mathbf{T}_{i_1 j_1},$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\mathbf{B}_k = \mathbf{T}_{i_{k-1} j_{k-1}}^T \ldots \mathbf{T}_{i_0 j_0}^T \mathbf{A} \mathbf{T}_{i_0 j_0} \ldots \mathbf{T}_{i_{k-1} j_{k-1}},$$

т.е. столбцы матрицы

$$\mathbf{T}_k := \mathbf{T}_{i_0 j_0} \mathbf{T}_{i_1 j_1} \cdot \ldots \cdot \mathbf{T}_{i_{k-1} j_{k-1}}$$

при некотором $k := K$. Разумеется, в реальном вычислительном процессе нужно обойтись без матричных умножений.

Сигналом для окончания процесса вращений Якоби может служить, например, достаточная малость одной из следующих величин: $\left\| \mathbf{V}_{\mathbf{B}_k} \right\|_F$, $\left| b_{ij}^{(k)} \right|$ (или обоих одновременно).

Пример 4.3. Методом вращений Якоби решим задачу нахождения всех собственных пар матрицы $\mathbf{A} := \begin{pmatrix} 1 & 1 & 3 \\ 1 & 5 & 1 \\ 3 & 1 & 1 \end{pmatrix}$.

К данной симметричной (незнакоопределенной) матрице $\mathbf{A}$ будем поэтапно применять записанный выше основной фрагмент алгоритма, переводящий ее в матрицу $\mathbf{B}$, являющуюся подобной $\mathbf{A}$ и имеющую меньшую сумму квадратов внедиагональных элементов. При этом условимся, что ключевой элемент a_{ij} будем брать в наддиагональной части матрицы $\mathbf{A}$ (можно и наоборот, ничего от этого принципиально не изменится) и что все вычисления будем проводить точно (т.е. рассматриваем идеальный процесс).

Этап 1-*й*. Выбираем ключевой элемент $a_{13} = 3$ (максимальный в условленной части матрицы), следовательно, фиксируем индексы $i := 1$, $j := 3$. Вычисляем

$$p := 2a_{13} = 6\,, \quad q = a_{33} - a_{11} = 1 - 1 = 0 \;\Rightarrow\; c = s = \frac{1}{\sqrt{2}}\,.$$

Далее находим элементы новой матрицы

$$b_{11} := c^2 a_{11} + s^2 a_{33} - 2csa_{13} = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 1 - 2\cdot\frac{1}{2}\cdot 3 = -2\,,$$

$$b_{33} := s^2 a_{11} + c^2 a_{33} + 2csa_{13} = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 1 + 2\cdot\frac{1}{2}\cdot 3 = 4$$

и при $m := 2$ ($\neq i$, $\neq j$)

$$b_{12} = b_{21} := ca_{21} - sa_{23} = \frac{1}{\sqrt{2}}(1-1) = 0\,,$$

$$b_{23} = b_{32} := ca_{21} + sa_{23} = \frac{1}{\sqrt{2}}(1+1) = \sqrt{2}\,.$$

Следовательно, данная матрица $\mathbf{A}$ подобна матрице

$$\mathbf{B} := \begin{pmatrix} -2 & 0 & 0 \\ 0 & 5 & \sqrt{2} \\ 0 & \sqrt{2} & 4 \end{pmatrix}.$$

Этап 2-*й*. Полученную на предыдущем этапе матрицу $\mathbf{B}$ принимаем за матрицу $\mathbf{A}$, т.е. будем считать по тем же формулам, положив $\mathbf{A} := \mathbf{B}$. Ключевой элемент здесь $a_{23} = \sqrt{2}$, т.е. $i := 2$, $j := 3$. Согласно алгоритму, имеем:

$$p := 2a_{23} = 2\sqrt{2}\,, \quad q := a_{33} - a_{22} = 4 - 5 = -1 \text{ (замечаем, что } pq < 0\text{)},$$

$$d := \sqrt{p^2 + q^2} = \sqrt{8+1} = 3, \quad r = \frac{|q|}{2d} = \frac{1}{6},$$

$$c := \sqrt{0.5 + r} = \sqrt{\frac{1}{2} + \frac{1}{6}} = \frac{2}{\sqrt{6}}, \quad s := \sqrt{0.5 - r} \cdot \mathrm{sgn}_+(pq) = -\sqrt{\frac{1}{2} - \frac{1}{6}} = -\frac{\sqrt{2}}{\sqrt{6}}.$$

Зная числа c и s, определяющие преобразование вращения, вычисляем

$$b_{22} := c^2 a_{22} + s^2 a_{33} - 2csa_{23} = \frac{4}{6} \cdot 5 + \frac{2}{6} \cdot 4 + \frac{8}{6} = 6,$$

$$b_{33} := s^2 a_{22} + c^2 a_{33} + 2csa_{23} = \frac{2}{6} \cdot 5 + \frac{4}{6} \cdot 4 - \frac{8}{6} = 3$$

и далее при $m := 1$

$$b_{13} = b_{31} := sa_{12} + ca_{13} = \left(-\frac{\sqrt{2}}{\sqrt{6}}\right) \cdot 0 + \frac{2}{\sqrt{6}} \cdot 0 = 0,$$

$$b_{12} = b_{21} := ca_{12} - sa_{13} = \frac{2}{\sqrt{6}} \cdot 0 + \frac{\sqrt{2}}{\sqrt{6}} \cdot 0 = 0.$$

Таким образом, уже после второго этапа (это не норма, а, будем считать, везение) получена диагональная матрица

$$\mathbf{B} := \begin{pmatrix} -2 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 3 \end{pmatrix},$$

подобная исходной матрице $\mathbf{A}$, в силу чего можно утверждать, что собственными значениями матрицы $\mathbf{A}$ являются числа

$$\lambda_1 := -2, \quad \lambda_2 := 6, \quad \lambda_3 := 3.$$

Чтобы найти отвечающие им собственные векторы, выпишем матрицы плоских вращений первого и второго этапов (в соответствии с их структурой вида (1.25)):

$$\mathbf{T}_{ij}^{(1)} := \mathbf{T}_{13} := \begin{pmatrix} c^{(1)} & 0 & s^{(1)} \\ 0 & 1 & 0 \\ -s^{(1)} & 0 & c^{(1)} \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \end{pmatrix},$$

$$\mathbf{T}_{ij}^{(2)} := \mathbf{T}_{23} := \begin{pmatrix} 1 & 0 & 0 \\ 0 & c^{(2)} & s^{(2)} \\ 0 & -s^{(2)} & c^{(2)} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2/\sqrt{6} & -\sqrt{2}/\sqrt{6} \\ 0 & \sqrt{2}/\sqrt{6} & 2/\sqrt{6} \end{pmatrix}.$$

Их произведение, т.е. матрица

$$\mathbf{T} := \mathbf{T}_{13} \cdot \mathbf{T}_{23} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{6} & \sqrt{2}/\sqrt{6} \\ 0 & 2/\sqrt{6} & -\sqrt{2}/\sqrt{6} \\ -1/\sqrt{2} & 1/\sqrt{6} & \sqrt{2}/\sqrt{6} \end{pmatrix},$$

согласно последним теоретическим выкладкам этого параграфа, имеет своими столбцами собственные векторы матрицы $\mathbf{A}$. Легко проверить по определению, что ее столбцы в естественном порядке, т.е. векторы

$$\mathbf{x}_1 := \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{x}_2 := \frac{1}{\sqrt{6}}\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \quad \mathbf{x}_3 := \frac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix},$$

образуют собственные пары с числами $\lambda_1 := -2$, $\lambda_2 := 6$ и $\lambda_3 := 3$, соответственно.

§ 4.5. МЕТОД БИСЕКЦИЙ

К выводу метода ***бисекций*** (***деления спектра*** [54]) имеются разные подходы, так или иначе базирующиеся на законе сохранения инерции симметричной матрицы. В основе одного подхода — подсчет инерции параметризованной матрицы приведением ее к диагональному виду $\mathbf{U}^T\mathbf{D}\mathbf{U}$- или $\mathbf{L}\mathbf{D}\mathbf{L}^T$-разложением [26], другого — с помощью анализа знаков последовательности ее главных миноров [17], третьего — построением систем многочленов Штурма [51]. Здесь будет развит первый из названных подходов.

Всякую симметричную матрицу ортогональными преобразованиями можно привести к подобной ей матрице трехдиагонального вида. Технология таких преобразований подобия показана далее (см. § 5.2 и, в частности, замечание 5.1). Поэтому без ограничения общности дальнейшие рассуждения о спектре симметричной матрицы будем проводить для матриц с трехдиагональной структурой, т.е. для таких $n \times n$-матриц $\mathbf{A} := (a_{ij})$, у которых

$$a_{ii} := \alpha_i \quad (i = 1, 2, ..., n),$$
$$a_{i,i+1} \equiv a_{i+1,i} := \beta_i \quad (i = 1, 2, ..., n-1),$$
$$a_{ij} := 0 \quad \text{при} \quad |i-j| > 1 .$$

Рассмотрим σ-параметризованную матрицу $\mathbf{A} - \sigma\mathbf{E}$ (определитель которой, очевидно, является для $\mathbf{A}$ характеристическим

многочленом):

$$\mathbf{A}-\sigma\mathbf{E}:=\begin{pmatrix}\alpha_1-\sigma & \beta_1 & 0 & & \\ \beta_1 & \alpha_2-\sigma & \beta_2 & & \mathbf{0} \\ 0 & \beta_2 & \alpha_3-\sigma & & \\ & & & \ddots & \\ & \mathbf{0} & & \beta_{n-2} & \alpha_{n-1}-\sigma \quad \beta_{n-1} \\ & & & 0 & \beta_{n-1} \quad \alpha_n-\sigma\end{pmatrix}. \qquad (4.31)$$

В соответствии с выкладками § 1.3 эта матрица может быть представлена в виде

$$\mathbf{A}-\sigma\mathbf{E}=\mathbf{U}^T\mathbf{D}(\sigma)\mathbf{U} \qquad (4.32)$$

с помощью формул (1.10), которые в данном случае выглядят так:

$$\begin{aligned} &d_1(\sigma):=\alpha_1-\sigma, \quad u_i=\frac{\beta_i}{d_i(\sigma)} \quad (i=1,2,...,n-1), \\ &d_i(\sigma)=(\alpha_i-\sigma)-u_{i-1}{}^2 d_{i-1}(\sigma) \quad (i=2,3,...,n). \end{aligned} \qquad (4.33)$$

Очевидно, $\mathbf{U}^T\mathbf{D}\mathbf{U}$-разложение симметричной матрицы подпадает под определение конгруэнтности 4.4; следовательно, для матриц $\mathbf{A}-\sigma\mathbf{E}$ и $\mathbf{D}(\sigma)$ справедливо утверждение теоремы Сильвестра о равенстве их инерций (см. свойство 4.10). Матрица $\mathbf{U}$ в разложении (4.32) нас не интересует, поэтому формулы (4.33) для последовательного вычисления значений $d_i(\sigma)$ можно упростить, исключив из них u_i :

$$d_1(\sigma):=\alpha_1-\sigma, \quad d_i(\sigma)=(\alpha_i-\sigma)-\frac{\beta_{i-1}{}^2}{d_{i-1}(\sigma)} \quad (i=2,3,...,n). \qquad (4.34)$$

Подсчитав число отрицательных, нулевых и положительных элементов матрицы $\mathbf{D}(\sigma):=\operatorname{diag}(d_1(\sigma);d_2(\sigma);...;d_n(\sigma))$, находим ее инерцию $(m_-(\sigma), m_0(\sigma), m_+(\sigma))$. Правда, на самом деле значение $m_0(\sigma)$ по *этим* формулам не получить, но в методе бисекций знание $m_0(\sigma)$ и не требуется.

Обычно находят число либо только отрицательных ($m_-(\sigma)$), либо только положительных ($m_+(\sigma)$) собственных значений.

Вычисление количества отрицательных собственных чисел матрицы $\mathbf{A}-\sigma\mathbf{E}$, имеющей вид (4.31), согласно формулам (4.34) можно проводить, например, с помощью следующего алгоритма:

1. $d := \alpha_1 - \sigma$;
2. *если* $d<0$, *положить* $m_-(\sigma) := 1$, *иначе* $m_-(\sigma) := 0$;
3. *для* $i := 2, 3, ..., n$:

$$d := (\alpha_i - \sigma) - \beta_{i-1}^2 / d\ ,$$

если $d<0$, *то* $m_-(\sigma) := m_-(\sigma) + 1$.

Выполнение алгоритма может прекратиться, если на каком-то этапе окажется $d=0$. В таком случае это значение d заменяется значением [54] $d := \varepsilon(|\alpha_i| + |\sigma| + \varepsilon)$ с малым $\varepsilon>0$ — допустимой погрешностью. Еще лучше начать цикл заново, слегка изменив значение σ, что в рамках рассматриваемого метода не должно играть существенной роли.

Числа $d_1(\sigma)$, $d_2(\sigma)$,..., $d_n(\sigma)$ — суть собственные значения матрицы $\mathbf{D}(\sigma)$ — не являются, вообще говоря, собственными значениями матрицы $\mathbf{A}-\sigma\mathbf{E}$, но совпадают с ними *по знаку*. Изменяя тем или иным образом значение параметра σ и выполняя подсчет значений $d_i(\sigma)$ по формулам (4.34), по изменению их знаков, т.е. по поведению индексов $m_-(\sigma)$ (и/или $m_+(\sigma)$) можно получать и уточнять информацию о собственных числах λ_i исходной матрицы $\mathbf{A}$. Например, положив $\sigma := 0$ и подсчитав значения $m_-(0)$ и $m_+(0)$, узнаем количество отрицательных и положительных собственных чисел невырожденной матрицы $\mathbf{A}$ (с учетом их кратностей).

Согласно свойству 4.2 на параметр σ в матрице $\mathbf{A}-\sigma\mathbf{E}$ можно смотреть как на значение смещения спектра матрицы $\mathbf{A}$ на величину σ, что уже использовалось ранее (и будет использоваться далее) для совершения сдвигов с целью ускорения сходимости итерационных процессов нахождения собственных пар. В соответствии с этим, если, например, при некотором значении $\sigma := a_1$ подсчитали значение индекса отрицательности $m_-(a_1)$, то по

отношению к исходной матрице $\mathbf{A}$ можно сказать, что это есть число ее собственных значений, меньших a_1. Подсчитав $m_-(a_2)$ при другом значении $\sigma := a_2 > a_1$, получаем число собственных значений матрицы $\mathbf{A}$, лежащих на числовой оси левее a_2. Таким образом, разность $m_-(a_2) - m_-(a_1)$ покажет количество собственных чисел матрицы $\mathbf{A}$, принадлежащих промежутку $[a_1, a_2)$. На этом факте и строят алгоритмы решения методом бисекций частичных проблем собственных значений симметричных матриц в разных постановках.

Так, отодвигая точку a_1 влево по оси, например, от нуля до тех пор, пока не окажется $m_-(a_1) = 0$, а точку a_2 — вправо, пока не станут все собственные числа смещенной на $a_2\mathbf{E}$ матрицы отрицательными, т.е. $m_-(a_2) = n$, можно получить границы спектра данной матрицы $\mathbf{A}$: любое ее собственное число λ_i должно лежать на промежутке $[a_1, a_2)$.

Можно наоборот, пользуясь известным неравенством $|\lambda_i| \le \|\mathbf{A}\|$, уточнять границы спектра $-\|\mathbf{A}\|$, $\|\mathbf{A}\|$ (при каких-то конкретных нормах, например $\|\cdot\|_\infty$), двигаясь вправо от значения $-\|\mathbf{A}\|$ и влево от значения $\|\mathbf{A}\|$. Таким способом могут быть, в частности, изолированы и затем приближенно найдены наименьшее и наибольшее собственные числа матрицы $\mathbf{A}$.

Одним из основных назначений метода бисекций является приближенное вычисление одного или нескольких собственных значений данной симметричной матрицы, принадлежащих заданному отрезку $[a, b]$.

Предположим, что на отрезке $[a, b]$ имеется одно простое (не кратное) собственное значение матрицы $\mathbf{A}$. Это значит, что $m_-(b) - m_-(a) = 1$. Положив $c := \dfrac{a+b}{2}$ и найдя $m_-(c)$, по разности $m_-(c) - m_-(a)$, равной 1 или 0, определяем полуинтервал для последующего исследования: $[a, c)$ или $[c, b)$ соответственно.

Такой процесс бисекции, а точнее деления отрезка пополам, продолжается до тех пор, пока длина промежутка, содержащего искомое собственное значение, не станет меньше 2ε, где ε — допустимая абсолютная погрешность. Получаем итерационный процесс, сходящийся со скоростью геометрической прогрессии со знаменателем $1/2$, каждый шаг которого требует выполнения $O(n)$ арифметических операций на реализацию вычислений по рекуррентной формуле (4.34).

Единственность собственного числа на заданном отрезке, очевидно, не играет существенной роли. Поэтому процесс дробления отрезка $[a, b]$ можно проводить до тех пор, пока не будут найдены все принадлежащие ему собственные числа, количество которых определяется величиной $m\_(b) - m\_(a)$. Препятствием не будет и наличие кратных собственных значений. Если, например, на отрезке $[a, b]$ имеется одно l-кратное собственное число, то в процессе деления отрезка пополам на одном из двух получаемых промежутков значение разности индексов отрицательности всегда будет нулем, на другом l.

Метод бисекций позволяет решать и такую задачу, как нахождение собственного числа с заданным номером в упорядоченном по возрастанию спектре данной симметричной матрицы. Для этого нужно сначала найти тем же методом такой полуинтервал $[a, b)$, для которого число $m\_(a)$ равно заданному номеру собственного значения, а $m\_(b)$ — на единицу больше (или на l, если это собственное значение l-кратное); далее это собственное значение уточняется описанным выше способом.

Поскольку метод бисекций позволяет вычислять только собственные числа, нахождение соответствующих им собственных векторов обычно осуществляют методом обратных итераций (§ 4.3). При этом в качестве сдвигов используют уже найденные приближенные значения собственных чисел, что максимально ускоряет сходимость итерационного процесса (зачастую бывает достаточно одной - двух итераций [26]).

Пример 4.4. Дана трехдиагональная матрица $\mathbf{A} := \begin{pmatrix} 2 & -1 & 0 \\ -1 & -3 & 1 \\ 0 & 1 & 2 \end{pmatrix}$.

Составим параметризованную матрицу $\mathbf{A} - \sigma\mathbf{E} := \begin{pmatrix} 2-\sigma & -1 & 0 \\ -1 & -3-\sigma & 1 \\ 0 & 1 & 2-\sigma \end{pmatrix}$.

В соответствии с ней основные расчетные формулы (4.34) метода бисекций для данного конкретного примера приобретают следующий вид:

$$d_1 := 2-\sigma, \qquad d_2 = -3-\sigma-\frac{1}{d_1}, \qquad d_3 = 2-\sigma-\frac{1}{d_2}. \tag{4.35}$$

Пользуясь формулами (4.35) при $\sigma := 0$, получаем:

$$d_1 = 2 \ \ (>0), \qquad d_2 = -3-\frac{1}{2} = -3{,}5 \ \ (<0), \qquad d_3 = 2-\frac{1}{-3{,}5} \ \ (>0).$$

Знаки этих чисел говорят о том, что спектр данной матрицы $\mathbf{A}$ состоит из одного отрицательного и двух положительных собственных значений. Рассмотрим применение метода бисекций для нахождения отрицательного собственного числа λ_1 с точностью до 0,05. Учитывая, что для любого собственного числа данной матрицы справедлива оценка $|\lambda_i| \le \|\mathbf{A}\|_\infty := 5$, за исходный промежуток возьмем $[-5, 0)$. На его концах имеем значения индекса отрицательности $m_-(-5) = 0$, $m_-(0) = 1$.

1-*я итерация*: Полагаем $\sigma := \dfrac{-5+0}{2} = -2{,}5$ и по формулам (4.35) подсчитываем

$$d_1 = 2-(-2{,}5) = 4{,}5, \quad d_2 = -3+2{,}5-\frac{1}{4{,}5} \approx -0{,}72, \quad d_3 \approx 2+2{,}5-\frac{1}{-0{,}72} \approx 5{,}89.$$

Так как $m_-(-2{,}5) = 1$, то $\lambda_1 \in [-5, -2{,}5)$.

2-*я итерация*: Середина промежутка, полученного на первой итерации, — $\dfrac{-5+(-2{,}5)}{2} = -3{,}75$. Округлим это значение и примем $\sigma := -3{,}7$ (пользуемся тем, что в принципе метод бисекций позволяет производить деление отрезка произвольным образом, т.е. не обязательно пополам). Вновь обращаемся к формулам (4.35):

$$d_1 = 2-(-3{,}7) = 5{,}7, \quad d_2 = -3+3{,}7-\frac{1}{5{,}7} \approx 0{,}52, \quad d_3 \approx 2+3{,}7-\frac{1}{0{,}52} \approx 3{,}78.$$

Отсюда получаем $m_-(-3{,}7) = 0$, следовательно, $\lambda_1 \in [-3{,}7, -2{,}5)$.

3-*я итерация*: $\sigma := \dfrac{-3,7+(-2,5)}{2} = -3,1$;

$$d_1 = 2-(-3,1) = 5,1, \quad d_2 = -3+3,1-\frac{1}{5,1} \approx -0,096, \quad d_3 \approx 2+3,1-\frac{1}{-0,096} \approx 15,5;$$

$$m\_(-3,1) = 1 \;\Rightarrow\; \lambda_1 \in [-3,7,\ -3,1).$$

4-*я итерация*: $\sigma := \dfrac{-3,7+(-3,1)}{2} = -3,4$;

$$d_1 = 2-(-3,4) = 5,4, \quad d_2 = -3+3,4-\frac{1}{5,4} \approx 0,21, \quad d_3 \approx 2+3,4-\frac{1}{0,21} \approx 0,64;$$

$$m\_(-3,4) = 0 \;\Rightarrow\; \lambda_1 \in [-3,4,\ -3,1).$$

5-*я итерация*: $\sigma := -3,3$;

$$d_1 = 2-(-3,3) = 5,3, \quad d_2 = -3+3,3-\frac{1}{5,3} \approx 0,11, \quad d_3 \approx 2+3,3-\frac{1}{0,11} \approx -3,79;$$

$$m\_(-3,3) = 1 \;\Rightarrow\; \lambda_1 \in [-3,4,\ -3,3).$$

Можно сделать заключение, что $\lambda_1 \approx -3,35\,(\pm 0,05)$ (точное значение $\lambda_1 = -0,5(1+\sqrt{33}) := -3,37228$).

УПРАЖНЕНИЯ

4.1. Сделайте два шага метода скалярных произведений для нахождения наибольшего по модулю собственного числа матрицы $\mathbf{A} := \begin{pmatrix} -2 & 4 \\ 4 & 1 \end{pmatrix}$.

4.2. Дана матрица $\mathbf{A} := \begin{pmatrix} 30 & -12 & 53 \\ -42 & 19 & -78 \\ -28 & 12 & -51 \end{pmatrix}$.

А) Степенным методом найдите несколько последовательных приближений к доминирующему собственному числу матрицы $\mathbf{A}$ и к соответствующему собственному вектору.

Б) Методом обратных итераций найдите младшую собственную пару $\{\lambda_3,\ \mathbf{x}_3\}$ данной матрицы $\mathbf{A}$. Можно ли утверждать, что $\lambda_3 = \Lambda + \lambda_1$, где Λ — наибольшее по модулю собственное число матрицы $\mathbf{A} - \lambda_1 \mathbf{E}$?

4.3. В условиях примера 4.1 (см. § 4.2) начните SP-алгоритм с вектора $\mathbf{y}^{(0)} := (1;\ 1)^T$. Что получено после выполнения одного полного цикла алгоритма? Дайте объяснение результату.

4.4. Дана матрица $\mathbf{A} := \begin{pmatrix} 5 & 2 & -3 \\ 4 & 5 & -4 \\ 6 & 4 & -4 \end{pmatrix}$.

Найдя грубые приближения к ее собственным числам степенным методом, уточните их обратными итерациями со сдвигами (см. (4.24)).

4.5. А) Проанализируйте сходимость степенного метода в случае, когда λ_1 — кратное вещественное наибольшее по модулю собственное число n-мерной матрицы простой структуры (см. формулы (4.10), (4.11)). Как можно найти все соответствующие ему собственные векторы в зависимости от показателя кратности?

Б) Что можно сказать о поведении последовательности отношений (4.10), если $\lambda_1 = -\lambda_2$ и $|\lambda_2| > |\lambda_3| \geq \ldots \geq |\lambda_n| \quad (\lambda_i \in \mathbb{R})$?

В) Рассмотрите и объясните поведение степенного метода в случае, когда данная матрица $\mathbf{A}$ — диагональная.

4.6. Найдите все собственные пары матрицы $\mathbf{A} := \begin{pmatrix} 4 & 2 & -1 \\ 2 & 4 & 1 \\ -1 & 1 & 3 \end{pmatrix}$:

а) методом скалярных произведений (для нахождения второй собственной пары используйте формулы (4.14), (4.15));

б) RQI-алгоритмом, начиная его с различных векторов.

4.7. Методом вращений Якоби найдите все собственные пары матрицы $\mathbf{A}$, если:

а) $\mathbf{A} := \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$; б) $\mathbf{A} := \begin{pmatrix} 6 & -2 & 2 \\ -2 & 5 & 0 \\ 2 & 0 & 7 \end{pmatrix}$.

4.8. Методом бисекций с точностью до 0,05 найдите второе по величине собственное число матрицы $\mathbf{A} := \begin{pmatrix} 1 & 2 & 0 \\ 2 & 5 & 1 \\ 0 & 1 & 4 \end{pmatrix}$.

Глава 5

QR-АЛГОРИТМ

§ 5.1. ПОНЯТИЕ ОБ LU-, U^TU- И QR-АЛГОРИТМАХ

Чаще (по крайней мере, в несимметричном случае) алгоритмы приближенного решения полных проблем собственных значений основываются на приведении данных матриц к подобным им матрицам не диагонального, а треугольного вида. Наиболее простой из таких алгоритмов вычисления собственных чисел опирается на LU-разложение матриц (см. § 1.2).

Пусть данная $n \times n$-матрица $\mathbf{A}$ представлена в виде $\mathbf{A} = \mathbf{LU}$, где $\mathbf{L}$ и $\mathbf{U}$ — соответственно нижняя и верхняя треугольные матрицы. Ввиду некоммутативности матричного умножения произведение $\mathbf{UL}$ дает другую матрицу. Обозначим ее $\mathbf{A}_1$.

Имеем $\mathbf{A}_1 := \mathbf{UL}$, откуда $\mathbf{U} = \mathbf{A}_1\mathbf{L}^{-1}$. Подставив это выражение матрицы $\mathbf{U}$ в равенство $\mathbf{A} = \mathbf{LU}$, получаем новое представление матрицы $\mathbf{A}$:

$$\mathbf{A} = \mathbf{L}\mathbf{A}_1\mathbf{L}^{-1}, \tag{5.1}$$

которое говорит о подобии матриц $\mathbf{A}$ и $\mathbf{A}_1$, т.е. о равенстве их собственных чисел $\lambda_{\mathbf{A}}$ и $\lambda_{\mathbf{A}_1}$ (см. определение 4.2 и свойство 4.7).

Если матрица $\mathbf{A}_1$, как и $\mathbf{A}$, представима в виде произведения нижней $\mathbf{L}_1$ и верхней $\mathbf{U}_1$ треугольных матриц, т.е. $\mathbf{A}_1 = \mathbf{L}_1\mathbf{U}_1$, то, положив $\mathbf{A}_2 := \mathbf{U}_1\mathbf{L}_1$ и выразив отсюда $\mathbf{U}_1 = \mathbf{A}_2\mathbf{L}_1^{-1}$, аналогично предыдущему получаем

$$\mathbf{A}_1 = \mathbf{L}_1\mathbf{A}_2\mathbf{L}_1^{-1}. \tag{5.2}$$

Следовательно, матрица $\mathbf{A}_1$ подобна матрице $\mathbf{A}_2$ и, значит, имеет место равенство $\lambda_{\mathbf{A}_1} = \lambda_{\mathbf{A}_2}$.

Суперпозиция этих двух преобразований, т.е. подстановка (5.2) в (5.1), дает выражение $\mathbf{A}$ через $\mathbf{A}_2$:

$$\mathbf{A} = \mathbf{L}\mathbf{L}_1\mathbf{A}_2\mathbf{L}_1^{-1}\mathbf{L}^{-1} = \mathbf{L}\mathbf{L}_1\mathbf{A}_2\left(\mathbf{L}\mathbf{L}_1\right)^{-1},$$

непосредственно утверждающее равенство собственных чисел $\lambda_{\mathbf{A}}$ и $\lambda_{\mathbf{A}_2}$.

Такой процесс построения теоретически бесконечной последовательности подобных матриц и составляет основу **LU-** (иначе, **LR-**) ***алгоритма****. Он определяется фактически двумя формулами:

$$\mathbf{A}_k = \mathbf{L}_k\mathbf{U}_k, \quad \mathbf{A}_{k+1} := \mathbf{U}_k\mathbf{L}_k, \tag{5.3}$$

где $\mathbf{A}_0 := \mathbf{A}$, $k = 0, 1, 2, \ldots$. При этом первая из формул (5.3) означает процедуру треугольной факторизации матрицы $\mathbf{A}_k$ на k-м шаге, а вторая — простое умножение верхней треугольной матрицы на нижнюю.

Доказано [31, 54, 67], что при ряде ограничений на данную матрицу **A** (простейшим из которых является, в частности, требование, чтобы все ее собственные числа были различны по модулю) итерационный процесс (5.3) осуществим, и формируемая им последовательность $(\mathbf{A}_k)$ сходится к треугольной матрице вида

$$\begin{pmatrix} \lambda_1 & * & * & \ldots & * \\ 0 & \lambda_2 & * & \ldots & * \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \lambda_n \end{pmatrix} \quad \text{или вида} \quad \begin{pmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ * & \lambda_2 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ * & * & * & \ldots & \lambda_n \end{pmatrix}$$

в зависимости от того, фиксируется единичная диагональ при **LU**-факторизации у матрицы **L** или у матрицы **U** соответственно. К сожалению, эти ограничения трудно назвать конструктивными, и реализующие **LU**-алгоритм программы больше опираются на эмпирику. Осуществимости, устойчивости и ускорения сходимости процесса (5.3) обычно добиваются (если это в принципе возможно) путем подходящих сдвигов матриц и перестановок их элементов; соответствующие исследования и рекомендации по этому поводу можно найти в [54, 67].

---

* Алгоритм предложен Рутисхаузером (1958).

Пример 5.1. Рассмотрим, как ведет себя **LU**-алгоритм (5.3), примененный к нахождению собственных чисел матрицы $\mathbf{A} := \begin{pmatrix} 2 & 1 \\ 6 & 1 \end{pmatrix}$.

Выполнив **LU**-разложение по формулам (1.5) (с фиксированием единичной диагонали у матрицы $\mathbf{U}$), получим

$$\mathbf{A}_0 := \mathbf{A} = \mathbf{L}_0\mathbf{U}_0 = \begin{pmatrix} 2 & 0 \\ 6 & -2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0{,}5 \\ 0 & 1 \end{pmatrix}.$$

Перемножая матрицы $\mathbf{L}_0$ и $\mathbf{U}_0$ в обратном порядке, строим матрицу $\mathbf{A}_1$ — результат первой итерации:

$$\mathbf{A}_1 := \mathbf{U}_0\mathbf{L}_0 = \begin{pmatrix} 5 & -1 \\ 6 & -2 \end{pmatrix}.$$

Факторизуя эту матрицу аналогично предыдущему, имеем ее представление вида

$$\mathbf{A}_1 = \mathbf{L}_1\mathbf{U}_1 = \begin{pmatrix} 5 & 0 \\ 6 & -0{,}8 \end{pmatrix} \cdot \begin{pmatrix} 1 & -0{,}2 \\ 0 & 1 \end{pmatrix},$$

используя которое, приходим к матрице

$$\mathbf{A}_2 := \mathbf{U}_1\mathbf{L}_1 = \begin{pmatrix} 3{,}8 & 0{,}16 \\ 6 & -0{,}8 \end{pmatrix}.$$

Следующий шаг дает

$$\mathbf{A}_2 = \mathbf{L}_2\mathbf{U}_2 = \begin{pmatrix} 3{,}8 & 0 \\ 6 & -1{,}0526... \end{pmatrix} \cdot \begin{pmatrix} 1 & 0{,}0421... \\ 0 & 1 \end{pmatrix},$$

$$\mathbf{A}_3 := \mathbf{U}_2\mathbf{L}_2 = \begin{pmatrix} 4{,}0526... & 0{,}0443... \\ 6 & -1{,}0526... \end{pmatrix}.$$

Как видим, диагональные элементы матрицы $\mathbf{A}_2$ отличаются от точных значений собственных чисел $\lambda_1 = 4$, $\lambda_2 = -1$ на 0,2, а матрица $\mathbf{A}_3$ позволяет указать значения λ_1 и λ_2 с погрешностью $\approx 0{,}05$.

Если в этом же примере фиксировать единичную диагональ у матриц $\mathbf{L}_k$, то процесс (5.3) будет развиваться следующим образом*:

$$\mathbf{A} = \tilde{\mathbf{L}}_0\tilde{\mathbf{U}}_0 = \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 \\ 0 & -2 \end{pmatrix}, \quad \tilde{\mathbf{A}}_1 := \tilde{\mathbf{U}}_0\tilde{\mathbf{L}}_0 = \begin{pmatrix} 5 & 1 \\ -6 & -2 \end{pmatrix};$$

---

* Для **LU**-разложения здесь используется совокупность формул (1.2), (1.3).

$$\tilde{\mathbf{A}}_1 = \tilde{\mathbf{L}}_1\tilde{\mathbf{U}}_1 = \begin{pmatrix} 1 & 0 \\ -1,2 & 1 \end{pmatrix}\cdot\begin{pmatrix} 5 & 1 \\ 0 & -0,8 \end{pmatrix}, \quad \tilde{\mathbf{A}}_2 := \tilde{\mathbf{U}}_1\tilde{\mathbf{L}}_1 = \begin{pmatrix} 3,8 & 1 \\ 0,96 & -0,8 \end{pmatrix};$$

$$\tilde{\mathbf{A}}_2 = \tilde{\mathbf{L}}_2\tilde{\mathbf{U}}_2 = \begin{pmatrix} 1 & 0 \\ 0,2526... & 1 \end{pmatrix}\cdot\begin{pmatrix} 3,8 & 1 \\ 0 & -1,0526... \end{pmatrix},$$

$$\tilde{\mathbf{A}}_3 := \tilde{\mathbf{U}}_2\tilde{\mathbf{L}}_2 = \begin{pmatrix} 4,0526... & 1 \\ -0,2659... & -1,0526... \end{pmatrix}.$$

Диагонали матриц $\mathbf{A}_k$ и $\tilde{\mathbf{A}}_k$, несущие приближения к собственным числам $\mathbf{A}$, при одних и тех же значениях k полностью совпадают, но во втором случае не так заметно стремление к нулю поддиагональных элементов, хотя относительная скорость убывания модулей наддиагональных элементов $\mathbf{A}_k$ и поддиагональных элементов $\tilde{\mathbf{A}}_k$ примерно одинакова.

В случае, когда данная матрица $\mathbf{A}$ является симметричной положительно определенной, имеет место представление

$$\mathbf{A} = \mathbf{U}^T\mathbf{U},$$

где элементы верхней треугольной матрицы $\mathbf{U}$ находят по формулам (1.6) – (1.7) § 1.3, причем ее диагональные элементы u_{ii} заведомо отличны от нуля. Следовательно, если образовать новую матрицу $\mathbf{A}_1$ перемножением найденных треугольных матриц в обратном порядке:

$$\mathbf{A}_1 := \mathbf{U}\mathbf{U}^T,$$

то с помощью равенства $\mathbf{U} = \left(\mathbf{U}^T\right)^{-1}\mathbf{A}$ получим представление матрицы $\mathbf{A}_1$ вида

$$\mathbf{A}_1 = \left(\mathbf{U}^T\right)^{-1}\mathbf{A}\mathbf{U}^T,$$

из которого заключаем, что матрицы $\mathbf{A}$ и $\mathbf{A}_1$ подобны. Построив последовательность матриц $(\mathbf{A}_k)$ по аналогичным (5.3) формулам

$$\mathbf{A}_k = \mathbf{U}_k{}^T\mathbf{U}_k, \quad \mathbf{A}_{k+1} := \mathbf{U}_k\mathbf{U}_k{}^T, \quad k = 0, 1, 2, \ldots; \quad \mathbf{A}_0 := \mathbf{A},$$

опирающимся на $\mathbf{U}^T\mathbf{U}$-разложение Холецкого, нетрудно прове-

рить, что все матрицы этой последовательности симметричны, положительно определены и подобны друг другу, т.е. гарантируется неизменность спектра:

$$\lambda_{\mathbf{A}_k} = \lambda_{\mathbf{A}} \quad \forall k \in \mathbb{N}_0 .$$

И если в **LU**-алгоритме последовательность матриц $\mathbf{A}_k$ сходится (при определенных условиях) к треугольной матрице, то в только что построенном $\mathbf{U}^T\mathbf{U}$-***алгоритме*** (называемом также ***методом Холецкого***), в силу симметрии матриц $\mathbf{A}_k$, имеет место сходимость к диагональной матрице:

$$\mathbf{A}_k \xrightarrow[k\to\infty]{} \Lambda := \operatorname{diag}(\lambda_1; \lambda_2; ...; \lambda_n) .$$

Скорость сходимости — линейная и существенно зависит от наличия в спектре матрицы $\mathbf{A}$ близких собственных чисел. Значительного ускорения сходимости в этом процессе добиваются за счет введения сдвигов (о чем уже шла речь в § 4.2, 4.3 и еще будет в § 5.4), а повышения вычислительной эффективности — с помощью предварительного преобразования исходной матрицы в подобную ей симметричную трехдиагональную матрицу (см. по этому поводу, например, замечание 5.1 в § 5.2). Все матрицы $\mathbf{A}_k$ при такой модификации сохраняют трехдиагональную структуру, а матрицы $\mathbf{U}_k$ — двухдиагональную.

Пример 5.2. Сделаем несколько шагов $\mathbf{U}^T\mathbf{U}$-алгоритма применительно к матрице $\mathbf{A} := \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$, использовавшейся в примерах 4.1, 4.2 (ее собственные числа $\lambda_1 = 3$, $\lambda_2 = 1$).

Ненулевые элементы матрицы $\mathbf{U} := \begin{pmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{pmatrix}$ при выполнении факторизации матрицы $\mathbf{A} := \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$ на каждом итерационном шаге будем подсчитывать по формулам:

$$u_{11} = \sqrt{a_{11}} , \qquad u_{12} = a_{12}/u_{11} , \qquad u_{22} = \sqrt{a_{22} - u_{12}^2} .$$

Имеем:

$$1)\ \mathbf{A}_1 := \mathbf{U}_0\mathbf{U}_0{}^T = \begin{pmatrix} \sqrt{2} & -\frac{1}{\sqrt{2}} \\ 0 & \frac{\sqrt{3}}{\sqrt{2}} \end{pmatrix} \cdot \begin{pmatrix} \sqrt{2} & 0 \\ -\frac{1}{\sqrt{2}} & \frac{\sqrt{3}}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} \frac{5}{2} & -\frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & \frac{3}{2} \end{pmatrix} \approx \begin{pmatrix} 2,5 & -0,87 \\ -0,87 & 1,5 \end{pmatrix};$$

$$2)\ \mathbf{A}_2 := \mathbf{U}_1\mathbf{U}_1{}^T = \begin{pmatrix} \frac{\sqrt{5}}{\sqrt{2}} & -\frac{\sqrt{3}}{\sqrt{10}} \\ 0 & \frac{\sqrt{6}}{\sqrt{5}} \end{pmatrix} \cdot \begin{pmatrix} \frac{\sqrt{5}}{\sqrt{2}} & 0 \\ -\frac{\sqrt{3}}{\sqrt{10}} & \frac{\sqrt{6}}{\sqrt{5}} \end{pmatrix} = \begin{pmatrix} \frac{14}{5} & -\frac{3}{5} \\ -\frac{3}{5} & \frac{6}{5} \end{pmatrix} \approx \begin{pmatrix} 2,8 & -0,6 \\ -0,6 & 1,2 \end{pmatrix};$$

$$3)\ \mathbf{A}_3 := \mathbf{U}_2\mathbf{U}_2{}^T = \begin{pmatrix} \frac{\sqrt{14}}{\sqrt{5}} & -\frac{3}{\sqrt{70}} \\ 0 & \frac{\sqrt{15}}{\sqrt{14}} \end{pmatrix} \cdot \begin{pmatrix} \frac{\sqrt{14}}{\sqrt{5}} & 0 \\ -\frac{3}{\sqrt{70}} & \frac{\sqrt{15}}{\sqrt{14}} \end{pmatrix} = \begin{pmatrix} \frac{41}{14} & -\frac{3\sqrt{3}}{14} \\ -\frac{3\sqrt{3}}{14} & \frac{15}{14} \end{pmatrix} \approx \begin{pmatrix} 2,93 & -0,37 \\ -0,37 & 1,07 \end{pmatrix};$$

$$4)\ \mathbf{A}_4 := \mathbf{U}_3\mathbf{U}_3{}^T = \begin{pmatrix} \sqrt{\frac{41}{14}} & -\sqrt{\frac{27}{574}} \\ 0 & \sqrt{\frac{42}{41}} \end{pmatrix} \cdot \begin{pmatrix} \sqrt{\frac{41}{14}} & 0 \\ -\sqrt{\frac{27}{574}} & \sqrt{\frac{42}{41}} \end{pmatrix} = \begin{pmatrix} \frac{122}{41} & -\frac{3\sqrt{3}}{41\sqrt{14}} \\ -\frac{3\sqrt{3}}{41\sqrt{14}} & \frac{42}{41} \end{pmatrix} \approx \begin{pmatrix} 2,98 & -0,03 \\ -0,03 & 1,02 \end{pmatrix}.$$

Наблюдаем процесс приближения соответствующих диагональных элементов к собственным числам:

$$2,5;\ \ 2.8;\ \ 2,93;\ \ 2,98;\ \ \ldots\ \ \rightarrow\ \ \lambda_1 := 3,$$
$$1,5;\ \ 1,2;\ \ 1,07;\ \ 1,02;\ \ \ldots\ \ \rightarrow\ \ \lambda_2 := 1,$$

а внедиагональных — к нулю:

$$-0,87;\ \ -0,6;\ \ -0,37;\ \ -0,03;\ \ \ldots\ \ \rightarrow\ \ 0.$$

Одним из серьезных факторов, ограничивающих сферу применения **LU**-алгоритмов, является их недостаточно хорошая численная устойчивость (улучшение этого параметра путем перестановок строк и столбцов сильно отражается на экономичности метода). Этот фактор может играть особенно существенную роль на фоне возможной неустойчивости самой несимметричной проблемы собственных значений.

Ярким примером матрицы, для которой задача нахождения собственных чисел является неустойчивой, служит следующая матрица [31]:

$$\mathbf{A} := \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

При размере $n \times n$ она имеет число 0 собственным значением n-й кратности.

Введем возмущение ε в левый нижний элемент матрицы $\mathbf{A}$. Характеристическим уравнением для возмущенной матрицы

$$\mathbf{A}_\varepsilon := \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ \varepsilon & 0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

служит уравнение

$$\begin{vmatrix} -\lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & -\lambda & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -\lambda & 1 \\ \varepsilon & 0 & 0 & \dots & 0 & -\lambda \end{vmatrix} = 0.$$

Раскрывая определитель по элементам первого столбца, получаем:

$$-\lambda \begin{vmatrix} -\lambda & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -\lambda & 1 \\ 0 & 0 & \dots & 0 & -\lambda \end{vmatrix} + (-1)^{n+1} \varepsilon \begin{vmatrix} 1 & 0 & \dots & 0 & 0 \\ -\lambda & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -\lambda & 1 \end{vmatrix} = 0 \Leftrightarrow$$

$$\Leftrightarrow \quad -\lambda(-\lambda)^{n-1} + (-1)^{n+1} \varepsilon = 0 \quad \Leftrightarrow \quad \lambda^n - \varepsilon = 0.$$

Следовательно, матрица $\mathbf{A}_\varepsilon$ имеет n различных, в общем, комплексных собственных значений $\lambda_i = \sqrt[n]{\varepsilon}$, $i = 1, 2, \dots, n$. Если взять, например, $n := 100$, а $\varepsilon := 10^{-100}$, то $|\lambda_i| = 0{,}1$, т.е. чрезвычайно малое, возможно, неощутимое для компьютера искажение всего одного элемента данной специфической матрицы приводит к существенному изменению ее спектра.

Разумеется, большинство важных в приложениях задач на собственные значения не так плохи. Однако, обозначив и, возможно, намеренно утрировав проблему, призовем читателя к осторожности в применениях уже рассмотренных методов и интерпретации их результатов, а также к пониманию необходимости построения более устойчивых методов численного решения несимметричных спектральных алгебраических задач.

Одним из таковых является достаточно универсальный метод, опирающийся на ортогональные преобразования матриц; его называют ***QR-алгоритмом***[*]. Детально этот метод будет рассмотрен в следующем параграфе, здесь же только обозначим его идею, очень близкую к той, которая заложена в **LU**-алгоритме.

Суть **QR**-алгоритма следующая. При $k = 0, 1, 2, \ldots$, начиная с $\mathbf{A}_0 := \mathbf{A}$, здесь строят последовательность матриц $(\mathbf{A}_k)$ по формулам

$$\mathbf{A}_k = \mathbf{Q}_k \mathbf{R}_k, \quad \mathbf{A}_{k+1} := \mathbf{R}_k \mathbf{Q}_k, \tag{5.4}$$

первая из которых означает разложение матрицы $\mathbf{A}_k$ в произведение ортогональной $\mathbf{Q}_k$ и правой треугольной $\mathbf{R}_k$ (такое разложение существует для любой квадратной матрицы, см. далее теорему 1.2), а вторая — перемножение полученных в результате факторизации $\mathbf{A}_k$ матриц $\mathbf{Q}_k$ и $\mathbf{R}_k$ в обратном порядке.

Аналогично предыдущему на основе свойства ортогональных матриц $\mathbf{Q}_k^T = \mathbf{Q}_k^{-1}$ в соответствии с (5.4) можно записать представление данной матрицы $\mathbf{A}$ в виде

$$\mathbf{A} = \mathbf{Q}_0 \mathbf{Q}_1 \ldots \mathbf{Q}_{k-1} \mathbf{Q}_k \mathbf{A}_{k+1} \mathbf{Q}_k^T \mathbf{Q}_{k-1}^T \ldots \mathbf{Q}_1^T \mathbf{Q}_0^T$$

или иначе

$$\mathbf{A} = (\mathbf{Q}_0 \mathbf{Q}_1 \ldots \mathbf{Q}_{k-1} \mathbf{Q}_k) \mathbf{A}_{k+1} (\mathbf{Q}_0 \mathbf{Q}_1 \ldots \mathbf{Q}_k)^{-1}. \tag{5.5}$$

Следовательно, любая из матриц последовательности $(\mathbf{A}_k)$ *ортогонально подобна* матрице $\mathbf{A}$.

---

[*] Этот метод предложен почти одновременно российским математиком В.Н. Кублановской (1961) и англичанином Дж. Фрэнсисом (1962). Его описание, более или менее подробное, можно обнаружить в книгах [1, 17, 18, 32, 36, 51, 54, 67, 68].

При определенных ограничениях, одним из которых опять выступает требование, чтобы матрица $\mathbf{A}$ не имела равных по модулю собственных значений, генерируемая процессом (5.4) последовательность матриц $(\mathbf{A}_k)$ сходится к матрице правой треугольной формы с диагональю из собственных чисел. Скорость аннулирования поддиагональных частей матриц $\mathbf{A}_k$ линейна и зависит, как и во многих других итерационных методах, в частности, описанных в предыдущей главе, от отношений $|\lambda_i|/|\lambda_j|$ при $i>j$ (традиционно полагаем $|\lambda_1|>|\lambda_2|>\ldots>|\lambda_n|$). Наличие комплексно сопряженных пар собственных чисел у данной вещественной матрицы $\mathbf{A}$ не является, вообще говоря, препятствием для применения **QR**-алгоритма; просто в этом случае предельной матрицей для последовательности $(\mathbf{A}_k)$ будет матрица квазитреугольного, иначе, блочно-треугольного, вида. Каждой комплексной паре собственных чисел в такой матрице будет соответствовать диагональный 2×2-блок, причем сходимость здесь наблюдается по форме матрицы, а не поэлементно (т.е. элементы внутри этих блоков могут изменяться без видимой зависимости от k при сохранении неизменными их собственных чисел).

Теоретическим фундаментом для **QR**-алгоритма служит следующая теорема о вещественном ***разложении Шура**** [17, 73].

Теорема 5.1. *Для любой вещественной $n\times n$-матрицы $\mathbf{A}$ найдется такая вещественная ортогональная $n\times n$-матрица $\mathbf{Q}$, что*

$$\mathbf{Q}^T\mathbf{A}\mathbf{Q}=\begin{pmatrix} R_{11} & R_{12} & \ldots & R_{1m} \\ 0 & R_{22} & \ldots & R_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & R_{mm} \end{pmatrix},$$

где R_{ii} — либо вещественное число, либо 2×2-матрица с комплексно-сопряженными собственными значениями.

---

* Шур Иссай (1875–1941) — немецкий математик.

§ 5.2. ПРИВЕДЕНИЕ МАТРИЦ К ФОРМЕ ХЕССЕНБЕРГА

Как следует из предыдущего параграфа, в процессе реализации QR-алгоритма приближенного вычисления всех собственных чисел произвольной вещественной квадратной матрицы $\mathbf{A}$ на основе формул (5.4) последовательность получающихся там матриц $\mathbf{A}_k$ должна сходиться либо к матрице треугольного вида с диагональю из собственных чисел матрицы $\mathbf{A}$, если все они вещественные, либо к матрице блочно-треугольного вида с максимальным размером блоков на диагонали 2×2, соответствующих парам комплексно-сопряженных собственных чисел. В любом случае, у матриц $\mathbf{A}_k$ при $k\to\infty$ должна аннулироваться треугольная часть, содержащая элементы a_{ij}, у которых $i>j+1$. Отсюда — особая важность матриц такой структуры, при которой элементы с индексами i, j, удовлетворяющие неравенству $i>j+1$, равны нулю. Эти ***матрицы почти треугольного вида*** [4, 18] называются иначе ***матрицами в форме Хессенберга*** или просто ***матрицами Хессенберга**** (*правыми*) [1, 18, 51, 67].

Пусть матрица

$$\mathbf{B}:=\begin{pmatrix} b_{11} & b_{12} & b_{13} & \cdots & b_{1,n-2} & b_{1,n-1} & b_{1n} \\ b_{21} & b_{22} & b_{23} & \cdots & b_{2,n-2} & b_{2,n-1} & b_{2n} \\ 0 & b_{32} & b_{33} & \cdots & b_{3,n-2} & b_{3,n-1} & b_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & b_{n-1,n-2} & b_{n-1,n-1} & b_{n-1,n} \\ 0 & 0 & 0 & \cdots & 0 & b_{n,n-1} & b_{nn} \end{pmatrix}$$

— некоторая матрица Хессенберга. Пользуясь материалом § 1.4, нетрудно представить, как с помощью $(n-2)$-х преобразований Хаусхолдера любую вещественную $n\times n$-матрицу $\mathbf{A}:=(a_{ij})$ привести к виду $\mathbf{B}$, иначе, разложить $\mathbf{A}$ в произведение ортогональной матрицы (композиции $n-2$ матриц отражения) и почти

---

* Хессенбе́рг Карл (1904—1959) — немецкий инженер. Метод построения матриц, получивших его имя, впервые изложен в его диссертации (1941).

треугольной матрицы $\mathbf{R}$. Очевидно, для этого нужно в процессе QR-факторизации, описанном ранее, роль диагонали преобразуемой матрицы отводить ее первой поддиагонали. Для первого этапа преобразований такой подход означает формирование матрицы Хаусхолдера из элементов a_{i1} первого столбца матрицы $\mathbf{A}$ по формулам

$$\beta_1 := \operatorname{sgn}_+(-a_{21})\sqrt{\sum_{k=2}^{n}(a_{k1})^2}\,, \quad \mu_1 := \frac{1}{\sqrt{2\beta_1^2 - 2\beta_1 a_{21}}},$$

$$\mathbf{w}_1 := \mu_1\left(0;\ a_{21}-\beta_1; a_{31}; \ldots; a_{n1}\right)^T, \quad \mathbf{H}_1 := \mathbf{E} - 2\mathbf{w}_1\mathbf{w}_1^T.$$

Будем считать, что на $(i-1)$-м этапе преобразований матрицы $\mathbf{A}$ к форме Хессенберга получена матрица $\mathbf{A}_{i-1} := (a_{ij}^{(i-1)})$ с нулями в нужных местах первых $i-1$ столбцов и, следовательно, имеют место два равносильных (в силу свойств матриц Хаусхолдера) равенства:

$$\mathbf{A} = \mathbf{H}_1\mathbf{H}_2\ldots\mathbf{H}_{i-1}\mathbf{A}_{i-1} \quad \text{и} \quad \mathbf{A}_{i-1} = \mathbf{H}_{i-1}\ldots\mathbf{H}_2\mathbf{H}_1\mathbf{A}.$$

Тогда i-й этап характеризуется равенствами

$$\mathbf{A} = \mathbf{H}_1\mathbf{H}_2\ldots\mathbf{H}_{i-1}\mathbf{H}_i\mathbf{A}_i\,, \quad \mathbf{A}_i = \mathbf{H}_i\mathbf{A}_{i-1}\,, \tag{5.6}$$

где (ср. (1.22)–(1.23))

$$\mathbf{H}_i := \mathbf{E} - 2\mathbf{w}_i\mathbf{w}_i^T,$$

$$\mathbf{w}_i^T := \mu_i\left(0;\ldots;\ 0;\ a_{i+1,i}^{(i-1)} - \beta_i;\ a_{i+2,i}^{(i-1)};\ldots; a_{ni}^{(i-1)}\right),$$

$$\beta_i := \operatorname{sgn}_+\left(-a_{i+1,i}^{(i-1)}\right)\sqrt{\sum_{k=i+1}^{n}\left(a_{ki}^{(i-1)}\right)^2}\,,\ \mu_i := \frac{1}{\sqrt{2\beta_i^2 - 2\beta_i a_{i+1,i}^{(i-1)}}}. \tag{5.7}$$

После реализации всех этапов получается представление

$$\mathbf{A} = \mathbf{Q}\mathbf{B}, \tag{5.8}$$

в котором, в развитие (5.6), обозначены:

$\mathbf{B} := \mathbf{A}_{n-2} = (a_{ij}^{(n-2)})$ — итоговая матрица Хессенберга,

$\mathbf{Q} := \mathbf{H}_1\mathbf{H}_2\ldots\mathbf{H}_{n-2}$ — результирующая ортогональная матрица.

Не забывая о главной цели всех выполняемых преобразований — получении спектра матрицы $\mathbf{A}$, нужно, чтобы преобразованная матрица была подобна исходной. Это будет в том случае, если построенную выше матрицу $\mathbf{B}$ умножить справа на также уже известную ортогональную матрицу $\mathbf{Q}$, поскольку из равносильного (5.8) равенства $\mathbf{B} = \mathbf{Q}^T\mathbf{A}$ следует

$$\mathbf{B}_0 := \mathbf{BQ} = \mathbf{Q}^T\mathbf{AQ} \tag{5.9}$$

— условие подобия матриц $\mathbf{B}_0$ и $\mathbf{A}$ (§ 4.1), в силу $\mathbf{Q}^T = \mathbf{Q}^{-1}$.

Отрадным является то обстоятельство, что форма Хессенберга инвариантна по отношению к используемым здесь ортогональным преобразованиям [18, 26, 51], т.е. матрица $\mathbf{B}_0$, получаемая умножением матрицы Хессенберга $\mathbf{B}$ на ортогональную матрицу $\mathbf{Q}$, тоже есть матрица в форме Хессенберга.

Переписывая (5.9) в виде

$$\mathbf{B}_0 = \mathbf{H}_{n-2}\dots\mathbf{H}_2\mathbf{H}_1\mathbf{A}\mathbf{H}_1\mathbf{H}_2\dots\mathbf{H}_{n-2},$$

приходим к выводу, что *подобную* $\mathbf{A}$ *матрицу Хессенберга* $\mathbf{B}_0$ *можно найти в результате конечного процесса ортогональных преобразований подобия в цикле вычислений по формулам*

$$\tilde{\mathbf{A}}_i = \mathbf{H}_i\tilde{\mathbf{A}}_{i-1}\mathbf{H}_i, \quad i = 1, 2, \dots, n-2, \quad \tilde{\mathbf{A}}_0 := \mathbf{A}, \ \mathbf{B}_0 := \tilde{\mathbf{A}}_{n-2}, \tag{5.10}$$

где $\mathbf{H}_i$ *— матрица Хаусхолдера, определяемая формулами* (5.7), *в которых под элементами* $a_{ij}^{(i-1)}$ *матрицы* $\mathbf{A}_{i-1}$ *следует понимать элементы* $\tilde{a}_{ij}^{(i-1)}$ *матрицы* $\tilde{\mathbf{A}}_{i-1}$.

Организуя процедуру поочередных умножений преобразуемой матрицы слева и справа на матрицы Хаусхолдера, легко увидеть, что получаемая в итоге подобная $\mathbf{A}$ матрица $\mathbf{B}_0$ действительно будет иметь верхнюю форму Хессенберга. Для этого целесообразно представить $n \times n$-матрицу Хаусхолдера i-го этапа в блочном виде: $\mathbf{H}_i := \begin{pmatrix} \mathbf{E}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{n-i} \end{pmatrix}$, где $\mathbf{E}_i$ — единичная матрица размера $i \times i$, а $\mathbf{M}_{n-i}$ — «существенная» $(n-i)\times(n-i)$-подматрица матрицы $\mathbf{H}_i$.

Ясно, что при умножении слева на матрицу $\mathbf{H}_1 := \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{n-1} \end{pmatrix}$ в матрице $\mathbf{H}_1\mathbf{A}$ сохранится неизменной первая строка матрицы $\mathbf{A}$, а при умножении $\mathbf{H}_1\mathbf{A}$ справа на ту же матрицу $\mathbf{H}_1$ в матрице $\mathbf{H}_1\mathbf{A}\mathbf{H}_1$ сохранится неизменным первый столбец матрицы $\mathbf{H}_1\mathbf{A}$ с нулями в позициях (3,1), (4,1), …, (n,1). При умножении слева на матрицу $\mathbf{H}_2 := \begin{pmatrix} \mathbf{E}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{n-2} \end{pmatrix}$ в матрице $\mathbf{H}_2(\mathbf{H}_1\mathbf{A}\mathbf{H}_1)$ сохранятся неизменными первые две строки предыдущей матрицы, а умножение справа на $\mathbf{H}_2$ не изменит первые два столбца с нулями под первой поддиагональю, и т.д.

Рассмотрим на простом числовом примере, как приводится матрица к подобной ей матрице в форме Хессенберга, когда для этого требуется только один шаг преобразований Хаусхолдера.

Пример 5.3. Дана матрица $\mathbf{A} := \begin{pmatrix} 5 & 1 & -3 \\ 3 & 0 & -2 \\ -4 & -1 & 1 \end{pmatrix}$. Найти матрицу $\mathbf{B}$, подобную матрице $\mathbf{A}$ и имеющую форму Хессенберга.

Решение проводим по формулам (5.10), (5.7) при $n := 3$, которые для данного конкретного случая можно записать так (в естественном для выполнения порядке):

$$\beta_1 := \operatorname{sgn}_+(-a_{21}) \cdot \sqrt{a_{21}^2 + a_{31}^2}\,; \quad \mu_1 := \frac{1}{\sqrt{2\beta_1(\beta_1 - a_{21})}};$$

$$\mathbf{w}_1^T := \mu_1 (0;\, a_{21} - \beta_1;\, a_{31}); \quad \mathbf{H}_1 := \mathbf{E} - 2\mathbf{w}_1\mathbf{w}_1^T; \quad \mathbf{B}_0 := \mathbf{H}_1\mathbf{A}\mathbf{H}_1.$$

Имеем:

$$\beta_1 = -\sqrt{3^2 + (-4)^2} = -5\,; \quad \mu_1 = \frac{1}{\sqrt{2(-5)(-5-3)}} = \frac{1}{4\sqrt{5}};$$

$$2\mathbf{w}_1\mathbf{w}_1^T = \frac{1}{40}\begin{pmatrix} 0 \\ 8 \\ -4 \end{pmatrix} \cdot (0;\, 8;\, -4) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1{,}6 & -0{,}8 \\ 0 & -0{,}8 & 0{,}4 \end{pmatrix};$$

$$\mathbf{H}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0{,}6 & 0{,}8 \\ 0 & 0{,}8 & 0{,}6 \end{pmatrix}; \quad \mathbf{H}_1\mathbf{A} = \begin{pmatrix} 5 & 1 & -3 \\ -5 & -0{,}8 & 2 \\ 0 & -0{,}6 & -1 \end{pmatrix}; \quad \mathbf{B}_0 = \begin{pmatrix} 5 & -3 & -1 \\ -5 & 2{,}08 & 0{,}56 \\ 0 & -0{,}44 & -1{,}08 \end{pmatrix}.$$

Замечание 5.1. При решении симметричных проблем собственных значений методом бисекций на первой стадии решения также часто применяют преобразования Хаусхолдера. Абсолютно те же формулы (5.10), (5.7) приведут симметричную матрицу $\mathbf{A}$ к подобной ей матрице $\mathbf{B}$ трехдиагонального вида (частному случаю формы Хессенберга). Это существенно повышает эффективность последующих преобразований в методах бисекций (§ 4.5) и Холецкого (§ 5.1) и менее существенно в методе вращений Якоби (§ 4.4).

§ 5.3. ФАКТОРИЗАЦИЯ МАТРИЦЫ ХЕССЕНБЕРГА

В принципе в **QR**-алгоритме, определяемом формулами (5.4), можно обойтись одними ортогональными преобразованиями Хаусхолдера, применяя их для требуемой этими формулами **QR**-факторизации или заполненных, как исходная матрица $\mathbf{A}$, матриц $\mathbf{A}_k$, или, что лучше, матриц $\mathbf{A}_k$ в форме Хессенберга с начальной матрицей $\mathbf{A}_0 := \mathbf{B}_0$, полученной процессом (5.10). Однако такой подход обычно не используется ввиду чрезвычайно медленной сходимости в первом случае (особенно при наличии близких по модулю собственных чисел) и редко используется во втором.

Если считать первой стадией **QR**-алгоритма приведение исходной матрицы к форме Хессенберга преобразованиями Хаусхолдера, то на второй стадии — построении последовательности подобных матриц — чаще привлекают преобразования Гивенса. При этом здесь матрица плоских вращений Гивенса, в отличие от представленного в (1.25) общего вида, характеризуется тем, что ее 2×2-подматрица с «существенными» элементами c и $\pm s$ определяется парой смежных индексов $(i, i+1)$ и потому может помечаться одним индексом i:

$$\mathbf{G}_i := \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & c & s & \cdots & 0 \\ 0 & \cdots & -s & c & \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{matrix} \\ \\ \leftarrow i \\ \leftarrow i+1 \\ \\ \\ \end{matrix} \cdot \qquad (5.11)$$

$$\begin{matrix} & & \uparrow & \uparrow & & \\ & & i & i+1 & & \end{matrix}$$

Имея возможность наложить еще одно условие на c и s в преобразовании Гивенса кроме условия нормировки (1.26), распорядимся этой свободой так, чтобы с помощью последовательности ортогональных преобразований матрицами $\mathbf{G}_1$, $\mathbf{G}_2$, ..., $\mathbf{G}_{n-1}$ типа (5.11) матрицу Хессенберга $\mathbf{B}$ удалось привести к правому треугольному виду, последовательно аннулируя поддиагональные элементы в первом, втором, ..., $(n-1)$-м столбцах.

С этой целью предположим, что уже сделаны $i-1$ таких шагов, в результате чего получена матрица

$$\mathbf{B}_{i-1} := \begin{pmatrix} b_{11}^{(1)} & \dots & b_{1,i-1}^{(1)} & b_{1i}^{(1)} & b_{1,i+1}^{(1)} & \dots & b_{1n}^{(1)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & b_{i-1,i-1}^{(i-1)} & b_{i-1,i}^{(i-1)} & b_{i-1,i+1}^{(i-1)} & \dots & b_{i-1,n}^{(i-1)} \\ 0 & \dots & 0 & b_{ii}^{(i-1)} & b_{i,i+1}^{(i-1)} & \dots & b_{in}^{(i-1)} \\ 0 & \dots & 0 & b_{i+1,i}^{(i-1)} & b_{i+1,i+1}^{(i-1)} & \dots & b_{i+1,n}^{(i-1)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & b_{nn}^{(i-1)} \end{pmatrix}.$$

Найдем произведение матрицы $\mathbf{G}_i$ на матрицу $\mathbf{B}_{i-1}$, полагая в $\mathbf{G}_i$ из (5.11) $c := c_i$, $s := s_i$. Приходим к матрице

$$\mathbf{B}_i := \begin{pmatrix} b_{11}^{(1)} & \dots & b_{1,i-1}^{(1)} & b_{1i}^{(1)} & b_{1,i+1}^{(1)} & \dots & b_{1n}^{(1)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & b_{i-1,i-1}^{(i-1)} & b_{i-1,i}^{(i-1)} & b_{i-1,i+1}^{(i-1)} & \dots & b_{i-1,n}^{(i-1)} \\ 0 & \dots & 0 & c_i b_{ii}^{(i-1)} + s_i b_{i+1,i}^{(i-1)} & c_i b_{i,i+1}^{(i-1)} + s_i b_{i+1,i+1}^{(i-1)} & \dots & c_i b_{in}^{(i-1)} + s_i b_{i+1,n}^{(i-1)} \\ 0 & \dots & 0 & c_i b_{i+1,i}^{(i-1)} - s_i b_{ii}^{(i-1)} & c_i b_{i+1,i+1}^{(i-1)} - s_i b_{i,i+1}^{(i-1)} & \dots & c_i b_{i+1,n}^{(i-1)} - s_i b_{in}^{(i-1)} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & b_{nn}^{(i-1)} \end{pmatrix}$$

с измененными элементами по сравнению с матрицей $\mathbf{B}_{i-1}$ только в i-й и в $(i+1)$-й строках. Потребуем, чтобы единственный нену-

левой поддиагональный элемент $b_{i+1,i}^{(i)}$ i-го столбца матрицы $\mathbf{B}_i := (b_{ij}^{(i)})$ обратился в нуль, т.е. подберем числа c_i и s_i, связанные, согласно (1.26), соотношением

$$s_i^2 + c_i^2 = 1 \tag{5.12}$$

так, что

$$b_{i+1,i}^{(i)} := c_i b_{i+1,i}^{(i-1)} - s_i b_{ii}^{(i-1)} = 0.$$

Отсюда получаем

$$t_i := \frac{s_i}{c_i} = \frac{b_{i+1,i}^{(i-1)}}{b_{ii}^{(i-1)}} \tag{5.13}$$

— значение тангенса угла θ_i поворота в плоскости, определяемой i-м и $(i+1)$-м ортами, если положить $s_i := \sin\theta_i$, $c_i := \cos\theta_i$.

Очевидно, что если на i-м шаге окажется $b_{ii}^{(i-1)} = 0$ (или значение $|b_{ii}^{(i-1)}|$ — существенно малая величина), то можно принять $c_i := 1$, $s_i := 0$, т.е. положить $\mathbf{G}_i := \mathbf{E}$. В противном случае, учитывая равенство (5.12), вычисляем c_i и s_i по формулам

$$c_i = \frac{1}{\sqrt{1+t_i^2}}, \quad s_i = \frac{t_i}{\sqrt{1+t_i^2}}, \tag{5.14}$$

предварительно подсчитав значение t_i с помощью (5.13). После пересчета диагональных и стоящих правее них элементов i-й и $(i+1)$-й строк по формулам

$$\begin{aligned} b_{ij}^{(i)} &= c_i b_{ij}^{(i-1)} + s_i b_{i+1,j}^{(i-1)} \quad \text{при } j = i, i+1, \ldots, n, \\ b_{i+1,j}^{(i)} &= c_i b_{i+1,j}^{(i-1)} - s_i b_{ij}^{(i-1)} \quad \text{при } j = i+1, \ldots, n \end{aligned} \tag{5.15}$$

матрица $\mathbf{B}_i$ будет готова к выполнению следующего, $(i+1)$-го шага преобразований Гивенса.

Таким образом, *совокупность формул* (5.13)–(5.15), *в которых натуральной переменной* i *последовательно придаются значения* $1, 2, \ldots, n-1$, *полностью определяет процесс приведения матрицы Хессенберга* $\mathbf{B} := \mathbf{B}_0 := (b_{ij}^{(0)})$ *к матрице правой треугольной формы* $\mathbf{B}_{n-1} := (b_{ij}^{(n-1)})$ *посредством преобразований* $\mathbf{B}_i = \mathbf{G}_i \mathbf{B}_{i-1}$.

Для формальной корректности сделанного утверждения следует лишь указать на необходимость переобозначений непересчитываемых элементов матриц $\mathbf{B}_i$, увеличивая их верхний индекс на единицу при каждом шаге. Однако практически при организации машинных вычислений в этом нет нужды. Более того, используя одни и те же массивы для хранения скаляров t_i, c_i, s_i и матриц $\mathbf{B}_i$ при всех i, можно в описанном процессе ортогональной триангуляризации матрицы Хессенберга преобразованиями плоских вращений Гивенса опустить индексы у величин t_i, c_i, s_i и верхние индексы у элементов матриц $\mathbf{B}_i$ (означающие номера шагов преобразований) и оформить цикл вычислений (присвоений) при i от 1 до $n-1$ со следующим телом цикла:

если $|b_{ii}| < \text{macheps}$, то $c := 1$, $s := 0$,

иначе $t := \dfrac{b_{i+1,i}}{b_{ii}}$, $c := \dfrac{1}{\sqrt{1+t^2}}$, $s := t \cdot c$;

для $j := i, i+1, \ldots, n$:

$$b_{ij} := cb_{ij} + sb_{i+1,j},$$

для $j := i+1, \ldots, n$:

$$b_{i+1,j} := cb_{i+1,j} - sb_{ij}.$$

Пример 5.4. Преобразованиями Гивенса выполним один шаг QR-алгоритма (5.4) для матрицы $\mathbf{B}_0 := \begin{pmatrix} 5 & -3 & -1 \\ -5 & 2{,}08 & 0{,}56 \\ 0 & -0{,}44 & -1{,}08 \end{pmatrix}$, полученной в результате преобразований Хаусхолдера в предыдущем примере.

При $i := 1$ по формулам (5.13) – (5.15) последовательно находим:

$$t_1 = \frac{-5}{5} = -1, \quad c_1 = \frac{1}{\sqrt{1+(-1)^2}} = \frac{1}{\sqrt{2}}, \quad s_1 = -\frac{1}{\sqrt{2}};$$

$$\mathbf{G}_1 = \begin{pmatrix} 0{,}5\sqrt{2} & -0{,}5\sqrt{2} & 0 \\ 0{,}5\sqrt{2} & 0{,}5\sqrt{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{B}_1 := \mathbf{G}_1\mathbf{B}_0 = \begin{pmatrix} 5\sqrt{2} & -2{,}54\sqrt{2} & -0{,}78\sqrt{2} \\ 0 & -0{,}46\sqrt{2} & -0{,}22\sqrt{2} \\ 0 & -0{,}44 & -1{,}08 \end{pmatrix}.$$

При $i := 2$ вычисляем (округляя до 10^{-3}):

$$t_2 = \frac{-0{,}44}{-0{,}46\sqrt{2}} = 0{,}676, \qquad c_2 = 0{,}828, \qquad s_2 = 0{,}560;$$

значит,

$$\mathbf{G}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0{,}828 & 0{,}560 \\ 0 & -0{,}560 & 0{,}828 \end{pmatrix}, \quad \mathbf{R}_1 := \mathbf{B}_2 := \mathbf{G}_2\mathbf{B}_1 = \begin{pmatrix} 7{,}071 & -3{,}592 & -1{,}103 \\ 0 & -0{,}785 & -0{,}863 \\ 0 & 0 & -0{,}720 \end{pmatrix}.$$

Следовательно,

$$\mathbf{Q}_1 := \mathbf{G}_1^T\mathbf{G}_2^T = \begin{pmatrix} 0{,}707 & 0{,}586 & -0{,}396 \\ -0{,}707 & 0{,}586 & -0{,}396 \\ 0 & 0{,}560 & 0{,}828 \end{pmatrix},$$

и, согласно (5.4), матрица

$$\mathbf{A}_2 := \mathbf{R}_1\mathbf{Q}_1 = \begin{pmatrix} 7{,}540 & 1{,}420 & -2{,}292 \\ 0{,}555 & -0{,}943 & -0{,}404, \\ 0 & -0{,}404 & -0{,}597 \end{pmatrix}$$

есть искомый результат первого полного шага **QR**-алгоритма. Эта матрица имеет те же собственные числа, что $\mathbf{B}_0$ и $\mathbf{A}$, и сохраняет форму Хессенберга. Модули ее поддиагональных элементов меньше, чем у матрицы $\mathbf{B}_0$, т.е. матрица $\mathbf{A}_2$ более близка к подобной $\mathbf{B}_0$ матрице треугольного вида, на диагонали которой должны быть собственные числа данной матрицы (именно, $\lambda_1 :\approx 7{,}639$, $\lambda_2 :\approx -1{,}205$, $\lambda_3 :\approx -0{,}435$).

Замечание 5.2. Весь **QR**-алгоритм можно было бы построить на базе одних только преобразований Гивенса, т.е. не приводя исходную матрицу $\mathbf{A}$ к форме Хессенберга (или к трехдиагональному виду, если $\mathbf{A}$ симметрична) другими преобразованиями. В таком случае стала бы заметной разница между преобразованиями Якоби и Гивенса. Суть этой разницы в следующем: если для

преобразований Якоби (см. § 4.4) понятия «ключевой элемент» и «обреченный элемент» совпадают, то для преобразований Гивенса это, вообще говоря, не так. В общем случае при вращениях Гивенса посредством матриц $\mathbf{G}_{ij}$ чуть более общего, чем (5.11), вида (1.25), угол поворота θ в фиксированной индексами i, j плоскости вращения подбирают так, чтобы аннулировать какой-нибудь элемент, стоящий либо в одном столбце, либо в одной строке с ключевым элементом a_{ij}. Такие преобразования теряют свойство минимальности суммы квадратов внедиагональных элементов, имеющее место в преобразованиях Якоби для симметричных матриц, но позволяют (Гивенс, 1954 г. [54]) привести симметричную матрицу к трехдиагональному виду существенно быстрее, чем это требуется для выполнения одного цикла преобразований в методе вращений Якоби. Приведение несимметричных матриц к форме Хессенберга методом Гивенса требует большего числа арифметических операций, чем это нужно для такого приведения методом Хаусхолдера, поэтому обычно для этих целей отдают предпочтение последнему. Хотя в случае слабозаполненных матриц у метода Гивенса есть определенные преимущества.

§ 5.4. СДВИГИ И ПОНИЖЕНИЕ РАЗМЕРНОСТИ В QR-АЛГОРИТМЕ

Пусть **QR**-алгоритм (его вторая стадия) с помощью ортогональных преобразований, например на основе преобразований Гивенса (5.13) – (5.15), применяемых к матрицам $\mathbf{A}_k$ в итерационном процессе (5.4), начинается с матрицы $\mathbf{A}_0$, которая подобна данной матрице $\mathbf{A} := (a_{ij})_{i,j=1}^{n}$ и уже имеет форму Хессенберга. (Это, как мы знаем, достигается за $n-2$ этапа преобразований Хаусхолдера (5.7), (5.10), т.е. первая стадия алгоритма — конечная).

Обсудим следующее приводимое без доказательства утверждение [51].

Теорема 5.2. *Если матрица* $\mathbf{A}$ *допускает строгое упорядочение модулей собственных чисел*

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|,$$

то порождаемая ***QR****-алгоритмом* (5.4) *последовательность*

матриц $\mathbf{A}_k$ *сходится к матрице вида*

$$\mathbf{U} = \begin{pmatrix} \lambda_1 & * & * & \cdots & * \\ 0 & \lambda_2 & * & \cdots & * \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

со скоростью аннулирования поддиагональных элементов, характеризуемой равенством

$$a_{i+1,i}^{(k)} = O\left(\left(|\lambda_{i+1}|/|\lambda_i|\right)^k\right). \qquad (5.16)$$

Из этой теоремы* видно, что даже предварительное приведение исходной матрицы $\mathbf{A}$ к форме Хессенберга гарантирует лишь линейную скорость сходимости **QR**-алгоритма с большим замедлением скорости аннулирования поддиагональных элементов матриц $\mathbf{A}_k$ в той их части, где на диагонали предельной матрицы $\mathbf{U}$ располагаются близкие по модулю собственные числа $\mathbf{A}$.

Предположим, что на каком-то k_0-м шаге **QR**-алгоритма (5.4), выполняемого в условиях теоремы 5.2, элемент $n \times n$-матрицы $\mathbf{A}_{k_0}$, находящийся в n-й строке на $(n-1)$-м месте, оказался равным нулю. В таком случае дальнейшие шаги в **QR**-алгоритме не могут затронуть последнюю строку и изменить элемент матрицы $a_{nn}^{(k_0)}$. Поэтому можно положить $\lambda_n := a_{nn}^{(k_0)}$ и исключить из последующих преобразований n-ю строку и n-й столбец, т.е. в дальнейшем обрабатывать матрицы размера $(n-1)\times(n-1)$.

Подмеченная возможность последовательного нахождения собственных чисел и понижения размера матриц, участвующих в выполняемых преобразованиях при нахождении каждого последующего собственного числа (такую процедуру называют ***исчерпыванием***), — не единственный способ повышения вы-

---

* В случае, если матрица $\mathbf{A}_0$ не имеет формы Хессенберга, то (5.16) в этой теореме заменяется формулой более общего вида

$$a_{ij}^{(k)} = O\left(\left(|\lambda_i|/|\lambda_j|\right)^k\right) \text{ при } i > j.$$

числительной эффективности **QR**-алгоритма (5.4). Существенного увеличения скорости сходимости этого процесса можно достигнуть посредством использования в нем удачного пошагового преобразования спектра.

Пусть на некотором шаге $k := \tilde{k}$ алгоритма (5.4) есть основание считать грубо реализованным приближенное равенство

$$\lambda_n :\approx a_{nn}^{(\tilde{k})}$$

(на практике наступление этого момента регистрируют, например, с помощью следующего критерия [18]:

$$\left|a_{nn}^{(k)} - a_{nn}^{(k-1)}\right| \le \frac{1}{3}\left|a_{nn}^{(k-1)}\right| \quad \Rightarrow \quad \tilde{k} := k). \tag{5.17}$$

Тогда, если сместить спектр матрицы $\mathbf{A}$ (а значит, и $\mathbf{A}_0$) на величину $a_{nn}^{(\tilde{k})}$ и применять **QR**-алгоритм к матрице $\tilde{\mathbf{A}}_0 := \mathbf{A}_0 - a_{nn}^{(\tilde{k})}\mathbf{E}$, то вместо характеристики скорости сходимости $O\left(\left(|\lambda_n|/|\lambda_{n-1}|\right)^k\right)$ в соответствии с утверждением теоремы 5.2 будем иметь характеристику $O\left(\left(\left|\lambda_n - a_{nn}^{(\tilde{k})}\right|\Big/\left|\lambda_{n-1} - a_{nn}^{(\tilde{k})}\right|\right)^k\right)$, что, в силу малости числителя в последнем выражении, может оказаться значительно лучше. Применяя такой ***сдвиг*** (называемый ***сдвигом по Рэлею*** [1, 54, 68]) при каждом k, начиная со значения $\tilde{k} \ge 1$, удовлетворяющего правилу (5.17), получают квадратично сходящийся процесс аннулирования $(n-1)$-го элемента в n-й строке. Чтобы выписать собственное число λ_n, теперь нужно учесть все смещения спектра, т.е. к последнему элементу диагонали после аннулирования элемента в позиции $(n, n-1)$ надо прибавить величины всех накопленных смещений (сдвигов).

Можно поступать и иначе: на каждом шаге алгоритма делать обратные сдвиги. Упрощенная запись такого варианта ***QR-алгоритма со сдвигами*** (прямыми и обратными) имеет следующий вид:

$$\mathbf{A}_k - \sigma_k \mathbf{E} = \mathbf{Q}_k \mathbf{R}_k \tag{5.18}$$

(***QR*-факторизация матрицы** $\mathbf{A}_k - \sigma_k \mathbf{E}$),

$$\mathbf{A}_{k+1} := \mathbf{R}_k \mathbf{Q}_k + \sigma_k \mathbf{E} \tag{5.19}$$

(*перемножение в обратном порядке и обратный сдвиг*), *где*

$$\sigma_k := \begin{cases} 0 & \text{при } k < \tilde{k}, \\ a_{nn}^{(k)} & \text{при } k \geq \tilde{k}, \end{cases} \qquad k = 0, 1, 2, \ldots,$$

а $\tilde{k}$ *определяется проверкой* (5.17).

При этом не следует забывать, что таким образом может быть найдено наименьшее по модулю в е щ е с т в е н н о е собственное число, после чего выполнение алгоритма продолжается с понижением размерности ($n := n-1$). Заметим, что наличие кратных вещественных собственных чисел не должно отражаться на работе алгоритма — все они могут быть им последовательно найдены.

Рассмотрим теперь случай возможного наличия у **A** комплексных собственных чисел.

Будем считать для удобства, что у матрицы **A** наименьший модуль имеют комплексно-сопряженные собственные числа λ_n и $\lambda_{n-1} = \overline{\lambda}_n$. В процессе работы **QR**-алгоритма, примененного к такой матрице (приведенной к форме Хессенберга), подобная **A** матрица приобретает вид

$$\begin{pmatrix} * & * & * & & & & & \\ * & * & * & & & * & & \\ 0 & * & * & \ddots & & & & \\ & & & & * & * & * & * \\ & & & & * & * & * & * \\ & 0 & & & 0 & \delta & a & b \\ & & & & 0 & 0 & c & d \end{pmatrix}.$$

Если абсолютная величина стоящего в позиции $(n-1, n-2)$ элемента δ пренебрежимо мала, то числа λ_n и $\overline{\lambda}_n$ — корни характеристического для 2×2-матрицы $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ уравнения

$$\lambda^2 - (a+d)\lambda + ad - bc = 0.$$

Сходимость процесса аннулирования δ можно ускорить, вводя сдвиги поочередно на величины, служащие приближениями к значениям λ_n и $\overline{\lambda}_n$. Очевидно, за таковые на k-м шаге **QR**-алгоритма естественно принять собственные числа τ_k и $\overline{\tau}_k$ последней на диагонали матрицы $\mathbf{A}_k$ 2×2-подматрицы $\tilde{\mathbf{A}}_k := \begin{pmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{nn}^{(k)} \end{pmatrix}$.

Следовательно, в рассматриваемом случае сдвиги с возвратом в **QR**-алгоритме могут быть организованы аналогично (5.18)–(5.19) по следующей схеме:

$$\begin{aligned} \mathbf{A}_k - \tau_k \mathbf{E} = \mathbf{Q}_k \mathbf{R}_k, \quad \mathbf{A}_{k+1} := \mathbf{R}_k \mathbf{Q}_k + \tau_k \mathbf{E}, \\ \mathbf{A}_{k+1} - \overline{\tau}_k \mathbf{E} = \mathbf{Q}_{k+1} \mathbf{R}_{k+1}, \quad \mathbf{A}_{k+2} := \mathbf{R}_{k+1} \mathbf{Q}_{k+1} + \overline{\tau}_k \mathbf{E}, \end{aligned} \tag{5.20}$$

где τ_k, $\overline{\tau}_k$ — корни квадратного уравнения

$$\tau^2 - \left(\operatorname{Sp} \tilde{\mathbf{A}}_k\right)\tau + \det \tilde{\mathbf{A}}_k = 0.$$

При этом сдвиги в таком процессе (называемые ***сдвигами по Уилкинсону*** [32, 54, 68] или ***правилом Фрэнсиса*** [26]) могут начинаться или с первого шага, или после выполнения подобного фигурирующему в (5.17) условия

$$\left\| \tilde{\mathbf{A}}_k - \tilde{\mathbf{A}}_{k-1} \right\| \le \frac{1}{3} \left\| \tilde{\mathbf{A}}_{k-1} \right\|$$

(что может положительно отразиться на точности вычисляемых собственных чисел с малыми модулями).

Присутствие комплексных чисел в формулах (5.20) указывает на то, что порождаемые вещественной матрицей $\mathbf{A}_k$ матрицы $\mathbf{Q}_k$, $\mathbf{R}_k$, $\mathbf{A}_{k+1}$, $\mathbf{Q}_{k+1}$, $\mathbf{R}_{k+1}$ уже не обязаны быть вещественными и, значит, реализация **QR**-алгоритма непосредственно по схеме (5.20) требует подключения арифметики комплексных чисел, а это снижает вычислительную эффективность сдвигов. К счастью, оказывается, что матрица $\mathbf{A}_{k+2}$, получающаяся на выходе алгоритма с двойным сдвигом (5.20), — вещественная. Покажем это.

Сначала убедимся, что имеет место представление

$$\mathbf{A}_{k+2} = (\mathbf{Q}_k\mathbf{Q}_{k+1})^* \mathbf{A}_k (\mathbf{Q}_k\mathbf{Q}_{k+1}) \tag{5.21}$$

(делающее сдвиги невидимыми, прячущимися в матрицы ортогональных преобразований). Для этого достаточно первое из равенств (5.20), переписанное в виде $\mathbf{A}_k = \mathbf{Q}_k\mathbf{R}_k + \tau_k\mathbf{E}$, умножить справа на матрицу $\mathbf{Q}_k$ и слева на сопряженную к ней матрицу $\mathbf{Q}_k^*$, что с учетом свойства $\mathbf{Q}_k^*\mathbf{Q}_k = \mathbf{E}$ и второго из равенств (5.20) дает

$$\mathbf{Q}_k^*\mathbf{A}_k\mathbf{Q}_k = \mathbf{A}_{k+1}.$$

Умножая, в свою очередь, это равенство справа на $\mathbf{Q}_{k+1}$ и слева на $\mathbf{Q}_{k+1}^*$, с учетом двух последних из равенств (5.20) приходим к выражению (5.21).

Далее из тех же определяющих двойной сдвиг формул (5.20) выводим

$$(\mathbf{Q}_k\mathbf{Q}_{k+1})(\mathbf{R}_{k+1}\mathbf{R}_k) = \mathbf{Q}_k(\mathbf{A}_{k+1} - \overline{\tau}_k\mathbf{E})\mathbf{R}_k =$$

$$= \mathbf{Q}_k(\mathbf{R}_k\mathbf{Q}_k + \tau_k\mathbf{E} - \overline{\tau}_k\mathbf{E})\mathbf{R}_k = (\mathbf{Q}_k\mathbf{R}_k\mathbf{Q}_k + \tau_k\mathbf{Q}_k - \overline{\tau}_k\mathbf{Q}_k)\mathbf{R}_k =$$

$$= (\mathbf{Q}_k\mathbf{R}_k + \tau_k\mathbf{E} - \overline{\tau}_k\mathbf{E})\mathbf{Q}_k\mathbf{R}_k = (\mathbf{A}_k - \overline{\tau}_k\mathbf{E})(\mathbf{A}_k - \tau_k\mathbf{E}). \tag{5.22}$$

Следовательно, произведение матриц

$$(\mathbf{Q}_k\mathbf{Q}_{k+1})(\mathbf{R}_{k+1}\mathbf{R}_k) = \mathbf{A}_k^2 - 2\operatorname{Re}\tau_k\mathbf{A}_k + |\tau_k|^2\mathbf{E}$$

является вещественной матрицей. Но вещественная матрица при определенной (например, описанной в § 1.4) процедуре может единственным образом быть ортогонально приведенной к верхней треугольной матрице с вещественными элементами (не отрицательными на диагонали). Значит, матрица $\mathbf{R} := \mathbf{R}_{k+1}\mathbf{R}_k$ — вещественная, тогда и матрица $\mathbf{Q} := \mathbf{Q}_k\mathbf{Q}_{k+1}$ — тоже вещественная. На основе этого и равенства (5.21) теперь можно утверждать, что матрица

$$\mathbf{A}_{k+2} := \mathbf{Q}^T\mathbf{A}_k\mathbf{Q} \tag{5.23}$$

в **QR**-алгоритме с двойными сдвигами должна быть вещественной.

Установленный факт означает, что, во-первых, парные комплексные сдвиги не должны влечь за собой появление комплексных чисел на последующих шагах алгоритма, а во-вторых, указывает на принципиальную возможность построения по матрице $\mathbf{A}_k$ сразу требуемой матрицы $\mathbf{A}_{k+2}$, минуя построение $\mathbf{A}_{k+1}$ и манипулируя только вещественными числами.

Параллельное рассмотрение равенства (5.23) и вытекающего из (5.22) с учетом сделанных обозначений равенства

$$\mathbf{QR} = \tilde{\mathbf{B}} := \left(\mathbf{A}_k - \overline{\tau}_k \mathbf{E}\right)\left(\mathbf{A}_k - \tau_k \mathbf{E}\right),$$

в котором $\mathbf{R} := \mathbf{Q}^T \tilde{\mathbf{B}}$ — треугольная матрица, показывает, что подходящая для подстановки в (5.23) вещественная ортогональная матрица $\mathbf{Q}$ должна быть такой, чтобы транспонированная к ней матрица $\mathbf{Q}^T$ приводила вещественную матрицу

$$\tilde{\mathbf{B}} := \mathbf{A}_k^2 - \left(\tau_k + \overline{\tau}_k\right)\mathbf{A}_k + \tau_k \overline{\tau}_k \mathbf{E} \tag{5.24}$$

к треугольному виду. При этом заметим, что матрица $\tilde{\mathbf{B}}$, вообще говоря, не является хессенберговой (появляются ненулевые элементы на второй поддиагонали), матрица же $\mathbf{A}_{k+2}$ должна иметь форму Хессенберга.

Требуемые ортогональные матрицы $\mathbf{Q}$ строятся как на базе преобразований отражений Хаусхолдера, так и на базе преобразований вращений Гивенса. В любом случае ключевую роль здесь играет п е р в ы й столбец матрицы $\tilde{\mathbf{B}}$, три первых (ненулевых) элемента которого, согласно выражению (5.24), есть

$$\begin{aligned} b_1 &:= a_{11}^2 + a_{12}a_{21} - \left(\tau_k + \overline{\tau}_k\right)a_{11} + \tau_k\overline{\tau}_k, \\ b_2 &:= a_{21}\left[a_{11} + a_{22} - \left(\tau_k + \overline{\tau}_k\right)\right], \\ b_3 &:= a_{32}a_{21}, \end{aligned} \tag{5.25}$$

где для упрощения записи опущен верхний индекс k.

Теоретической подоплекой этому служит утверждение, которое называют *теоремой единственности* [32] или *теоремой о*

неявном $\mathbf{Q}$ [23]. Суть утверждения состоит в том, что однозначность представления матрицы $\mathbf{A}$ в верхней форме Хессенберга $\mathbf{B}$ равенством $\mathbf{B} = \mathbf{Q}^T\mathbf{A}\mathbf{Q}$ обеспечивается фиксированием первого столбца ортогональной матрицы $\mathbf{Q}$, но при условии, что $\mathbf{B}$ является ***неприводимой*** [23, 67] или, иначе, ***неразложимой матрицей*** [32] (т.е. что у $\mathbf{B}$ нет нулевых элементов на поддиагонали)*. Однозначность здесь трактуется как наличие связи между двумя неприводимыми хессенберговыми матрицами $\mathbf{B}$ и $\tilde{\mathbf{B}}$ вида

$$\tilde{\mathbf{B}} = \mathbf{D}^{-1}\mathbf{B}\mathbf{D} \quad \text{с} \quad \mathbf{D} := \operatorname{diag}(\pm 1; \ldots; \pm 1),$$

если они определяются равенствами $\mathbf{B} := \mathbf{Q}^T\mathbf{A}\mathbf{Q}$ и $\tilde{\mathbf{B}} := \mathbf{Z}^T\mathbf{A}\mathbf{Z}$ с ортогональными матрицами $\mathbf{Q}$ и $\mathbf{Z}$, имеющими один и тот же первый столбец.

Примем за основу в процессе двойных сдвигов преобразование Хаусхолдера** и применительно к матрице $\mathbf{A}_k$ (имеющей форму Хессенберга) выполним сначала один шаг преобразования подобия

$$\mathbf{A}_k^{(0)} = \mathbf{H}_k^{(0)}\mathbf{A}_k\mathbf{H}_k^{(0)},$$

где в матрице Хаусхолдера $\mathbf{H}_k^{(0)} := \mathbf{E} - 2\mathbf{w}^T\mathbf{w}$ вектор $\mathbf{w}$ полагаем коллинеарным вектору $(b_1; b_2; b_3; 0; \ldots; 0)^T$ с элементами b_1, b_2, b_3, определенными равенствами (5.25). При этом $n \times n$-матрица $\mathbf{A}_k^{(0)}$ приобретает структуру вида [67]

---

* В книге [26] дана следующая формулировка упомянутой теоремы: «*Пусть* $\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{B}$ — *неразложимая верхняя матрица Хессенберга. Тогда первый столбец матрицы* $\mathbf{Q}$ *однозначно (с точностью до знаков) определяет ее 2-й, ..., n-й столбцы*»

** В [67] можно найти описание такого процесса и на основе преобразования Гивенса, но там же отмечено, что при двойных сдвигах преобразования Хаусхолдера несколько экономичнее.

$$\begin{pmatrix} * & * & * & * & * & * & \dots \\ * & * & * & * & * & * & \dots \\ * & * & * & * & * & * & \dots \\ * & * & * & * & * & * & \dots \\ 0 & 0 & 0 & * & * & * & \dots \\ 0 & 0 & 0 & 0 & * & * & \dots \\ . & . & . & . & . & . & . \end{pmatrix}.$$

Следующие $n-2$ матрицы Хаусхолдера $\mathbf{H}_k^{(1)}, \dots, \mathbf{H}_k^{(n-2)}$ строят по матрице $\mathbf{A}_k^{(0)}$ так, чтобы привести ее к форме Хессенберга (как это было показано ранее в § 5.2, только здесь учитывают ее заведомую близость к нужному виду, что значительно сокращает число выполняемых операций — ***флопов*** в современной терминологии*). В результате матрица

$$\mathbf{A}_k^{(n-2)} := \left(\mathbf{H}_k^{(1)}\mathbf{H}_k^{(2)}\dots\mathbf{H}_k^{(n-2)}\right)^T \mathbf{A}_k^{(0)} \left(\mathbf{H}_k^{(1)}\mathbf{H}_k^{(2)}\dots\mathbf{H}_k^{(n-2)}\right)$$

будет ортогонально подобной $\mathbf{A}_k^{(0)}$ и иметь форму Хессенберга. Но тогда, согласно всем предыдущим рассуждениям и с учетом того, что матрицы $\mathbf{H}_k^{(1)}, \dots, \mathbf{H}_k^{(n-2)}$ имеют те же первые строки, что и определенная выше матрица $\mathbf{H}_k^{(0)}$, матрица $\mathbf{A}_k^{(n-2)}$ может далее играть роль матрицы $\mathbf{A}_{k+2}$ в **QR**-алгоритме с неявными двойными сдвигами, основанном на формулах (5.20).

Залогом справедливости последнего положения должна служить неприводимость получающихся в алгоритме матриц Хессенберга. Если на каком-то этапе алгоритма в матрице Хессенберга появится нулевой поддиагональный элемент, то отсюда будет следовать ее разложимость, т.е. алгоритм далее можно будет продолжать с матрицами меньшего размера.

Заметим, что ни на каком этапе рассуждений не использова-

---

* Флоп (от англ. *floating point operation*) — одна операция с плавающей точкой [23, 26].

лась комплексная сопряженность фигурировавших в них чисел τ_k и $\overline{\tau}_k$ — параметров сдвигов. Это означает *допустимость техники двойных неявных сдвигов и в случае, когда* τ_k и $\overline{\tau}_k$ *— вещественные числа*, если, например, весь **QR**-алгоритм основывать на такой технике (в соответствующем варианте его иногда называют ***двухшаговым QR-алгоритмом Фрэнсиса*** [68]).

Проверку на допустимость считать нулями поддиагональные элементы $a_{i+1,i}^{(k)}$ в **QR**-алгоритме, необходимую для решения вопроса о возможности расщепления решаемой подзадачи или о том, что уже определилась очередная пара собственных чисел, рекомендуется выполнять с помощью неравенства

$$\left|a_{i+1,i}^{(k)}\right| \le \left(\left|a_{ii}^{(k)}\right| + \left|a_{i+1,i+1}^{(k)}\right|\right)\beta,$$

в котором $\beta > 0$ — малое число, связываемое с точностью используемой компьютерной арифметики (например, можно взять $\beta = 2^{-l}$, где l — число двоичных разрядов, выделяемых под мантиссу компьютерного числа).

Практика применения **QR**-алгоритмов с двойными сдвигами показывает их высокую эффективность и широкую применимость. Как правило, нахождение каждой пары собственных чисел требует не более 2–5 итераций [18, 68]. Однако (ввиду отсутствия завершенной теории сходимости таких алгоритмов) не исключены ситуации, когда происходят, мягко говоря, задержки с получением очередной пары собственных чисел*. В таких случаях рекомендуют переопределять параметры сдвигов. Например, в [68] дан следующий практический совет: если за 10 итераций не найдена пара собственных чисел, то параметры τ_k, $\overline{\tau}_k$ в алгоритме с неявными сдвигами следует попытаться найти из системы

---

* Примером матрицы в форме Хессенберга, для которой не эффективны ни основной **QR**-алгоритм, ни рассмотренные сдвиги, может служить матрица со всеми нулевыми элементами, кроме первой поддиагонали с единицами на ней и одним ненулевым элементом в правой верхней позиции [32].

$$\begin{cases} \tau_k + \overline{\tau}_k = 1.5\left(\left|a^{(k)}_{n,n-1}\right| + \left|a^{(k)}_{n-1,n-2}\right|\right), \\ \tau_k \overline{\tau}_k = \left(\left|a^{(k)}_{n,n-1}\right| + \left|a^{(k)}_{n-1,n-2}\right|\right)^2 \end{cases}$$

(для простоты здесь указаны индексы, соответствующие последним собственным числам λ_{n-1} и λ_n в спектре $\mathbf{A}$, что, в общем, естественно, поскольку в алгоритме с двойными сдвигами и исчерпыванием должно быть предусмотрено переприсваивание $n := n-2$ после вычисления каждой очередной пары комплексных собственных чисел или $n := n-1$ после получения вещественного собственного числа).

QR-алгоритм нахождения собственных чисел матрицы с двойными сдвигами (упрощенный вариант)

Входные данные: *матрица в форме Хессенберга* $\mathbf{A} := (a_{ij})_{i,j=1}^{n}$,

ε — *точность.*

Выходные данные: *вектор собственных чисел* $\boldsymbol{\lambda} := (\lambda_1; \ldots; \lambda_n)$.

1. Вычислить собственные значения $u_1 + vi$, $u_2 - vi$ правой нижней 2×2-подматрицы $\mathbf{A}$ (если они вещественны, то $v := 0$, иначе $u_1 := u_2$);
2. $t_1 := a_{11}^2 - a_{11}(u_1 + u_2) + \left(u_1 u_2 + v^2\right) + a_{12}a_{21}$;
3. $t_2 := a_{21}\left(a_{11} + a_{22} - u_1 - u_2\right)$;
4. $t_3 := a_{32}a_{21}$;
5. если $t_1 < 0$, то $s := \sqrt{t_1^2 + t_2^2 + t_3^2}$; иначе $s := -\sqrt{t_1^2 + t_2^2 + t_3^2}$;
6. $\mu := \dfrac{1}{\sqrt{2s(s - t_1)}}$;
7. $w_1 := \mu\left(t_1 - s\right)$;
8. $w_2 := \mu t_2$;
9. $w_3 := \mu t_3$;
10. для $i := 4, \ldots, n$: $w_i := 0$;

11. построить матрицу Хаусхолдера **H** с использованием вектора $\mathbf{w} := (w_1; w_2; \ldots; w_n)^T$;
12. $\mathbf{A} := \mathbf{HAH}$.
13. Для $i := 2, \ldots, n-1$:
    13.1. для $j := 1, \ldots, n$: $w_j := 0$;
    13.2. $s := \sum_{j=i}^{n} a_{j,i-1}^2$;
    13.3. если $a_{i,i-1} < 0$, то $s := \sqrt{s}$; иначе $s := -\sqrt{s}$;
    13.4. $\mu := \dfrac{1}{\sqrt{2s(s - a_{i,i-1})}}$;
    13.5. $w_i := \mu(a_{i,i-1} - s)$;
    13.6. для $j := i+1, \ldots, n$: $w_j := \mu a_{j,i-1}$;
    13.7. построить матрицу Хаусхолдера **H** с использованием вектора **w**;
    13.8. $\mathbf{A} := \mathbf{HAH}$.
14. Если $|a_{n,n-1}| < \varepsilon$, то:
    14.1. $\lambda_n := a_{nn}$;
    14.2. $n := n-1$;
15. если $|a_{n-1,n-2}| < \varepsilon$ или $n = 2$, то:
    15.1. вычислить собственные значения $u_1 + vi$, $u_2 - vi$ правой нижней 2×2-подматрицы **A**;
    15.2. $\lambda_{n-1} := u_1 + vi$;
    15.3. $\lambda_n := u_2 - vi$;
    15.4. $n := n-2$;
16. если $n = 1$, то:
    16.1. $\lambda_n := a_{nn}$;
    16.2. $n := 0$;
17. если $n \neq 0$, то переход к пункту 1, иначе выход.

Пример 5.5. Продолжив выполнение предписаний **QR**-алгоритма (5.4), один полный шаг которого на основе преобразований Гивенса над 3×3-матрицей Хессенберга $\mathbf{B}_0$ продемонстрирован в примере 5.4, ее собственные числа $\lambda_1 :\approx 7{,}639$, $\lambda_2 :\approx -1{,}205$, $\lambda_3 :\approx -0{,}435$ с точностью до 0,001 получим за 8 таких шагов. **QR**-алгоритм (5.18) – (5.19) с одинарными сдвигами (без использования условия (5.17)) приводит к тому же результату за 4 полных шага. Неявными двойными сдвигами (т.е. двухшаговым **QR**-алгоритмом Фрэнсиса) находим те же числа всего за 2 итерации. Эффект от сдвигов налицо, причем здесь двойные сдвиги применены в ситуации, когда все собственные числа — вещественные.

Рассмотренный выше **QR**-алгоритм в разных вариациях настроен на нахождение всех собственных чисел произвольных вещественных $n\times n$-матриц и, в отличие от метода вращений Якоби, ориентированного на симметричные матрицы, мало приспособлен для вычисления собственных векторов. Хотя в рамках этого метода, разумеется, можно путем значительных вычислительных затрат, связанных с преобразованиями самих матриц ортогональных преобразований, находить и собственные векторы (как это делать, можно узнать, например, из [32]), но чаще здесь прибегают к использованию обратного степенного метода (см. § 4.3), тем более что задача нахождения *всех* собственных векторов — сравнительно редкая. При этом обратные итерации обычно применяют не к исходной матрице $\mathbf{A}$, а к матрице $\mathbf{B}$, подобной $\mathbf{A}$ и имеющей форму Хессенберга. Если приведение матрицы $\mathbf{A}$ к виду $\mathbf{B}$ выполнялось преобразованиями Хаусхолдера, то $\mathbf{B}=\mathbf{H}^T\mathbf{A}\mathbf{H}$, где $\mathbf{H}$ — результирующая матрица $n-2$ элементарных отражений, и значит, согласно свойству 4.7 собственных пар матрицы, найдя собственный вектор $\mathbf{y}$ матрицы $\mathbf{B}$, искомый собственный вектор $\mathbf{x}$ данной матрицы $\mathbf{A}$ получаем равенством $\mathbf{x}=\mathbf{H}\mathbf{y}$.

§ 5.5. ПРИМЕНЕНИЕ QR-АЛГОРИТМА К НАХОЖДЕНИЮ КОРНЕЙ МНОГОЧЛЕНА

Как уже отмечалось в § 4.1, классический подход к решению полной алгебраической проблемы собственных значений для $n\times n$-матрицы $\mathbf{A}$ предполагает последовательное решение трех

подзадач: развертывание векового определителя $\det(\mathbf{A}-\lambda\mathbf{E})$, вычисление корней λ_i полученного многочлена n-й степени, нахождение всех линейно независимых нетривиальных решений вырожденных СЛАУ $(\mathbf{A}-\lambda\mathbf{E})\mathbf{x}=\mathbf{0}$ при каждом конкретном значении $\lambda := \lambda_i$. Один из способов решения первой из перечисленных подзадач опирается на построение по данной матрице $\mathbf{A}$ ***сопровождающей матрицы***, или ***матрицы Фробениуса***, [18], которая может иметь, например, следующий вид:

$$\mathbf{C} := \begin{pmatrix} -c_1 & -c_2 & -c_3 & \cdots & -c_{n-2} & -c_{n-1} & -c_n \\ 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \end{pmatrix}. \qquad (5.26)$$

По первой строке этой матрицы можно записать характеристический для матрицы $\mathbf{A}$ многочлен

$$P_n(\lambda) := \lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \ldots + c_{n-1}\lambda + c_n \ , \qquad (5.27)$$

с которого должно стартовать решение второй подзадачи.

Убедимся в том, что этот многочлен является характеристическим не только для исходной матрицы $\mathbf{A}$, но и для самой сопровождающей матрицы $\mathbf{C}$. С этой целью запишем выражение $\det(\mathbf{C}-\lambda\mathbf{E})$ и разложим этот определитель по элементам первой строки. Имеем:

$$\det(\mathbf{C}-\lambda\mathbf{E}) := \begin{vmatrix} -c_1-\lambda & -c_2 & -c_3 & \cdots & -c_{n-1} & -c_n \\ 1 & -\lambda & 0 & \cdots & 0 & 0 \\ 0 & 1 & -\lambda & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -\lambda & 0 \\ 0 & 0 & 0 & \cdots & 1 & -\lambda \end{vmatrix} =$$

$$= (-c_1 - \lambda)\begin{vmatrix} -\lambda & 0 & \cdots & 0 & 0 \\ 1 & -\lambda & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & -\lambda & 0 \\ 0 & 0 & \cdots & 1 & -\lambda \end{vmatrix} + c_2 \begin{vmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & -\lambda & 0 \\ 0 & 0 & \cdots & 1 & -\lambda \end{vmatrix} -$$

$$-c_3 \begin{vmatrix} 1 & -\lambda & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & -\lambda & 0 \\ 0 & 0 & \cdots & 1 & -\lambda \end{vmatrix} + \ldots + (-1)^{n-1} c_{n-1} \begin{vmatrix} 1 & -\lambda & 0 & \cdots & 0 \\ 0 & 1 & -\lambda & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & -\lambda \end{vmatrix} +$$

$$+ (-1)^n c_n \begin{vmatrix} 1 & -\lambda & 0 & \cdots & 0 \\ 0 & 1 & -\lambda & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -\lambda \\ 0 & 0 & 0 & \cdots & 1 \end{vmatrix} = .$$

$$= (-1)^n (c_1 + \lambda)\lambda^{n-1} + (-1)^n c_2 \lambda^{n-2} +$$

$$+ (-1)^n c_3 \lambda^{n-3} + \ldots + (-1)^n c_{n-1}\lambda + (-1)^n c_n =$$

$$= (-1)^n (\lambda^n + c_1 \lambda^{n-1} + c_2 \lambda^{n-2} + \ldots + c_{n-1}\lambda + c_n).$$

Как видим, с точностью до знака получен тот же характеристический многочлен $P_n(\lambda)$.

Найденное соответствие между корнями многочлена (5.27) и собственными числами матрицы (5.26) позволяет обратить считающуюся здесь основной задачу на собственные значения и применять эффективные методы ее решения для вычисления корней алгебраических уравнений.

Именно, если нужно найти *все* корни уравнения n-й степени канонического вида

$$a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \ldots + a_{n-1} x + a_n = 0, \qquad (5.28)$$

то альтернативой известным методам решения алгебраических

уравнений, таким, как, например, достаточно сложный и не всегда успешный метод Лобачевского–Греффе [6], может служить следующий прием. (Упоминание о таком подходе можно встретить, например, в [47].)

По заданным коэффициентам уравнения (5.28) строят матрицу **C** вида (5.27) с элементами первой строки

$$c_i := a_i / a_0 \quad (i = 1, 2, ..., n).$$

Далее к этой матрице применяют **QR**-алгоритм. Найденные этим алгоритмом собственные числа и будут искомыми корнями уравнения (5.28). Заметим, что матрица **C** заведомо имеет форму Хессенберга, т.е. не требуется выполнять преобразования первого этапа алгоритма.

Пример 5.6. К многочлену $P(x) := x^6 - 7x^5 + 12x^4 - x^2 + 7x - 12$, составленному по корням 4, 3, ± 1, $\pm i$, применен **QR**-алгоритм Фрэнсиса с двойными сдвигами с задаваемой точностью $\varepsilon := 0{,}00001$ (действия производились в числовой среде типа `real`). Результат следующий: за 6 итераций получен корень $\xi_1 :\approx 1{,}00000000000$, следующая итерация дала $\xi_2 :\approx -1{,}00000000000$, еще одна итерация привела к паре комплексно-сопряженных корней $\xi_{3,4} :\approx 0{,}00000000000 \pm 0{,}99999999998i$ и из оставшейся 2×2-матрицы найдены два оставшихся корня $\xi_5 :\approx 3{,}00000000330$ и $\xi_6 :\approx 3{,}9999935760$.

Замечание 5.3. Использование двойных сдвигов при вычислении корней многочленов **QR**-алгоритмом ускоряет их нахождение. Однако в тех случаях, когда заведомо известно, что многочлен имеет только вещественные корни (соответствующая матрица не имеет комплексных собственных чисел), может оказаться более целесообразным применение одинарных сдвигов.

УПРАЖНЕНИЯ

5.1. Для нахождения собственных пар симметричных положительно определенных матриц постройте **LU**-алгоритм на базе $\mathbf{U}^T\mathbf{U}$ (или $\mathbf{L}\mathbf{L}^T$-разложения Холецкого). Опробуйте его на матрицах

$$\mathbf{A} := \begin{pmatrix} 5 & -2 \\ -2 & 2 \end{pmatrix} \quad \text{и} \quad \mathbf{B} := \begin{pmatrix} 6 & 1 & 0 \\ 1 & 5 & -2 \\ 0 & -2 & 1 \end{pmatrix}.$$

Сохраняют ли получаемые на каждом шаге такого алгоритма подобные **B** матрицы трехдиагональную структуру?

5.2. К матрице $\mathbf{A} := \begin{pmatrix} 6 & -2 \\ 4 & 0 \end{pmatrix}$ примените **LU**-алгоритм и сравните результат третьего шага с найденными точно собственными числами.

5.3. Матрицу $\mathbf{A} := \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$ приведите к подобной ей матрице трехдиагонального вида преобразованием Гивенса.

5.4. Выведите формулы (типа формул (1.24)) для непосредственного подсчета элементов матрицы i-го этапа преобразований Хаусхолдера, приводящих произвольную квадратную матрицу $\mathbf{A}$ к подобной ей матрице $\mathbf{B}_0$ (см. (5.9)) в форме Хессенберга.

5.5. Дана матрица $\mathbf{A} := \begin{pmatrix} 5 & 2 & -3 \\ 4 & 5 & -4 \\ 6 & 4 & -4 \end{pmatrix}$.

А) Приведите ее к матрице верхней формы Хессенберга, подобной данной.

Б) К матрице, полученной в п. А), примените 2 – 3 шага **QR**-алгоритма (без сдвигов и со сдвигами). Сравните результаты этих шагов с точными значениями собственных чисел $\lambda_1 = 3$, $\lambda_2 = 2$, $\lambda_3 = 1$.

5.6. Сформируйте матрицу, применением к которой **QR**-алгоритма можно найти все корни уравнения

$$5x^5 - 4x^4 + 3x^3 - 2x^2 + 1 = 0 \ .$$

Глава 6

СИНГУЛЯРНОЕ РАЗЛОЖЕНИЕ ПРЯМОУГОЛЬНЫХ МАТРИЦ

§ 6.1. СИНГУЛЯРНЫЕ ЧИСЛА И СИНГУЛЯРНОЕ РАЗЛОЖЕНИЕ

Обратимся к рассмотрению матриц, которые в общем случае могут быть неквадратными или вырожденными. Для разных алгебраических задач с такими матрицами возникают проблемы, связанные с отсутствием у них классических обратных матриц и собственных значений. Обобщим понятие собственного числа на случай прямоугольной матрицы и далее рассмотрим вычисление и применения таких чисел, называемых сингулярными.

Отправной точкой для обобщения понятия собственного числа матрицы служит тот очевидный факт, что для любой вещественной $m \times n$-матрицы $\mathbf{A}$ порождаемые ею матрицы $\mathbf{A}^T\mathbf{A}$ и $\mathbf{A}\mathbf{A}^T$ являются квадратными симметричными $n \times n$- и $m \times m$-матрицами соответственно, а также менее очевидный факт совпадения спектров этих матриц в их общей части (по минимальному из размеров m и n). При этом собственные числа матриц $\mathbf{A}^T\mathbf{A}$ и $\mathbf{A}\mathbf{A}^T$ заведомо вещественны и неотрицательны. Для квадратной матрицы $\mathbf{A}$ вырожденность равносильна наличию нулевого собственного числа у матриц $\mathbf{A}^T\mathbf{A}$ и $\mathbf{A}\mathbf{A}^T$. Приняв эти факты [18, 46], дадим следующее определение.

Определение 6.1. ***Сингулярными числами**** *вещественной $m \times n$-матрицы $\mathbf{A}$ называются арифметические квадратные корни из собственных чисел $\lambda_1, \lambda_2, \ldots, \lambda_k$ матриц $\mathbf{A}^T\mathbf{A}$ и $\mathbf{A}\mathbf{A}^T$, где $k := \min\{m, n\}$.*

---

* *Singularis* (лат.) — отдельный, особый.

Традиционно сингулярные числа обозначаются буквой σ и нумеруются в порядке убывания:

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k \quad (\geq 0).$$

Количество ненулевых сингулярных чисел матрицы $\mathbf{A}$ совпадает с ее рангом (а также с рангом матриц $\mathbf{A}^T\mathbf{A}$ и $\mathbf{A}\mathbf{A}^T$). Иногда сингулярными числами матрицы называют только ненулевые числа σ_i, т.е. в определении 6.1 под k подразумевается общий ранг указанных матриц [31, 64]. У квадратных матриц произведение всех сингулярных чисел равно модулю определителя. Отсюда — естественность критерия невырожденности квадратной матрицы, состоящего в отличии от нуля всех ее сингулярных чисел.

Непосредственная связь между собственными и сингулярными числами одной и той же матрицы $\mathbf{A}$ обнаруживается в случае, когда она симметричная. Симметричные матрицы образуют подмножество так называемых ***нормальных матриц*** — матриц, обладающих свойством $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T$ [18, 46 и др.]. Модули собственных чисел нормальных матриц равны их сингулярным числам. И если собственные числа симметричных матриц неотрицательны, то для них наборы сингулярных и собственных чисел полностью совпадают. Таким образом, понятие сингулярного числа в определенном смысле действительно обобщает понятие собственного числа матрицы, что обусловливает достаточно обширные применения сингулярных чисел (некоторые из них представлены в следующей главе).

Пример 6.1. А) Пусть $\mathbf{A} := \begin{pmatrix} 3 & 2 \\ 2 & 0 \end{pmatrix}$. Очевидно, $\mathbf{A}^T = \mathbf{A}$, следовательно, $\mathbf{A}$ — нормальная матрица. Ее собственные числа $\lambda_1 := 4$, $\lambda_2 := -1$. Собственные числа матрицы $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \begin{pmatrix} 13 & 6 \\ 6 & 4 \end{pmatrix}$ равны 16 и 1, а значит, сингулярные числа $\mathbf{A}$ есть $\sigma_1 := 4$, $\sigma_2 := 1$. Убеждаемся, что $|\lambda_1| = \sigma_1$, $|\lambda_2| = \sigma_2$.

Б) Собственные числа матрицы $\mathbf{B} := \begin{pmatrix} 4 & -2 \\ -2 & 1 \end{pmatrix}$ (тоже симметричной) суть $\lambda_1 := 5$, $\lambda_2 := 0$, а собственные числа матрицы $\mathbf{B}^T\mathbf{B} = \mathbf{B}\mathbf{B}^T = \begin{pmatrix} 20 & -10 \\ -10 & 5 \end{pmatrix}$ равны 25 и 0. Видим, что для матрицы $\mathbf{B}$ имеют место равенства $\sigma_1 := \sqrt{25} = \lambda_1$, $\sigma_2 := \sqrt{0} = \lambda_2$.

Для вычисления сингулярных чисел прямоугольной матрицы, а также для получения некоторой другой дополнительной информации о ней используют особую определяемую ниже факторизацию матрицы.

Определение 6.2. *Представление вещественной* $m \times n$*-матрицы* $\mathbf{A}$ *в виде*

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}, \tag{6.1}$$

где Σ *— диагональная* $m \times n$*-матрица с диагональю из невозрастающих сингулярных чисел* $\sigma_1, \ldots, \sigma_k$, *а* $\mathbf{U}$ *и* $\mathbf{V}$ *— ортогональные, соответственно,* $m \times m$*- и* $n \times n$*-матрицы, называется* ***сингулярным разложением*** *матрицы* $\mathbf{A}$ *или, иначе,* **SVD-*разложением***[*].

Заметим, что часто сингулярные числа сразу вводятся как диагональные элементы матрицы Σ в SVD-разложении (6.1) (см., например, [23, 68, 71]). С этим хорошо согласуется еще одно определение сингулярного числа σ посредством совокупности равенств

$$\mathbf{A}\mathbf{x} = \sigma\mathbf{y}, \quad \mathbf{A}^T\mathbf{y} = \sigma\mathbf{x}, \tag{6.2}$$

где n-мерный вектор $\mathbf{x}$ и m-мерный вектор $\mathbf{y}$ — *правый* и *левый* ***сингулярные векторы*** соответственно [17, 23]. Эти векторы образуют ортогональные базисы, т.е. существуют такие ортогональные $n \times n$-матрица $\mathbf{X}$ и $m \times m$-матрица $\mathbf{Y}$ из сингулярных векторов-столбцов, что векторно-матричные равенства (6.2) мож-

---

[*] От англ. *Singular Value Decomposition.*

но переписать в матричном виде так:

$$\mathbf{AX} = \mathbf{Y\Sigma}, \quad \mathbf{A}^T\mathbf{Y} = \mathbf{X\Sigma}. \tag{6.3}$$

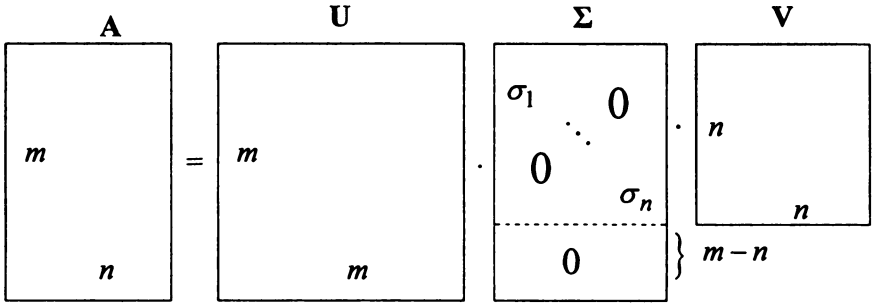
В силу ортогональности матрицы $\mathbf{X}$ первое из равенств (6.3) равносильно равенству

$$\mathbf{A} = \mathbf{Y\Sigma X}^T, \tag{6.4}$$

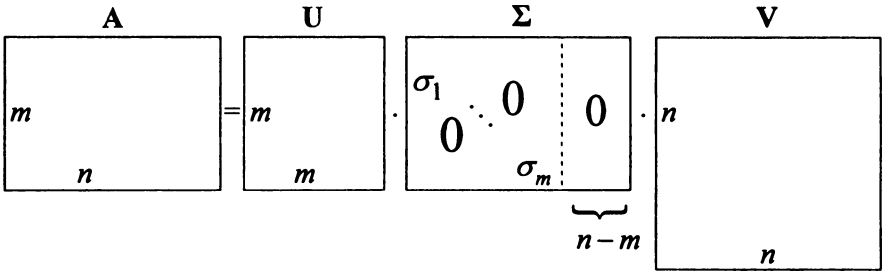
совпадающему с (6.1), если отождествить $\mathbf{U}$ с $\mathbf{Y}$, а $\mathbf{V}$ с $\mathbf{X}^T$.

Равенство (6.1) при разном соотношении чисел строк и столбцов фигурирующих в нем матриц можно символически отобразить следующим образом:

1) при $m > n$



2) при $n > m$



Доказано (см., например, [23, 71]), что *всякая вещественная матрица имеет вещественное сингулярное разложение*. Для квадратных матриц сингулярное разложение было введено и обосновано Сильвестром (1889) [73].

§ 6.2. СТРАТЕГИЯ ПОЛУЧЕНИЯ SVD-РАЗЛОЖЕНИЯ. ЭТАП ДВУХДИАГОНАЛИЗАЦИИ

Если не требуется знание ортогональных матриц $\mathbf{U}$ и $\mathbf{V}$, присутствующих в сингулярном разложении (6.1), то сингулярные числа σ_i матрицы $\mathbf{A}$, в принципе, можно найти по определению, привлекая какой-либо алгоритм вычисления вещественных собственных чисел симметричной матрицы $\mathbf{A}^T\mathbf{A}$ (если $n \le m$) или $\mathbf{A}\mathbf{A}^T$ (если $n > m$). Однако даже в этом, более ограниченном в плане применений случае такой подход нежелателен из-за возможной большой потери точности. Свидетельством этому может служить следующий простой пример [68].

Пример 6.2. Пусть $\mathbf{A} := \begin{pmatrix} 1 & 1 \\ \delta & 0 \\ 0 & \delta \end{pmatrix}$, где δ — малое число. Матрица $\mathbf{A}^T\mathbf{A} = \begin{pmatrix} 1+\delta^2 & 1 \\ 1 & 1+\delta^2 \end{pmatrix}$ имеет собственные числа $\lambda_1 := 2+\delta^2$, $\lambda_2 := \delta^2$. При значениях δ очень маленьких, но значимых в расчетах, может оказаться, что величина δ^2 будет заменена машинным нулем. Тогда вместо истинных сингулярных чисел $\sigma_1 := \sqrt{2+\delta^2}$, $\sigma_2 := |\delta|$ матрицы $\mathbf{A}$ через построение матрицы $\mathbf{A}^T\mathbf{A}$ будут получены числа $\tilde{\sigma}_1 := \sqrt{2}$, $\tilde{\sigma}_2 := 0$ — квадратные корни из собственных чисел матрицы $\widetilde{\mathbf{A}^T\mathbf{A}} := \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ (что искажает, например, значение ранга матрицы $\mathbf{A}$).

Поскольку фигурирующие в сингулярном разложении (6.1) матрицы $\mathbf{U}$ и $\mathbf{V}$ — ортогональные, вполне естественно получить их в результате выполнения последовательностей простых ортогональных преобразований.

Весь процесс сингулярного разложения можно условно разбить на два этапа. Первый этап состоит в приведении исходной $m \times n$-матрицы $\mathbf{A}$ к двухдиагональной форме (*этап двухдиагонализации*), что обычно реализуется с помощью преобразований Хаусхолдера. На втором этапе двухдиагональная матрица — итог первого этапа — приводится к диагональному виду (*этап диаго-*

нализации), что достигается применением к ней **QR**-алгоритма на основе преобразований Гивенса. Названные ортогональные преобразования и **QR**-алгоритм достаточно подробно описаны в предыдущих параграфах применительно к задаче нахождения собственных чисел. Здесь они должны быть использованы так, чтобы учитывались специфика данного объекта и поставленная цель.

Рассмотрим преобразования первого этапа в процессе сингулярного разложения.

Обращаясь к равенству (6.4), придающему содержательный смысл представлению матрицы **A** в виде (6.1), замечаем, что если у матрицы **U** должны быть ортогональными ее столбцы — левые сингулярные векторы, то матрица **V** должна формироваться так, чтобы были ортогональны ее строки — правые сингулярные векторы. Это позволяет приблизиться к пониманию первого шага на пути двухдиагонализации матрицы **A**. Он состоит в том, что матрица **A** умножается слева на матрицу Хаусхолдера $\mathbf{P}_1$, определяемую элементами a_{i1} первого столбца **A** так, чтобы аннулировать все его элементы, находящиеся под диагональю, а справа — тоже на матрицу Хаусхолдера $\mathbf{Q}_1$, определяемую элементами $\tilde{a}_{1j}$ первой строки матрицы $\mathbf{P}_1\mathbf{A}$ так, чтобы аннулировать все ее элементы, стоящие правее элемента $\tilde{a}_{12}$. Следовательно, результат первого шага — это матрица

$$\mathbf{A}_1 := \mathbf{P}_1\mathbf{A}\mathbf{Q}_1, \tag{6.5}$$

где

$$\mathbf{P}_1 := \mathbf{E} - 2\mathbf{w}_1\mathbf{w}_1^T, \qquad \mathbf{w}_1 := \mu_1\left(a_{11} - p_1; a_{21}; \ldots; a_{m1}\right)^T,$$

$$p_1 := \operatorname{sgn}_+\left(-a_{11}\right)\sqrt{\sum_{i=1}^{m} a_{i1}^2}, \qquad \mu_1 := \frac{1}{\sqrt{2p_1^2 - 2p_1a_{11}}},$$

а

$$\mathbf{Q}_1 := \mathbf{E} - 2\mathbf{z}_1\mathbf{z}_1^T, \qquad \mathbf{z}_1 := \nu_1\left(0; \tilde{a}_{12} - q_2; \tilde{a}_{13}; \ldots; \tilde{a}_{1n}\right)^T,$$

$$q_2 := \operatorname{sgn}_+(-\tilde{a}_{12})\sqrt{\sum_{j=2}^{n}\tilde{a}_{1j}^2}\,, \qquad \nu_1 := \frac{1}{\sqrt{2q_2^2 - 2q_2\tilde{a}_{12}}}\,.$$

Согласно смыслу и свойствам преобразований Хаусхолдера, рассмотренным в § 1.4, матрица $\mathbf{P}_1\mathbf{A}$ имеет первым столбцом вектор $p_1\mathbf{e}_1$, который не изменяется при умножении ее справа на матрицу $\mathbf{Q}_1$, в результате чего вторым элементом первой строки оказывается число q_2. Таким образом, полученная в (6.5) матрица $\mathbf{A}_1$ имеет вид

$$\mathbf{A}_1 := \begin{pmatrix} p_1 & q_2 & 0 & \dots & 0 \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & a_{m2}^{(1)} & a_{m3}^{(1)} & \dots & a_{mn}^{(1)} \end{pmatrix}.$$

Беря за основу теперь подматрицу матрицы $\mathbf{A}_1$, получающуюся вычеркиванием первых строки и столбца, аналогичными преобразованиями Хаусхолдера — левым $\mathbf{P}_2$ и правым $\mathbf{Q}_2$, матрицу $\mathbf{A}_1$ приводим к виду

$$\mathbf{A}_2 := \mathbf{P}_2\mathbf{A}_1\mathbf{Q}_2 = \mathbf{P}_2\mathbf{P}_1\mathbf{A}\mathbf{Q}_1\mathbf{Q}_2 = \begin{pmatrix} p_1 & q_2 & 0 & 0 & \dots & 0 \\ 0 & p_2 & q_3 & 0 & \dots & 0 \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & \dots & a_{3n}^{(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & a_{m3}^{(2)} & a_{m4}^{(2)} & \dots & a_{mn}^{(2)} \end{pmatrix}.$$

Нетрудно представить, что в результате конечного процесса таких преобразований данная $m \times n$-матрица $\mathbf{A}$ будет ортогонально приведена к двухдиагональной форме. Если матрица $\mathbf{A}$ — квадратная ($m = n$), то итоговая матрица при таких преобразованиях (обозначим ее $\mathbf{B}$) приобретет вид

$$\mathbf{B} := \begin{pmatrix} p_1 & q_2 & 0 & 0 & \dots & 0 & 0 \\ 0 & p_2 & q_3 & 0 & \dots & 0 & 0 \\ 0 & 0 & p_3 & q_4 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & p_{n-1} & q_n \\ 0 & 0 & 0 & 0 & \dots & 0 & p_n \end{pmatrix}. \qquad (6.6)$$

Для этого потребуется $n-1$ матриц Хаусхолдера типа $\mathbf{P}_i$ и $n-2$ — типа $\mathbf{Q}_j$ (заметим, что здесь элемент q_n — это просто соответствующий элемент $a_{n-1,n}^{(n-2)}$ предыдущего шага). Если у $\mathbf{A}$ число строк превышает число столбцов ($m > n$), то нужно выполнить еще одно левое преобразование $\mathbf{P}_n$, чтобы создать нули под элементом p_n, в результате чего матрица (6.6) дополняется снизу нулевыми строками в количестве $m-n$. При $m < n$ по сравнению со случаем $m = n$ нужно делать лишнее правое преобразование $\mathbf{Q}_{n-1}$.

Независимо от того, какой из случаев $m > n$, $m = n$ или $m < n$ имеет место, сингулярные числа у данной матрицы $\mathbf{A}$ будут те же, что и у квадратной матрицы $\mathbf{B}$ вида (6.6).

Действительно, считая для определенности, что $m > n$, обозначим через $\begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix}$ матрицу (6.6), дополненную снизу $m-n$ нулевыми строками. Тогда результат рассмотренных ортогональных преобразований можно записать в виде

$$\begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix} = \mathbf{P}_n \dots \mathbf{P}_2 \mathbf{P}_1 \mathbf{A} \mathbf{Q}_1 \mathbf{Q}_2 \dots \mathbf{Q}_{n-2}. \qquad (6.7)$$

Отсюда, полагая

$$\mathbf{P} := \mathbf{P}_1 \mathbf{P}_2 \dots \mathbf{P}_n, \quad \mathbf{Q} := \mathbf{Q}_1 \mathbf{Q}_2 \dots \mathbf{Q}_{n-2}, \qquad (6.8)$$

получаем равносильные (6.7) равенства

$$\mathbf{P} \begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix} = \mathbf{A}\mathbf{Q} \quad \text{и} \quad \mathbf{A} = \mathbf{P} \begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix} \mathbf{Q}^T. \qquad (6.9)$$

Если же $m < n$, то такие же преобразования производим с уча-

стием матрицы $(\mathbf{B}|\mathbf{0})$ с $n-m$ нулевыми столбцами справа. В этом случае в равенствах (6.9) матрица $\begin{pmatrix}\mathbf{B}\\ \mathbf{0}\end{pmatrix}$ заменяется матрицей $(\mathbf{B}|\mathbf{0})$, а вместо (6.8) нужно принять

$$\mathbf{P} := \mathbf{P}_1\mathbf{P}_2 \ldots \mathbf{P}_{n-1}, \quad \mathbf{Q} := \mathbf{Q}_1\mathbf{Q}_2 \ldots \mathbf{Q}_{n-1}.$$

Ясно, что эти нулевые $m-n$ строк (при $m > n$) или $n-m$ столбцов (при $m < n$) можно отбросить и продолжать процесс сингулярного разложения с двухдиагональной *квадратной* матрицей $\mathbf{B}$.

При описанном преобразовании матрица $\mathbf{B}$ имеет те же сингулярные числа, что и $\mathbf{A}$, поскольку из итогового равенства $\mathbf{B} = \mathbf{P}^T\mathbf{A}\mathbf{Q}$ и вытекающего из него равенства $\mathbf{B}^T = \mathbf{Q}^T\mathbf{A}^T\mathbf{P}$ следует равенство $\mathbf{B}^T\mathbf{B} = \mathbf{Q}^T\mathbf{A}^T\mathbf{P}\mathbf{P}^T\mathbf{A}\mathbf{Q} = \mathbf{Q}^T(\mathbf{A}^T\mathbf{A})\mathbf{Q}$, означающее $\lambda_{\mathbf{B}^T\mathbf{B}} = \lambda_{\mathbf{A}^T\mathbf{A}}$, т.е. $\sigma_{\mathbf{B}} = \sigma_{\mathbf{A}}$.

Матрица $\mathbf{B}^T\mathbf{B}$ является Хессенберговой и должна быть ортогонально приведена к диагональному виду. Естественно попытаться приспособить для этого **QR**-алгоритм. Реализации такого подхода посвящен следующий параграф.

Пример 6.3. На примере матрицы $\mathbf{A} := \begin{pmatrix} 2 & -1 & -2 \\ 0 & 3 & 1 \\ -1 & 1 & 0 \\ 2 & 0 & -1 \end{pmatrix}$ проследим за описанным процессом двухдиагонализации с помощью правых и левых преобразований Хаусхолдера.

Имеем следующую последовательность промежуточных результатов (округленных до 0,001):

$$p_1 = -3, \quad \mathbf{P}_1 = \frac{1}{15}\begin{pmatrix} -10 & 0 & 5 & -10 \\ 0 & 15 & 0 & 0 \\ 5 & 0 & 14 & 2 \\ -10 & 0 & 2 & 11 \end{pmatrix}, \quad \mathbf{P}_1\mathbf{A} = \begin{pmatrix} -3 & 1 & 2 \\ 0 & 3 & 1 \\ 0 & 0{,}6 & -0{,}8 \\ 0 & 0{,}8 & 0{,}6 \end{pmatrix};$$

$$q_2 = -\sqrt{5} \approx -2{,}236, \quad \mathbf{Q}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0{,}447 & -0{,}894 \\ 0 & -0{,}894 & 0{,}447 \end{pmatrix},$$

$$\mathbf{A}_1 := (\mathbf{P}_1\mathbf{A})\mathbf{Q}_1 \approx \begin{pmatrix} -3 & -2,236 & 0 \\ 0 & -2,236 & -2,236 \\ 0 & 0,447 & -0,894 \\ 0 & -0,894 & -0,447 \end{pmatrix};$$

$$p_2 = \sqrt{6} \approx 2,450, \qquad \mathbf{P}_2 \approx \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -0,913 & 0,183 & -0,365 \\ 0 & 0,183 & 0,982 & 0,035 \\ 0 & -0,365 & 0,035 & 0,930 \end{pmatrix},$$

$$\mathbf{A}_2 := \mathbf{P}_2\mathbf{A}_1 \approx \begin{pmatrix} -3 & -2,236 & 0 \\ 0 & 2,450 & 2,041 \\ 0 & 0 & -1,303 \\ 0 & 0 & 0,369 \end{pmatrix};$$

$$p_3 \approx 1,354, \qquad \mathbf{P}_3 \approx \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -0,962 & 0,271 \\ 0 & 0 & 0,271 & 0,962 \end{pmatrix},$$

$$\mathbf{A}_3 := \mathbf{P}_3\mathbf{P}_2\mathbf{P}_1\mathbf{A}\mathbf{Q}_1 \approx \begin{pmatrix} -3 & -2,236 & 0 \\ 0 & 2,450 & 2,041 \\ 0 & 0 & 1,354 \\ 0 & 0 & 0 \end{pmatrix}.$$

Последняя матрица имеет нужный вид. Следовательно, справедливо представление

$$\mathbf{A} = \mathbf{P}\mathbf{A}_3\mathbf{Q}^T$$

данной матрицы $\mathbf{A}$ через двухдиагональную матрицу $\mathbf{A}_3$ и ортогональные матрицы

$$\mathbf{Q} := \mathbf{Q}_1 \quad \text{и} \quad \mathbf{P} := \mathbf{P}_1\mathbf{P}_2\mathbf{P}_3 = \begin{pmatrix} -0,667 & 0,304 & -0,459 & -0,502 \\ 0 & -0,913 & -0,275 & -0,302 \\ 0,333 & 0,122 & -0,844 & 0,402 \\ -0,667 & -0,243 & 0,037 & 0,704 \end{pmatrix}.$$

Замечание 6.1. Результат первого этапа сингулярного разложения, т.е. представление матрицы $\mathbf{A}$ в виде

$$\mathbf{A} = \mathbf{PBQ}^T,$$

где $\mathbf{B}$ — двухдиагональная матрица (6.6), можно использовать для решения СЛАУ

$$\mathbf{Ax} = \mathbf{b}.$$

Эта система в рассматриваемом случае равносильна системе

$$\mathbf{BQ}^T\mathbf{x} = \mathbf{P}^T\mathbf{b},$$

которая введением переменного вектора $\mathbf{y} := \mathbf{Q}^T\mathbf{x} = (y_1; \ldots; y_n)^T$ и фиксированного вектора $\mathbf{c} := \mathbf{P}^T\mathbf{b} = (c_1; \ldots; c_n)^T$ превращается в систему

$$\begin{cases} p_1 y_1 + q_2 y_2 = c_1, \\ \qquad p_2 y_2 + q_3 y_3 = c_2, \\ \qquad\qquad \cdots\cdots\cdots\cdots \\ \qquad\qquad\qquad p_{n-1} y_{n-1} + q_n y_n = c_{n-1}, \\ \qquad\qquad\qquad\qquad p_n y_n = c_n. \end{cases}$$

Найдя отсюда компоненты вспомогательного вектора $\mathbf{y}$ по формулам

$$y_n = \frac{c_n}{p_n}, \quad y_k = \frac{c_k - q_{k+1} y_{k+1}}{p_k}, \qquad \text{где} \quad k := n-1, n-2, \ldots, 1,$$

приходим к искомому решению данной СЛАУ: $\mathbf{x} = \mathbf{Qy}$. Ясно, что это единственное решение может быть получено подобным способом лишь в случае, когда значения p_k при всех $k \in \{1, 2, \ldots, n\}$ отличны от нуля, т.е. если $\operatorname{rank} \mathbf{A} = n$.

Пример 6.4. Рассмотрим переопределенную систему

$$\begin{cases} 2x_1 - x_2 - 2x_3 = 2, \\ \quad 3x_2 + x_3 = 5, \\ -x_1 + x_2 = 1, \\ 2x_1 - x_3 = 3 \end{cases} \tag{6.10}$$

с матрицей системы $\mathbf{A}$, фигурирующей в предыдущем примере. Используем полученное там представление

$$\mathbf{A} \approx \begin{pmatrix} -0{,}667 & 0{,}304 & -0{,}459 & -0{,}502 \\ 0 & -0{,}913 & -0{,}275 & -0{,}302 \\ 0{,}333 & 0{,}122 & -0{,}844 & 0{,}402 \\ -0{,}667 & -0{,}243 & 0{,}037 & 0{,}704 \end{pmatrix} \cdot \begin{pmatrix} p_1 & q_2 & 0 \\ 0 & p_2 & q_3 \\ 0 & 0 & p_3 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0{,}447 & -0{,}894 \\ 0 & -0{,}894 & 0{,}447 \end{pmatrix}$$

с $p_1 := -3$, $p_2 := \sqrt{6} \approx 2{,}450$, $p_3 :\approx 1{,}354$, $q_2 := -\sqrt{5} \approx -2{,}236$, $q_3 :\approx 2{,}041$.

В соответствии с замечанием 6.1 имеем:

$$\mathbf{c} := \mathbf{P}^{\mathbf{T}}\mathbf{b} \approx \begin{pmatrix} -0,667 & 0 & 0,333 & -0,667 \\ 0,304 & -0,913 & 0,122 & -0,243 \\ -0,459 & -0,275 & -0,844 & 0,037 \\ -0,502 & -0,302 & 0,402 & 0,704 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 5 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} -3,002 \\ -4,564 \\ -3,026 \\ 0,000 \end{pmatrix};$$

$$y_3 := \frac{c_3}{p_3} \approx \frac{-3,026}{1,354} \approx -2,235\ ,$$

$$y_2 := \frac{c_2 - q_3 y_3}{p_2} \approx \frac{-4,564 - 2,041 \cdot (-2,235)}{2,450} \approx 0,001\ ,$$

$$y_1 := \frac{c_1 - q_2 y_2}{p_1} \approx \frac{-3,002 - (-2,236) \cdot 0,001}{-3} \approx 1,000;$$

$$\mathbf{x} := \mathbf{Q}\mathbf{y} \approx \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0,447 & -0,894 \\ 0 & -0,894 & 0,447 \end{pmatrix} \cdot \begin{pmatrix} 1,000 \\ 0,001 \\ -2,235 \end{pmatrix} \approx \begin{pmatrix} 1,000 \\ 1,998 \\ -1,000 \end{pmatrix}.$$

Получаем вектор, отличие которого от точного решения $\mathbf{x}^* := (1;\ 2;\ -1)^T$ системы (6.10) обусловлено лишь погрешностями округлений.

§ 6.3. РАЗЛОЖЕНИЕ ДВУХДИАГОНАЛЬНОЙ МАТРИЦЫ

Как отмечалось ранее, нахождение SVD-разложения $m \times n$-матрицы $\mathbf{A}$ можно продолжить, применяя к *квадратной* матрице $\mathbf{B}$ двухдиагональной структуры (6.6) (условно размера $n \times n$) некоторую модификацию **QR**-алгоритма. Опишем первый шаг итерационного процесса, определяющий одну их возможных таких модификаций.

Пусть $\mathbf{S}$ и $\mathbf{T}$ — ортогональные матрицы, требования к которым будут сформированы позже. Построим матрицу $\mathbf{B}_1$ по правилу

$$\mathbf{B}_1 := \mathbf{S}\mathbf{B}\mathbf{T}^T\ . \tag{6.11}$$

Тогда из равенств

$$\mathbf{B}_1^T\mathbf{B}_1 = \left(\mathbf{S}\mathbf{B}\mathbf{T}^T\right)^T \mathbf{S}\mathbf{B}\mathbf{T}^T = \mathbf{T}\mathbf{B}^T\mathbf{S}^T\mathbf{S}\mathbf{B}\mathbf{T}^T = \mathbf{T}\left(\mathbf{B}^T\mathbf{B}\right)\mathbf{T}^T$$

следует, что матрицы $\mathbf{B}_1^T\mathbf{B}_1$ и $\mathbf{B}^T\mathbf{B}$ ортогонально подобны*. Значит, матрицы $\mathbf{B}_1$ и $\mathbf{B}$, связанные соотношением (6.11), имеют одни и те же сингулярные числа, согласно определению 6.1.

Так как матрица $\mathbf{B}$ — двухдиагональная, то матрица $\mathbf{F} := \mathbf{B}^T\mathbf{B}$ — трехдиагональная. Хотя номинально QR-алгоритм должен применяться именно к этой матрице для нахождения квадратов сингулярных чисел, такой подход, связанный с *явным* построением матрицы $\mathbf{F}$, как было замечено, нежелателен (см. пример 6.2). Отсюда — интерес к преобразованиям вида (6.11), в которых матрица $\mathbf{F}$ в явном виде не участвует.

Преобразования (6.11) совершаются на основе плоских вращений Гивенса (§ 1.5). При этом и матрица $\mathbf{T}$, и матрица $\mathbf{S}$ представляют собой произведения $n-1$ матриц вращений: $\mathbf{T} := \mathbf{T}_n\mathbf{T}_{n-1}...\mathbf{T}_2$, $\mathbf{S} := \mathbf{S}_2\mathbf{S}_3...\mathbf{S}_n$, однако роль элементарных матриц здесь разная. Только матрицы $\mathbf{T}_2^T$ строят так, чтобы обеспечивалась сходимость последовательностей трехдиагональных матриц типа $\mathbf{F}$ к диагональному виду в процессе итераций QR-алгоритма. На остальные же матрицы $\mathbf{S}_2$, $\mathbf{T}_3^T$, $\mathbf{S}_3$, $\mathbf{T}_4^T$,...,$\mathbf{S}_n$, $\mathbf{T}_n^T$ возлагается обязанность поддерживать двухдиагональную структуру матриц типа матриц $\mathbf{B}$ на каждом полном промежуточном шаге алгоритма. Схема основных действий здесь следующая.

Сначала двухдиагональную матрицу

$$\mathbf{B} := \begin{pmatrix} p_1 & q_2 & 0 & \cdots & 0 \\ 0 & p_2 & q_3 & \cdots & 0 \\ 0 & 0 & p_3 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & p_n \end{pmatrix}$$

---

* Если определить матрицу $\mathbf{B}_1$ равенством $\mathbf{B}_1 := \mathbf{S}^T\mathbf{B}\mathbf{T}$, получим аналогичное соотношение подобия $\mathbf{B}_1^T\mathbf{B}_1 = \mathbf{T}^T\left(\mathbf{B}^T\mathbf{B}\right)\mathbf{T}$. Также подобны между собой матрицы $\mathbf{B}_1\mathbf{B}_1^T$ и $\mathbf{B}\mathbf{B}^T$.

умножают справа на матрицу плоских вращений

$$\mathbf{T}_2^T := \begin{pmatrix} \cos\varphi_2 & -\sin\varphi_2 & 0 & \cdots & 0 \\ \sin\varphi_2 & \cos\varphi_2 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \qquad (6.12)$$

угол φ_2 в которой определим позже (см. формулу (6.20)). Результат умножения $\mathbf{B}$ на $\mathbf{T}_2^T$ — это матрица

$$\mathbf{BT}_2^T := \begin{pmatrix} p_1\cos\varphi_2 + q_2\sin\varphi_2 & q_2\cos\varphi_2 - p_1\sin\varphi_2 & 0 & \cdots & 0 \\ p_2\sin\varphi_2 & p_2\cos\varphi_2 & q_3 & \cdots & 0 \\ 0 & 0 & p_3 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & p_n \end{pmatrix}$$

с одним ненулевым поддиагональным элементом.

Затем делаем преобразование Гивенса, умножая матрицу $\mathbf{BT}_2^T$ слева на матрицу вращений

$$\mathbf{S}_2 := \begin{pmatrix} \cos\theta_2 & \sin\theta_2 & 0 & \cdots & 0 \\ -\sin\theta_2 & \cos\theta_2 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \qquad (6.13)$$

где угол θ_2 подбирают так, чтобы оказался нулевым поддиагональный элемент матрицы $\mathbf{S}_2\mathbf{BT}_2^T$, имеющей вид

$$\begin{pmatrix} d_1\cos\theta_2 + h_2\sin\theta_2 & h_1\cos\theta_2 + d_2\sin\theta_2 & q_3\sin\theta_2 & 0 & \cdots & 0 \\ h_2\cos\theta_2 - d_1\sin\theta_2 & d_2\cos\theta_2 - h_1\sin\theta_2 & q_3\cos\theta_2 & 0 & \cdots & 0 \\ 0 & 0 & p_3 & q_4 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & p_n \end{pmatrix}.$$

В ней использованы следующие обозначения элементов верхнего главного 2×2-минора матрицы $\mathbf{BT}_2^T$:

$$d_1 := p_1\cos\varphi_2 + q_2\sin\varphi_2, \quad h_1 := q_2\cos\varphi_2 - p_1\sin\varphi_2,$$
$$h_2 := p_2\sin\varphi_2, \quad d_2 := p_2\cos\varphi_2 .$$

Поддиагональный элемент $h_2\cos\theta_2 - d_1\sin\theta_2$ аннулируем выбором угла θ_2 такого, что

$$\operatorname{tg}\theta_2 = \frac{h_2}{d_1},$$

т.е. полагая

$$\cos\theta_2 = \frac{1}{\sqrt{1+\operatorname{tg}^2\theta_2}} := \frac{d_1}{\sqrt{d_1^2+h_2^2}},$$
$$\sin\theta_2 = \frac{\operatorname{tg}\theta_2}{\sqrt{1+\operatorname{tg}^2\theta_2}} := \frac{h_2}{\sqrt{d_1^2+h_2^2}}. \qquad (6.14)$$

Однако двухдиагональная структура матрицы $\mathbf{S}_2\mathbf{BT}_2^T$ еще не восстановлена, так как в ней присутствует ненулевой элемент $q_3\sin\theta_2$ на в т о р о й наддиагонали. Этот элемент уничтожается применением к этой матрице правого преобразования Гивенса, определяемого матрицей

$$\mathbf{T}_3^T := \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cos\varphi_3 & -\sin\varphi_3 & 0 & \cdots & 0 \\ 0 & \sin\varphi_3 & \cos\varphi_3 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Подобрав угол φ_3 так, чтобы появившийся в матрице $\left(\mathbf{S}_2\mathbf{BT}_2^T\right)\mathbf{T}_3^T$ на позиции $(1,3)$ элемент стал нулевым, т.е. из равенства $\left(q_3\sin\theta_2\right)\cos\varphi_3 - \left(h_1\cos\theta_2 + d_2\sin\theta_2\right)\sin\varphi_3 = 0$, имеем ненулевой поддиагональный элемент на позиции $(3,2)$. Он удаляется с помощью левого преобразования Гивенса $\mathbf{S}_3$ и т.д.

Такой процесс, когда ненулевой элемент перегоняется из первой поддиагонали в первую наддиагональ и обратно с последовательным смещением его на одну позицию вдоль главной диагонали к правому нижнему углу, называется ***процессом преследования***.

Описанный процесс преследования завершается после применения $(n-1)$-го левого преобразования Гивенса с матрицей $\mathbf{S}_n$, условно говоря, вытесняющего ненулевой элемент за пределы матрицы вправо. Получающуюся при этом двухдиагональную матрицу

$$\mathbf{B}_1 := \mathbf{S}_n \ldots \mathbf{S}_3 \mathbf{S}_2 \mathbf{B} \mathbf{T}_2^T \mathbf{T}_3^T \ldots \mathbf{T}_n^T \tag{6.15}$$

(ср. с (6.11)) принимают за результат первой итерации **QR**-алгоритма, т.е. далее следует выполнять такие же действия с матрицей

$$\mathbf{B} := \mathbf{B}_1 . \tag{6.16}$$

Теперь нужно выяснить, как начинать этот процесс итераций, по каким критериям его останавливать и какие ситуации могут встретиться в ходе его реализации.

Ответ на вопрос о начале процесса (6.15) построения матрицы $\mathbf{B}_1$ связан с отложенным ранее требованием к выбору угла φ_2 в матрице $\mathbf{T}_2$ плоских вращений (6.12). Пользуясь свободой задания φ_2, этот параметр на каждой итерации подбирают так, чтобы строящийся процесс был эквивалентен **QR**-алгоритму с неявными сдвигами для нахождения собственных чисел матрицы

$$\mathbf{F} := \mathbf{B}^T \mathbf{B} = \begin{pmatrix} p_1^2 & p_1 q_2 & 0 & \cdots & 0 & 0 \\ p_1 q_2 & p_2^2 + q_2^2 & p_2 q_3 & \cdots & 0 & 0 \\ 0 & p_2 q_3 & p_3^2 + q_3^2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & p_{n-1}^2 + q_{n-1}^2 & p_{n-1} q_n \\ 0 & 0 & 0 & \cdots & p_{n-1} q_n & p_n^2 + q_n^2 \end{pmatrix}. \tag{6.17}$$

Согласно изложенному в § 5.4, эти неявные сдвиги τ целесообразно брать равными собственным числам 2×2-матрицы, стоящей

в правом нижнем углу матрицы (6.17), т.е. находить их из характеристического уравнения

$$\begin{vmatrix} p_{n-1}^2 + q_{n-1}^2 - \tau & p_{n-1}q_n \\ p_{n-1}q_n & p_n^2 + q_n^2 - \tau \end{vmatrix} = 0 . \qquad (6.18)$$

Решая уравнение (6.18), приходим к выражению (выкладки опущены)

$$\tau := p_n^2 + q_n^2 + q_n p_{n-1}\left(f \pm g\right), \qquad (6.19)$$

где применены обозначения

$$f := \frac{q_{n-1}^2 - q_n^2 + p_{n-1}^2 - p_n^2}{2q_n p_{n-1}}, \quad g := \sqrt{1+f^2} .$$

Благодаря симметричности матрицы $\mathbf{B}^T\mathbf{B}$ и соответственно матрицы $\mathbf{F}-\tau\mathbf{E}$, а также наличию всего одного ненулевого внедиагонального элемента в первой строке и в первом столбце матрицы $\mathbf{B}\mathbf{T}_2^T$, можно одним ортогональным преобразованием поворота посредством матрицы $\mathbf{T}_2$ совершить требуемый **QR**-алгоритмом переход к новой матрице, характеризуемой сдвигом (одинарным!) на величину τ. Для этого, согласно упоминавшейся в § 5.4 теореме единственности, достаточно матрицу $\mathbf{T}_2^T$ зафиксировать так, чтобы ее первый столбец был пропорционален первому столбцу матрицы $\mathbf{F}-\tau\mathbf{E}$. Следовательно, для угла поворота φ_2 можно записать условие

$$\frac{p_1^2 - \tau}{\cos\varphi_2} = \frac{p_1 q_2}{\sin\varphi_2},$$

приводящее к значению

$$\operatorname{ctg}\varphi_2 = \frac{p_1^2 - \tau}{p_1 q_2}. \qquad (6.20)$$

Из (6.20) получаем нужные для формирования матрицы вращений (6.12) элементы

$$\sin\varphi_2 := \frac{1}{\sqrt{1+\operatorname{ctg}^2\varphi_2}}, \quad \cos\varphi_2 := \operatorname{ctg}\varphi_2 \sin\varphi_2 . \qquad (6.21)$$

Таким образом, процесс преобразований двухдиагональной матрицы $\mathbf{B}$ к матрице диагональной (в пределе) структуры начинаем с построения матрицы $\mathbf{T}_2^T$ вида (6.12), привлекая для этого последовательно формулы (6.19), (6.20), (6.21). Далее строим матрицу $\mathbf{S}_2$ вида (6.13) с элементами, вычисляемыми по формуле (6.14), затем матрицу $\mathbf{T}_3^T$ (ее элементы читатель легко найдет сам, пользуясь требованиями к $\mathbf{T}_3^T$ и образцом построения $\mathbf{S}_2$) и т.д.

§ 6.4. ПОНИЖЕНИЕ РАЗМЕРНОСТИ, СБОРКА РЕЗУЛЬТИРУЮЩИХ МАТРИЦ SVD-РАЗЛОЖЕНИЯ

Различают три ситуации, которые могут иметь место при реализации итерационного процесса (6.15), (6.16) с точностью, определяемой некоторым малым числом $\varepsilon > 0$. Поскольку процесс имеет кубическую скорость сходимости [68], такие ситуации обычно обнаруживаются достаточно быстро (возможно, за 2–3 полных итерации).

Ситуация 1. В матрице $\mathbf{B} := \mathbf{B}_1$ элемент q_n, стоящий над последним элементом p_n диагонали, существенно мал: $|q_n| < \varepsilon$. Это означает, что можно считать найденным n-е сингулярное число, т.е. положить*

$$\sigma_n := |p_n| .$$

Далее включается ***процедура исчерпывания***: n заменяют на $n-1$ и снова производят преобразования по тем же формулам с матрицами меньшего размера.

Ситуация 2. В k-м столбце матрицы $\mathbf{B}$ при $k < n$ над диагональю находится практически нулевой элемент, т.е. $|q_k| < \varepsilon$.

---

* Рассмотренное выше построение ортогональной матрицы $\mathbf{T}_2^T$ по первому столбцу матрицы $\mathbf{B}^T\mathbf{B} - \tau\mathbf{E}$ в силу неполной однозначности, отмечавшейся в § 5.4, допускает появление в результате описанной процедуры чисел, отличающихся знаком от искомых сингулярных чисел. Отсюда — появление знака абсолютной величины.

Приняв $q_k := 0$, приходим к выводу, что матрица **B** является разложимой. Следовательно, процесс сингулярного разложения можно продолжать с двумя матрицами размеров $(k-1)\times(k-1)$ и $(n-k+1)\times(n-k+1)$, диагонали которых совпадают с соответствующими частями диагонали матрицы **B**.

Ситуация 3. Если k-й диагональный элемент матрицы **B** пренебрежимо мал ($|p_k| < \varepsilon$), то это означает, что данная матрица имеет нуль своим сингулярным числом. В таком случае, если $k < n$, матрица **B** должна быть подвергнута дополнительным преобразованиям, цель которых — подготовить матрицу к разложению, чтобы можно было дальше продолжить процесс с матрицами меньших размеров.

Сначала двухдиагональная матрица **B** умножается слева на матрицу плоских вращений с активными элементами на пересечении k-х и $(k+1)$-х строк и столбцов:

$$
\begin{array}{c}
\begin{matrix} & & (k) & (k+1) & & \\ & & \downarrow & \downarrow & & \end{matrix} \\
\begin{matrix} (k) \to \\ (k+1) \to \end{matrix}
\begin{pmatrix}
1 & & & & & 0 \\
& \ddots & & & & \\
& & c & -s & & \\
& & s & c & & \\
& & & & \ddots & \\
0 & & & & & 1
\end{pmatrix}
\end{array}
\cdot
\begin{array}{c}
\begin{matrix} (k) & (k+1) & (k+2) & & \\ \downarrow & \downarrow & \downarrow & & \end{matrix} \\
\begin{pmatrix}
p_1 & q_2 & & & 0 \\
& \ddots & & & \\
& p_k & q_{k+1} & & \\
& & p_{k+1} & q_{k+2} & \\
& & & \ddots & q_n \\
0 & & & & p_n
\end{pmatrix}
\end{array}
=
$$

$$
=
\begin{array}{c}
\begin{pmatrix}
p_1 & q_2 & & & & 0 \\
& \ddots & & & & \\
& & cp_k & cq_{k+1} - sp_{k+1} & -sq_{k+2} & \\
& & sp_k & sq_{k+1} + cp_{k+1} & cq_{k+2} & \\
& & & \ddots & & q_n \\
0 & & & & & p_n
\end{pmatrix}
\begin{matrix} \\ \\ \leftarrow (k) \\ \leftarrow (k+1) \\ \\ \\ \end{matrix}
\\
\begin{matrix} & & \uparrow & \uparrow & \uparrow & \\ & & (k) & (k+1) & (k+2) & \end{matrix}
\end{array}
.
$$

Элементы c и s здесь выбираем так, чтобы аннулировать элемент результата, стоящий на позиции $(k,\ k+1)$. Из равенств

$$cq_{k+1} - sp_{k+1} = 0, \quad c^2 + s^2 = 1$$

находим соответствующие значения c и s:

$$c := \frac{p_{k+1}}{\sqrt{p_{k+1}^2 + q_{k+1}^2}}, \quad s := \frac{q_{k+1}}{\sqrt{p_{k+1}^2 + q_{k+1}^2}}.$$

Полученную матрицу, в которой для неизменяющихся далее (кроме $p_{k(1)}$) элементов введем обозначения

$$p_{k(1)} := cp_k, \quad \delta_{k+1} := sp_k, \quad \tilde{p}_{k+1} := sq_{k+1} + cp_{k+1}, \quad \tilde{q}_{k+2} := cq_{k+2},$$

опять умножим слева на матрицу поворота в плоскости, определяемой теперь k-ми и $(k+2)$-ми строками и столбцами. Покажем лишь «рабочие» фрагменты матриц, участвующих в этом преобразовании:

$$\begin{array}{c} \\ (k) \\ (k+1) \\ (k+2) \end{array}\!\!\begin{array}{c} \begin{array}{ccccc} (k) & (k+1) & (k+2) & & \end{array} \\ \begin{pmatrix} \vdots & \vdots & \vdots & & \\ \cdots & c_1 & 0 & -s_1 & \cdots \\ \cdots & 0 & 1 & 0 & \cdots \\ \cdots & s_1 & 0 & c_1 & \cdots \\ & \vdots & \vdots & \vdots & \end{pmatrix} \end{array} \cdot \begin{array}{c} \begin{array}{ccccc} (k) & (k+1) & (k+2) & (k+3) & \end{array} \\ \begin{pmatrix} \vdots & \vdots & \vdots & \vdots & \\ \cdots & p_{k(1)} & 0 & -sq_{k+2} & 0 & \cdots \\ \cdots & \delta_{k+1} & \tilde{p}_{k+1} & \tilde{q}_{k+2} & 0 & \cdots \\ \cdots & 0 & 0 & p_{k+2} & q_{k+3} & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \end{array} =$$

$$= \begin{pmatrix} \vdots & \vdots & \vdots & \vdots & \\ \cdots & c_1 p_{k(1)} & 0 & -s_1 p_{k+2} - c_1 sq_{k+2} & -s_1 q_{k+3} & \cdots \\ \cdots & \delta_{k+1} & \tilde{p}_{k+1} & \tilde{q}_{k+2} & 0 & \cdots \\ \cdots & s_1 p_{k(1)} & 0 & c_1 p_{k+2} - s_1 sq_{k+2} & c_1 q_{k+3} & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \begin{array}{c} \\ (k) \\ (k+1) \\ (k+2) \\ \\ \end{array}.$$

$$\begin{array}{cccc} (k) & (k+1) & (k+2) & (k+3) \end{array}$$

Подставим в последнюю матрицу значения

$$c_1 := \frac{p_{k+2}}{\sqrt{p_{k+2}^2 + s^2 q_{k+2}^2}}, \quad s_1 := -\frac{sq_{k+2}}{\sqrt{p_{k+2}^2 + s^2 q_{k+2}^2}}$$

и сделаем обозначения

$$p_{k(2)} := c_1 p_{k(1)}, \quad \delta_{k+2} := s_1 p_{k(1)},$$

$$\tilde{p}_{k+2} := c_1 q_{k+2} - s_1 s p_{k+2}, \quad \tilde{q}_{k+3} := c_1 q_{k+3}.$$

Тогда она (фрагмент) примет вид

$$\begin{array}{c} \\ (k) \\ (k+1) \\ (k+2) \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccc} (k) & (k+1) & (k+2) & (k+3) & \end{array} \\ \begin{pmatrix} \vdots & \vdots & \vdots & \vdots & \\ \cdots & p_{k(2)} & 0 & -s_1 q_{k+3} & \cdots \\ \cdots & \delta_{k+1} & \tilde{p}_{k+1} & \tilde{q}_{k+2} & 0 & \cdots \\ \cdots & \delta_{k+2} & 0 & \tilde{p}_{k+2} & \tilde{q}_{k+3} & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \end{array},$$

отражающий появление еще одного нуля в k-й строке справа от диагонали (и еще одного ненулевого элемента в k-м столбце).

В результате $n-k$ таких преобразований Гивенса правый нижний блок матрицы будет выглядеть следующим образом:

$$\begin{pmatrix} \tilde{p}_k & 0 & 0 & 0 & \cdots & 0 \\ \delta_{k+1} & \tilde{p}_{k+1} & \tilde{q}_{k+2} & 0 & \cdots & 0 \\ \delta_{k+2} & 0 & \tilde{p}_{k+2} & \tilde{q}_{k+3} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \delta_{n-1} & 0 & 0 & 0 & \cdots & \tilde{q}_n \\ \delta_n & 0 & 0 & 0 & \cdots & \tilde{p}_n \end{pmatrix}.$$

Приведем два довода в пользу того, что первую строку в этом блоке (а значит, k-ю строку в преобразованной матрице $\mathbf{B}$) можно считать нулевой. Во-первых, в ней единственный могущий быть отличным от нуля элемент $\tilde{p}_k$ $(:= p_{k(n-k)})$ — это продукт последовательного умножения существенно малого по условию числа p_k на значения косинусов c, c_1, ..., которые могут только уменьшить значение $|p_k|$. Во-вторых, ортогональные преобразования, сохраняя евклидову норму столбца матрицы, гарантируют справедливость равенства

$$\tilde{p}_k^2 + \delta_{k+1}^2 + \ldots + \delta_n^2 = p_k^2,$$

из чего становится ясным, что если $|p_k| < \varepsilon$, то тем более $|\tilde{p}_k| < \varepsilon$ и $|\delta_i| < \varepsilon$ (при всех $i \in \{k+1, \ldots, n\}$).

Таким образом, в любой из рассмотренных ситуаций итерационный процесс завершается получением диагональной (с некоторой точностью) или блочно-диагональной матрицы такой, что диагональная матрица из модулей элементов этой матрицы может быть принята за искомую матрицу $\mathbf{\Sigma}$ определенного равенством (6.1) SVD-разложения. Остается выполнить правильную сборку всех матриц левых и правых ортогональных преобразований для фиксирования матриц $\mathbf{U}$ и $\mathbf{V}$ в разложении (6.1).

Вспомним, что на первом этапе (двухдиагонализации) с помощью преобразований Хаусхолдера были получены представления (см. (6.7)–(6.9)):

$$\mathbf{A} = \mathbf{P}\mathbf{B}\mathbf{Q}^T, \quad \text{если } m = n, \tag{6.22}$$

$$\mathbf{A} = \mathbf{P}\begin{pmatrix}\mathbf{B}\\ \mathbf{0}\end{pmatrix}\mathbf{Q}^T, \quad \text{если } m > n, \tag{6.23}$$

и

$$\mathbf{A} = \mathbf{P}\left(\mathbf{B}\middle|\,\mathbf{0}\right)\mathbf{Q}^T, \text{ если } m < n. \tag{6.24}$$

Обозначим

$$\mathbf{S} := \mathbf{S}_n^{(l)} \ldots \mathbf{S}_3^{(l)}\mathbf{S}_2^{(l)}, \quad \mathbf{T} := \mathbf{T}_n^{(l)} \ldots \mathbf{T}_3^{(l)}\mathbf{T}_2^{(l)}. \tag{6.25}$$

Эти ортогональные матрицы $\mathbf{S}$ и $\mathbf{T}$ являются окончательными результатами последовательных применений соответственно левых и правых преобразований вращения в описанном выше процессе преследования и в основном итерационном процессе приведения матрицы $\mathbf{B}$ к диагональной форме. Значение n здесь уменьшается в соответствии с процедурой исчерпывания и возможными распадениями матриц, а $l \in \{1, 2, \ldots\}$ — заранее неизвестные числа фактически осуществляемых шагов основного алгоритма. Тогда, согласно рассмотренному процессу приведения, можно записать

$$\tilde{\mathbf{\Sigma}} = \mathbf{S}\mathbf{B}\mathbf{T}^T, \quad \text{т.е.} \quad \mathbf{B} = \mathbf{S}^T\tilde{\mathbf{\Sigma}}\mathbf{T}, \tag{6.26}$$

где $\tilde{\mathbf{\Sigma}} := \operatorname{diag}(\pm\sigma_1; \pm\sigma_2; \ldots; \pm\sigma_n)$.

Совмещая (6.22) и (6.26) при $m = n$, имеем:

$$\mathbf{A} = \mathbf{P}\mathbf{S}^T\tilde{\mathbf{\Sigma}}\mathbf{T}\mathbf{Q}^T \tag{6.27}$$

— разложение типа (6.1), если принять

$$\mathbf{U} := \mathbf{P}\mathbf{S}^T, \quad \mathbf{V} := \mathbf{T}\mathbf{Q}^T. \tag{6.28}$$

Заметим, что, *превращая полученную в* (6.26) *диагональную матрицу* $\tilde{\mathbf{\Sigma}}$ *в искомую матрицу* $\mathbf{\Sigma}$ *заменой отрицательных элементов положительными, следует изменить знаки элементов у соответствующих строк матрицы* $\mathbf{T}$, *а значит и* $\mathbf{V}$.

В случаях $m > n$ и $m < n$ в соответствии с представлениями (6.23) и (6.24) матрица $\tilde{\mathbf{\Sigma}}$ в (6.27) должна быть дополнена нулевыми строками или, соответственно, столбцами, а матрица $\mathbf{S}^T$ или матрица $\mathbf{T}$ в (6.28) расширяются снизу соответствующими фрагментами единичной матрицы до требуемых равенством (6.1) размеров.

Пример 6.5 (продолжение примера 6.3).

Сначала покажем округленные результаты промежуточных шагов на первой итерации **QR**-алгоритма с неявными сдвигами, применяемого для сингулярного разложения квадратной двухдиагональной матрицы $\mathbf{B}$.

Возьмем в качестве матрицы $\mathbf{B}$ полученную в итоге преобразований в примере 6.3 матрицу $\mathbf{A}_3$ без последней (нулевой) строки.

Итак, имеем

$$\mathbf{B} := \begin{pmatrix} -3 & -2{,}236 & 0 \\ 0 & 2{,}450 & 2{,}041 \\ 0 & 0 & 1{,}354 \end{pmatrix}, \quad \mathbf{F} := \mathbf{B}^T\mathbf{B} = \begin{pmatrix} 9 & 6{,}708 & 0 \\ 6{,}708 & 11 & 5 \\ 0 & 5 & 6 \end{pmatrix}.$$

По формулам (6.19) при $n := 3$ находим значения*:

$$f = 0{,}5, \quad g = -1{,}118, \quad \tau = 14{,}09.$$

С помощью формул (6.20) и (6.21) составляем матрицу вращения (6.22)

---

* При подсчете значения τ в формуле (6.19) взят знак «+». Использование здесь знака «–» изменит промежуточные, но не окончательные результаты. Имеются ли аргументированные предпочтения при выборе знака для определения величины сдвига τ, автору неизвестно.

$$\mathbf{T}_2^T := \begin{pmatrix} -0,604 & -0,797 & 0 \\ 0,797 & -0,604 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{и вычисляем} \quad \mathbf{BT}_2^T = \begin{pmatrix} 0,032 & 3,741 & 0 \\ 1,951 & -1,480 & 2,041 \\ 0 & 0 & 1,354 \end{pmatrix}.$$

Далее, привлекая формулы (6.14), строим матрицу (6.13)

$$\mathbf{S}_2 := \begin{pmatrix} 0,016 & 0,999 & 0 \\ -0,999 & 0,016 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{и вычисляем} \quad \mathbf{S}_2\mathbf{BT}_2^T = \begin{pmatrix} 1,951 & -1,419 & 2,041 \\ 0 & -3,765 & 0,033 \\ 0 & 0 & 1,354 \end{pmatrix}.$$

Продолжая процесс преследования, подсчитываем:

$$\mathbf{T}_3^T := \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0,571 & -0,821 \\ 0 & 0,821 & -0,571 \end{pmatrix}, \quad \left(\mathbf{S}_2\mathbf{BT}_2^T\right)\mathbf{T}_3^T = \begin{pmatrix} 1,951 & 2,486 & 0 \\ 0 & 2,177 & 3,072 \\ 0 & 1,112 & -0,773 \end{pmatrix};$$

$$\mathbf{S}_3 := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0,890 & 0,455 \\ 0 & -0,455 & 0,890 \end{pmatrix}, \quad \mathbf{B}_1 := \mathbf{S}_3\mathbf{S}_2\mathbf{BT}_2^T\mathbf{T}_3^T = \begin{pmatrix} 1,951 & 2,486 & 0 \\ 0 & 2,444 & 2,385 \\ 0 & 0 & -2,086 \end{pmatrix}.$$

На второй итерации аналогичные действия совершаются над матрицей $\mathbf{B} := \mathbf{B}_1$, в результате чего приходим к матрице

$$\mathbf{B}_1 := \begin{pmatrix} -1,289 & -1,741 & 0 \\ 0 & -1,868 & -0,860 \\ 0 & 0 & -4,133 \end{pmatrix},$$

и т.д.

Итогами пятой и шестой итераций служат соответственно матрицы

$$\begin{pmatrix} -0,873 & 0,003 & 0 \\ 0 & 4,213 & -0,423 \\ 0 & 0 & 2,704 \end{pmatrix} \quad \text{и} \quad \begin{pmatrix} -0,873 & 0.000 & 0 \\ 0 & 2,681 & 0,000 \\ 0 & 0 & 4,248 \end{pmatrix}.$$

Последняя матрица может быть принята за матрицу $\tilde{\mathbf{\Sigma}}$ в равенстве (6.26), а фигурирующие в нем ортогональные матрицы $\mathbf{S}^T$ и $\mathbf{T}$ — результаты последовательных левых и правых преобразований вращения — в данном случае таковы:

$$\mathbf{S}^T := \begin{pmatrix} 0,148 & 0,581 & 0,800 \\ 0,357 & 0,723 & -0,591 \\ -0,922 & 0,373 & -0,101 \end{pmatrix}, \quad \mathbf{T} := \begin{pmatrix} 0,508 & -0,623 & 0,595 \\ -0,650 & 0,176 & 0,739 \\ -0,565 & 0,762 & -0,316 \end{pmatrix}.$$

Следовательно, сингулярные числа матрицы $\mathbf{B}$ — это числа

$$\sigma_1 := \approx 4{,}248\,, \quad \sigma_2 := \approx 2{,}681\,, \quad \sigma_3 := \approx 0{,}873\,.$$

Таким образом, имеет место следующее представление полученной в примере 6.3 матрицы $\mathbf{A}_3$:

$$\mathbf{A}_3 = \begin{pmatrix} 0{,}148 & 0{,}581 & 0{,}800 & 0 \\ 0{,}357 & 0{,}723 & -0{,}591 & 0 \\ -0{,}922 & 0{,}373 & -0{,}101 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0{,}873 & 0 & 0 \\ 0 & 2{,}681 & 0 \\ 0 & 0 & 4{,}248 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} -0{,}508 & 0{,}623 & -0{,}595 \\ -0{,}650 & 0{,}176 & 0{,}739 \\ -0{,}565 & 0{,}762 & -0{,}316 \end{pmatrix}.$$

Обращаясь теперь к результатам первого этапа, т.е. используя найденное в примере 6.3 разложение $\mathbf{A} = \mathbf{P}\mathbf{A}_3\mathbf{Q}^T$ заданной там матрицы $\mathbf{A}$, приходим к искомому сингулярному разложению

$$\begin{pmatrix} 2 & -1 & -2 \\ 0 & 3 & 1 \\ -1 & 1 & 0 \\ 2 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 0{,}433 & -0{,}339 & -0{,}667 & -0{,}503 \\ -0{,}072 & -0{,}763 & 0{,}568 & -0{,}302 \\ 0{,}871 & -0{,}033 & 0{,}280 & 0{,}402 \\ -0{,}219 & -0{,}550 & -0{,}393 & 0{,}704 \end{pmatrix} \times$$

$$\times \begin{pmatrix} 0{,}873 & 0 & 0 \\ 0 & 2{,}681 & 0 \\ 0 & 0 & 4{,}248 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} -0{,}508 & 0{,}253 & -0{,}823 \\ -0{,}650 & -0{,}739 & 0{,}173 \\ -0{,}565 & 0{,}624 & 0{,}540 \end{pmatrix}.$$

Замечание 6.2. В реальных алгоритмах SVD-разложений [23, 68] обычно исходят из того же представления $m \times n$-матрицы $\mathbf{A}$ в виде (6.1), но в этом представлении диагональная матрица $\mathbf{\Sigma}$ — всегда квадратная, а одна из матриц $\mathbf{U}$ и $\mathbf{V}$ может быть неквадратной (в случае $m \neq n$); в такой ситуации, естественно, матрицы $\mathbf{U}$ и $\mathbf{V}$ удовлетворяют равенствам $\mathbf{U}^T\mathbf{U} = \mathbf{V}\mathbf{V}^T = \mathbf{E}$. Данный подход к сингулярному разложению можно расценивать как сознательное отсечение его несущественной части, позволяющее сэкономить машинные ресурсы без потери информации, важной для приложений.

Кроме того, *как правило, игнорируется фигурирующее в определении* 6.2 *требование к расположению сингулярных чисел на диагонали матрицы* $\mathbf{\Sigma}$ *по невозрастанию*, что связано с их многократной «перетасовкой» в процессе выполнения предписаний **QR**-алгоритма со сдвигами. Как видим из рассмотренного выше примера, здесь тоже нарушается это требование.

УПРАЖНЕНИЯ

6.1. Найдите сингулярные числа матрицы $\mathbf{A} := \begin{pmatrix} 1 & -2 \\ 0 & -1 \\ 3 & 0 \end{pmatrix}$ по определению.

6.2. Пользуясь определением, найдите сингулярные числа и сингулярные векторы матрицы $\mathbf{A} := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$.

6.3. Выполните приведение матрицы $\mathbf{A} := \begin{pmatrix} 1 & 3 \\ -2 & 1 \\ 2 & -1 \end{pmatrix}$ к верхнему двухдиагональному виду:

а) преобразованиями Хаусхолдера;

б) преобразованиями Гивенса.

6.4. Проверьте справедливость формулы (6.19) для определения значения сдвига τ из уравнения (6.18).

6.5. Для матрицы $\mathbf{B} := \begin{pmatrix} 2 & 1 & 0 \\ 0 & 3 & -1 \\ 0 & 0 & 1 \end{pmatrix}$ найдите значение первого сдвига τ, знание которого требуется при выполнении SVD-разложения на основе **QR**-алгоритма с неявными сдвигами:

а) по формуле (6.19) ;

б) непосредственно из уравнения типа (6.18).

Глава 7

ПРИМЕНЕНИЯ СИНГУЛЯРНЫХ РАЗЛОЖЕНИЙ

§ 7.1. РАНГ МАТРИЦЫ, МОДУЛЬ ОПРЕДЕЛИТЕЛЯ, ЧИСЛО ОБУСЛОВЛЕННОСТИ

Предположим, что у вещественной $m \times n$-матрицы $\mathbf{A}$ известно сингулярное разложение

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}. \tag{7.1}$$

Какую из этого можно извлечь полезную информацию? Какие алгебраические задачи можно решить, используя такое специфическое представление прямоугольной в общем случае матрицы?

Отдельные ответы на поставленные вопросы очевидны, другие требуют определенного обсуждения. Рассмотрим несколько возможных точек приложения SVD-разложений (см. также [17, 26, 36, 55, 64 и др.]).

Ранг матрицы. Ранее уже отмечалось, что ***ранг матрицы*** совпадает с числом ее ненулевых сингулярных чисел (это хорошо видно и из дальнейших рассуждений, см. § 7.2). Поэтому, зная, что в мультипликативном представлении (7.1)

$$\boldsymbol{\Sigma} = \operatorname{diag}\left(\sigma_1; \sigma_2; \ldots; \sigma_r; 0; \ldots; 0\right), \tag{7.2}$$

где $\sigma_i > 0$ при всех $i \in \{1, \ldots, r\}$, принимаем

$$\operatorname{rank} \mathbf{A} = r \quad \left(\leq \min\{m, n\}\right).$$

Модуль определителя. Пусть $\mathbf{A} \in \mathbb{R}_{n \times n}$ и матрица $\boldsymbol{\Sigma}$ в разложении (7.1) имеет вид

$$\boldsymbol{\Sigma} = \operatorname{diag}\left(\sigma_1; \sigma_2; \ldots; \sigma_n\right),$$

где $\sigma_i \geq 0$. Тогда для определителя матрицы $\mathbf{A}$ справедливо равенство

$$\left|\det \mathbf{A}\right| = \sigma_1 \sigma_2 \ldots \sigma_n,$$

и если $r < n$, то $\det \mathbf{A} = 0$.

Действительно,

$$|\det \mathbf{A}| = |\det \mathbf{U}| \cdot |\det \mathbf{\Sigma}| \cdot |\det \mathbf{V}| = 1 \cdot |\sigma_1 \sigma_2 \ldots \sigma_n| \cdot 1,$$

поскольку модули определителей ортогональных матриц равны единице (см. доказательство леммы 2.1 в § 2.5).

Число обусловленности. Знание набора сингулярных чисел прямоугольной $m \times n$-матрицы $\mathbf{A}$ позволяет определить для нее ***число обусловленности*** $\operatorname{cond} \mathbf{A}$ равенством [23, 55]

$$\operatorname{cond} \mathbf{A} = \frac{\sigma_1}{\sigma_k},$$

в котором σ_1 — наибольшее, а σ_k — наименьшее *ненулевые* сингулярные числа, т.е. $\operatorname{rank} \mathbf{A} = k$. (Другое определение числа обусловленности для произвольной матрицы см. далее в § 7.3)

Для нормальных матриц, в силу отмечавшейся для них в § 6.1 связи между собственными и сингулярными числами, данное здесь определение $\operatorname{cond} \mathbf{A}$ совпадает с определением числа обусловленности Тодда через модуль отношения собственных чисел — границ спектра [13]. Тем более очевидно такое совпадение $\operatorname{cond} \mathbf{A}$ для симметричных положительно определенных матриц в силу возможности отождествить их собственные и сингулярные числа.

Поскольку сингулярные числа матрицы $\mathbf{A}$ могут интерпретироваться как полуоси гиперэллипсоида $\{\mathbf{A}\mathbf{x} \mid \|\mathbf{x}\|_2 = 1\}$, величина определенного здесь числа $\operatorname{cond} \mathbf{A}$ характеризует степень вытянутости этого гиперэллипсоида, т.е. служит мерой деформации единичного шара матрицей $\mathbf{A}$. Как эта деформация сказывается, например, при решении СЛАУ, показано, в частности, в [13]. Интересно отметить, что сам процесс вычисления собственных и сингулярных значений ортогональными преобразованиями подобия считается прекрасно обусловленным, так как в этих задачах существенную роль играет обусловленность не самой матрицы $\mathbf{A}$, а тех матриц, на которых строится переход к подобной ей диагональной (или треугольной) матрице [64].

Замечание 7.1. Как видно из материала этого параграфа, при использовании SVD-разложений не всегда востребованы все три составляющие этих разложений. Учет этого обстоятельства может уменьшить объем вычислений при решении конкретных частных задач, связанных с привлечением таких разложений.

§ 7.2. РЕШЕНИЕ ОДНОРОДНЫХ И НЕОДНОРОДНЫХ СЛАУ

Сначала рассмотрим однородную систему линейных алгебраических уравнений

$$\mathbf{A}\mathbf{x} = \mathbf{0} \tag{7.3}$$

с $m \times n$-матрицей $\mathbf{A}$ ранга r, представленной сингулярным разложением вида (7.1). В таком случае имеем:

$$\mathbf{A}\mathbf{x} = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{U}\Sigma\mathbf{V}\mathbf{x} = \mathbf{0} \quad \Leftrightarrow \quad \Sigma\mathbf{V}\mathbf{x} = \mathbf{0}. \tag{7.4}$$

Введя вектор $\mathbf{y} := (y_1; y_2; \ldots; y_n)^T$ равенством

$$\mathbf{y} := \mathbf{V}\mathbf{x}, \tag{7.5}$$

последнюю систему в (7.4) с диагональной матрицей (7.2) представляем в виде

$$\begin{cases} \sigma_1 y_1 = 0, \\ \ldots\ldots \\ \sigma_r y_r = 0, \\ 0 y_{r+1} = 0, \\ \ldots\ldots \\ 0 y_n = 0 \end{cases} \tag{7.6}$$

(здесь не записаны $m-n$ равенств $0=0$, если $m > n$). Решением системы (7.6) служит вектор

$$\mathbf{y} := (0; \ldots; 0; c_1; \ldots; c_{n-r})^T,$$

содержащий $n-r$ параметров $c_1, \ldots, c_{n-r}$, могущих принимать произвольные вещественные значения. Следовательно, искомое решение $\mathbf{x}$ данной системы (7.3), согласно (7.5), учитывая ортогональность матрицы $\mathbf{V}$, можно представить так:

$$\mathbf{x} := \mathbf{V}^T\mathbf{y} = \begin{pmatrix} v_{11} & \cdots & v_{r1} & v_{r+1,1} & \cdots & v_{n1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ v_{1r} & \cdots & v_{rr} & v_{r+1,r} & \cdots & v_{nr} \\ v_{1,r+1} & \cdots & v_{r,r+1} & v_{r+1,r+1} & \cdots & v_{n,r+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ v_{1n} & \cdots & v_{rn} & v_{r+1,n} & \cdots & v_{nn} \end{pmatrix} \cdot \begin{pmatrix} 0 \\ \vdots \\ 0 \\ c_1 \\ \vdots \\ c_{n-r} \end{pmatrix} =$$

$$= \begin{pmatrix} 0\cdot v_{11} + & \ldots & + 0\cdot v_{r1} & + c_1 v_{r+1,1} + & \ldots & + c_{n-r} v_{n1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0\cdot v_{1r} + & \ldots & + 0\cdot v_{rr} & + c_1 v_{r+1,r} + & \ldots & + c_{n-r} v_{nr} \\ 0\cdot v_{1,r+1} + & \ldots & +0\cdot v_{r,r+1} & + c_1 v_{r+1,r+1} + & \ldots & +c_{n-r} v_{n,r+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0\cdot v_{1n} + & \ldots & + 0\cdot v_{rn} & + c_1 v_{r+1,n} + & \ldots & + c_{n-r} v_{nn} \end{pmatrix} =$$

$$= 0\cdot\begin{pmatrix} v_{11} \\ \vdots \\ v_{1n} \end{pmatrix} + \ldots + 0\cdot\begin{pmatrix} v_{r1} \\ \vdots \\ v_{rn} \end{pmatrix} + c_1\cdot\begin{pmatrix} v_{r+1,1} \\ \vdots \\ v_{r+1,n} \end{pmatrix} + \ldots + c_{n-r}\cdot\begin{pmatrix} v_{n1} \\ \vdots \\ v_{nn} \end{pmatrix}.$$

Получено ***общее решение однородной системы*** (7.3) в виде линейной комбинации строк правой ортогональной матрицы $\mathbf{V}$ в SVD-разложении (7.1) (иначе, столбцов матрицы $\mathbf{V}^T$, играющих роль правых сингулярных векторов, согласно (6.2)–(6.4)). Такое представление общего решения достаточно информативно. Оно позволяет решать вопросы о существовании нетривиального решения, наличии линейно зависимых уравнений в системе (7.3), базисе ядра матричного оператора A и т.п. [23, 55].

Заметим, что в полученном с помощью сингулярного разложения представлении общего решения не участвуют сами сингулярные числа. Однако при этом важно четко знать количество *ненулевых* сингулярных чисел и их позиции в диагональной матрице $\mathbf{\Sigma}$ (применение сдвигов в процессе SVD-разложения, ускоряющее этот процесс, не исключает возможность нарушения естественного порядка в появлении этих чисел на диагонали формируемой матрицы $\mathbf{\Sigma}$).

Обобщим систему (7.3) введением в правую часть вектора $\mathbf{b}$:

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{7.7}$$

Тогда аналогом (7.4) будут служить преобразования

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \Leftrightarrow \quad \mathbf{U}\Sigma\mathbf{V}\mathbf{x} = \mathbf{b} \quad \Leftrightarrow \quad \Sigma\mathbf{y} = \mathbf{z},$$

где $\mathbf{y}$ — определенный уже вектор (7.5), а

$$\mathbf{z} := \mathbf{U}^T\mathbf{b} = \left(z_1; z_2; \ldots; z_m\right)^T.$$

В развернутом виде эта преобразованная к новым переменным система выглядит так:

$$\begin{cases} \sigma_1 y_1 = z_1, \\ \cdots\cdots \\ \sigma_r y_r = z_r, \\ 0 y_{r+1} = z_{r+1}, \\ \cdots\cdots \\ 0 y_n = z_n. \end{cases} \tag{7.8}$$

Подобная запись соответствует случаю, когда $m = n$. Если $m > n$, то здесь добавляются $m - n$ равенств

$$0 = z_{n+1}, \quad \ldots, \quad 0 = z_m,$$

а если $m < n$, то последнее в (7.8) равенство заменяется равенством $0 y_m = z_m$ (переменные $y_{m+1}, \ldots, y_n$ при этом оказываются заведомо свободными).

Ясно, что в отличие от системы (7.6), являющейся частным случаем системы (7.8) при $\mathbf{b} := \mathbf{0}$, система (7.8) разрешима не всегда. Рассмотрим следующие ситуации (для определенности в записях считаем $m \geq n$).

1) Ранг матрицы $\mathbf{A}$ равен n. Условие $r = n$ означает, что в системе (7.8) не может быть нулевых множителей при неизвестных y_i ни при каких значениях $i \in \{1, 2, \ldots, n\}$ и, следовательно, все компоненты вектора $\mathbf{y}$ можно однозначно найти по формуле

$$y_i = \frac{z_i}{\sigma_i} \quad \forall\, i \in \{1, 2, \ldots, n\}. \tag{7.9}$$

Тогда вектор $\mathbf{x} := \mathbf{V}^T\mathbf{y}$ с найденным посредством формулы (7.9) вектором $\mathbf{y}$ — единственное решение данной системы (7.7).

2) $\operatorname{rank}\mathbf{A} = r < n$. Первые r компонент вектора $\mathbf{y}$ в этом случае, как и в предыдущем, можно найти по формуле (7.9) из первых r равенств системы (7.8). Насчет остальных $n-r$ равенств этой системы сделаем два предположения.

А) Пусть $z_{r+1} = \ldots = z_m = 0$. В этом случае значения переменных $y_{r+1}, \ldots, y_n$ могут быть любыми, и, значит, вектор $\mathbf{y}$ — решение системы (7.8) — имеет вид

$$\mathbf{y} = \left(\frac{z_1}{\sigma_1}; \ldots; \frac{z_r}{\sigma_r}; c_1; \ldots; c_{n-r} \right)^T. \qquad (7.10)$$

Следовательно, решение исходной неоднородной системы (7.7) (обобщающее решение однородной системы (7.3)) — это вектор

$$\mathbf{x} = \mathbf{V}^T\mathbf{y} = \frac{z_1}{\sigma_1}\begin{pmatrix} v_{11} \\ \vdots \\ v_{1n} \end{pmatrix} + \ldots + \frac{z_r}{\sigma_r}\begin{pmatrix} v_{r1} \\ \vdots \\ v_{rn} \end{pmatrix} + c_1\begin{pmatrix} v_{r+1,1} \\ \vdots \\ v_{r+1,n} \end{pmatrix} + \ldots + c_{n-r}\begin{pmatrix} v_{n1} \\ \vdots \\ v_{nn} \end{pmatrix}. \qquad (7.11)$$

Полученный результат соответствует известному принципу суперпозиции: ***общее решение неоднородной линейной системы складывается из общего решения однородной системы и частного решения неоднородной системы*** (за последнее принимают вектор (7.11) с нулевыми значениями параметров $c_1, \ldots, c_{n-r}$, см. далее формулу (7.12)).

Б) Пусть среди чисел $z_{r+1}, \ldots, z_m$ есть хотя бы одно ненулевое. В таком случае система (7.8) противоречива, а значит, данная неоднородная система (7.7) не имеет классического решения. Однако может оказаться полезным вычисление ***псевдорешения*** данной системы, что бывает особенно естественным при $m > n$. Оно может быть получено путем преобразования системы транспонированной матрицей*. Сделаем такое преобразование с систе-

---

* Именно, псевдорешение системы $\mathbf{A}\mathbf{y} = \mathbf{f}$ — это решение симметричной системы $\mathbf{A}^T\mathbf{A}\mathbf{y} = \mathbf{A}^T\mathbf{f}$.

мой (7.8), т.е. будем вместо нее рассматривать систему

$$\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{y} = \mathbf{\Sigma}^T\mathbf{z}\,.$$

Эта система при указанных предположениях относительно элементов матрицы $\mathbf{\Sigma}$ и вектора $\mathbf{z}$ имеет вид

$$\begin{cases} \sigma_1^2 y_1 = \sigma_1 z_1, \\ \ldots\ldots \\ \sigma_r^2 y_r = \sigma_r z_r, \\ 0\,y_{r+1} = 0, \\ \ldots\ldots \\ 0\,y_n = 0, \end{cases}$$

из чего заключаем, что ее решение совпадает с вектором (7.10) и приводит к псевдорешению исходной системы (7.7) в форме (7.11). ***Нормальное псевдорешение*** данной системы, т.е. *псевдорешение с наименьшей евклидовой нормой*, получается подстановкой в (7.11) нулевых значений произвольных постоянных $c_1, \ldots, c_{n-r}$:

$$\mathbf{x}_0 := \frac{z_1}{\sigma_1}\begin{pmatrix} v_{11} \\ \vdots \\ v_{1n} \end{pmatrix} + \ldots + \frac{z_r}{\sigma_r}\begin{pmatrix} v_{r1} \\ \vdots \\ v_{rn} \end{pmatrix}. \qquad (7.12)$$

Очевидно равенство $\mathbf{x}_0 = \mathbf{V}^T\mathbf{y}_0$, где $\mathbf{y}_0 := \left(\frac{z_1}{\sigma_1}; \ldots; \frac{z_r}{\sigma_r}; 0; \ldots; 0\right)^T$.

Пример 7.1. Согласно результатам примера 6.4, 4×3-матрица $\mathbf{A}$ системы 6.10 (см. пример 6.3) имеет полный ранг: $r = n = 3$. Следовательно, решение этой системы (оно же и нормальное псевдорешение) можно найти по формулам (7.9) (или (7.12) при $r := n$).

Привлекая итоговое равенство примера 6.4 — реализованное разложение $\mathbf{A} = \mathbf{U\Sigma V}$, получим решение системы (6.10). Имеем:

$$\mathbf{z} := \mathbf{U}^{\mathbf{T}}\mathbf{b} \approx \begin{pmatrix} 0{,}433 & -0{,}072 & 0{,}871 & -0{,}219 \\ -0{,}339 & -0{,}763 & -0{,}033 & -0{,}550 \\ -0{,}667 & 0{,}568 & 0{,}280 & -0{,}393 \\ -0{,}503 & -0{,}302 & 0{,}402 & 0{,}704 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 5 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 0{,}720 \\ -6{,}176 \\ 0{,}607 \\ -0{,}002 \end{pmatrix};$$

$$\mathbf{y} := \left(\frac{z_i}{\sigma_i}\right)_{i=1}^{3} \approx \begin{pmatrix} 0{,}720/0{,}873 \\ -6{,}176/2{,}681 \\ 0{,}607/4{,}248 \end{pmatrix} \approx \begin{pmatrix} 0{,}825 \\ -2{,}304 \\ 0{,}143 \end{pmatrix};$$

$$\mathbf{x} := \mathbf{V}^{\mathrm{T}}\mathbf{y} \approx \begin{pmatrix} -0{,}508 & -0{,}650 & -0{,}565 \\ 0{,}253 & -0{,}739 & 0{,}624 \\ -0{,}823 & 0{,}173 & 0{,}540 \end{pmatrix} \cdot \begin{pmatrix} 0{,}825 \\ -2{,}304 \\ 0{,}143 \end{pmatrix} \approx \begin{pmatrix} 0{,}998 \\ 2{,}001 \\ -1{,}000 \end{pmatrix}.$$

Обратим внимание на маленькое значение $z_4 \approx -0{,}002$. Если его отождествить с нулем (учитывая его сравнимость с величиной ошибок округления), то полученное значение $\mathbf{x}$ можно считать единственным *решением* системы (6.10). В противном случае найденный вектор следует воспринимать как *нормальное псевдорешение*. Разумеется, если при этом «забыть» факт $r = n$. Анализ значений вектора $\mathbf{z}$ может сыграть существенную роль в трактовке получаемого решения в случае $r < n$.

§ 7.3. ПСЕВДООБРАТНАЯ МАТРИЦА

По данной $m \times n$-матрице построим $n \times m$-матрицу

$$\mathbf{A}^{+} := \mathbf{V}^{T}\boldsymbol{\Sigma}^{+}\mathbf{U}^{T}, \tag{7.13}$$

в которой ортогональные $n \times n$-матрица $\mathbf{V}$ и $m \times m$-матрица $\mathbf{U}$ — это матрицы из сингулярного разложения (7.1), а $n \times m$-матрица $\boldsymbol{\Sigma}^{+}$ определяется через сингулярные числа матрицы $\mathbf{A}$ ранга r равенством

$$\boldsymbol{\Sigma}^{+} := \operatorname{diag}\left(\frac{1}{\sigma_1}; \ldots; \frac{1}{\sigma_r}; 0; \ldots; 0\right).$$

Применим линейное преобразование посредством матрицы $\mathbf{A}^{+}$ к вектору $\mathbf{b} \in \mathbb{R}_m$. Имеем выражение

$$\mathbf{A}^{+}\mathbf{b} = \mathbf{V}^{T}\boldsymbol{\Sigma}^{+}\mathbf{U}^{T}\mathbf{b} = \mathbf{V}^{T}\boldsymbol{\Sigma}^{+}\mathbf{z}, \tag{7.14}$$

где через $\mathbf{z}$, как и в предыдущем параграфе, обозначен m-мерный вектор $\mathbf{U}^{T}\mathbf{b}$. Легко убедиться, что произведение матрицы $\boldsymbol{\Sigma}^{+}$ на вектор $\mathbf{z} := (z_1; z_2; \ldots; z_m)^{T}$ дает вектор

$$\boldsymbol{\Sigma}^{+}\mathbf{z} := \left(\frac{z_1}{\sigma_1}; \ldots; \frac{z_r}{\sigma_r}; 0; \ldots; 0\right)^{T} \in \mathbb{R}_n.$$

Умножив матрицу $\mathbf{V}^T := \left(v_{ij}\right)^T$ на этот вектор $\mathbf{\Sigma}^+\mathbf{z}$, в соответствии с (7.14) находим представление вектора $\mathbf{A}^+\mathbf{b}$ в виде

$$\mathbf{A}^+\mathbf{b} = \frac{z_1}{\sigma_1}\begin{pmatrix} v_{11} \\ \vdots \\ v_{1n} \end{pmatrix} + \ldots + \frac{z_r}{\sigma_r}\begin{pmatrix} v_{r1} \\ \vdots \\ v_{rn} \end{pmatrix}.$$

Сравнив последнее равенство с равенством (7.12), приходим к выводу, что с помощью матрицы (7.13), построенной на основе SVD-разложения (7.1), получено *нормальное псевдорешение* системы (7.7), т.е.

$$\mathbf{x}_0 = \mathbf{A}^+\mathbf{b}.$$

Таким образом, матрица $\mathbf{A}^+$ играет роль обратной матрицы и называется ***псевдообратной матрицей*** для произвольной прямоугольной матрицы $\mathbf{A}$. Нетрудно проверить, что если $m = n = r$, то $\mathbf{A}^+$ — это просто обратная матрица $\mathbf{A}^{-1}$. В общем же случае она совпадает с ***матрицей Мура–Пенроуза*** $\mathbf{X}$, определяемой совокупностью следующих соотношений [18, 23 и др.]:

$$\mathbf{AXA} = \mathbf{A}, \quad \mathbf{XAX} = \mathbf{X}, \quad (\mathbf{AX})^T = \mathbf{AX}, \quad (\mathbf{XA})^T = \mathbf{XA}. \quad (7.15)$$

Зная какое-нибудь ***скелетное разложение*** $m \times n$-матрицы $\mathbf{A}$, т.е. ее представление в виде $\mathbf{A} = \mathbf{BC}$, где $\mathbf{B}$ — $m \times r$-матрица, $\mathbf{C}$ — $r \times n$-матрица, $r := \operatorname{rank}\mathbf{A} = \operatorname{rank}\mathbf{B} = \operatorname{rank}\mathbf{C}$, псевдообратную $n \times m$-матрицу $\mathbf{A}^+$ можно однозначно найти по явной формуле [15]

$$\mathbf{A}^+ := \mathbf{C}^T(\mathbf{CC}^T)^{-1}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T.$$

С помощью псевдообратной матрицы можно дать более широкое определение ***числа обусловленности прямоугольной матрицы*** $\mathbf{A}$ по сравнению с тем, как его определили в § 7.1, а именно [17]:

$$\operatorname{cond}\mathbf{A} := \|\mathbf{A}\| \cdot \|\mathbf{A}^+\|.$$

Это определение позволяет использовать разные матричные нормы и является естественным развитием понятия числа обусловленности квадратной невырожденной матрицы, определенного как

$$\operatorname{cond} \mathbf{A} := \| \mathbf{A} \| \cdot \| \mathbf{A}^{-1} \|$$

(см. далее § 8.3).

Кроме подробно рассмотренного здесь способа построения псевдообратных матриц с использованием сингулярных разложений имеются и другие конструктивные методы решения этой задачи. К таковым можно отнести технически более простой ***метод Гревилля.*** Построение $n \times m$-матрицы $\mathbf{A}^+$ для заданной $m \times n$-матрицы $\mathbf{A}$ этим методом сводится к следующему [20].

Введем обозначения:

$\mathbf{a}_k$ – k-й столбец матрицы $\mathbf{A}$;

$\mathbf{A}_k := (\mathbf{a}_1; \mathbf{a}_2; ...; \mathbf{a}_k)$ – матрица, образованная первыми k столбцами матрицы $\mathbf{A}$;

$\mathbf{b}_k$ – последняя строка в псевдообратной $k \times m$- матрице $\mathbf{A}_k^+$ ($k = 1, 2, ..., n$).

Сначала вычисляется матрица-строка по формуле

$$\mathbf{A}_1^+ := \mathbf{a}_1^+ = \frac{\mathbf{a}_1^T}{\mathbf{a}_1^T \mathbf{a}_1}. \qquad (7.16)$$

При этом, если $\mathbf{a}_1 = \mathbf{0}$, то $\mathbf{A}_1^+ := \mathbf{0}$.

Далее при $k = 2, 3, ..., n$ применяется совокупность рекуррентных формул:

$$\mathbf{A}_k^+ := \begin{pmatrix} \mathbf{B}_k \\ \mathbf{b}_k \end{pmatrix}, \qquad \mathbf{B}_k := \mathbf{A}_{k-1}^+ - \mathbf{d}_k \mathbf{b}_k,$$

$$\mathbf{d}_k := \mathbf{A}_{k-1}^+ \mathbf{a}_k, \qquad \mathbf{c}_k := \mathbf{a}_k - \mathbf{A}_{k-1} \mathbf{d}_k.$$

Фигурирующая в ней вектор-строка $\mathbf{b}_k$ находится с помощью псевдообращения вектора-столбца $\mathbf{c}_k$, т.е. равенством

$$\mathbf{b}_k := \mathbf{c}_k^+, \qquad \text{если} \quad \mathbf{c}_k \neq \mathbf{0}$$

(для этого можно привлечь формулу (7.16)), или подсчитывается по формуле

$$\mathbf{b}_k := \left(1 + \mathbf{d}_k^T \mathbf{d}_k\right)^{-1} \mathbf{d}_k^T \mathbf{A}_{k-1}^+, \quad \text{если} \quad \mathbf{c}_k = \mathbf{0}.$$

Идея метода Гревилля последовательного вычисления элементов псевдообратной матрицы $\mathbf{A}^+$ строка за строкой заключается в поиске наилучшего приближенного решения (по методу наименьших квадратов, см. § 7.5, 7.6) матричного уравнения $\mathbf{AX} = \mathbf{E}$. Наилучшее приближенное решение $\mathbf{X}^0$ определяют из условия

$$\left\|\mathbf{E} - \mathbf{AX}^0\right\| = \min_{\mathbf{X}} \left\|\mathbf{E} - \mathbf{AX}\right\|,$$

с дополнительным требованием минимальности величины $\|\mathbf{X}^0\|$, где в качестве матричной нормы используется норма Фробениуса, т.е.

$$\|\mathbf{A}\|^2 = \sum_{i,k} |a_{ik}|^2 = \sum_{k=1}^{n} \left|\mathbf{A}_{\cdot,k}\right|^2 = \sum_{i=1}^{m} \left|\mathbf{A}_{i,\cdot}\right|^2.$$

Таким образом, каждый столбец искомой матрицы должен быть наилучшим в указанном смысле приближенным решением СЛАУ $\mathbf{AX}_{\cdot,k} = \mathbf{E}_{\cdot,k}$, что позволяет отождествить $\mathbf{X}^0$ с $\mathbf{A}^+$.

Встает вопрос о сравнительных характеристиках конструктивных способов построения псевдообратных матриц: метода сингулярного разложения и метода Гревилля. Результаты сравнения вычислительных затрат при реализации этих двух методов показывают, что они близки по порядку как числа требуемых арифметических операций, так и по объему требуемой оперативной памяти компьютера. Некоторые преимущества в этом на стороне метода Гревилля. Однако в смысле численной устойчивости предпочтение следует отдать SVD-разложению. Его можно считать безусловно устойчивым, в то время как метод Гревилля заведомо неприемлем, например, при псевдообращении матриц с первым столбцом, близким к нуль-вектору, что легко увидеть непосредственно из формулы (7.16).

Замечание 7.2. Последнему методу более близок еще один способ определения псевдообратной матрицы $\mathbf{A}^+$, который, как и набор формул (7.15), непосредственно не порождает алгоритма вычисления ее элементов.

Пусть прямоугольная матрица $\mathbf{A}$ рассматривается в контексте решения СЛАУ

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{7.17}$$

Эту систему можно погрузить в семейство систем вида

$$\left(\mathbf{A}^T\mathbf{A} + \alpha\mathbf{E}\right)\mathbf{x} = \mathbf{A}^T\mathbf{b}\ , \tag{7.18}$$

в котором ей соответствует значение параметра $\alpha = 0$. При любом $\alpha > 0$ спектр матрицы $\mathbf{A}^T\mathbf{A} + \alpha\mathbf{E}$ строго положителен и, следовательно, решение $\mathbf{x}_\alpha$ параметризованной системы (7.18) существует и имеет представление

$$\mathbf{x}_\alpha := \left(\mathbf{A}^T\mathbf{A} + \alpha\mathbf{E}\right)^{-1}\mathbf{A}^T\mathbf{b}\ .$$

Так как в случае обратимости матрицы $\mathbf{A}$

$$\lim_{\alpha \to 0+} \mathbf{x}_\alpha = \mathbf{x}^*$$

— решение исходной системы (7.17), то естественно считать выражение

$$\mathbf{A}^+ := \lim_{\alpha \to 0+} \left(\mathbf{A}^T\mathbf{A} + \alpha\mathbf{E}\right)^{-1}\mathbf{A}^T \tag{7.19}$$

псевдообратной матрицей, а вектор $\mathbf{A}^+\mathbf{b}$ — *псевдорешением* системы (7.17). Можно проверить, что эти матрица и вектор совпадают с введенными выше псевдообратной матрицей и нормальным псевдорешением соответственно.

В качестве содержательного примера, где стоит использовать псевдообращение, рассмотрим фрагмент динамической теории «затраты – выпуск» В. В. Леонтьева. Согласно ей, ***модель межотраслевого баланса*** описывается системой обыкновенных дифференциальных уравнений вида

$$\mathbf{A}\dot{\mathbf{x}}(t) - \left(\mathbf{E} - \mathbf{B}\right)\mathbf{x}(t) + \mathbf{y}(t) = \mathbf{0} \tag{7.20}$$

с постоянными матрицами $\mathbf{A} := (a_{ij})$ и $\mathbf{B} := (b_{ij})$. Элемент a_{ij} матрицы $\mathbf{A}$ интерпретируется как запас продукции i-й отрасли, нужной для производства продукции j-й отрасли, а b_{ij} — число единиц продукции i-й отрасли, требуемое для производства единицы продукции j-й отрасли. Векторные функции $\mathbf{x}(t)$ и $\mathbf{y}(t)$ характеризуют соответственно валовой выпуск и конечный спрос.

Учитывая тот факт, что численные методы решения дифференциальных уравнений настроены на их нормализованный вид, т.е. разрешенный относительно производной, уравнение (7.20) должно быть преобразовано к виду

$$\dot{\mathbf{x}}(t) = \mathbf{A}^{-1}(\mathbf{E} - \mathbf{B})\mathbf{x}(t) - \mathbf{A}^{-1}\mathbf{y}(t). \qquad (7.21)$$

Такое преобразование может быть выполнено, если матрица $\mathbf{A}^{-1}$ существует и может быть устойчиво вычислена, что в практике применения рассматриваемой модели редко имеет место в силу ряда объективных причин. Поэтому вполне разумно от уравнения (7.20) переходить не к уравнению (7.21), а к уравнению

$$\dot{\mathbf{x}}(t) = \mathbf{A}^{+}(\mathbf{E} - \mathbf{B})\mathbf{x}(t) - \mathbf{A}^{+}\mathbf{y}(t),$$

в котором псевдообратная матрица $\mathbf{A}^{+}$ должна тем или иным способом надежно вычисляться. При выборе метода численного интегрирования полученного уравнения следует принимать во внимание число жесткости матрицы $\mathbf{A}^{+}(\mathbf{E} - \mathbf{B})$.

§ 7.4. НЕКОТОРЫЕ ДРУГИЕ ПРИМЕНЕНИЯ SVD-РАЗЛОЖЕНИЙ

1. Решение систем нелинейных уравнений

Одним из наиболее популярных методов решения систем вида

$$\begin{cases} f_1(x_1, ..., x_n) = 0, \\ \cdots\cdots\cdots\cdots \\ f_n(x_1, ..., x_n) = 0, \end{cases} \qquad (7.22)$$

где $f_i : \mathbb{R}_n \to \mathbb{R}_1$ — заданные нелинейные функции, является ***метод Ньютона*** [12, 13]. Используя обозначения

$$\mathbf{x} := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{f}(\mathbf{x}) := \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{pmatrix}, \quad \mathbf{0} := \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

т.е. представляя систему (7.22) одним уравнением $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, этот

метод можно определить явной формулой

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - [\mathbf{f}'(\mathbf{x}^{(k)})]^{-1}\mathbf{f}(\mathbf{x}^{(k)}), \qquad (7.23)$$

где $\mathbf{f}'(\mathbf{x})$ — матрица Якоби векторной функции $\mathbf{f}: \mathbb{R}_n \to \mathbb{R}_n$. Придавая в (7.23) индексу k последовательно значения 0, 1, 2,..., по заданному начальному приближению $\mathbf{x}^{(0)}$ получают последовательность векторов $(\mathbf{x}^{(k)})$, при определенных условиях квадратично сходящуюся к искомому решению $\mathbf{x}^*$ системы (7.22). Чаще метод Ньютона реализуют в неявной форме: при каждом значении $k := 0, 1, 2, ...$ из линейной системы

$$\mathbf{f}'(\mathbf{x}^{(k)})\mathbf{p}^{(k)} = -\mathbf{f}(\mathbf{x}^{(k)}) \qquad (7.24)$$

находят вектор поправок $\mathbf{p}^{(k)}$, прибавление которого к текущему приближению дает следующее приближение:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{p}^{(k)}. \qquad (7.25)$$

Поскольку СЛАУ вида (7.24) при каких-то значениях k может не иметь единственного решения, есть смысл на этих итерациях в качестве вектора $\mathbf{p}^{(k)}$ в (7.25) использовать нормальное псевдорешение системы (7.24), которое может быть найдено, например, с помощью SVD-разложения матрицы $\mathbf{f}'(\mathbf{x}^{(k)})$. Так как нормальное псевдорешение должно совпадать с обычным решением СЛАУ, если таковое существует, то в принципе метод Ньютона можно основывать на одних только сингулярных разложениях. Однако в силу их сравнительной (например, с LU-разложением) неэкономичности более целесообразно строить гибридные алгоритмы, в которых SVD-разложение привлекается только в случае вырожденности матриц Якоби. Последнее легко обнаруживается в процессе постолбцового выбора главного элемента в ходе решения СЛАУ (7.24) методом Гаусса.

Очевидна лишь одна ситуация, когда явно напрашивается применение SVD-разложения на каждом итерационном шаге процесса (7.24), (7.25) — это в случае, если матрица Якоби не является квадратной, т.е. при решении систем $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ с $\mathbf{f}: \mathbb{R}_n \to \mathbb{R}_m$.

Пример 7.2. Рассмотрим переопределенную систему $\begin{cases} x_1^2 + x_2^2 - 1 = 0, \\ x_1^2 + x_2 - 1 = 0, \\ x_1 - 1 = 0 \end{cases}$

с матрицей Якоби $\begin{pmatrix} 2x_1 & 2x_2 \\ 2x_1 & 1 \\ 1 & 0 \end{pmatrix}$ и единственным решением $\mathbf{x}^* := (1;\ 0)$ (рис. 7.1).

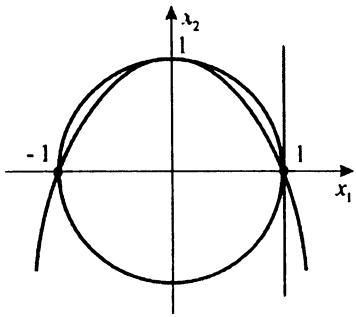


Рис. 7.1

Ясно, что к этой простой системе применить непосредственно классический метод Ньютона нельзя. Использование псевдорешений в возникающих в методе Ньютона линейных задачах позволяет за 20 итераций из точки $\mathbf{x}^{(0)} := (0{,}5;\ -1)$ получить решение $\mathbf{x}^* :\approx (1{,}0000000030;\ -0{,}0000000059)$.

Описанный здесь подход можно использовать и в сочетании с другими основанными на линеаризации итерационными методами решения систем нелинейных уравнений, в частности систем, заведомо переопределенных, как в приведенном примере 7.2.

2. Решение интегральных уравнений первого рода

Линейное ***интегральное уравнение*** Фредгольма первого рода

$$\int_a^b Q(t,s)x(s)ds = f(t), \quad t \in [c,d], \tag{7.26}$$

является собой классический пример некорректно поставленной

задачи, имеющей в то же время достаточно обширные приложения. Решения в стандартном понимании у такого уравнения, как правило, не существует. Поэтому в тех или иных предположениях об исходных сведениях о задаче (7.26) возникают вопросы о том, что можно назвать решением данной задачи и как его получить. В рамках охватываемой здесь тематики подойдем к ответам на эти вопросы следующим образом.

Применив к интегралу в (7.26) какую-нибудь квадратурную формулу с упорядоченным набором узлов $s_1, s_2, ..., s_n$ на промежутке $[a, b]$ и с соответствующими им весовыми коэффициентами $c_1, c_2, ..., c_n$, от данного интегрального уравнения (7.26) приходим к промежуточному уравнению

$$\sum_{j=1}^{n} c_j Q(t, s_j) x_j = f(t), \quad t \in [c, d],$$

относительно n неизвестных чисел $x_j \approx x(s_j)$. Это уравнение полностью дискретизируется введением на отрезке $[c, d]$ своей сетки для переменной t: $t_1 < t_2 < ... < t_m$, в результате чего получаем систему ***сеточных уравнений***

$$\sum_{j=1}^{n} c_j Q(t_i, s_j) x_j = f(t_i) \quad , \qquad i = 1, 2, ..., m . \qquad (7.27)$$

Полагая $a_{ij} := c_j Q(t_i, s_j)$, $b_i := f(t_i)$, видим, что (7.27) можно отнести к стандартному виду СЛАУ (7.7) с $m \times n$-матрицей $\mathbf{A} := (a_{ij})$ и m-мерным вектором $\mathbf{b} := (b_i)$ относительно n-мерного вектора $\mathbf{x} := (x_j)$. Такая система, как это следует из предыдущего, всегда имеет единственное нормальное псевдорешение, которое может быть найдено, например, с помощью SVD-разложения (см. формулу (7.12)). Получив это нормальное псевдорешение $\mathbf{x}^* := (x_j^*)$, остается совершить ***восполнение*** (например, интерполяцией),

т.е. подобрать такую функцию $x^*(s)$, $s \in [a, b]$ (возможно, удовлетворяющую каким-либо заранее заданным требованиям), что $x^*(s_j) \approx x_j^* \quad \forall j \in \{1, 2, ..., n\}$.

К сожалению, нормальное псевдорешение системы (7.7) *неустойчиво* по отношению к возмущениям матрицы $\mathbf{A}$ (этот факт на простейшем примере демонстрируется в § 8.5, см. также [13]). Поэтому описанный подход к построению обобщенных решений интегральных уравнений первого рода следует рассматривать не как альтернативу применению специальных методов, основанных на регуляризации (см. по этому поводу, например, [10, 15]), а, скорее, как дополнение к ним.

Заметим, что неустойчивость нормального псевдорешения повышает требования к устойчивости самого процесса его получения. Можно считать, что организация такого процесса на основе сингулярных разложений, опирающихся на ортогональные преобразования, таким свойством обладает.

3. Сжатие информации

Предположим, что некоторая информация представлена матрицей $\mathbf{X} := (x_{ij})$. Это может быть, например, таблица значений какого-либо параметра плоского изображения или матрица, составленная из определенного фиксированного числа «отрезков» дискретного временного ряда. Для простоты будем считать, что эта матрица — квадратная ($i, j = 1, 2, ..., n$). Выполнив сингулярное разложение матрицы $\mathbf{X}$, имеем:

$$\mathbf{X} = \mathbf{U\Sigma V} = \begin{pmatrix} u_{11} & \dots & u_{1n} \\ \dots & \ddots & \dots \\ u_{n1} & \dots & u_{nn} \end{pmatrix} \cdot \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix} \cdot \begin{pmatrix} v_{11} & \dots & v_{1n} \\ \dots & \ddots & \dots \\ v_{n1} & \dots & v_{nn} \end{pmatrix} =$$

$$= \begin{pmatrix} \sum\limits_{k=1}^{n} u_{1k}\sigma_k v_{k1} & \dots & \sum\limits_{k=1}^{n} u_{1k}\sigma_k v_{kn} \\ \dots & \ddots & \dots \\ \sum\limits_{k=1}^{n} u_{nk}\sigma_k v_{k1} & \dots & \sum\limits_{k=1}^{n} u_{nk}\sigma_k v_{kn} \end{pmatrix}.$$

Следовательно, каждый элемент x_{ij} матрицы **X** характеризуется своим разложением по сингулярным значениям, имеющем вид

$$x_{ij} = \sigma_1 u_{i1} v_{1j} + \sigma_2 u_{i2} v_{2j} + ... + \sigma_n u_{in} v_{nj},$$

где сингулярные числа $\sigma_1, \sigma_2, ..., \sigma_n$, напомним, считаются упорядоченными по убыванию. Обычно значимость слагаемых в подобном разложении убывает достаточно быстро с увеличением их номеров. Поэтому отбрасывание нескольких последних слагаемых должно мало изменить информацию, содержащуюся в x_{ij}, т.е. без большого ущерба можно положить

$$x_{ij} \approx \hat{x}_{ij} := \sigma_1 u_{i1} v_{1j} + \sigma_2 u_{i2} v_{2j} + ... + \sigma_k u_{ik} v_{kj},$$

где $k < n$.

Таким образом, *неполное сингулярное разложение*, требуя для своего хранения меньше памяти, может быть использовано в качестве способа сжатия информации (с некоторой потерей). Имеющиеся данные по экспериментальному тестированию такого метода показывают сорокакратное сжатие при пятипроцентной потере информации и десятикратное — при двухпроцентной.

Похожие на проведенные рассуждения используют также при построении методов реставрации пространственных изображений [56], алгоритмов кодирования, при анализе временных рядов.

§ 7.5. ДВА ИСТОЧНИКА ЛИНЕЙНЫХ ЗАДАЧ НАИМЕНЬШИХ КВАДРАТОВ (ЛЗНК)

Задача нахождения нормальных псевдорешений вырожденных, как правило, переопределенных систем линейных алгебраических уравнений возникает при решении некоторых других, более содержательных математических задач, в частности, связанных с обработкой эмпирических данных. Рассмотрим две такие задачи, решение которых легло в основу широко распространенного ***метода наименьших квадратов*** **(МНК).** Этот метод появился в результате ряда геодезических исследований знаменитых

ученых XVIII –XIX вв. Лежандра* и Гаусса. Характерной особенностью таких задач является то обстоятельство, что исходный материал для выявления тех или иных закономерностей — заведомо приближенный, в силу неточности измерительных приборов, неповторимости условий наблюдений, случайных ошибок и пр. Отсюда естественность стремления исследователей к накоплению как можно большего числа экспериментальных данных с тем, чтобы, усредняя их каким-либо образом, получить требуемые значения параметров изучаемых закономерностей наиболее точно.

Нахождение недоступных непосредственному измерению величин

Предположим, что нужно знать значения n различных величин $x_1, x_2, \ldots, x_n$, которые нельзя измерить непосредственно, но известно, что они находятся в линейной связи, и коэффициенты этой связи могут быть приближенно получены в результате измерений. Проведя серию из m таких измерений, имеем систему линейных алгебраических уравнений

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m. \end{cases} \qquad (7.28)$$

Чтобы эта система могла быть однозначно разрешимой, т.е. чтобы можно было найти единственный набор значений переменных $x_1, x_2, \ldots, x_n$, интересующих исследователя, число измерений m должно быть не меньшим числа неизвестных n, и при этом ранг матрицы системы должен в точности совпадать с n. В силу того что все числа a_{ij} и b_i в (7.28) (или часть из них) являются экспериментальными данными, получаемое таким путем одно-

* Лежа́ндр Адриен Мари (1752–1833) — французский математик. Обосновал и развил теорию геодезических измерений. Внес большой вклад в развитие математического анализа, вариационного исчисления, теории чисел.

значное решение может сильно отличаться от однозначного же решения, находимого таким же образом с помощью другой серии экспериментов или наблюдений.

Поэтому число m в системе (7.28) берут заведомо бóльшим, чем n, и рассматривают переопределенную, как правило, несовместную СЛАУ. Поскольку такая система в общем случае не имеет решения в привычном понимании, в то время как по смыслу исходной задачи решение должно существовать, следует каким-то образом обобщить понятие решения СЛАУ и получить метод его нахождения.

С этой целью будем искать набор $x_1^*, x_2^*, \ldots, x_n^*$ значений переменных $x_1, x_2, \ldots, x_n$ не такой, который обращал бы каждое из уравнений системы (7.28) в тождество, а при котором была бы минимальной сумма квадратов отклонений значений левых частей уравнений (7.28) от правых (иначе, квадратов невязок). Это означает, что вектор $\left(x_1^*; x_2^*; \ldots; x_n^*\right)$ должен быть решением задачи минимизации

$$r_1^2 + r_2^2 + \ldots + r_m^2 \to \min,$$

где $r_i := a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n - b_i$.

Очевидно, минимизируемая функция

$$\Phi(x_1, x_2, \ldots, x_n) := \sum_{i=1}^{m} \left(a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n - b_i\right)^2,$$

являясь неотрицательной квадратичной, имеет глобальный минимум. Он может быть найден из необходимых условий гладкого экстремума, т.е. в результате приравнивания нулю частных производных функции $\Phi(x_1, x_2, \ldots, x_n)$, взятых по каждому из ее аргументов.

Выполняя дифференцирование функции $\Phi(x_1, x_2, \ldots, x_n)$ по переменным $x_1, x_2, \ldots, x_n$, приходим к линейной алгебраической системе

$$\begin{cases} 2\sum\limits_{i=1}^{m}(a_{i1}x_1+a_{i2}x_2+\ldots+a_{in}x_n-b_i)a_{i1}=0, \\ 2\sum\limits_{i=1}^{m}(a_{i1}x_1+a_{i2}x_2+\ldots+a_{in}x_n-b_i)a_{i2}=0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ 2\sum\limits_{i=1}^{m}(a_{i1}x_1+a_{i2}x_2+\ldots+a_{in}x_n-b_i)a_{in}=0. \end{cases}$$

Полученной системе n уравнений с n неизвестными может быть придан стандартный вид СЛАУ:

$$\begin{cases} \left(\sum\limits_{i=1}^{m}a_{i1}^2\right)x_1+\left(\sum\limits_{i=1}^{m}a_{i2}a_{i1}\right)x_2+\ldots+\left(\sum\limits_{i=1}^{m}a_{in}a_{i1}\right)x_n=\sum\limits_{i=1}^{m}b_ia_{i1}, \\ \left(\sum\limits_{i=1}^{m}a_{i1}a_{i2}\right)x_1+\left(\sum\limits_{i=1}^{m}a_{i2}^2\right)x_2+\ldots+\left(\sum\limits_{i=1}^{m}a_{in}a_{i2}\right)x_n=\sum\limits_{i=1}^{m}b_ia_{i2}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \left(\sum\limits_{i=1}^{m}a_{i1}a_{in}\right)x_1+\left(\sum\limits_{i=1}^{m}a_{i2}a_{in}\right)x_2+\ldots+\left(\sum\limits_{i=1}^{m}a_{in}^2\right)x_n=\sum\limits_{i=1}^{m}b_ia_{in}. \end{cases} \quad (7.29)$$

Если эта система (называемая ***нормальной системой МНК***, или ***системой нормальных уравнений***) имеет полный ранг, то ее решение — вектор $\left(x_1^*; x_2^*; \ldots; x_n^*\right)$ — в принципе может быть получено каким-либо методом решения однозначно определенных линейных систем, в частности, учитывающим симметрию СЛАУ (7.29) (см. гл. 2). Очевидно, этот вектор суть *нормальное псевдорешение* исходной системы (7.28). В случае вырожденности или плохой обусловленности матрицы системы (7.29) обозначенную проведенными выше рассуждениями так называемую ***линейную задачу наименьших квадратов***[*] следует решать какими-то иными методами, выбор которых обсуждается в следующем параграфе.

---

[*] Или чуть иначе, она называется ***линейная задача о наименьших квадратах***. В любом случае используем аббревиатуру **ЛЗНК**.

Пример 7.3. Пусть определению подлежат две величины: x_1 и x_2, связь между которыми задается тремя равенствами:

$$\begin{cases} x_1 + 2x_2 = 8{,}5, \\ 2x_1 + x_2 = 6{,}7, \\ x_1 + 3x_2 = 11{,}1, \end{cases} \tag{7.30}$$

и пусть правые части уравнений (7.30) — заведомо приближенные числа. Совершенно очевидно, что эта система противоречива, т.е. не имеет решения в обычном смысле. Для получения ее нормального псевдорешения, т.е. решения в смысле метода наименьших квадратов, составим новую систему по типу (7.29) при $n = 2$, $m = 3$:

$$\begin{cases} 6x_1 + 7x_2 = 33, \\ 7x_1 + 14x_2 = 57. \end{cases}$$

Решив последнюю, получаем пару $x_1^* := 1{,}8$, $x_2^* :\approx 3{,}171$, не удовлетворяющую никакому отдельно взятому уравнению данной системы (7.30) и в то же время устраивающую систему в целом в оговоренном смысле. Геометрическое представление о найденном решении можно получить из рис. 7.2.

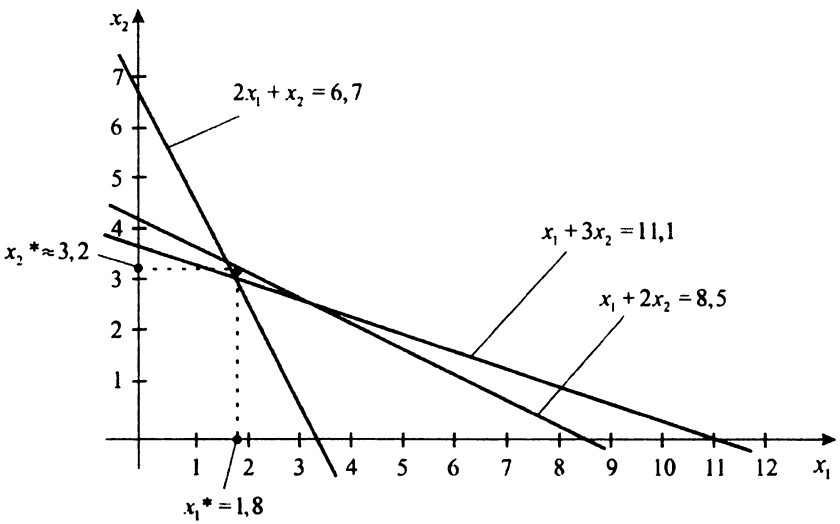


Рис. 7.2

Замечание 7.3. Геометрические образы уравнений системы (7.30) не изменяются при умножении левых и правых частей уравнений на одинаковые отличные от нуля числа, т.е. можно считать, что преобразованная таким способом система в этом смысле эквивалентна исходной. Однако находимые методом

наименьших квадратов нормальные псевдорешения эквивалентных в указанном смысле систем будут различаться. Например, если второе уравнение системы (7.30) умножить на 100, т.е. заменить уравнением

$$200x_1 + 100x_2 = 670,$$

а первое и третье оставить неизменными, то система нормальных уравнений вида (7.29) для такого случая есть

$$\begin{cases} 40002x_1 + 20005x_2 = 134019,6, \\ 20005x_1 + 10013x_2 = 67050,3. \end{cases}$$

Решая ее, получаем псевдорешение $\tilde{x}_1 \approx 1,756$, $\tilde{x}_2 \approx 3,188$, близкое к найденному выше, но все же отличное от него.

Невязки r_i двумерных линейных уравнений можно интерпретировать как отклонения точки от прямых в тех случаях, когда их уравнения приводятся к нормальному виду. Деля каждое уравнение данной системы на квадратный корень из суммы квадратов его коэффициентов, приходим сначала к нормализованной системе

$$\begin{cases} \frac{1}{\sqrt{5}}x_1 + \frac{2}{\sqrt{5}}x_2 = \frac{8,5}{\sqrt{5}}, \\ \frac{2}{\sqrt{5}}x_1 + \frac{1}{\sqrt{5}}x_2 = \frac{6,7}{\sqrt{5}}, \\ \frac{1}{\sqrt{10}}x_1 + \frac{3}{\sqrt{10}}x_2 = \frac{11,1}{\sqrt{10}}, \end{cases}$$

из нее, как и выше, получаем однозначно разрешимую систему

$$\begin{cases} 1,1x_1 + 1,1x_2 = 5,49, \\ 1,1x_1 + 1,9x_2 = 8,07, \end{cases}$$

решение которой $\bar{x}_1 :\approx 1,766$, $\bar{x}_2 :\approx 3,225$ имеет хорошую геометрическую интерпретацию: сумма квадратов расстояний от точки $(\bar{x}_1; \bar{x}_2)$ до всех прямых, определяемых уравнениями данной системы (7.30), меньше, чем от любой другой точки $(x_1; x_2)$.

Ответ на вопрос о том, стоит ли искать такое оптимальное псевдорешение с помощью нормализации уравнений данной системы, или составлять систему вида (7.29) непосредственно из данной системы, или проводить перед ее составлением некоторое предварительное масштабирование (усиливающее или уменьшающее роль отдельных связей между искомыми величинами), в конкретных случаях зависит от содержательного смысла коэффициентов и неизвестных СЛАУ (7.28).

Подбор параметров эмпирических функций

Пусть между независимой переменной x и зависимой переменной y имеется некая неизвестная функциональная связь $y = f(x)$. Эта связь отображается таблицей

x	x_0	x_1	...	x_m
y	y_0	y_1	...	y_m

приближенных значений $y_i :\approx f(x_i)$, получаемых в ходе наблюдений или экспериментов. Требуется дать приближенное аналитическое описание этой связи, т.е. подобрать функцию $\varphi(x)$ такую, которая представляла бы на отрезке $[x_0, x_m]$ заданную отдельными приближенными значениями y_i функцию $f(x)$.

В основу подбора типа искомой зависимости могут быть положены графические или иные соображения, в частности, учитывающие содержательный смысл изучаемой связи. Подобрав подходящее семейство функций $\varphi(x, a_0, a_1, \ldots, a_n)$ для описания неизвестной функции $f(x)$, приближенно представленной заданной таблицей, ставим вопрос о том, какими следует зафиксировать параметры $a_0 := a_0^*$, $a_1 := a_1^*, \ldots, a_n := a_n^*$, чтобы имело место

$$\varphi\left(x, a_0^*, a_1^*, \ldots, a_n^*\right) \approx f(x) \quad \forall x \in [x_0, x_m] .$$

В силу нескольких объективных причин, в частности наличия случайных ошибок при получении экспериментальных данных y_i, нецелесообразно приравнивать число $n+1$ искомых параметров a_i числу $m+1$ известных значений y_i и находить эти параметры из *условия интерполяции*

$$\varphi(x_i, a_0, a_1, \ldots, a_n) = y_i \quad \forall i \in \{0, 1, \ldots, m\}$$

хотя бы потому, что в построенной таким путем математической модели изучаемого явления заведомо будут присутствовать все ошибки эксперимента.

Чтобы уменьшить влияние ошибок эксперимента и к тому же обойтись небольшим числом параметров, число m берут существенно большим, чем n, и параметры a_i фиксируют такими, чтобы с оптимальным набором значений $a_0^*, a_1^*, \ldots, a_n^*$ функция $\varphi\left(x, a_0^*, a_1^*, \ldots, a_n^*\right)$ «в среднем» хорошо описывала поведение функции $f(x)$. Одна из наиболее употребительных расшифровок термина «в среднем» состоит в том, что этот набор параметров должен находиться как решение экстремальной задачи

$$\Phi(a_0, a_1, \ldots, a_n) := \sum_{i=0}^{m} \left(\varphi(x_i, a_0, a_1, \ldots, a_n) - y_i\right)^2 \to \min. \quad (7.31)$$

При таком подходе, геометрически означающем минимизацию суммы квадратов расстояний от каждой исходно заданной точки до соответствующей точки графика $y = \varphi\left(x, a_0^*, a_1^*, \ldots, a_n^*\right)$, нивелируется роль отдельных случайных ошибок («выбросов»), т.е. происходит сглаживание результатов в процессе их аналитического описания.

Решение задачи (7.31) можно найти из необходимого условия экстремума функции $\Phi(a_0, a_1, \ldots, a_n)$, очевидно, являющегося здесь и достаточным:

$$\begin{cases} \sum_{i=0}^{m} \left(\varphi(x_i, a_0, a_1, \ldots, a_n) - y_i\right) \dfrac{\partial \varphi}{\partial a_0}\Big|_{x = x_i} = 0, \\ \sum_{i=0}^{m} \left(\varphi(x_i, a_0, a_1, \ldots, a_n) - y_i\right) \dfrac{\partial \varphi}{\partial a_1}\Big|_{x = x_i} = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum_{i=0}^{m} \left(\varphi(x_i, a_0, a_1, \ldots, a_n) - y_i\right) \dfrac{\partial \varphi}{\partial a_n}\Big|_{x = x_i} = 0. \end{cases} \quad (7.32)$$

В общем случае полученная в итоге система МНК будет нелинейной. Если же в функцию $\varphi(x, a_0, a_1, \ldots, a_n)$ параметры входят линейным образом, то и система (7.32) — линейная.

Рассмотрим представление искомой функции $f(x)$ по заданным выше ее приближенным значениям y_i функцией $\varphi(x)$ вида

$$\varphi(x) := a_0\varphi_0(x) + a_1\varphi_1(x) + \ldots + a_n\varphi_n(x), \qquad (7.33)$$

где $\{\varphi_j(x)\}_{j=0}^{n}$ — некоторая система линейно независимых функций, определенных, по меньшей мере, на отрезке $[x_0, x_m]$ (причем, напомним, $n \le m$; как правило, $n \ll m$).

Так как $\dfrac{\partial\varphi}{\partial a_j} = \varphi_j(x)$, то результат подстановки функции (7.33) и ее производных по параметрам в систему (7.32) приводит ее к виду

$$\begin{cases} \sum\limits_{i=0}^{m} \varphi_0(x_i)\left(a_0\varphi_0(x_i) + a_1\varphi_1(x_i) + \ldots + a_n\varphi_n(x_i) - y_i\right) = 0, \\ \sum\limits_{i=0}^{m} \varphi_1(x_i)\left(a_0\varphi_0(x_i) + a_1\varphi_1(x_i) + \ldots + a_n\varphi_n(x_i) - y_i\right) = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum\limits_{i=0}^{m} \varphi_n(x_i)\left(a_0\varphi_0(x_i) + a_1\varphi_1(x_i) + \ldots + a_n\varphi_n(x_i) - y_i\right) = 0. \end{cases}$$

Последняя система легко может быть представлена в стандартной форме симметричной СЛАУ относительно искомых параметров $a_0, a_1, \ldots, a_n$ — коэффициентов линейной комбинации (7.33):

$$\begin{cases} (\varphi_0,\varphi_0)a_0 + (\varphi_0,\varphi_1)a_1 + \ldots + (\varphi_0,\varphi_n)a_n = (\varphi_0, y), \\ (\varphi_1,\varphi_0)a_0 + (\varphi_1,\varphi_1)a_1 + \ldots + (\varphi_1,\varphi_n)a_n = (\varphi_1, y), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ (\varphi_n,\varphi_0)a_0 + (\varphi_n,\varphi_1)a_1 + \ldots + (\varphi_n,\varphi_n)a_n = (\varphi_n, y), \end{cases}$$

где использованы обозначения скалярного произведения

$$(\varphi_k,\varphi_j) := \sum_{i=0}^{m} \varphi_k(x_i)\,\varphi_j(x_i)\ ,\ \ (\varphi_k, y) := \sum_{i=0}^{m} \varphi_k(x_i)\,y_i\,.$$

Таким образом, задача, в исходной постановке отличающаяся от предыдущей, в итоге применения метода наименьших квадратов свелась к решению корректно определенной СЛАУ с квадратной симметричной матрицей коэффициентов такой же, как и СЛАУ (7.29).

§ 7.6. ОСОБЕННОСТИ И МЕТОДЫ РЕШЕНИЯ ЛЗНК

Формальная постановка линейной задачи о наименьших квадратах, частные случаи которой приведены в предыдущем параграфе, следующая [45].

Для данных вещественных m-мерного вектора $\mathbf{b}$ *и* $m \times n$-*матрицы* $\mathbf{A}$ *ранга* $k \le \min\{m, n\}$ *требуется найти вещественный* n-*мерный вектор* $\mathbf{x}$, *минимизирующий евклидову длину вектора* $\mathbf{Ax} - \mathbf{b}$.

В зависимости от того, как сочетаются значения натуральных параметров m, n и k, можно предусмотреть шесть случаев:

А) $k = m = n$;

Б) $k < m = n$;

В) $k = m < n$;

Г) $k = n < m$;

Д) $k < m < n$;

Е) $k < n < m$.

Наиболее существенную роль в приложениях играют переопределенные системы полного ранга (случай Г). Объединяя этот случай со случаем А, будем считать, что в исходной для ЛЗНК системе

$$\mathbf{Ax} = \mathbf{b} \tag{7.34}$$

$\operatorname{rank} \mathbf{A} = n$ и число m приближенных уравнений $\sum_{j=1}^{n} a_{ij} x_j = b_i$ в этой системе не меньше числа n неизвестных x_j (см., например, систему (7.28)) .

Как следует из рассмотренного выше материала, формальным решением поставленной ЛЗНК является псевдорешение системы

(7.34) (оно же здесь, т.е. при $k = n$, — нормальное псевдорешение). Выбор способа численного нахождения этого решения, т.е. такого вектора $\mathbf{x}^+$, что

$$\mathbf{x}^+ := \arg \min_{\mathbf{x} \in \mathbb{R}_n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 , \tag{7.35}$$

зависит от свойств матрицы $\mathbf{A}^T\mathbf{A}$ ***нормальной системы МНК***

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b} . \tag{7.36}$$

Если матрица $\mathbf{A}^T\mathbf{A}$ *хорошо обусловлена, то определяемое посредством* (7.35) *решение ЛЗНК может быть успешно найдено применением к системе* (7.36) *какого-либо из наименее затратных методов решения однозначно определенных СЛАУ, например, учитывающего симметрию матрицы* $\mathbf{A}^T\mathbf{A}$ *метода Холецкого* (см. § 2.3).

В противном случае следует искать альтернативный способ получения псевдорешения $\mathbf{x}^+$ (или нормального псевдорешения $\mathbf{x}_0$), каковым может быть метод, использующий SVD-разложение исходной матрицы $\mathbf{A}$.

Заметим, что операция симметризации Гаусса системы (7.34), приводящая к системе (7.36), сильно ухудшает ее обусловленность. Действительно, пусть $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k$ — ненулевые сингулярные числа матрицы $\mathbf{A}$ ранга $k = n$, т.е. ее мера обусловленности (см. § 7.1)

$$\operatorname{cond} \mathbf{A} := \frac{\sigma_1}{\sigma_k} .$$

Так как соответствующие ненулевые собственные числа матрицы $\mathbf{A}^T\mathbf{A}$ — это $\lambda_1 := \sigma_1^2$, $\lambda_2 := \sigma_2^2$, ... , $\lambda_k := \sigma_k^2$ (они же — ее сингулярные числа), то

$$\operatorname{cond} \mathbf{A}^T\mathbf{A} = \frac{\lambda_1}{\lambda_k} = \frac{\sigma_1^2}{\sigma_k^2} = (\operatorname{cond} \mathbf{A})^2 .$$

Так, обращаясь к примеру 6.1, где найдены сингулярные числа $\sigma_1 := \sqrt{2+\delta^2}$, $\sigma_2 := |\delta|$ матрицы $\mathbf{A} := \begin{pmatrix} 1 & 1 \\ \delta & 0 \\ 0 & \delta \end{pmatrix}$ через собственные числа $\lambda_1 := 2+\delta^2$, $\lambda_2 := \delta^2$ матрицы $\mathbf{A}^T\mathbf{A} = \begin{pmatrix} 1+\delta^2 & 1 \\ 1 & 1+\delta^2 \end{pmatrix}$, при малых δ имеем:

$$\operatorname{cond}\mathbf{A} = \frac{\sqrt{2+\delta^2}}{\delta} = \sqrt{\frac{2}{\delta^2}+1} = O\left(\frac{1}{\delta}\right),$$

$$\operatorname{cond}\mathbf{A}^T\mathbf{A} = \frac{2+\delta^2}{\delta^2} = \frac{2}{\delta^2}+1 = O\left(\frac{1}{\delta^2}\right),$$

и если, например, $\delta := 10^{-10}$, то $\operatorname{cond}\mathbf{A} :\approx 10^{10}$, а $\operatorname{cond}\mathbf{A}^T\mathbf{A} :\approx 10^{20}$.

Учитывая ту роль, которую играет число обусловленности в воздействии ошибок исходных данных на результат решения СЛАУ (по крайней мере, при $m = n = k$ легко показать, что $\operatorname{cond}\mathbf{A}$ служит коэффициентом связи между относительными ошибками матрицы $\mathbf{A}$ и вектора $\mathbf{b}$ и относительной ошибкой решения системы, см. § 8.3), понимаем, насколько критичным может оказаться возведение в квадрат числа обусловленности матрицы системы при стандартном подходе к решению ЛЗНК.

Как было продемонстрировано в § 6.2, 6.3, процедура выполнения сингулярного разложения $m \times n$-матрицы $\mathbf{A}$ может быть выполнена на основе ортогональных преобразований и без непосредственного построения матрицы $\mathbf{A}^T\mathbf{A}$, а в § 7.2 выведена формула (7.12) получения нормального псевдорешения $\mathbf{x}_0$ системы вида (7.34) с использованием этого разложения (см. также представление вектора $\mathbf{x}_0$ через псевдообратную матрицу $\mathbf{A}^+$ в § 7.3).

Таким образом, *в случае плохой обусловленности системы* (7.36) *целесообразно обратиться к более затратному, но более устойчивому методу получения решения ЛЗНК* (7.35) *с помощью SVD-разложения.*

Вернемся на время к более общему случаю и проанализируем ЛЗНК в предположении, что $m\times n$-матрица $\mathbf{A}$ системы (7.34) имеет ранг k и представлена в виде

$$\mathbf{A}=\mathbf{PBS}, \tag{7.37}$$

где $\mathbf{P}$ и $\mathbf{S}$ — ортогональные матрицы размера $m\times m$ и $n\times n$ соответственно, а $m\times n$-матрица $\mathbf{B}$ имеет блочную структуру вида

$\mathbf{B}:=\begin{pmatrix}\mathbf{B}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}$ с $k\times k$-подматрицей $\mathbf{B}_{11}$.

Переписав систему (7.34) с учетом факторизации (7.37) как $\mathbf{BSx}=\mathbf{P}^T\mathbf{b}$, придаем ей вид

$$\mathbf{By}=\mathbf{g}, \tag{7.38}$$

где $\mathbf{y}:=\mathbf{Sx}$, $\mathbf{g}:=\mathbf{P}^T\mathbf{b}$. Далее, в соответствии с заданным разбиением на клетки матрицы $\mathbf{B}$ полагаем

$$\mathbf{y}:=\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{pmatrix},\ \mathbf{g}:=\begin{pmatrix}\mathbf{g}_1\\ \mathbf{g}_2\end{pmatrix},\ \text{где}\ \mathbf{y}_1,\mathbf{g}_1\in\mathbb{R}_k,\ \mathbf{y}_2\in\mathbb{R}_{n-k},\ \mathbf{g}_2\in\mathbb{R}_{m-k}.$$

Тогда вместо (7.38) можно записать

$$\begin{pmatrix}\mathbf{B}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\cdot\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{pmatrix}=\begin{pmatrix}\mathbf{g}_1\\ \mathbf{g}_2\end{pmatrix},\ \text{т.е.}\ \begin{cases}\mathbf{B}_{11}\cdot\mathbf{y}_1+\mathbf{0}\cdot\mathbf{y}_2=\mathbf{g}_1,\\ \mathbf{0}\cdot\mathbf{y}_1+\mathbf{0}\cdot\mathbf{y}_2=\mathbf{g}_2.\end{cases}$$

Предположение о том, что $\operatorname{rank}\mathbf{A}=k$ и матрицы $\mathbf{P}$ и $\mathbf{S}$ — ортогональные, в силу (7.37) и клеточного вида матрицы $\mathbf{B}$, позволяет считать, что матрица $\mathbf{B}_{11}$ не вырождена и, значит, система

$$\mathbf{B}_{11}\mathbf{y}_1=\mathbf{g}_1 \tag{7.39}$$

имеет единственное решение, которое обозначим $\mathbf{y}_1^*$.

Покажем, что *вектор*

$$\mathbf{x}^+:=\mathbf{S}^T\mathbf{y}^+:=\mathbf{S}^T\begin{pmatrix}\mathbf{y}_1^*\\ \mathbf{y}_2\end{pmatrix} \tag{7.40}$$

при произвольном $\mathbf{y}_2\in\mathbb{R}_{n-k}$ *является псевдорешением системы* (7.34), *т.е. решением соответствующей ей ЛЗНК.*

Действительно, в силу того что* $\|\mathbf{P}\|_2 = \|\mathbf{P}^T\|_2 = 1$, имеем

$$\|\mathbf{Ax}-\mathbf{b}\|_2^2 = \|\mathbf{PBSx}-\mathbf{b}\|_2^2 = \|\mathbf{BSx}-\mathbf{P}^T\mathbf{b}\|_2^2 =$$

$$= \|\mathbf{By}-\mathbf{g}\|_2^2 = \|\mathbf{B}_{11}\mathbf{y}_1-\mathbf{g}_1\|_2^2 + \|\mathbf{g}_2\|_2^2,$$

откуда следует

$$\min_{\mathbf{x}\in\mathbb{R}_n} \|\mathbf{Ax}-\mathbf{b}\|_2 = \min_{\mathbf{y}\in\mathbb{R}_n} \|\mathbf{By}-\mathbf{g}\|_2 = \|\mathbf{g}_2\|_2$$

при $\mathbf{y}^+ := \begin{pmatrix} \mathbf{y}_1^* \\ \mathbf{y}_2 \end{pmatrix}$ с любым $\mathbf{y}_2 \in \mathbb{R}_{n-k}$.

Очевидно, единственным псевдорешением с наименьшей нормой для системы (7.38) служит псевдорешение $\mathbf{y}^+$ с $\mathbf{y}_2 := \mathbf{0}$, т.е. ее нормальное псевдорешение $\mathbf{y}_0 := \begin{pmatrix} \mathbf{y}_1^* \\ \mathbf{0} \end{pmatrix}$. Соответственно решением ЛЗНК — нормальным псевдорешением системы (7.34) — будет вектор

$$\mathbf{x}_0 := \mathbf{S}^T\mathbf{y}_0 = \mathbf{S}^T \cdot \begin{pmatrix} \mathbf{y}_1^* \\ \mathbf{0} \end{pmatrix}. \qquad (7.41)$$

Заметим, что *вектор невязки один и тот же для любого псевдорешения:*

$$\mathbf{r} := \mathbf{b}-\mathbf{Ax}^+ = \mathbf{Pg}-\mathbf{PBSS}^T\mathbf{y}^+ = \mathbf{P}\cdot\begin{pmatrix}\mathbf{g}_1\\ \mathbf{g}_2\end{pmatrix} - \mathbf{P}\cdot\begin{pmatrix}\mathbf{B}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\cdot\begin{pmatrix}\mathbf{y}_1^*\\ \mathbf{y}_2\end{pmatrix} = \mathbf{P}\cdot\begin{pmatrix}\mathbf{0}\\ \mathbf{g}_2\end{pmatrix};$$

при этом имеет место следующая оценка сверху его нормы:

$$\|\mathbf{r}\|_2 = \|\mathbf{b}-\mathbf{Ax}^+\|_2 \le \|\mathbf{P}\|_2 \cdot \|\mathbf{g}_2\|_2 = \|\mathbf{g}_2\|_2. \qquad (7.42)$$

---

* Из ортогональности $\mathbf{P}$ следует, что $\sigma_{\mathbf{P}} := \sqrt{\lambda_{\mathbf{P}^T\mathbf{P}}} = \sqrt{\lambda_{\mathbf{E}}} = 1$, т.е. равенство единице нормы любой ортогональной матрицы. Попутно обратим внимание на идеальную обусловленность ортогональных матриц: $\operatorname{cond}\mathbf{P} := \dfrac{\max\sigma_{\mathbf{P}}}{\min\sigma_{\mathbf{P}}} = \dfrac{1}{1} = 1$.

Сингулярное разложение (7.1) матрицы $\mathbf{A}$ есть частный случай разложения (7.37). Перенесем на него полученные выше результаты.

Отождествим $m\times n$-матрицу $\mathbf{B}:=\begin{pmatrix}\mathbf{B}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}$ с матрицей такого же размера (см. (7.2))

$$\mathbf{\Sigma}:=\operatorname{diag}(\sigma_1;\sigma_2;\ldots;\sigma_k;0;\ldots;0):=\begin{pmatrix}\operatorname{diag}(\sigma_1;\sigma_2;\ldots;\sigma_k) & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix},$$

а также ортогональные $m\times m$-матрицу $\mathbf{P}$ с $\mathbf{U}$, $n\times n$-матрицу $\mathbf{S}$ с $\mathbf{V}$. Правую часть в промежуточной системе (7.38) выражаем через матрицу $\mathbf{U}$ разложения (7.1) и вектор $\mathbf{b}$ исходной системы (7.34):

$$\mathbf{g}=\mathbf{U}^T\mathbf{b}:=\begin{pmatrix}u_{11} & u_{21} & \cdots & u_{m1}\\ u_{12} & u_{22} & \cdots & u_{m2}\\ \cdots & \cdots & \cdots & \cdots\\ u_{1m} & u_{2m} & \cdots & u_{mm}\end{pmatrix}\cdot\begin{pmatrix}b_1\\ b_2\\ \cdots\\ b_m\end{pmatrix}=\begin{pmatrix}u_{11}b_1+u_{21}b_2+\ldots+u_{m1}b_m\\ u_{12}b_1+u_{22}b_2+\ldots+u_{m2}b_m\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ u_{1m}b_1+u_{2m}b_2+\ldots+u_{mm}b_m\end{pmatrix};$$

иначе, в блочном виде,

$$\mathbf{g}:=\begin{pmatrix}\mathbf{g}_1\\ \mathbf{g}_2\end{pmatrix}:=\begin{pmatrix}\begin{pmatrix}u_{11}b_1+u_{21}b_2+\ldots+u_{m1}b_m\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ u_{1k}b_1+u_{2k}b_2+\ldots+u_{mk}b_m\end{pmatrix}\\ \begin{pmatrix}u_{1,k+1}b_1+u_{2,k+1}b_2+\ldots+u_{m,k+1}b_m\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ u_{1m}b_1+u_{2m}b_2+\ldots+u_{mm}b_m\end{pmatrix}\end{pmatrix}.$$

Невырожденная система (7.39) в таком случае — это

$$\begin{pmatrix}\sigma_1 & 0 & \cdots & 0\\ 0 & \sigma_2 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & \cdots & \sigma_k\end{pmatrix}\cdot\begin{pmatrix}y_1\\ y_2\\ \cdots\\ y_k\end{pmatrix}=\begin{pmatrix}u_{11}b_1+u_{21}b_2+\ldots+u_{m1}b_m\\ u_{12}b_1+u_{22}b_2+\ldots+u_{m2}b_m\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ u_{1k}b_1+u_{2k}b_2+\ldots+u_{mk}b_m\end{pmatrix}$$

с единственным решением

$$\mathbf{y}_1^* := \begin{pmatrix} y_1^* \\ y_2^* \\ \cdots \\ y_k^* \end{pmatrix} := \begin{pmatrix} \frac{1}{\sigma_1}\sum_{i=1}^{m} u_{i1} b_i \\ \frac{1}{\sigma_2}\sum_{i=1}^{m} u_{i2} b_i \\ \cdots \\ \frac{1}{\sigma_k}\sum_{i=1}^{m} u_{ik} b_i \end{pmatrix}. \qquad (7.43)$$

Тогда в соответствии с формулой (7.40) псевдорешение системы (7.34) ЛЗНК есть вектор

$$\mathbf{x}^+ := \mathbf{V}^T \begin{pmatrix} \mathbf{y}_1^* \\ \mathbf{y}_2 \end{pmatrix} := \begin{pmatrix} v_{11} & v_{21} & \cdots & v_{k1} & v_{k+1,1} & \cdots & v_{n1} \\ v_{12} & v_{22} & \cdots & v_{k2} & v_{k+1,2} & \cdots & v_{n2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ v_{1n} & v_{2n} & \cdots & v_{kn} & v_{k+1,n} & \cdots & v_{nn} \end{pmatrix} \cdot \begin{pmatrix} \begin{pmatrix} y_1^* \\ \cdots \\ y_k^* \end{pmatrix} \\ \begin{pmatrix} y_{k+1} \\ \cdots \\ y_n \end{pmatrix} \end{pmatrix} =$$

$$= \begin{pmatrix} \sum_{i=1}^{k} v_{i1} y_i^* + \sum_{i=k+1}^{n} v_{i1} y_i \\ \sum_{i=1}^{k} v_{i2} y_i^* + \sum_{i=k+1}^{n} v_{i2} y_i \\ \cdots \\ \sum_{i=1}^{k} v_{in} y_i^* + \sum_{i=k+1}^{n} v_{in} y_i \end{pmatrix}$$

со значениями $y_1^*, y_2^*, \ldots, y_k^*$ из (7.43) и произвольными значениями $y_{k+1}, y_{k+2}, \ldots, y_n$. Решением ЛЗНК с наименьшей евклидовой нормой, т.е. нормальным псевдорешением системы (7.34) будет вектор (см. (7.41))

$$\mathbf{x}_0 := \begin{pmatrix} \sum_{i=1}^{k} v_{i1} y_i^* \\ \sum_{i=1}^{k} v_{i2} y_i^* \\ \cdots \\ \sum_{i=1}^{k} v_{in} y_i^* \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{k} v_{i1} \cdot \frac{1}{\sigma_i} \sum_{j=1}^{m} u_{ji} b_j \\ \sum_{i=1}^{k} v_{i2} \cdot \frac{1}{\sigma_i} \sum_{j=1}^{m} u_{ji} b_j \\ \cdots \\ \sum_{i=1}^{k} v_{in} \cdot \frac{1}{\sigma_i} \sum_{j=1}^{m} u_{ji} b_j \end{pmatrix}. \qquad (7.44)$$

Невязкой любого псевдорешения $\mathbf{x}^+$ (в том числе и нормального $\mathbf{x}_0$) служит вектор

$$\mathbf{r} := \mathbf{b} - \mathbf{A}\mathbf{x}^+ = \mathbf{U} \cdot \begin{pmatrix} \mathbf{0} \\ \mathbf{g}_2 \end{pmatrix} =$$

$$= \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1k} & u_{1,k+1} & \cdots & u_{1m} \\ u_{21} & u_{22} & \cdots & u_{2k} & u_{2,k+1} & \cdots & u_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ u_{m1} & u_{m2} & \cdots & u_{mk} & u_{m,k+1} & \cdots & u_{mm} \end{pmatrix} \cdot \begin{pmatrix} \begin{pmatrix} 0 \\ \cdots \\ 0 \end{pmatrix} \\ \begin{pmatrix} u_{1,k+1} b_1 + \ldots + u_{m,k+1} b_m \\ \cdots \\ u_{1m} b_1 + \ldots + u_{mm} b_m \end{pmatrix} \end{pmatrix} =$$

$$= \begin{pmatrix} u_{1,k+1} \sum_{i=1}^{m} u_{i,k+1} b_i + \ldots + u_{1m} \sum_{i=1}^{m} u_{im} b_i \\ u_{2,k+1} \sum_{i=1}^{m} u_{i,k+1} b_i + \ldots + u_{2m} \sum_{i=1}^{m} u_{im} b_i \\ \cdots \\ u_{m,k+1} \sum_{i=1}^{m} u_{i,k+1} b_i + \ldots + u_{mm} \sum_{i=1}^{m} u_{im} b_i \end{pmatrix} = \begin{pmatrix} \sum_{j=k+1}^{m} u_{1j} \sum_{i=1}^{m} u_{ij} b_i \\ \sum_{j=k+1}^{m} u_{2j} \sum_{i=1}^{m} u_{ij} b_i \\ \cdots \\ \sum_{j=k+1}^{m} u_{mj} \sum_{i=1}^{m} u_{ij} b_i \end{pmatrix}.$$

Согласно (7.42) норма невязки может быть оценена величиной

$$\|\mathbf{g}_2\|_2 = \sqrt{\sum_{j=k+1}^{m} \left(\sum_{i=1}^{m} u_{ij} b_i \right)^2}. \qquad (7.45)$$

Оценим снизу нормальное псевдорешение (7.44). Имеем:

$$\|\mathbf{x}_0\|_2 = \left\| \mathbf{V}^T \begin{pmatrix} \mathbf{y}_1^* \\ \mathbf{0} \end{pmatrix} \right\|_2 \geq \left\| \mathbf{V}^T \begin{pmatrix} 0 \\ 0 \\ \cdots \\ y_k^* \\ \mathbf{0} \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} v_{k1} y_k^* \\ v_{k2} y_k^* \\ \cdots \\ v_{kn} y_k^* \end{pmatrix} \right\|_2 = \left| y_k^* \right| \cdot \left\| \begin{pmatrix} v_{k1} \\ v_{k2} \\ \cdots \\ v_{kn} \end{pmatrix} \right\|_2 = \left| y_k^* \right|$$

(см. (7.41), (7.43)). Таким образом, приходим к неравенству

$$\|\mathbf{x}_0\|_2 \geq \left| y_k^* \right| := \frac{1}{\sigma_k} \left| \sum_{i=1}^{m} u_{ik} b_i \right| , \qquad (7.46)$$

из которого видим, что при $\sigma_k \to 0$ нормальное псевдорешение может неограниченно расти (что вполне ожидаемо, в силу $\operatorname{cond} \mathbf{A} := \dfrac{\sigma_1}{\sigma_k} \xrightarrow[\sigma_k \to 0]{} \infty$).

Предположим, что имеет место ситуация, когда

$$\sigma_1 \geq \ldots \geq \sigma_{k-1} \gg \sigma_k \approx 0 ,$$

т.е. наименьшее сингулярное число σ_k, индекс которого определяет ранг матрицы $\mathbf{A}$ системы (7.34), очень близко к нулю со всеми вытекающими отсюда последствиями (нечетким установлением значения ранга, плохой обусловленностью, неустойчивым вычислением $\mathbf{x}_0$). В таком случае можно поступить следующим образом.

Заменим малое значение σ_k нулем. Тогда ранг матрицы $\mathbf{A}$ понизится на единицу, а число обусловленности (теперь это σ_1 / σ_{k-1}) уменьшится. Разумеется, то и другое относится уже к матрице $\tilde{\mathbf{A}}$, близкой к данной. Пусть указанная замена $\sigma_k := 0$, произведенная в *процессе* SVD-разложения, привела в итоге к равенству

$$\mathbf{A} \approx \tilde{\mathbf{A}} = \tilde{\mathbf{U}} \tilde{\mathbf{\Sigma}} \tilde{\mathbf{V}} ,$$

в котором обозначаем $\tilde{\mathbf{\Sigma}} := \mathrm{diag}(\sigma_1; \sigma_2; \ldots; \sigma_{k-1}; 0; \ldots; 0)$ и полагаем $\tilde{\mathbf{U}} :\approx \mathbf{U}$, $\tilde{\mathbf{V}} :\approx \mathbf{V}$.

Делая допущение, что подмена одного, близкого к нулю значения σ_k, нулем коснется только соответствующих ему правых и левых сингулярных векторов, в соответствии с формулами (7.44), (7.46), (7.45) получаем следующее:

нормальное псевдорешение

$$\tilde{\mathbf{x}}_0 := \begin{pmatrix} \sum_{i=1}^{k-1} v_{i1} y_i^* \\ \ldots \\ \sum_{i=1}^{k-1} v_{in} y_i^* \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{k-1} v_{i1} \cdot \frac{1}{\sigma_i} \sum_{j=1}^{m} u_{ji} b_j \\ \ldots \\ \sum_{i=1}^{k-1} v_{in} \cdot \frac{1}{\sigma_i} \sum_{j=1}^{m} u_{ji} b_j \end{pmatrix}$$

с оценкой снизу

$$\left\| \tilde{\mathbf{x}}_0 \right\|_2 \geq \left| y_{k-1}^* \right| = \frac{1}{\sigma_{k-1}} \left| \sum_{i=1}^{m} u_{i,k-1} b_i \right|$$

и невязку

$$\tilde{\mathbf{r}} := \begin{pmatrix} \sum_{j=k}^{m} u_{1j} \sum_{i=1}^{m} u_{ij} b_i \\ \ldots \\ \sum_{j=k}^{m} u_{mj} \sum_{i=1}^{m} u_{ij} b_i \end{pmatrix}$$

с оценкой сверху

$$\left\| \tilde{\mathbf{r}} \right\|_2 \leq \left\| \mathbf{g}_2 \right\|_2 = \sqrt{\sum_{j=k}^{m} \left(\sum_{i=1}^{m} u_{ij} b_i \right)^2} \ .$$

Анализируя последние результаты, можно резюмировать, что при сознательной замене нулем малого сингулярного числа (сравнимого с точностью задания исходных данных и/или точностью

арифметических вычислений) снижается количество информации, участвующей в получении нормального псевдорешения $\mathbf{x}_0$ — изымается левый сингулярный вектор, соответствующий отбрасываемому сингулярному числу σ_k; однако за счет уменьшения числа обусловленности это решение находится более устойчиво и граница его нормы снижается. В то же время в выражении невязки (и в оценке ее нормы) появляется дополнительное слагаемое, привлекающее соответствующий отбрасываемому сингулярному числу правый сингулярный вектор, что означает возможное увеличение рассогласованности системы на полученном таким способом решении.

Ясно, что в процессе решения ЛЗНК подобным методом в целях повышения устойчивости можно заметно занижать ранг, заменяя нулями целые группы существенно малых сингулярных чисел (меньших некоторого наперед задаваемого допуска — параметра метода, выбор подходящего значения которого зависит от ряда конкретных обстоятельств). SVD-разложение с искусственным занижением числа ненулевых сингулярных чисел называют ***усеченным сингулярным разложением*** [26].

Обратившись к началу § 1.4, понимаем, что при построении ортогональной матрицы, посредством которой осуществляется один шаг преобразования при **QR**-факторизации квадратной матрицы, имеет значение лишь длина вектора-столбца преобразуемой матрицы, а число столбцов сказывается лишь на числе элементарных шагов. Поэтому описанное в § 1.4, 1.5 **QR**-разложение квадратных матриц тривиально переносится на случай прямоугольных матриц.

Пусть в исходной $m \times n$-матрице $\mathbf{A} := (a_{ij})$

$$m > n \quad \text{и} \quad k := \operatorname{rank} \mathbf{A} = n \ . \tag{7.47}$$

Выполнив ее **QR**-факторизацию с помощью преобразований Хаусхолдера или Гивенса с результирующими матрицами $\mathbf{Q} := (q_{ij})$ размера $m \times m$ и $\mathbf{R} := (r_{ij})$ размера $m \times n$, запишем равенство

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nn} \\ a_{n+1,1} & \cdots & a_{n+1,n} \\ \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} = \begin{pmatrix} q_{11} & \cdots & q_{1n} & q_{1,n+1} & \cdots & q_{1m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ q_{n1} & \cdots & q_{nn} & q_{n,n+1} & \cdots & q_{nm} \\ q_{n+1,1} & \cdots & q_{n+1,n} & q_{n+1,n+1} & \cdots & q_{n+1,m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ q_{m1} & \cdots & q_{mn} & q_{m,n+1} & \cdots & q_{mm} \end{pmatrix} \begin{pmatrix} r_{11} \cdots r_{1n} \\ \cdots\cdots\cdots \\ 0 \ldots r_{nn} \\ \cdots\cdots\cdots \\ 0 \ldots 0 \\ 0 \ldots 0 \end{pmatrix} \begin{pmatrix} 1 \cdots 0 \\ \vdots \ddots \vdots \\ 0 \cdots 1 \end{pmatrix}.$$

Ортогональность матрицы $\mathbf{Q}$ и формально добавленной единичной матрицы $\mathbf{E}$ размера $n \times n$ позволяет считать это равенство разложением вида (7.37), в котором

$$\mathbf{P} := \mathbf{Q}, \quad \mathbf{B} := \mathbf{R}, \quad \mathbf{S} := \mathbf{E},$$

причем в предположении (7.37) матрица $\mathbf{B}$ здесь состоит только из двух блоков с существенным блоком $\mathbf{B}_{11} := \mathbf{R}_{11} := (r_{ij})_{i,j=1}^{n}$.

В таком случае система ЛЗНК (7.34) представляется как

$$\mathbf{QREx} = \mathbf{b}, \tag{7.48}$$

и, значит, ищется n-мерный вектор $\mathbf{x}$ такой, что

$$\mathbf{Rx} = \mathbf{Q}^T\mathbf{b}. \tag{7.49}$$

После соответствующего разбиения m-мерного вектора $\mathbf{Q}^T\mathbf{b}$ на блоки $\mathbf{g}_1 \in \mathbb{R}_n$, $\mathbf{g}_2 \in \mathbb{R}_{m-n}$ приходим к системе

$$\begin{cases} \mathbf{R}_{11}\mathbf{x} = \mathbf{g}_1, \\ \mathbf{0}\,\mathbf{x} = \mathbf{g}_2, \end{cases} \tag{7.50}$$

откуда легко получаем нормальное псевдорешение

$$\mathbf{x}_0^+ = \mathbf{R}_{11}^{-1}\mathbf{g}_1$$

(если все $r_{ii} \neq 0$) с оценкой невязки

$$\left\|\mathbf{b} - \mathbf{A}\mathbf{x}_0^+\right\|_2 \leq \|\mathbf{g}_2\|_2.$$

Если $m \times n$-матрица $\mathbf{A}$ при $m > n$ не является полноранговой, т.е. в отличие от (7.47) $k := \operatorname{rank} \mathbf{A} < n$, то матрица $\mathbf{R}$ в представлении системы (7.48) имеет вид

$$\mathbf{R} := \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

где $\mathbf{R}_{11} := \begin{pmatrix} r_{11} & \cdots & r_{1k} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & r_{kk} \end{pmatrix}$ (с $r_{ii} \neq 0$), а $\mathbf{R}_{12} := \begin{pmatrix} r_{1,k+1} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{k,k+1} & \cdots & r_{kn} \end{pmatrix}$.

Полагая $\mathbf{x} := \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$, $\mathbf{Q}^T\mathbf{b} := \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{pmatrix}$, где $\mathbf{x}_1, \mathbf{g}_1 \in \mathbb{R}_k$, $\mathbf{x}_2 \in \mathbb{R}_{n-k}$, $\mathbf{g}_2 \in \mathbb{R}_{m-k}$, получаем систему

$$\begin{cases} \mathbf{R}_{11}\mathbf{x}_1 + \mathbf{R}_{12}\mathbf{x}_2 = \mathbf{g}_1, \\ \mathbf{0}\mathbf{x}_1 + \mathbf{0}\mathbf{x}_2 = \mathbf{g}_2, \end{cases}$$

из которой находим вектор

$$\mathbf{x}_1^* := \mathbf{R}_{11}{}^{-1}\left(\mathbf{g}_1 - \mathbf{R}_{12}\mathbf{x}_2\right).$$

Тогда псевдорешением для данной задачи может служить вектор $\mathbf{x}^+ := \begin{pmatrix} \mathbf{x}_1^* \\ \mathbf{x}_2 \end{pmatrix}$ с произвольной составляющей $\mathbf{x}_2 \in \mathbb{R}_{n-k}$, а ее нормальным псевдорешением — вектор $\mathbf{x}_0 := \begin{pmatrix} \mathbf{x}_1^* \\ \mathbf{0} \end{pmatrix}$.

Пример 7.4 (другое решение ЛЗНК (7.30) примера 7.3).

Система (7.30), записанная в векторно-матричном виде (7.34), имеет матрицу $\mathbf{A}$ и вектор $\mathbf{b}$ соответственно

$$\mathbf{A} := \begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 1 & 3 \end{pmatrix}, \quad \mathbf{b} := \begin{pmatrix} 8{,}5 \\ 6{,}7 \\ 11{,}1 \end{pmatrix}.$$

Применим к поиску нормального псевдорешения этой системы последний подход.

Сначала выполним **QR**-разложение матрицы **A**, для чего воспользуемся преобразованиями Гивенса (§ 1.5). Имеем:

$$c_1 = \frac{1}{\sqrt{1+4}} = \frac{1}{\sqrt{5}}, \quad s_1 = \frac{2}{\sqrt{5}}, \quad \mathbf{G}_1 = \begin{pmatrix} 1/\sqrt{5} & 2/\sqrt{5} & 0 \\ -2/\sqrt{5} & 1/\sqrt{5} & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{A}_1 := \mathbf{G}_1\mathbf{A} = \begin{pmatrix} \sqrt{5} & 4/\sqrt{5} \\ 0 & -3/\sqrt{5} \\ 1 & 3 \end{pmatrix};$$

$$c_2 = \frac{\sqrt{5}}{\sqrt{5+1}} = \frac{\sqrt{5}}{\sqrt{6}}, \quad s_2 = \frac{1}{\sqrt{6}}, \quad \mathbf{G}_2 = \begin{pmatrix} \sqrt{5}/\sqrt{6} & 0 & 1/\sqrt{6} \\ 0 & 1 & 0 \\ -1/\sqrt{6} & 0 & \sqrt{5}/\sqrt{6} \end{pmatrix},$$

$$\mathbf{A}_2 := \mathbf{G}_2\mathbf{A}_1 = \begin{pmatrix} \sqrt{6} & 7/\sqrt{6} \\ 0 & -3/\sqrt{5} \\ 0 & 11/\sqrt{30} \end{pmatrix};$$

$$c_3 = \frac{-3/\sqrt{5}}{\sqrt{9/5+121/30}} = -\frac{3\sqrt{6}}{5\sqrt{7}}, \quad s_3 = \frac{11/\sqrt{30}}{\sqrt{175/30}} = \frac{11}{5\sqrt{7}},$$

$$\mathbf{G}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -3\sqrt{6}/(5\sqrt{7}) & 11/(5\sqrt{7}) \\ 0 & -11/(5\sqrt{7}) & -3\sqrt{6}/(5\sqrt{7}) \end{pmatrix}, \quad \mathbf{R} := \mathbf{A}_3 := \mathbf{G}_3\mathbf{A}_2 = \begin{pmatrix} \sqrt{6} & 7/\sqrt{6} \\ 0 & \sqrt{35/6} \\ 0 & 0 \end{pmatrix}.$$

При этом

$$\mathbf{G}_2\mathbf{G}_1 = \begin{pmatrix} 1/\sqrt{6} & 2/\sqrt{6} & 1/\sqrt{6} \\ -2/\sqrt{5} & 1/\sqrt{5} & 0 \\ -1/\sqrt{30} & -2/\sqrt{30} & 5/\sqrt{30} \end{pmatrix},$$

$$\mathbf{Q} := (\mathbf{G}_2\mathbf{G}_1)^T\mathbf{G}_3^T = \begin{pmatrix} 1/\sqrt{6} & \sqrt{5/42} & \sqrt{5/7} \\ 2/\sqrt{6} & -8/\sqrt{210} & -1/\sqrt{35} \\ 1/\sqrt{6} & 11/\sqrt{210} & -3/\sqrt{35} \end{pmatrix}.$$

Подсчитаем теперь правую часть ортогонально преобразованной системы вида (7.49):

$$\mathbf{Q}^T\mathbf{b} = \begin{pmatrix} 1/\sqrt{6} & 2/\sqrt{6} & 1/\sqrt{6} \\ \sqrt{5/42} & -8/\sqrt{210} & 11/\sqrt{210} \\ \sqrt{5/7} & -1/\sqrt{35} & -3/\sqrt{35} \end{pmatrix} \cdot \begin{pmatrix} 8,5 \\ 6,7 \\ 11,1 \end{pmatrix} = \begin{pmatrix} 33/\sqrt{6} \\ 111/\sqrt{210} \\ 2.5/\sqrt{35} \end{pmatrix}.$$

Следовательно, требуемое нормальное псевдорешение можно найти из треугольной системы

$$\begin{cases} \sqrt{6}\,x_1 + \dfrac{7}{\sqrt{6}}\,x_2 = \dfrac{33}{\sqrt{6}}, \\ \sqrt{\dfrac{35}{6}}\,x_2 = \dfrac{111}{\sqrt{210}} \end{cases} \tag{7.51}$$

(см. (7.50)) с оценкой евклидовой нормы невязки $\|\mathbf{Ax} - \mathbf{b}\|_2 \le \dfrac{5}{2\sqrt{35}} :\approx 0,42$.

Решив систему (7.51), получаем решение поставленной ЛЗНК, совпадающее с найденным ранее с помощью симметризации.

Замечание 7.4. Не исключено, что в процессе **QR**-разложения даже в случае, когда $m \times n$-матрица **A** является полноранговой, на каком-то шаге преобразований окажется $r_{ii} = 0$ (см. замечание 1.3 и, в подтверждение ему, пример 1.3). Чтобы избежать этого, очевидно, при написании работоспособного алгоритма следует предусмотреть возможность выполнения перестановок хотя бы только строк матрицы, реализуя принцип частичного выбора главного элемента в преобразуемой части ведущего столбца подобно тому, как это делается в схеме единственного деления (§ 2.1). Более подробно об этом можно прочесть в монографии [23].

Учитывая свойство ортогональных преобразований сохранять евклидову норму преобразуемых векторов (см. замечание 1.6), использование **QR**-разложения может дать более точное решение ЛЗНК, чем, например, применение метода Холецкого; тем более что здесь преобразования совершаются над исходной матрицей системы (7.34), а не над симметризованной. Однако, по крайней мере, в случае неполноранговых матриц можно рассчитывать на более хороший результат при использовании сингулярного разложения.

УПРАЖНЕНИЯ

7.1. Найдите множество решений и нормальное (относительно нулевого вектора) решение системы

$$\begin{cases} 2x_1 + 3x_2 + x_3 = 13, \\ x_1 + x_2 + x_3 = 6, \\ 3x_1 + 5x_2 + x_3 = 20. \end{cases}$$

7.2. Методом сингулярного разложения решите систему $\begin{cases} x_1 + 3x_2 = -2, \\ -2x_1 + x_2 = 3, \\ 2x_1 - x_2 = -1 \end{cases}$.

7.3. Пользуясь сингулярным разложением матрицы $\mathbf{A} := \begin{pmatrix} 2 & -1 & -2 \\ 0 & 3 & 1 \\ -1 & 1 & 0 \\ 2 & 0 & -1 \end{pmatrix}$, полученным в примере 6.4 (§ 6.4), найдите:

а) псевдообратную матрицу $\mathbf{A}^+$;

б) число обусловленности $\operatorname{cond} \mathbf{A}$ (двумя способами);

в) нормальное псевдорешение системы

$$\mathbf{Ax} = \mathbf{b} \quad \text{при} \quad \mathbf{b} := (1; 2; -1; -2)^T$$

(двумя способами). Что можно сказать об общем решении данной системы?

7.4. Запишите алгоритм решения нелинейной системы $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ с $\mathbf{F} : \mathbb{R}_m \to \mathbb{R}_n$ (где $m \geq n$) каким-либо вариантом метода секущих ([12, 13]). Пользуясь этим, найдите нормальное псевдорешение системы примера 7.2 (§ 7.4).

7.5. Найдите и изобразите графически нормальное псевдорешение системы

$$\begin{cases} 2x + 3y = 2,4, \\ 3x + 4y = 3,3, \\ 4x + 5y = 4,3. \end{cases}$$

Глава 8

ФАКТОРЫ, ВЛИЯЮЩИЕ НА ВЫБОР МЕТОДА

§ 8.1. АРИФМЕТИЧЕСКАЯ СЛОЖНОСТЬ МЕТОДА

При выборе метода численного решения той или иной алгебраической задачи в условиях, когда эта задача имеет умеренную размерность (по отношению к используемому вычислительному средству) и достаточно хорошо обусловлена (об этом см. далее § 8.3), наиболее существенную роль играет число арифметических операций, требуемых для получения решения. Особенно важно это при решении больших серий однотипных задач, что обычно и имеет место. Что касается прямых численных методов, то эта характеристика зависит лишь от размерности решаемой задачи и может быть заранее подсчитана; точность получаемого решения при этом зависит от того, по какому закону происходит накопление ошибок округления, т.е. от вычислительной устойчивости метода. Для итерационных методов можно заранее подсчитать арифметические затраты только на один шаг итерации, а общее число требуемых арифметических операций определяется заказываемой точностью и требуемым числом итерационных шагов для ее достижения, что зависит не только от используемого метода (его скорости сходимости), но и от свойств самой задачи. Поэтому проводить корректно теоретическое сравнение по вычислительным затратам прямых методов с итерационными не всегда возможно.

Число требуемых арифметических операций для реализации метода (иначе, ***арифметическая сложность метода***) есть функция размерности решаемой задачи, представляющая собой, как правило, линейную комбинацию степенных функций. Можно встретить три подхода к подсчету (и учету) арифметической сложности: 1) подсчет всех арифметических операций; 2) подсчет только мультипликативных операций (умножений и делений) как более «дорогих»; 3) нахождение только старшего члена упомянутой функции затрат для отражения асимптотического поведения

арифметических затрат с ростом размерности задачи. Отдельно или вместе с мультипликативными операциями учитывается нередко встречающаяся трансцендентная операция извлечения квадратного корня.

Следует иметь в виду, что объем арифметических затрат, указываемый в разных литературных источниках, может несколько различаться. Это касается, прежде всего, более-менее сложных методов, допускающих разные алгоритмические реализации.

Ниже приведены данные об арифметической сложности многих из рассмотренных здесь методов. Часть этих результатов сопровождается демонстрацией соответствующих расчетов, другие даны без вывода, но, как правило, с надлежащими ссылками.

Наиболее легко подсчитать арифметические затраты, необходимые для реализации метода прогонки, поскольку этот способ решения СЛАУ с трехдиагональными матрицами коэффициентов определяется всего несколькими тривиальными расчетными формулами, не предполагающими вложения циклов.

Так, классический метод правой прогонки решения n-мерной системы вида (2.20) требует (см. формулы (2.21), (2.23)):

для вычисления знаменателей прогоночных коэффициентов

$$\Delta_i := c_i + b_i\delta_{i-1} \quad \text{при } i = 2, 3, \ldots, n$$

$n-1$ аддитивных и $n-1$ мультипликативных операций;

для вычисления самих прогоночных коэффициентов

$$\delta_1 := -\frac{d_1}{c_1}, \qquad \delta_i := -\frac{d_i}{\Delta_i} \quad \text{при } i = 2, 3, \ldots, n-1,$$

$$\lambda_1 := \frac{r_1}{c_1}, \qquad \lambda_i := \frac{r_i - b_i\lambda_{i-1}}{\Delta_i} \quad \text{при } i = 2, 3, \ldots, n$$

$n-1$ аддитивных и $3n-2$ мультипликативных операций;

для нахождения компонент вектора-решения

$$x_n := \lambda_n, \quad x_i := \delta_i x_{i+1} + \lambda_i \quad \text{при } i = n-1, \ldots, 2, 1$$

$n-1$ аддитивных и $n-1$ мультипликативных операций.

Итого для получения решения СЛАУ (2.20) методом правой прогонки нужно выполнить $8n-7$ арифметических операций,

из них мультипликативных $5n-4$ и аддитивных $3n-3$. Очевидно, абсолютно такой же объем вычислений требуется и для реализации метода левой прогонки. Такой же порядок арифметической сложности $O(8n)$ имеет и метод встречных прогонок*. Немонотонная прогонка в худшем случае потребует $O(12n)$ арифметических действий [62], а циклическая — $O(14n)$.

Рассмотрим теперь вопрос об арифметической сложности решения n-мерных СЛАУ с заполненными матрицами коэффициентов методом Гаусса. Будем вести подсчет числа затрачиваемых арифметических операций на реализацию описанной в § 2.1 схемы единственного деления.

Анализируя процесс вывода итоговых формул (2.4), (2.5), по которым можно получить решение системы (2.1), последовательно выясняем, что преобразование исходной системы (2.1) к виду (2.2) требует $n-1$ делений, $n(n-1)$ умножений, $n(n-1)$ сложений (вычитаний) — итого $(2n+1)(n-1)$ арифметических операций. Пользуясь этим, подсчитываем число операций, нужное для преобразования системы (2.2) к трапециевидной форме (2.3)**:

$$\sum_{k=2}^{n}(2k+1)(k-1)=\sum_{i=1}^{n-1}i(2i+3)=2\sum_{i=1}^{n-1}i^2+3\sum_{i=1}^{n-1}i=$$

$$=2\cdot\frac{1}{6}(n-1)n(2n-1)+3\cdot\frac{1}{2}(n-1)n=\frac{2}{3}n^3+\frac{1}{2}n^2-\frac{7}{6}n\ .$$

На обратный ход метод Гаусса затрачивает n делений, умножений $1+2+...+(n-1)=(n-1)n/2$, аддитивных операций тоже $(n-1)n/2$ — итого n^2 операций.

---

* Запись $O(cn^p)$, используемая для обозначения ***порядка арифметической сложности*** того или иного численного метода, указывает на значение *старшего члена* в формуле, характеризующей число требуемых арифметических операций.

** Наряду с формулой суммы членов арифметической прогрессии здесь применяется известная формула (см. [19, 24] и др.) суммы квадратов конечной последовательности натуральных чисел $\sum_{i=1}^{k}i^2=\frac{1}{6}k(k+1)(2k+1)$.

Складывая результаты затрат на прямой и обратный ходы, приходим к выводу, что арифметическая сложность схемы единственного деления, реализующей метод Гаусса, определяется формулой $Q := \frac{2}{3}n^3 + \frac{3}{2}n^2 - \frac{7}{6}n$, причем аддитивных и мультипликативных операций здесь примерно поровну.

При $n \to \infty$ бесконечно большая величина $\frac{2}{3}n^3 + \frac{3}{2}n^2 - \frac{7}{6}n$ эквивалентна величине $2n^3/3$, поэтому можно говорить, что арифметическая сложность метода Гаусса есть $O(2n^3/3)$. О том, что арифметические затраты на реализацию метода составляют величину порядка n^3, грубо можно судить по максимальному числу вложений циклов в соответствующем ему алгоритме (см. § 2.1).

Затраты на вычисление определителя n-го порядка методом Гаусса имеют ту же асимптотику $O(2n^3/3)$; обращение $n \times n$-матрицы с использованием схемы единственного деления потребует $O(8n^3/3)$ арифметических операций [19].

Ясно, что решение симметричных СЛАУ методом квадратных корней требует вдвое меньшего числа арифметических операций ($O(n^3/3)$), чем непосредственное применение метода Гаусса, но при этом нужно выполнить еще n трансцендентных операций извлечения корня.

Замечание 8.1. При анализе вычислительной эффективности численных методов линейной алгебры не следует забывать, что кроме арифметических операций метод может требовать выполнения того или иного числа сравнений элементов. Поскольку обычно машинное сравнение по времени выполнения можно приравнять к аддитивной операции, то присутствие в численном методе большого числа сравнений может существенно отразиться на его реальной эффективности. Например, постолбцовый выбор ведущего элемента в методе Гаусса (§ 2.1) требует $n(n-1)/2$ сравнений, что не нарушает асимптотики основного метода Гаусса $O(2n^3/3)$, в то время как методу главных элементов, предполагающему поиск наибольшего по модулю элемента во всей матрице, нужно $(n-1)n(2n-1)/6 = O(n^3/3)$ сравнений, и значит, при оценке его вычислительной эффективности следует исходить из асимптотики $O(n^3)$.

Если сравнения, привлекаемые в прямом методе, служат для повышения точности результата и устойчивости процесса его получения, то в итерационном методе основная цель таких сравнений — ускорение сходимости приближений к результату заданной точности. Так, один шаг метода вращений Якоби нахождения всех собственных чисел симметричной $n \times n$ - матрицы (§ 4.4) требует $8n+5$ мультипликативных, $4n+3$ аддитивных и 3 операции извлечения корня, т.е. $O(12n)$ операций независимо от того, о какой модификации этого метода идет речь. Но если в циклическом методе вращений сравнения не нужны, то в классическом варианте при поиске оптимального «обреченного» элемента делается $n(n-1)/2 = O(n^2/2)$ сравнений, что существенно «удорожает» итерационный шаг, значительно уменьшая при этом число итераций. Компромисс — циклический метод вращений Якоби с барьерами.

Убедимся теперь, что другая реализация метода Гаусса — компактная схема, опирающаяся на $\mathbf{LU}$-разложение, имеет точно такую же арифметическую сложность, как и схема единственного деления.

Подсчитаем сначала вычислительные затраты на выполнение треугольной факторизации $n \times n$-матрицы по формулам (1.2)–(1.3). Будем при этом ориентироваться на заполнение представленной в (1.1) матрицы $\mathbf{L}+\mathbf{U}-\mathbf{E}$ компактного хранения разложения в оговоренном ранее порядке: строка, столбец, строка,.... Очевидно, на вычисление элементов строк (по формуле (1.2)) нужно затратить

$$1\cdot(n-1)+2\cdot(n-2)+\ldots+(n-2)\cdot 2+(n-1)\cdot 1=\sum_{k=1}^{n-1}k(n-k)$$

мультипликативных и столько же аддитивных операций, а на вычисление элементов столбцов (по формуле (1.3)) —

$$\sum_{k=1}^{n-1}k(n-k) \text{ и } 1\cdot(n-2)+2\cdot(n-3)+\ldots+(n-2)\cdot 1=\sum_{k=1}^{n-2}k(n-k-1)$$

мультипликативных и аддитивных операций соответственно. Итого, $\mathbf{LU}$-разложение $n \times n$-матрицы требует следующее число операций:

$$3\sum_{k=1}^{n-1}k(n-k)+\sum_{k=1}^{n-2}k(n-k-1)=$$

$$= 3n\sum_{k=1}^{n-1} k + (n-1)\sum_{k=1}^{n-2} k - 3\sum_{k=1}^{n-1} k^2 - \sum_{k=1}^{n-2} k^2 =$$

$$= \frac{3}{2}n^2(n-1) + \frac{1}{2}(n-1)^2(n-2) - \frac{1}{2}(n-1)n(2n-1) -$$

$$- \frac{1}{6}(n-2)(n-1)(2n-3) = \frac{2}{3}n^3 - \frac{1}{2}n^2 - \frac{1}{6}n .$$

Если сюда добавить $2\sum_{k=1}^{n-1} k = n^2 - n$ операций, которые нужно затратить на преобразование правой части по формуле (2.8) при решении системы (2.1) с помощью **LU**-разложения, то в итоге на прямой ход метода Гаусса таким способом будет затрачено $\frac{2}{3}n^3 + \frac{1}{2}n^2 - \frac{7}{6}n$ операций, как и в схеме единственного деления.

Обратный ход (с затратами n^2 операций) одинаков не только в разных реализациях метода Гаусса, но и в других прямых методах решения СЛАУ с заполненными матрицами произвольной структуры, каковыми являются, например, методы отражений и вращений (§ 2.5). Используемые ими **QR**-разложения, основанные на ортогональных преобразованиях отражения (Хаусхолдера, § 1.4) и вращения (Гивенса, § 1.5), более затратны по сравнению с **LU**-разложением: **QR**-разложение преобразованиями Хаусхолдера требует $O(4n^3/3)$ операций, а преобразованиями Гивенса — $O(2n^3)$ операций* [18].

Вычислительная эффективность применения различных факторизаций при решении других задач (в **QR**-алгоритме нахождения собственных чисел, в процессе сингулярного разложения и т.п.) существенно зависит от того, с какими матрицами проводятся преобразования, как организуется итерационный процесс и с

---

* Считается, что время выполнения каждой из участвующих в этих преобразованиях трансцендентных операций не более чем в 6 раз превосходит время выполнения арифметических операций, и их присутствие не отражается на асимптотике арифметической сложности ортогональных факторизаций.

какой скоростью он сходится, и др. Без сомнения, наиболее дорогостоящим в смысле арифметических затрат является процесс сингулярного разложения — $O(n^3mk + n^2m^2k)$ операций для $m \times n$-матрицы при k итерациях на диагонализацию двухдиагональной матрицы.

§ 8.2. ЧИСЛЕННАЯ УСТОЙЧИВОСТЬ МЕТОДА

Применяя те или иные численные методы, нельзя забывать о том, что их реальное поведение может существенно отличаться от ожидаемого. Только в исключительных ситуациях все вычисления, необходимые для решения задачи выбранным методом, выполняются точно. При реальных компьютерных расчетах неизбежно появляются ***ошибки округления***, величина которых определяется величиной разрядной сетки, выделяемой под запись числа с плавающей точкой ($fl(\cdot)$, см. приложение). Ошибки округления появляются и на стадии ввода в компьютер исходных данных, например при переводе обыкновенных дробей в числа типа $fl(\cdot)$. То, как поведут себя ошибки округления в процессе массовых вычислений при использовании конкретного численного метода, составляет предмет исследования его на устойчивость к таким ошибкам, иначе, вычислительную, или ***численную, устойчивость***. *Метод считается численно устойчивым, если в процессе его применения не происходит катастрофического накопления погрешностей округления, существенно искажающих результат (рост их ограничен).*

Изучим поведение погрешностей округления при реализации простейшего итерационного процесса решения систем линейных алгебраических уравнений — *метода простых итераций.* Этот метод, определяемый формулой

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{c}, \quad k = 0, 1, 2, \ldots , \tag{8.1}$$

генерирует последовательность векторов $\left(\mathbf{x}^{(k)}\right)$, которая, как было показано в § 3.2, при условии $\rho(\mathbf{B}) < 1$ из любого начального вектора $\mathbf{x}^{(0)}$ сходится к вектору $\mathbf{x}^*$ такому, что

$$\mathbf{x}^* = \mathbf{B}\mathbf{x}^* + \mathbf{c} . \tag{8.2}$$

Это в идеале. Реальный же МПИ — это процесс вида

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{c} + \boldsymbol{\gamma}^{(k)}, \quad k = 0, 1, 2, \ldots, \tag{8.3}$$

где $\boldsymbol{\gamma}^{(k)}$ — вектор ошибок округления, совершаемых на каждой k-й итерации.

Посмотрим, как ведет себя последовательность векторов

$$\boldsymbol{\mu}_k := \tilde{\mathbf{x}}^{(k)} - \mathbf{x}^*$$

— ошибок приближений $\tilde{\mathbf{x}}^{(k)}$, получаемых реальным МПИ (8.3).

Вычитая (8.2) из (8.3), имеем

$$\tilde{\mathbf{x}}^{(k+1)} - \mathbf{x}^* = \mathbf{B}\left(\tilde{\mathbf{x}}^{(k)} - \mathbf{x}^*\right) + \boldsymbol{\gamma}^{(k)},$$

следовательно,

$$\boldsymbol{\mu}_{k+1} = \mathbf{B}\boldsymbol{\mu}_k + \boldsymbol{\gamma}^{(k)} = \mathbf{B}\left(\mathbf{B}\boldsymbol{\mu}_{k-1} + \boldsymbol{\gamma}^{(k-1)}\right) + \boldsymbol{\gamma}^{(k)} =$$

$$= \mathbf{B}^2\left(\mathbf{B}\boldsymbol{\mu}_{k-2} + \boldsymbol{\gamma}^{(k-2)}\right) + \mathbf{B}\boldsymbol{\gamma}^{(k-1)} + \boldsymbol{\gamma}^{(k)} = \ldots =$$

$$= \mathbf{B}^{k+1}\boldsymbol{\mu}_0 + \left(\mathbf{B}^k\boldsymbol{\gamma}^{(0)} + \mathbf{B}^{k-1}\boldsymbol{\gamma}^{(1)} + \ldots + \mathbf{B}\boldsymbol{\gamma}^{(k-1)} + \boldsymbol{\gamma}^{(k)}\right). \tag{8.4}$$

Первое слагаемое в последнем выражении отвечает за погрешность идеального МПИ (8.1) и может быть сделано сколь угодно малым в процессе итерирования при условии $\rho(\mathbf{B}) < 1$ (см. лемму 3.1). Чтобы оценить второе слагаемое, предположим, что порог абсолютных погрешностей округлений, допускаемых на каждой итерации, определяется величиной γ, т.е.

$$\left\|\boldsymbol{\gamma}^{(k)}\right\| \le \gamma \qquad \forall k \in \mathbb{N}_0.$$

Тогда

$$\left\|\mathbf{B}^k\boldsymbol{\gamma}^{(0)} + \mathbf{B}^{k-1}\boldsymbol{\gamma}^{(1)} + \ldots + \mathbf{B}\boldsymbol{\gamma}^{(k-1)} + \boldsymbol{\gamma}^{(k)}\right\| \le \gamma\left\|\mathbf{E} + \mathbf{B} + \cdots + \mathbf{B}^k\right\|,$$

и, если $\|\mathbf{B}\| \le q < 1$, то второе слагаемое в (8.4), хотя и не стремится к нулю, но ограничено по норме величиной $\gamma\dfrac{1-q^k}{1-q} < \dfrac{\gamma}{1-q}$.

При условии же $\rho(\mathbf{B})<1$, теоретически обеспечивающем сходимость идеального МПИ (8.1), малость этого второго слагаемого отнюдь не гарантируется, что означает допустимость ситуаций, когда в ходе реальных итераций погрешность округлений будет накапливаться вплоть до переполнения используемого множества компьютерных чисел.

Более детальный анализ влияния ошибок округления на итерационный процесс с попыткой пролить свет на природу этого влияния можно найти, например, в [4, 54]. Здесь же ограничимся напоминанием о том, что необходимо с осторожностью применять процессы, когда для них нет эффективных оценок погрешности, и, по возможности, учитывать влияние ошибок округления, если такие оценки есть. Например, применительно к МПИ решения СЛАУ выше фактически доказано следующее утверждение.

Теорема 8.1.[*] *Пусть* $\|\mathbf{B}\| \le q < 1$ *и приближения* $\tilde{\mathbf{x}}^{(k)}$ *к решению* $\mathbf{x}^*$ *системы* $\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{c}$ *получаются посредством равенства* (8.3), *где* $\boldsymbol{\gamma}^{(k)}$ *— вектор ошибок округлений таких, что* $\left\|\boldsymbol{\gamma}^{(k)}\right\| \le \gamma$. *Тогда погрешность* k*-го приближения при любом* $k \in \mathbb{N}$ *можно оценить посредством неравенства*

$$\left\|\mathbf{x}^* - \tilde{\mathbf{x}}^{(k)}\right\| \le \frac{q}{1-q}\left\|\tilde{\mathbf{x}}^{(k)} - \tilde{\mathbf{x}}^{(k-1)}\right\| + \frac{\gamma}{1-q}. \qquad (8.5)$$

Действительно, для последовательности $\left(\mathbf{x}^{(k)}\right)$, получаемой МПИ (8.1), справедливо равенство

$$\mathbf{x}^* - \mathbf{x}^{(k+1)} = \mathbf{B}^{k+1}\left(\mathbf{x}^* - \mathbf{x}^{(0)}\right).$$

Следовательно, считая, что процессы (8.1) и (8.3) начинаются с одного начального приближения $\mathbf{x}^{(0)} \equiv \tilde{\mathbf{x}}^{(0)}$, в идентичном (8.4) равенстве

$$\mathbf{x}^* - \tilde{\mathbf{x}}^{(k+1)} = \mathbf{B}^{k+1}\left(\mathbf{x}^* - \tilde{\mathbf{x}}^{(0)}\right) - \left(\mathbf{B}^k\boldsymbol{\gamma}^{(0)} + \mathbf{B}^{k-1}\boldsymbol{\gamma}^{(1)} + \ldots + \mathbf{B}\boldsymbol{\gamma}^{(k-1)} + \boldsymbol{\gamma}^{(k)}\right)$$

[*] Аналогичное утверждение см. в [1].

можно заменить $\mathbf{B}^{k+1}\left(\mathbf{x}^* - \tilde{\mathbf{x}}^{(0)}\right)$ на $\mathbf{x}^* - \mathbf{x}^{(k+1)}$. Таким образом, погрешности $(k+1)$-х приближений реального (8.3) и идеального (8.1) методов различаются лишь слагаемым, оцененным выше по норме величиной $\gamma/(1-q)$, т.е. и для процесса (8.3) можно воспользоваться оценкой, выведенной в теореме 3.2.

Отметим, что как видно из оценки (8.5) при значениях q, приближающихся к единице, роль ошибок округления в образовании общей погрешности тем сильнее, чем медленнее сходимость итерационного процесса.

Алгоритмы прямых методов решения СЛАУ, а также практически все методы решения задач на собственные значения, как можно заметить, намного сложнее рассмотренного выше на предмет численной устойчивости метода простых итераций. Поэтому к ним труднее применить ***прямой анализ ошибок*** подобно тому, как это показано выше. Чаще здесь ограничиваются способом исследования вычислительной устойчивости методов, который называют ***обратным анализом ошибок*** (предложен Уилкинсоном в середине XX в.). Суть этого способа состоит в том, чтобы неизбежные погрешности вычислений, накапливаемые в процессе применения метода, «относить назад», на вход решаемой задачи, условно считая, что все вычислительные операции реализуются точно, без погрешностей, но при этом решается некоторая другая, ***возмущенная задача*** (по отношению к данной). В результате усилия исследователя сосредоточиваются на том, чтобы оценить величину этого, так называемого ***эквивалентного возмущения***, характеризующего близость возмущенной задачи к исходной, что определяет приемлемость метода.

Например, приближенную $\mathbf{LU}$-факторизацию $n \times n$-матрицы $\mathbf{A}$ (§ 1.2), выполняемую в некоторой системе компьютерных чисел с плавающей точкой и уровнем погрешностей округления γ и приводящую в итоге к некоторым матрицам $\tilde{\mathbf{L}}$ и $\tilde{\mathbf{U}}$, приближенно реализующим равенство $\mathbf{LU} = \mathbf{A}$, можно интерпретировать как точную $\mathbf{LU}$-факторизацию возмущенной матрицы $\mathbf{A} + \Delta\mathbf{A}$, т.е. как результат решения уравнения $\mathbf{LU} = \mathbf{A} + \Delta\mathbf{A}$. Показано [26], что при этом ***матрица эквивалентного возмущения*** $\Delta\mathbf{A}$

(иначе, ***обратная ошибка*** факторизации) удовлетворяет неравенству

$$\| \Delta \mathbf{A} \| \le n\gamma \| \mathbf{L} \| \cdot \| \mathbf{U} \| . \qquad (8.6)$$

Обратный анализ ошибок решения системы $\mathbf{Ax} = \mathbf{b}$ методом Гаусса на базе $\mathbf{LU}$-разложения в тех же условиях приближенных вычислений состоит в оценке матрицы эквивалентных возмущений $\delta\mathbf{A}$ в равенстве (возмущенная система)

$$(\mathbf{A} + \delta\mathbf{A})\mathbf{x} = \mathbf{b}. \qquad (8.7)$$

Такую оценку можно получить, например, как следствие оценки (8.6) и оценок эквивалентных возмущений в процессе приближенного решения треугольных систем $\mathbf{Ly} = \mathbf{b}$ и $\mathbf{Ux} = \mathbf{y}$ (см. § 2.2). Именно для возмущенной системы $(\mathbf{L} + \delta\mathbf{L})\mathbf{y} = \mathbf{b}$ устанавливается неравенство

$$| \delta\mathbf{L} | \le n\gamma | \mathbf{L} | , \qquad (8.8)$$

аналогично для системы $(\mathbf{U} + \delta\mathbf{U})\mathbf{x} = \mathbf{y}$ — неравенство

$$| \delta\mathbf{U} | \le n\gamma | \mathbf{U} | , \qquad (8.9)$$

где символ $|\bullet|$ используется для обозначения матрицы из модулей соответствующих элементов. Из оценок (8.6), (8.8) и (8.9) выводится оценка для фигурирующего в (8.7) возмущения $\delta\mathbf{A}$ следующего вида:

$$\| \delta\mathbf{A} \|_\infty \le 3gn^3\gamma \| \mathbf{A} \|_\infty .$$

В последнем неравенстве участвует некоторая величина g, называемая ***коэффициентом роста***. Она является функцией размерности n и зависит как от свойств самой матрицы $\mathbf{A}$, так и от способа реализации метода Гаусса. Так, например, показано, что при постолбцовом выборе главного элемента теоретически g может достигать значения 2^{n-1}, однако на практике величина g редко превосходит n [26].

Имеется много других, в частности, более детализированных оценок эквивалентных возмущений для разных методов решения задач линейной алгебры в разных постановках, но почти все они страдают тем недостатком, что либо слишком грубы (т.е. сильно

завышают действительную погрешность), либо слишком дороги (т.е. требуют существенно больших затрат на их получение, чем на вычисление собственно решения).

§ 8.3. ОБУСЛОВЛЕННОСТЬ ЗАДАЧИ

Наряду с погрешностями численного метода решения задачи, которые, как было показано в предыдущем параграфе, имеет смысл интерпретировать как эквивалентные возмущения исходной задачи, сама исходная задача может содержать погрешности, связанные с рядом причин. Наличие этих погрешностей задачи и метода означает, что фактически, имея цель решить одну задачу, получают решение другой, возмущенной задачи. Насколько могут быть близки или далеки решения этих задач, во многом зависит от свойств самих решаемых задач. Одним из важнейших свойств задачи (как с позиции принципиальной возможности получения приемлемого численного решения, так и с позиции интерпретации полученного результата) является ***обусловленность задачи***, т.е. способность задачи реагировать на искажение исходных данных. Эта качественная характеристика задачи может быть выражена количественно с помощью так называемого числа (меры) обусловленности. Рассмотрим подробнее и уточним это понятие на примере задачи о нахождении единственного решения СЛАУ

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{8.10}$$

с обратимой матрицей $\mathbf{A}$ размера $n \times n$.

Предположим, что правая часть уравнения (8.10) получила приращение (возмущение) $\Delta\mathbf{b}$, т.е. вместо истинного вектора $\mathbf{b}$ используется приближенный вектор $\mathbf{b} + \Delta\mathbf{b}$. Реакцией решения $\mathbf{x}$ на возмущение $\Delta\mathbf{b}$ правой части будет вектор поправок $\Delta\mathbf{x}$, т.е. если вектор $\mathbf{x}$ — решение (8.10), то $\mathbf{x} + \Delta\mathbf{x}$ — решение уравнения

$$\mathbf{A}(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}. \tag{8.11}$$

Понимая под абсолютной погрешностью приближенного вектора норму разности между точным и приближенным векторами, а под относительной погрешностью — отношение абсолютной погрешности к норме вектора (точного или приближенного), выясним связь между относительными погрешностями вектора

свободных членов и вектора-решения. Иначе говоря, получим оценку вида

$$\frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x}\|} \le (?)\frac{\|\Delta \mathbf{b}\|}{\|\mathbf{b}\|}, \tag{8.12}$$

где $\|\cdot\|$ — какая-либо векторная норма, а $(?)$ — неизвестный пока коэффициент связи.

Подставляя (8.10) в (8.11), видим, что поправка $\Delta \mathbf{x}$ связана с возмущением $\Delta \mathbf{b}$ таким же, как и (8.10), равенством

$$\mathbf{A}\Delta \mathbf{x} = \Delta \mathbf{b},$$

из которого находим ее явное выражение

$$\Delta \mathbf{x} = \mathbf{A}^{-1}\Delta \mathbf{b}. \tag{8.13}$$

Нормируя равенства (8.10) и (8.13), имеем

$$\|\mathbf{b}\| \le \|\mathbf{A}\|\cdot\|\mathbf{x}\| \quad \text{и} \quad \|\Delta \mathbf{x}\| \le \|\mathbf{A}^{-1}\|\cdot\|\Delta \mathbf{b}\|,$$

где матричная норма должна быть согласованной с выбранной векторной нормой. Эти два числовых неравенства одинакового смысла можно перемножить:

$$\|\mathbf{b}\|\cdot\|\Delta \mathbf{x}\| \le \|\mathbf{A}\|\cdot\|\mathbf{x}\|\cdot\|\mathbf{A}^{-1}\|\cdot\|\Delta \mathbf{b}\|.$$

Отсюда делением на $\|\mathbf{b}\|\cdot\|\mathbf{x}\|$ получаем искомую связь заданного вида (8.12):

$$\frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x}\|} \le \|\mathbf{A}\|\cdot\|\mathbf{A}^{-1}\|\cdot\frac{\|\Delta \mathbf{b}\|}{\|\mathbf{b}\|}. \tag{8.14}$$

Положительное число $\|\mathbf{A}\|\cdot\|\mathbf{A}^{-1}\|$ — коэффициент этой связи — называют ***числом (мерой) обусловленности*** матрицы $\mathbf{A}$ и обозначают $\operatorname{cond}\mathbf{A}$. Распространены также обозначения $\nu(\mathbf{A})$, $\chi(\mathbf{A})$, $\mu(\mathbf{A})$.

Легко показать, что то же самое число

$$\operatorname{cond}\mathbf{A} := \|\mathbf{A}\|\cdot\|\mathbf{A}^{-1}\| \tag{8.15}$$

служит коэффициентом роста относительных погрешностей при неточном задании элементов матрицы $\mathbf{A}$ в (8.10). А именно, если матрица $\mathbf{A}$ получила возмущение $\Delta \mathbf{A}$, в результате чего вместо $\mathbf{x}$

получено $\mathbf{x}+\Delta\mathbf{x}$ — решение возмущенной системы

$$(\mathbf{A}+\Delta\mathbf{A})(\mathbf{x}+\Delta\mathbf{x})=\mathbf{b},$$

то справедливы неравенства

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|}\le \operatorname{cond}\mathbf{A}\cdot\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}+\Delta\mathbf{A}\|} \quad \text{и} \quad \frac{\|\Delta\mathbf{x}\|}{\|x+\Delta\mathbf{x}\|}\le \operatorname{cond}\mathbf{A}\cdot\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}. \qquad (8.16)$$

Замечание 8.2. Можно получить оценки, объединяющие оценки (8.14) и (8.16) в случае одновременного учета возмущений матрицы $\mathbf{A}$ системы (8.10) и ее правой части $\mathbf{b}$. Более того, величину $\nu(\mathbf{A}):=\|\mathbf{A}\|\cdot\|\mathbf{A}^{-1}\|$ можно считать также мерой обусловленности линейного обратимого оператора $\mathbf{A}$ в произвольном нормированном пространстве. Справедливо утверждение [65]:

Теорема 8.2. *Пусть* $\mathbf{A}\mathbf{x}=\mathbf{b}$ — *данное, а* $\tilde{\mathbf{A}}\tilde{\mathbf{x}}=\tilde{\mathbf{b}}$ — *возмущенное линейные операторные уравнения с относительными уровнями возмущений* $\delta_{\mathbf{A}}\ge\dfrac{\|\mathbf{A}-\tilde{\mathbf{A}}\|}{\|\mathbf{A}\|}$ *и* $\delta_{\mathbf{b}}\ge\dfrac{\|\mathbf{b}-\tilde{\mathbf{b}}\|}{\|\mathbf{b}\|}$. *Тогда, если* $\delta_{\mathbf{A}}\nu(\mathbf{A})<1$, *то эти уравнения одновременно однозначно разрешимы и справедлива оценка относительной погрешности решения, имеющая вид*

$$\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|}\le\frac{\nu(\mathbf{A})}{1-\delta_{\mathbf{A}}\cdot\nu(\mathbf{A})}\cdot\delta_{\mathbf{b}}+\frac{\nu^2(\mathbf{A})}{1-\delta_{\mathbf{A}}\cdot\nu(\mathbf{A})}\cdot\delta_{\mathbf{A}}.$$

Итак, неравенства (8.14) и (8.16) показывают, что *чем больше число обусловленности, тем сильнее сказывается на решении линейной системы ошибка в исходных данных.* Грубо говоря, если $\operatorname{cond}\mathbf{A}=O(10^p)$ и исходные данные имеют погрешность в l-м знаке после запятой, то независимо от способа решения системы (8.10) в результате можно гарантировать не более $l-p$ знаков после запятой.

Если число $\operatorname{cond}\mathbf{A}$ велико, то система считается плохо обусловленной. Говорить о том, «что такое хорошо и что такое плохо», в отрыве от контекста решаемой задачи почти бессмысленно, так как здесь могут играть роль размерность задачи, точность, с которой должно быть найдено ее решение, точность выполнения арифметических операций и т.п. Однако можно дать оценку числа обусловленности снизу. А именно, если используются подчинен-

ные матричные нормы (для которых норма единичной матрицы есть единица), то, очевидно,

$$\operatorname{cond} \mathbf{A} := \| \mathbf{A} \| \cdot \| \mathbf{A}^{-1} \| \geq \| \mathbf{A} \cdot \mathbf{A}^{-1} \| = \| \mathbf{E} \| = 1,$$

т.е. *число обусловленности не может быть меньше единицы.* Можно также указать верхнюю границу для чисел обусловленности, превышение которой при решении линейных систем в конкретной вычислительной среде может привести к заведомо ложным результатам. Так, решение считается ненадежным, если окажется $\operatorname{cond} \mathbf{A} \geq (\text{macheps})^{-1}$ или даже $\operatorname{cond} \mathbf{A} \geq (\text{macheps})^{-0.5}$ [28]. При этом заметим, что умножение матрицы $\mathbf{A}$ на скаляр α не изменяет ее числа обусловленности, ибо

$$\operatorname{cond}(\alpha \mathbf{A}) = \| \alpha \mathbf{A} \| \cdot \| (\alpha \mathbf{A})^{-1} \| = \| \mathbf{A} \| \cdot \| \mathbf{A}^{-1} \| = \operatorname{cond} \mathbf{A}.$$

Классическим примером плохо обусловленной матрицы является так называемая ***матрица Гильберта***[*]

$$\mathbf{H}_n := \left(\frac{1}{i + j - 1} \right)_{i, j=1}^{n},$$

демонстрирующая катастрофическое возрастание числа обусловленности с ростом размерности [51, 71]. Так, уже при $n = 8$ $\operatorname{cond} \mathbf{H}_8 > 10^{10}$, и обратная матрица $\mathbf{H}_8^{-1}$, полученная на машине с точностью представления чисел порядка 10^{-8}, может не содержать ни одного верного знака.

Очевидно, число обусловленности зависит от выбора матричной нормы (индуцированной, как правило, той или иной векторной нормой, в терминах которой характеризуется относительная погрешность решения алгебраической системы). Однако нетрудно получить оценку числа обусловленности через собственные числа матрицы. Действительно, пусть собственные числа λ_i матрицы $\mathbf{A}$ упорядочены по модулю:

$$|\lambda_1| \geq |\lambda_2| \geq \ \ldots \ \geq |\lambda_n|,$$

---

[*] Эта матрица не просто плод воображения. Ее появление закономерно, например, при приближении функции многочленом канонического вида методом наименьших квадратов [13].

т.е. ***спектральный радиус*** матрицы $\mathbf{A}$ есть $\rho_{\mathbf{A}} := |\lambda_1|$. Тогда, в силу известного неравенства $\rho_{\mathbf{A}} \le \|\mathbf{A}\|$ и соотношения между собственными числами прямой и обратной матриц, имеем:

$$\|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \ge \rho_{\mathbf{A}} \cdot \rho_{\mathbf{A}^{-1}} = |\lambda_1| \cdot \frac{1}{|\lambda_n|}.$$

Таким образом, оценкой снизу меры обусловленности матрицы $\mathbf{A}$ может служить величина $|\lambda_1|/|\lambda_n|$ (называемая иногда ***числом обусловленности Тодда***). Для симметричных матриц эта оценка и в самом деле является числом обусловленности, соответствующим спектральной норме матрицы (индуцированной евклидовой нормой вектора). Учитывая смысл собственных чисел матрицы, можно сказать, что число обусловленности показывает величину отношения наибольшего коэффициента растяжения вектора посредством линейного преобразования $\mathbf{A}$ к наименьшему.

Можно встретить и более общие определения меры обусловленности матрицы, пригодные для прямоугольных и вырожденных матриц. Одно из них, введенное выше через отношение сингулярных чисел (см. § 7.1), естественным образом обобщает число обусловленности Тодда. Обобщением числа обусловленности вида (8.15) служит следующее представление [74]:

$$\operatorname{cond} \mathbf{A} := \sup_{\mathbf{x} \ne 0} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} \Big/ \inf_{\mathbf{y} \ne 0} \frac{\|\mathbf{A}\mathbf{y}\|}{\|\mathbf{y}\|}.$$

Здесь $\operatorname{cond} \mathbf{A}$ задается через *векторные* нормы и может быть применено к вырожденным и к неквадратным матрицам $\mathbf{A}$. В случае обратимых матриц $\mathbf{A}$ при использовании согласованных матричных норм отсюда получается (8.15). Через псевдообратные матрицы число обусловленности, аналогичное (8.15), дано в § 7.3.

Следует отметить, что непосредственный подсчет чисел обусловленности, особенно при больших размерах матриц, является весьма дорогостоящим делом из-за необходимости обращать матрицы или находить их собственные значения. Поэтому зачастую о приемлемости порядка возможного роста относительной погреш-

ности результата решения какой-либо алгебраической задачи относительно данной матрицы судят либо по каким-то достаточным признакам (например, по доминированию элементов главной диагонали матрицы), либо на основе теоретического изучения матрицы [51, 59], либо путем применения специальных алгоритмов приближенного оценивания cond **A** [26, 28, 48, 51, 70]. Исследование матриц на обусловленность может быть естественным образом увязано с процессом решения той или иной алгебраической задачи относительно данной матрицы.

Пример 8.1. Линейная система

$$\begin{cases} x + 10y = 11, \\ 100x + 1001y = 1101 \end{cases} \tag{8.17}$$

имеет единственное решение $x := 1$, $y := 1$. Допустив абсолютную погрешность величиной 0,01 в правой части одного уравнения системы (8.17), получим возмущенную систему

$$\begin{cases} x + 10y = 11{,}01, \\ 100x + 1001y = 1101 \end{cases} \tag{8.18}$$

с единственным решением $x := 11{,}01$, $y := 0$. Очевидно, последнее трудно назвать близким к решению исходной системы.

Причина такого большого различия в решениях близких систем — в их плохой обусловленности. Матрица коэффициентов систем (8.17) и (8.18) $\mathbf{A} := \begin{pmatrix} 1 & 10 \\ 100 & 1001 \end{pmatrix}$ имеет обратную $\mathbf{A}^{-1} := \begin{pmatrix} 1001 & -10 \\ -100 & 1 \end{pmatrix}$. Следовательно, число обусловленности в матричной норме, индуцированной векторной нормой-максимум (иначе, нормой l_∞), есть

$$\mathrm{cond}_\infty(\mathbf{A}) = 1101 \cdot 1011 = 1113111 > 10^6.$$

Учитывая, что в данном примере $\mathbf{b} := \begin{pmatrix} 11 \\ 1101 \end{pmatrix}$, $\Delta\mathbf{b} := \begin{pmatrix} 0{,}01 \\ 0 \end{pmatrix}$, на основе (8.14) получаем следующую оценку относительной погрешности решения в l_∞-нормах:

$$\delta_{\mathbf{x}} \le \mathrm{cond}_\infty(\mathbf{A}) \cdot \delta_{\mathbf{b}} = 1113111 \cdot \frac{0{,}01}{1101} = 10{,}11.$$

Так как норма-максимум решения $\mathbf{x} := \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ равна 1, то оценка абсолютной погрешности $\|\Delta\mathbf{x}\|$ решения суть

$$\|\Delta\mathbf{x}\| = \|\mathbf{x}\| \cdot \delta_{\mathbf{x}} \le 10{,}11.$$

Как видим, решение $\mathbf{x}+\Delta\mathbf{x}:=\begin{pmatrix}11{,}01\\0\end{pmatrix}$ возмущенной системы (8.18) вписывается в оценку

$$\|\mathbf{x}+\Delta\mathbf{x}\|\le\|\mathbf{x}\|+\|\Delta\mathbf{x}\|\le 1+10{,}11=11{,}11.$$

Аналогичный результат может быть получен через число обусловленности Тодда. Решая характеристическое уравнение

$$\lambda^2-1002\lambda+1=0,$$

находим собственные числа матрицы $\mathbf{A}$: $\lambda_1\approx 1002$ и $\lambda_2\approx 0{,}000998$, приводящие к оценке

$$\operatorname{cond}\mathbf{A}\ge\lambda_1/\lambda_2\approx 1004000>10^6.$$

На данном примере также можно наглядно убедиться в том, что *малость* ***невязки*** $\mathbf{r}:=\mathbf{b}-\mathbf{A}\tilde{\mathbf{x}}$ *плохо обусловленной системы еще не говорит о близости приближенного решения* $\tilde{\mathbf{x}}$ *к точному* $\mathbf{x}$. Действительно, невязки $\mathbf{r}_1$ и $\mathbf{r}_2$ векторов $\tilde{\mathbf{x}}_1:=\begin{pmatrix}11{,}01\\0\end{pmatrix}$ и $\tilde{\mathbf{x}}_2:=\begin{pmatrix}1\\1{,}1\end{pmatrix}$ для основной (невозмущенной) системы имеют нормы соответственно $\|\mathbf{r}_1\|_\infty=0{,}01$ и $\|\mathbf{r}_2\|_\infty=100{,}1$. Вектор $\tilde{\mathbf{x}}_2$, явно более близкий к точному решению $\mathbf{x}:=\begin{pmatrix}1\\1\end{pmatrix}$, чем $\tilde{\mathbf{x}}_1$, имеет существенно бóльшую невязку! Объяснение этому, парадоксальному, на первый взгляд, факту можно найти, например, в книге [5].

Двумерный случай допускает простую геометрическую трактовку понятия обусловленности. Плохая обусловленность системы двух уравнений с двумя неизвестными означает, что прямые, являющиеся геометрическими образами уравнений, пересекаются на координатной плоскости под очень острым углом. В этом случае небольшое искажение в данных, интерпретируемое как параллельный перенос (при возмущении свободного члена) или поворот прямых (при возмущении матрицы коэффициентов), приводит к значительному перемещению их точки пересечения, т.е. геометрического образа решения. Более того, пересекающиеся прямые при небольшом их повороте могут стать параллельными и наоборот.

В многомерном случае и интерпретация обусловленности сложнее, и оценки погрешности, естественно, грубее.

§ 8.4. СПОСОБЫ УЛУЧШЕНИЯ ОБУСЛОВЛЕННОСТИ

В какой-то мере ситуацией с обусловленностью задачи допустимо управлять. Можно попытаться исходную задачу перед началом решения заменить эквивалентной ей, но лучше обусловленной задачей. Один из способов достижения такой цели заключается в применении ***масштабирования матрицы*** или, иначе, ее ***уравновешивания***.

Для заданной матрицы $\mathbf{A}$ строят такие невырожденные диагональные матрицы $\mathbf{D}_1$ и $\mathbf{D}_2$, что умножение $\mathbf{D}_1$ на $\mathbf{A}$ дает матрицу, *равновесную по строкам* (определенная норма каждого вектора-строки матрицы $\mathbf{D}_1\mathbf{A}$ находится в заданных границах, например модули всех элементов принадлежат отрезку [0,1, 1]) , а умножением $\mathbf{A}$ на $\mathbf{D}_2$ достигается *равновесность матрицы* $\mathbf{AD}_2$ *по столбцам.* Применительно к СЛАУ (8.10) первое из этих преобразований равносильно умножению каждого уравнения системы на некоторое ненулевое число (возможно, 1), что влечет за собой соответствующую замену вектора правой части, а второе преобразование соответствует замене переменных, что не изменяет правую часть системы. Таким образом, из системы (8.10) путем подбора подходящих матриц $\mathbf{D}_1$ и $\mathbf{D}_2$ может быть получена новая система $\mathbf{D}_1\mathbf{AD}_2\tilde{\mathbf{x}} = \mathbf{D}_1\mathbf{b}$ с ***равновесной матрицей*** $\mathbf{D}_1\mathbf{AD}_2$ (более подробно об этом см., например, в [71]).

К сожалению, во-первых, нет универсальных алгоритмов, эффективно решающих проблему масштабирования произвольных матриц в произвольных нормах (хотя имеются библиотечные программы, как-то выполняющие уравновешивание матриц [26]). Во-вторых, следует заметить, что подобная процедура предварительного масштабирования задачи, уменьшая число обусловленности, может улучшить показатели численного метода ее решения в отношении накопления погрешностей округления, но не может восполнить недостаток информации в исходных данных, т.е. уменьшить неопределенность в получаемом решении, в чем можно убедиться на следующем простом примере.

Пример 8.2. Посмотрим, что дает масштабирование возмущенной системы (8.18).

Взяв матрицу $\mathbf{D}_1 := \begin{pmatrix} 1 & 0 \\ 0 & 0{,}001 \end{pmatrix}$, найдем преобразованные матрицу и вектор, соответственно, $\mathbf{D}_1\mathbf{A} = \begin{pmatrix} 1 & 10 \\ 0{,}1 & 1{,}001 \end{pmatrix}$ и $\mathbf{D}_1(\mathbf{b}+\Delta\mathbf{b}) = \begin{pmatrix} 11{,}01 \\ 1{,}101 \end{pmatrix}$, что равносильно получению системы

$$\begin{cases} x + 10y = 11{,}01, \\ 0{,}1x + 1{,}001y = 1{,}101 \end{cases} \tag{8.19}$$

из системы (8.18) умножением второго уравнения на 0,001. Посредством матрицы $\mathbf{D}_2 := \begin{pmatrix} 1 & 0 \\ 0 & 0{,}1 \end{pmatrix}$ матрица системы (8.19) преобразуется в полностью равновесную матрицу $\mathbf{D}_1\mathbf{A}\mathbf{D}_2 = \begin{pmatrix} 1 & 1 \\ 0{,}1 & 0{,}1001 \end{pmatrix}$, что применительно к самой системе (8.19) означает приведение ее к виду

$$\begin{cases} x + \tilde{y} = 11{,}01, \\ 0{,}1x + 0{,}1001\tilde{y} = 1{,}101 \end{cases} \tag{8.20}$$

заменой $\tilde{y} := 10y$. Обусловленность последней системы лучше, чем исходной ($\mathrm{cond}_\infty(\mathbf{D}_1\mathbf{A}\mathbf{D}_2) = 22001 \approx 0{,}02\,\mathrm{cond}_\infty\mathbf{A}$). Этот факт может положительно отразиться на численной устойчивости процесса получения ее решения, но не изменяет ситуации с отклонением решения возмущенной системы от решения исходной системы: как видим, система (8.20) и, естественно, промежуточная система (8.19) имеют решением те же значения $x := 11{,}01$, $y\ (= 0{,}1\tilde{y}) := 0$, что и система (8.18).

Близким к масштабированию является процесс ***предобуславливания матриц***. Проще об этом проводить рассуждения, опять привязываясь к задаче получения решения СЛАУ и даже имея в виду конкретный численный метод, каковым будем считать метод сопряженных градиентов.

Пусть матрица $\mathbf{A}$ в системе (8.10) является симметричной положительно определенной и пусть $\mathbf{M}$ — некоторая другая такая же матрица, которую будем считать в каком-то смысле близкой к данной матрице $\mathbf{A}$. Наряду с уравнением (8.10) рассмотрим эквивалентное ему уравнение

$$\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \mathbf{M}^{-1}\mathbf{b}. \tag{8.21}$$

Ставим цель: подобрать матрицу $\mathbf{M}$ так, чтобы, с одной стороны, матрица $\mathbf{M}^{-1}\mathbf{A}$ новой системы (8.21) была лучше обусловлена, с другой стороны, чтобы вычислительные затраты при переходе к новой задаче сильно не возросли.

Очевидны крайние случаи.

Если взять $\mathbf{M} := \mathbf{E}$, ничего не изменится: $\operatorname{cond}(\mathbf{M}^{-1}\mathbf{A}) = \operatorname{cond}\mathbf{A}$ и затраты останутся прежними. Если положить $\mathbf{M} := \mathbf{A}$, то $\mathbf{M}^{-1}\mathbf{A} = \mathbf{E}$ и, значит, $\operatorname{cond}(\mathbf{M}^{-1}\mathbf{A}) = 1$. Обусловленность становится наилучшей, но при этом потребуется обращать данную матрицу $\mathbf{A}$, что еще хуже, чем решать исходную систему. Ясно, что нужно стараться по матрице $\mathbf{A}$ построить такую симметричную матрицу $\mathbf{M}$, чтобы $\mathbf{M} \approx \mathbf{A}$ и чтобы она была легко обращаема.

Изучение ситуации с предобусловливанием в процессе применения метода сопряженных градиентов к решению системы (8.10) показывает [50], что на каждом шаге этого метода требуется решать СЛАУ, постоянной матрицей коэффициентов в которых является как раз матрица предобусловливания $\mathbf{M}$. На основе описанного в § 3.4 метода сопряженных градиентов, примененного в данном случае к системе (8.21), после выполнения преобразований, связанных с заменой переменных, приходят к следующему алгоритму, определяющему ***предобусловленный метод сопряженных градиентов.***

1. Задать $\mathbf{x}^{(0)}$, $\varepsilon > 0$; вычислить вектор $\xi^{(0)} := \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$. Построить матрицу $\mathbf{M}$.
2. Решить систему $\mathbf{M}\tilde{\xi}^{(0)} = \xi^{(0)}$. Положить $\mathbf{p}^{(0)} := \tilde{\xi}^{(0)}$.
3. Для $k := 0, 1, 2, \ldots$:

   вычислить

   $$\alpha_k = -\left(\tilde{\xi}^{(k)},\ \xi^{(k)}\right) \Big/ \left(\mathbf{p}^{(k)},\ \mathbf{A}\mathbf{p}^{(k)}\right),$$

   $$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \mathbf{p}^{(k)},$$

   $$\xi^{(k+1)} = \xi^{(k)} + \alpha_k \mathbf{A}\mathbf{p}^{(k)};$$

   проверить на сходимость

   $$\left(\left(\xi^{(k+1)}, \xi^{(k+1)}\right) \le \varepsilon \Rightarrow \text{stop},\quad \mathbf{x}^* := \mathbf{x}^{(k+1)}\right);$$

   решить систему $\mathbf{M}\tilde{\xi}^{(k+1)} = \xi^{(k+1)}$;

   вычислить

   $$\beta_k = \left(\tilde{\xi}^{(k+1)}, \xi^{(k+1)}\right) \Big/ \left(\tilde{\xi}^{(k)},\ \xi^{(k)}\right),$$

   $$\mathbf{p}^{(k+1)} = \tilde{\xi}^{(k+1)} + \beta_k \mathbf{p}^{(k)}.$$

Конкретизируется данный метод выбором способа построения ***матрицы предобусловливания*** $\mathbf{M}$. Наиболее известны два таких способа: неполная факторизация Холецкого исходной матрицы и усеченное разложение в ряд Неймана обратной матрицы.

О неполном разложении Холецкого упоминалось в § 1.3 (см. замечание 1.3). За матрицу $\mathbf{M}$ здесь принимают матрицу $\mathbf{U}^T\mathbf{U}$ в представлении

$$\mathbf{A} = \mathbf{U}^T\mathbf{U} + \mathbf{B},$$

где в процессе заполнения матрицы $\mathbf{U}$ считаются заведомо нулевыми элементы в позициях, соответствующих нулевым элементам матрицы $\mathbf{A}$. В случае разреженной матрицы $\mathbf{A}$ выполнение такого разложения сравнительно недорого и решение систем $\mathbf{M}\tilde{\xi}^{(k)} = \xi^{(k)}$ с такой факторизованной матрицей $\mathbf{M}$ тоже необременительно. Соединение указанного выбора матрицы предобусловливания и метода сопряженных градиентов называют ***методом неполного разложения Холецкого – сопряженных градиентов*** (зарубежная аббревиатура **ICCG** — *incomplete Choleski conjugate gradient* [50]).

Другой способ построения процесса предобусловливания опирается на приближенное представление обратной матрицы отрезком матричного ряда.

Допустим, что исходная невырожденная матрица $\mathbf{A}$ не имеет на диагонали нулей. Тогда, выполнив расщепление этой матрицы $\mathbf{A} = \mathbf{D} - \mathbf{C}$, где $\mathbf{D} := \mathrm{diag}(a_{11}; ...; a_{nn})$, представим ее в виде

$$\mathbf{A} = \mathbf{D}(\mathbf{E} - \mathbf{D}^{-1}\mathbf{C}),$$

откуда при условии, что спектральный радиус матрицы $\mathbf{B} := \mathbf{D}^{-1}\mathbf{C}$ меньше единицы, по лемме Неймана (§ 3.2) следует

$$\mathbf{A}^{-1} = (\mathbf{E} - \mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{D}^{-1} = (\mathbf{E} + \mathbf{B} + \mathbf{B}^2 + ...)\mathbf{D}^{-1}. \qquad (8.22)$$

Примем за $\mathbf{M}^{-1}$ усеченное до m членов разложение (8.22):

$$\mathbf{M}^{-1} := (\mathbf{E} + \mathbf{B} + \mathbf{B}^2 + ... + \mathbf{B}^{m-1})\mathbf{D}^{-1},$$

т.е. решением систем требуемого алгоритмом вида $\mathbf{M}\tilde{\xi} = \xi$ с матрицей $\mathbf{M} := \mathbf{D}(\mathbf{E} + \mathbf{B} + \mathbf{B}^2 + ... + \mathbf{B}^{m-1})^{-1}$ считаем вектор

$$\tilde{\xi} := \mathbf{M}^{-1}\xi = (\mathbf{E} + \mathbf{B} + \mathbf{B}^2 + ... + \mathbf{B}^{m-1})\mathbf{D}^{-1}\xi \ . \qquad (8.23)$$

Легко видеть, что если организовать итерации по формуле

$$\xi^{(k)} = \mathbf{B}\,\xi^{(k-1)} + \mathbf{D}^{-1}\xi\,; \quad k = 1, 2, \ldots; \quad \xi^{(0)} := 0\,,$$

то при $k := m$ получим

$$\xi^{(m)} = \mathbf{B}\,\xi^{(m-1)} + \mathbf{D}^{-1}\xi = \mathbf{B}(\mathbf{B}\,\xi^{(m-2)} + \mathbf{D}^{-1}\xi) + \mathbf{D}^{-1}\xi =$$

$$= \mathbf{B}^2(\mathbf{B}\,\xi^{(m-3)} + \mathbf{D}^{-1}\xi) + (\mathbf{B} + \mathbf{E})\mathbf{D}^{-1}\xi = (\mathbf{B}^{m-1} + \ldots + \mathbf{B} + \mathbf{E})\mathbf{D}^{-1}\xi\,.$$

Это означает, что результат (8.23) можно считать продуктом выполнения m шагов метода Якоби (3.20); отсюда — название предобусловленного метода, привлекающего на каждой итерации итерационный же метод Якоби: ***m-шаговый метод Якоби – сопряженных градиентов***.

Заметим, что число итераций m здесь служит подбираемым параметром. При невыполнении условий сходимости метода Якоби следует ограничиться значением $m = 1$.

Завершая параграф, обратим внимание на тот факт, что все рассмотренные здесь подходы к улучшению обусловленности задач опираются на эквивалентные преобразования и формально не изменяют их решений. Однако иногда оказывается целесообразным нарушать это правило и от одной задачи переходить к другой с заведомо другим решением, к которому предъявляются следующие основные требования: оно должно устойчиво вычисляться и в определенном смысле должно быть близким к решению исходной задачи. Таким способам улучшения обусловленности посвящен следующий параграф.

§ 8.5. НЕУСТОЙЧИВОСТЬ РЕШЕНИЯ И РЕГУЛЯРИЗАЦИЯ

Рассмотрим простейшую двумерную систему

$$\begin{cases} ax_1 + ax_2 = b_1\,, \\ ax_1 + ax_2 = b_2\,, \quad a \neq 0 \end{cases} \tag{8.24}$$

с вырожденной матрицей $\mathbf{A} := \begin{pmatrix} a & a \\ a & a \end{pmatrix}$. Очевидно, в случае $b_1 = b_2 := \beta$ эта система имеет бесчисленное множество решений

вида $\mathbf{x}^* := (x_1;\ \beta/a - x_1)^T$. Нормальным (относительно нуля) решением системы (8.24) является такой элемент $\mathbf{x}_0$ этого множества, который реализует минимум расстояния от нуля до $\mathbf{x}^*$ в некоторой фиксированной метрике. Принимая за основу евклидову метрику, имеем:

$$\sqrt{(x_1 - 0)^2 + (\beta/a - x_1 - 0)^2} = \min \quad \Leftrightarrow$$

$$\Leftrightarrow \quad x_1{}^2 + \beta^2/a^2 - 2x_1\beta/a + x_1{}^2 = \min \quad \Leftrightarrow$$

$$\Leftrightarrow \quad 4x_1 - 2\beta/a = 0 \quad \Leftrightarrow \quad x_1 = \beta/2a\ ,$$

т.е. $\mathbf{x}_0 := (\beta/2a;\ \beta/2a)^T$ — нормальное решение системы (8.24). При $b_1 \neq b_2$ система (8.24) противоречива, но имеет псевдорешение $\mathbf{x}^+$, которое, как мы знаем, может быть найдено из уравнения $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. В данном случае получаем систему

$$\begin{cases} 2a^2x_1 + 2a^2x_2 = a(b_1 + b_2), \\ 2a^2x_1 + 2a^2x_2 = a(b_1 + b_2) \end{cases}$$

с множеством псевдорешений вида $\mathbf{x}^+ := \left(x_1;\ \dfrac{b_1 + b_2}{2a} - x_1\right)^T$, из которых аналогично предыдущему выделяем нормальное псевдорешение $\mathbf{x}_0^+ := \arg\min \rho_{\mathbb{R}_2}(\mathbf{x}^+, \mathbf{0}) = \left(\dfrac{b_1 + b_2}{4a};\ \dfrac{b_1 + b_2}{4a}\right)^T$, при $b_1 = b_2$ совпадающее с полученным ранее нормальным решением $\mathbf{x}_0$.

Исказим один элемент матрицы $\mathbf{A}$ на величину ε. Возмущенная матрица $\mathbf{A}_\varepsilon := \begin{pmatrix} a & a \\ a & a + \varepsilon \end{pmatrix}$ при $\varepsilon \neq 0$ имеет отличный от нуля определитель: $\det \mathbf{A}_\varepsilon = a\varepsilon$, и значит, существует обратная к $\mathbf{A}_\varepsilon$ матрица $\mathbf{A}_\varepsilon^{-1} := \dfrac{1}{a\varepsilon}\begin{pmatrix} a + \varepsilon & -a \\ -a & a \end{pmatrix}$. Следовательно, решение $\mathbf{x}_\varepsilon^*$

возмущенной системы

$$\begin{cases} ax_1 + ax_2 = b_1, \\ ax_1 + (a+\varepsilon)x_2 = b_2 \end{cases} \tag{8.25}$$

можно представить так:

$$\mathbf{x}_\varepsilon^* := \mathbf{A}_\varepsilon^{-1}\mathbf{b} = \begin{pmatrix} b_1/a + b_1/\varepsilon - b_2/\varepsilon \\ -b_1/\varepsilon + b_2/\varepsilon \end{pmatrix}. \tag{8.26}$$

Из выражения (8.26) видим, что в случае $b_1 = b_2$ получаемое таким образом решение параметризованной системы (8.25) есть $\mathbf{x}^* := (\beta/a;\ 0)^T$, т.е. не зависит от величины возмущения — параметра ε. В случае же $b_1 \neq b_2$ каждая компонента решения (8.26) при $\varepsilon \to 0$ стремится к бесконечности, а вовсе не к нормальному псевдорешению исходной системы (8.24), предельной для возмущенной.

Рассмотренный пример говорит о том, что нормальное псевдорешение СЛАУ не обладает устойчивостью по отношению к возмущению матрицы, и в определенных ситуациях следует принимать меры по повышению такой устойчивости. Обычно этого добиваются подменой данной неустойчивой задачи другой задачей, точнее, семейством других задач, которые в каком-то смысле близки к данной задаче и в то же время могут быть устойчиво решены. Такая процедура называется ***регуляризацией***.

Одним из наиболее развитых и достаточно широко известных способов регуляризации является ***метод α-регуляризации Тихонова****, теория которого разработана для наиболее общего случая некорректных задач в операторной постановке и классическим объектом применения которого являются интегральные уравнения Фредгольма первого рода. Упрощенная суть этого метода состоит в следующем.

Пусть вместо некоторого «точного» уравнения

$$Ay = f, \tag{8.27}$$

---

* Ти́хонов Андрей Николаевич (1906–1993) — российский математик и геофизик. Наряду со многими другими своими научными достижениями положил начало бурному развитию теории и практики методов решения некорректных задач (1943).

где A — линейный вполне непрерывный оператор, действующий из гильбертова пространства Y в гильбертово пространство F, известно «приближенное» уравнение

$$\tilde{A}y = \tilde{f} \tag{8.28}$$

и оценки близости

$$\|\tilde{A} - A\| \le \xi, \quad \|\tilde{f} - f\| \le \delta.$$

При поиске решения уравнения (8.27) на базе уравнения (8.28) вместо минимизации невязки $\|\tilde{A}y - \tilde{f}\|$, которая на неустойчивых задачах может вести себя нерегулярно (см. пример 8.1), предлагается минимизировать так называемый ***сглаживающий функционал*** (***функционал Тихонова***)

$$\Phi_\alpha[y, \tilde{A}, \tilde{f}] := \|\tilde{A}y - \tilde{f}\|^2 + \alpha\|y\|^2.$$

Здесь $\alpha > 0$ — ***параметр регуляризации***, а $\Omega[y] := \|y\|^2$ — ***стабилизирующий функционал*** (***стабилизатор***). Доказано, что функционал Тихонова всегда имеет и притом единственный элемент y_α такой, что

$$\Phi_\alpha[y_\alpha, \tilde{A}, \tilde{f}] = \inf_{y \in Y} \Phi_\alpha[y, \tilde{A}, \tilde{f}].$$

При всяком фиксированном $\alpha > 0$ последнюю задачу можно решать как подходящими численными методами минимизации, так и с помощью решения ***уравнения Тихонова***

$$\alpha y_\alpha + \tilde{A}^* \tilde{A} y_\alpha = \tilde{A}^* \tilde{f}, \tag{8.29}$$

которое представляет собой запись необходимого условия экстремума функционала $\Phi_\alpha[y, \tilde{A}, \tilde{f}]$ ($\tilde{A}^*$ — оператор, сопряженный с оператором $\tilde{A}$).

Параметр регуляризации α следует подбирать так, чтобы принимаемое за решение задачи (8.27) ***регуляризованное решение*** y_α было, по возможности, оптимальным. Это означает, что оно должно достаточно хорошо удовлетворять уравнению (8.28) (что определяется малостью невязки $\|\tilde{A}y_\alpha - \tilde{f}\|^2$) и достаточно

надежно вычисляться (что связано с величиной стабилизатора $\|y_\alpha\|^2$). При этом, чем меньше величины δ и ξ, тем меньшим должно быть значение α, т.е. бóльшим относительный вес первого слагаемого в функционале Тихонова. Таким образом, в методе α-регуляризации Тихонова (впрочем, как и при других способах регуляризации) главной проблемой является проблема выбора параметра регуляризации, призванного сбалансировать возможность получения регуляризованного решения y_α, как можно более устойчивого и как можно более близкого (в условиях оговоренной неопределенности) к решению именно «нужной» задачи. Имеется несколько способов выбора параметра регуляризации [10, 15, 49 и др.].

В нашем случае, когда уравнение (8.27) — это СЛАУ (8.10) с вещественной матрицей $\mathbf{A}$ (не обязательно квадратной), уравнение Тихонова (8.29) относительно регуляризованного решения $\mathbf{x}_\alpha$ имеет вид

$$\alpha\mathbf{x}_\alpha + \tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\,\mathbf{x}_\alpha = \tilde{\mathbf{A}}^T\tilde{\mathbf{b}}\ , \qquad (8.30)$$

где $\tilde{\mathbf{A}}$ и $\tilde{\mathbf{b}}$ — матрица и вектор реально известной вместо (8.10) системы

$$\tilde{\mathbf{A}}\mathbf{x} = \tilde{\mathbf{b}}\,. \qquad (8.31)$$

Ясно, что неотрицательно определенная матрица $\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}$ за счет подходящего сдвига подбором положительного скаляра $\alpha > 0$ может быть превращена в положительно определенную устойчиво обратимую матрицу $\alpha\mathbf{E} + \tilde{\mathbf{A}}^T\tilde{\mathbf{A}}$, позволяющую получить регуляризованное нормальное псевдорешение

$$\mathbf{x}_\alpha^* = (\alpha\mathbf{E} + \tilde{\mathbf{A}}^T\tilde{\mathbf{A}})^{-1}\tilde{\mathbf{A}}^T\tilde{\mathbf{b}}\,. \qquad (8.32)$$

Как правило, без привлечения дополнительной информации о решаемой задаче (8.10) указать оптимальное (в оговоренном выше смысле) значение параметра α не представляется возможным. Однако некоторые рекомендации по выбору подходящих значений α, требующие лишь знания о том, имеет ли исходная система решение, существуют.

Так, в книге [17] показан процесс оценивания погрешности $\| \mathbf{x}_0^+ - \mathbf{x}_\alpha^* \|_2$ в предположении (не отражающемся, как утверждается, на общности результатов), что в исходной системе (8.10) и в приближенной системе (8.31) матрицы $\mathbf{A}$ и $\tilde{\mathbf{A}}$ соответственно являются диагональными матрицами Σ и $\tilde{\Sigma}$ с сингулярными числами на диагонали. Итогом служит следующий вывод.

Пусть норма расхождения в точных и приближенных данных $\mathbf{A}, \mathbf{b}$ *и* $\tilde{\mathbf{A}}, \tilde{\mathbf{b}}$ *оценивается одной и той же величиной* ε. *Тогда, если система* (8.10) *имеет решение, следует взять в* (8.30) $\alpha := \varepsilon^{2/3}$, *и ошибка результата* (8.32) *при этом составит величину* $O(\varepsilon^{2/3})$; *в противном случае нужно положить* $\alpha := \varepsilon^{1/2}$ *с ошибкой в решении* $O(\varepsilon^{1/2})$.

Пример 8.3. Еще раз обратимся к системам (8.17), (8.18) примера 8.1, числовые данные которых различаются на величину $\varepsilon = 0{,}01$. Для упрощения счета допустим, что именно система (8.17) является приближенно известной, а истинная система, точные данные которой неизвестны, может быть любой из таких систем, матрицы коэффициентов и правые части которых могут отличаться по норме от фигурирующих в системе (8.17) не более чем на 0,01; в частности, это может оказаться система (8.18). Тогда в уравнении Тихонова (8.30), имеющем в данном случае вид

$$\alpha \begin{pmatrix} x \\ y \end{pmatrix} + \tilde{\mathbf{A}}^T \tilde{\mathbf{A}} \begin{pmatrix} x \\ y \end{pmatrix} = \tilde{\mathbf{A}}^T \mathbf{b},$$

полагаем $\tilde{\mathbf{A}} := \begin{pmatrix} 1 & 10 \\ 100 & 1001 \end{pmatrix}$, $\tilde{\mathbf{b}} := \begin{pmatrix} 11 \\ 1101 \end{pmatrix}$ и приходим к α-параметризованной системе

$$\begin{cases} (10001+\alpha)x + 100110y = 110111, \\ 100110x + (1002101+\alpha)y = 1102211. \end{cases} \tag{8.33}$$

Предположив, что истинная система непротиворечива, в соответствии с приведенным выше выводом зафиксируем в (8.33) $\alpha := \sqrt[3]{\varepsilon^2} \approx 0{,}05$. С этим значением параметра α получаем регуляризованное нормальное псевдорешение $x_\alpha^* \approx 0{,}109$, $y_\alpha^* \approx 1{,}089$. Если же исходить из предположения, что исходная система решений не имеет, то в систему (8.33) подставляем $\alpha := \sqrt{\varepsilon} := 0{,}1$. Это приводит к несколько иному регуляризованному нормальному псевдорешению $x_\alpha^* \approx 0{,}099$, $y_\alpha^* \approx 1{,}089$. Заметим, что определитель матрицы симметризованной системы (8.33) при этом изменяется от значения 1 в случае

$\alpha := 0$, т.е. в отсутствие регуляризации, до значений $O(10^6)$ при $\alpha := 0{,}1$ и $O(5 \cdot 10^5)$ при $\alpha := 0{,}05$.

Еще один путь улучшения обусловленности линейных систем связан с их решением наиболее трудоемким способом — посредством сингулярного разложения. Здесь следует отметить два момента. Во-первых, проблемы могут возникнуть в процессе выполнения самого SVD-разложения при наличии близких к нулю сингулярных чисел (и тем более, групп таких чисел). Во-вторых, наличие очень маленьких сингулярных чисел, как правило, означает плохую обусловленность матрицы системы (см. соответствующее определение $\mathrm{cond}\,\mathbf{A}$ в § 7.1), что влечет неустойчивость ее решения. Выход состоит в использовании ***усеченного сингулярного разложения***, или иначе, ***сингулярного разложения неполного ранга***. Такой способ регуляризации, состоящий в искусственном занижении ранга решаемой системы, достаточно подробно проанализирован ранее при рассмотрении линейной задачи наименьших квадратов (§ 7.6). Как и при α-регуляризации Тихонова, здесь так же остро стоит вопрос о выборе подходящего значения параметра регуляризации — в данном случае величины допуска, устанавливающего границу тех сингулярных чисел, которые следует заменить нулем. Без привлечения дополнительной информации о задаче и вычислительной среде, в которой она решается, численные значения параметра указаны быть не могут.

УПРАЖНЕНИЯ

8.1. Выполните подсчет арифметической сложности:

а) метода встречных прогонок;

б) процесса $\mathbf{U}^T\mathbf{D}\mathbf{U}$-разложения симметричных матриц и основанного на нем метода решения симметричных систем;

в) обращения матриц на основе треугольной факторизации.

8.2. Каковы вычислительные затраты, приходящиеся на один итерационный шаг метода Зейделя решения n-мерной СЛАУ?

8.3. Оцените число арифметических действий, достаточное для решения n-мерной СЛАУ методом сопряженных градиентов, рассматривая его как прямой метод.

8.4. Пусть методом Якоби решение системы

$$b_i x_{i-1} + c_i x_i + d_i x_{i+1} = r_i \quad (i = 1,\ 2,\ \ldots,\ n; \qquad b_1 = d_n := 0)$$

с нужной точностью достижимо за k шагов. Существуют ли такие значения k и n, при которых применение метода Якоби в этой ситуации эффективнее метода прогонки по числу арифметических операций?

8.5. Предположим, что некоторая n-мерная система вида $\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{c}$, у которой $\|\mathbf{B}\| \approx 0{,}5$, решается методом простых итераций с уровнем абсолютных погрешностей арифметических операций порядка 10^{-6}. Допустим, что при этом $\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\| \approx 1$. Каким числом следует ограничить количество итераций, чтобы вычислительная погрешность не стала существенно превышать погрешность метода?

8.6. Докажите справедливость неравенств (8.15).

8.7. Известно, что правая часть системы

$$\begin{cases} 3x - 2y = 2{,}99, \\ 301x - 201y = 299{,}99 \end{cases}$$

содержит ошибку в последнем знаке. Как она может отразиться на решении системы?

8.8. В условиях упр.8.7 составьте уравнение Тихонова с подходящим конкретным значением параметра регуляризации.

8.9. Приведите систему

$$\begin{cases} 100x_1 + 10x_2 + x_3 = 111, \\ 1000x_1 + 101x_2 + 10x_3 = 1111, \\ 10010x_1 + 1000x_2 + 101x_3 = 11111 \end{cases}$$

к эквивалентной системе с равновесной матрицей.

Сравните числа обусловленности исходной и преобразованной матриц.

Приложение

НЕКОТОРЫЕ ВСПОМОГАТЕЛЬНЫЕ СВЕДЕНИЯ

1. О векторных и матричных нормах. Изучение свойств и обоснование численных методов решения задач линейной алгебры немыслимо без привлечения отдельных элементарных понятий функционального анализа, конкретней, анализа конечномерных пространств. Одним из наиболее подходящих инструментов как при исследовании, так и в процессе применения численных методов является норма. Норма обобщает широко употребительное понятие длины вектора на плоскости и в трехмерном пространстве и позволяет сравнивать между собой элементы неупорядоченных множеств, каковыми являются множества векторов. Как и многие другие математические понятия, норма определяется аксиоматически.

Пусть $\mathbb{R}_n$ — множество всех n-мерных вещественных векторов, образующее вместе с определенными в нем естественными операциями сложения и умножения на число линейное пространство.

Определение 1. ***Нормой вектора*** $\mathbf{x} \in \mathbb{R}_n$ *называется такое действительное число, обозначаемое* $\|\mathbf{x}\|$, *что*:

1) $\|\mathbf{x}\| \geq 0$, *причем* $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$;

2) $\|\lambda \mathbf{x}\| = |\lambda| \cdot \|\mathbf{x}\| \quad \forall\, \lambda \in \mathbb{R}_1$ (*аксиома однородности*);

3) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad \forall\, \mathbf{y} \in \mathbb{R}_n$ (*неравенство треугольника*).

Линейное пространство $\mathbb{R}_n$ с введенной в нем нормой называют ***нормированным пространством*** (точнее, вещественным нормированным пространством). Вектор, норма которого равна единице, называют ***нормированным вектором*** в соответствующем нормированном пространстве.

Легко видеть, что обычное понятие длины (модуля) вектора удовлетворяет всем аксиомам, определяющим норму, т.е. длина

есть норма. Обратное неверно: определение 1 норму вектора однозначно не задает. Можно указать бесчисленное множество конструкций $\|\mathbf{x}\|$, удовлетворяющих всем трем аксиомам нормы вектора. Например, нетрудно проверить, что нормой вектора $\mathbf{x} := (x_1; \ldots; x_n)$ имеет право называться выражение

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \qquad (1)$$

при любом $p \in [1, \infty)$. Это так называемая l_p-***норма***, или ***норма Гёльдера*** [18]. Фиксированием значений параметра p отсюда можно получать конкретные привлекательные по тем или иным соображениям нормы.

Так, при $p = 2$ из общей формулы (1) получаем ***евклидову норму вектора***

$$\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^{n} |x_i|^2}, \qquad (2)$$

являющуюся наиболее естественным расширением понятия длины вектора на n-мерный случай. Множество $\mathbb{R}_n$ всех n-мерных векторов с введенной в нем евклидовой нормой называют ***евклидовым пространством*** и часто обозначают через $\mathbb{E}_n$. Это пространство характерно тем, что в нем определено ***скалярное произведение*** векторов $\mathbf{x} := (x_1; \ldots; x_n)$ и $\mathbf{y} := (y_1; \ldots; y_n)$ равенством

$$(\mathbf{x}, \mathbf{y}) := \sum_{i=1}^{n} x_i y_i$$

и при этом, очевидно, имеет место связь

$$\|\mathbf{x}\|_2 = \sqrt{(\mathbf{x}, \mathbf{x})}.$$

При $p = 1$ из l_p-нормы получается ***норма-сумма*** (иначе, ***первая норма***)

$$\|\mathbf{x}\|_1 := \sum_{i=1}^{n} |x_i|,$$

а переход в выражении (1) к пределу при $p \to \infty$ дает так называемую ***норму-максимум*** (иначе, ***норму-бесконечность***)

$$\|\mathbf{x}\|_\infty := \max_{i=1,\,n} |x_i| .$$

Указанные наиболее распространенные на практике три нормы $\|\mathbf{x}\|_2$, $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_\infty$ называют еще соответственно *сферической*, *октаэдрической* и *кубической* по названию поверхности, определяемой уравнением $\|\mathbf{x}\| = \text{const}$, если $\mathbf{x}$ считать переменным трехмерным радиус-вектором.

Понятие нормы тесно связано с несколько более широким понятием ***метрики***, характеризующим расстояние между элементами множества (не обязательно являющегося линейным пространством) и также, подобно норме, вводимым аксиоматически. В этом контексте можно сказать, что норма элемента определяет расстояние от него до нуля соответствующего пространства.

Пусть теперь рассматривается множество всех $n \times n$-матриц с вещественными элементами, также образующих линейное пространство $\mathbb{R}_{n\times n}$.

Определение 2. ***Нормой матрицы*** $\mathbf{A}$ *называется действительное число* $\|\mathbf{A}\|$, *удовлетворяющее условиям*:

1) $\|\mathbf{A}\| \ge 0$, *причем* $\|\mathbf{A}\| = 0 \;\Leftrightarrow\; \mathbf{A} = \mathbf{0}$;

2) $\|\lambda\mathbf{A}\| = |\lambda| \cdot \|\mathbf{A}\| \quad \forall\, \lambda \in \mathbb{R}_1$;

3) $\|\mathbf{A} + \mathbf{B}\| \le \|\mathbf{A}\| + \|\mathbf{B}\|$;

4) $\|\mathbf{AB}\| \le \|\mathbf{A}\| \cdot \|\mathbf{B}\|$ ($\mathbf{B}$ — *произвольная* $n \times n$-*матрица*).

Норму матрицы, определяемую только с помощью первых трех условий (аксиом), называют ***аддитивной***, или ***обобщенной***, ***матричной нормой*** [18]. С участием четвертого условия определение 2 задает в пространстве матриц ***мультипликативную норму***.

Как и норму вектора, норму матрицы, отвечающую определению 2, можно вводить далеко не единственным способом. Из множества всевозможных норм матрицы наибольший интерес пред-

ставляют такие, которые определенным образом соотносятся с векторными нормами, поскольку чаще всего матрицы и векторы рассматриваются в комплексе. Так, при умножении матрицы $\mathbf{A}$ на вектор $\mathbf{x}$ получается вектор $\mathbf{Ax}$, и естественно потребовать, чтобы матричная норма удовлетворяла ***условию согласованности***

$$\|\mathbf{Ax}\| \le \|\mathbf{A}\| \cdot \|\mathbf{x}\|$$

(в развитие аксиомы 4 определения 2 его можно рассматривать как некоторое смешанное условие мультипликативности).

Для матрицы $\mathbf{A} := (a_{ij})_{i,j=1}^{n}$ норму можно задать, например, следующим образом:

$$\|\mathbf{A}\|_F := \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2}\,. \qquad (3)$$

Легко проверить, что она удовлетворяет всем четырем аксиомам определения 2 и согласована с евклидовой нормой вектора. Такую норму называют ***нормой Фробениуса*** [23, 54], ***евклидовой***, или ***шуровской нормой*** [46, 67], ***сферической нормой*** [16] а также ***нормой Э. Шмидта*** [39] или ***Гильберта–Шмидта*** [54].

Более сильным требованием к норме матрицы, чем условие согласованности, является условие подчиненности. А именно: норма $n \times n$-матрицы $\mathbf{A}$ называется ***подчиненной*** норме n-мерного вектора $\mathbf{x}$ (или ***индуцированной*** ею), если она задана равенством

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

Таким образом, при заданной векторной норме за подчиненную ей норму матрицы $\mathbf{A}$ принимают максимум норм векторов $\mathbf{Ax}$, когда $\mathbf{x}$ пробегает множество векторов, норма которых равна единице.

Очевидно, при всякой подчиненной норме для единичной матрицы $\mathbf{E}$ должно быть справедливым следующее:

$$\|\mathbf{E}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{Ex}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{x}\| = 1\,.$$

Так как $\|\mathbf{E}\|_F = \sqrt{\sum_{i=1}^{n} 1^2} = \sqrt{n}$, то введенная выше норма Фробениу-

са, будучи согласованной с евклидовой нормой вектора, не является подчиненной ей. Другим примером согласованной с евклидовой нормой вектора, но ей не подчиненной, может служить так называемая M-*норма* $n\times n$-матрицы $\mathbf{A}$, определяемая как $M(\mathbf{A}) := n \cdot \max\limits_{i,j} |a_{ij}|$; очевидно, M-норма матрицы $\mathbf{E}$ равна n.

Нормой вещественной матрицы $\mathbf{A}$, подчиненной евклидовой норме вектора, служит ***спектральная норма***

$$\|\mathbf{A}\|_2 := \sqrt{\Lambda},$$

где Λ — наибольшее собственное число матрицы $\mathbf{A}^T\mathbf{A}$. В иной терминологии $\|\mathbf{A}\|_2$ — это есть наибольшее сингулярное число матрицы $\mathbf{A}$.

Нормами матрицы $\mathbf{A} := (a_{ij})_{i,j=1}^{n}$, подчиненными другим введенным выше векторным нормам, являются:

$$\|\mathbf{A}\|_1 := \max_j \sum_{i=1}^{n} |a_{ij}| \quad \text{— для } \|\mathbf{x}\|_1,$$

$$\|\mathbf{A}\|_\infty := \max_i \sum_{j=1}^{n} |a_{ij}| \quad \text{— для } \|\mathbf{x}\|_\infty.$$

В тех случаях, когда безразлично, какую из норм следует употребить, пишут просто $\|\mathbf{x}\|$ или $\|\mathbf{A}\|$, понимая под этим любые нормы (или любые подчиненные, или любые согласованные, что бывает ясно из рассмотрения объектов, с которыми оперируют). Так, например, при любой мультипликативной норме матрицы очевидным следствием четвертой аксиомы является неравенство

$$\|\mathbf{A}^k\| \le \|\mathbf{A}\|^k \quad \forall\, k \in \mathbb{N}.$$

Кроме того, с любой матричной нормой имеет место оценка

$$\|\mathbf{A}\| \ge |\lambda_{\mathbf{A}}|,$$

где $\lambda_{\mathbf{A}}$ — произвольное собственное число матрицы $\mathbf{A}$. Иначе, любая норма матрицы не меньше, чем ее ***спектральный радиус*** $\rho_{\mathbf{A}} := \max |\lambda_{\mathbf{A}}|$.

Пусть имеется последовательность $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(k)}, \ldots$, члены которой — n-мерные векторы $\mathbf{x}^{(k)} := (x_1^{(k)}; x_2^{(k)}; \ldots; x_n^{(k)})$. Если все последовательности $(x_i^{(k)})$ соответствующих компонент этих векторов имеют пределы, т.е. при всех $i \in \{1, 2, \ldots, n\}$ существуют значения $x_i^* := \lim\limits_{k\to\infty} x_i^{(k)}$, то вектор $\mathbf{x}^* := (x_1^*; x_2^*; \ldots; x_n^*)$ считают *пределом векторной последовательности* $(\mathbf{x}^{(k)})_{k=1}^{\infty}$. Этот факт сходимости оформляют обычным образом:

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}^* \quad \text{или} \quad \mathbf{x}^{(k)} \xrightarrow[k\to\infty]{} \mathbf{x}^*.$$

Так же покоординатно (поэлементно) понимают и сходимость последовательностей матриц. А именно, *пределом последовательности матриц* $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(k)}, \ldots$ с элементами, соответственно, $a_{ij}^{(1)}, a_{ij}^{(2)}, \ldots, a_{ij}^{(k)}, \ldots$ называется такая матрица $\mathbf{A}^*$, элементами которой являются числа $a_{ij}^* := \lim\limits_{k\to\infty} a_{ij}^{(k)}$.

Необходимым и достаточным условием *поэлементной* сходимости последовательности векторов $\mathbf{x}^{(k)}$ к вектору $\mathbf{x}^*$, последовательности матриц $\mathbf{A}^{(k)}$ к матрице $\mathbf{A}^*$ является условие ***сходимости по норме***:

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \to 0, \quad \|\mathbf{A}^{(k)} - \mathbf{A}^*\| \to 0,$$

соответственно. Оценка близости вектора $\mathbf{x}^{(k)}$ к вектору $\mathbf{x}^*$, матрицы $\mathbf{A}^{(k)}$ к матрице $\mathbf{A}^*$ также производится с помощью соответствующих норм: величины $\|\mathbf{x}^* - \mathbf{x}^{(k)}\|$, $\|\mathbf{A}^* - \mathbf{A}^{(k)}\|$ характеризуют ***абсолютную погрешность*** в приближенных равенствах $\mathbf{x}^* \approx \mathbf{x}^{(k)}$, $\mathbf{A}^* \approx \mathbf{A}^{(k)}$, а величины $\|\mathbf{x}^* - \mathbf{x}^{(k)}\| / \|\mathbf{x}^*\|$, $\|\mathbf{A}^* - \mathbf{A}^{(k)}\| / \|\mathbf{A}^*\|$ — ***относительную погрешность***.

Важно отметить, что все нормы в конечномерных пространствах (каковыми являются векторные и матричные пространства) эквивалентны. Здесь данный факт трактуется так: если доказана сходимость векторной или матричной последовательности в одной нормировке, то сходимость к тому же пределу обеспечена и при любой другой нормировке.

Более общо и конкретно ***эквивалентность*** двух разных норм $\|\cdot\|^{(1)}$ и $\|\cdot\|^{(2)}$ в одном пространстве определяется посредством установления между ними неравенств вида

$$c_1 \cdot \|\cdot\|^{(1)} \le \|\cdot\|^{(2)} \le c_2 \cdot \|\cdot\|^{(1)}$$

с некоторыми постоянными c_1, c_2. Например, имеют место следующие соотношения между вышеупомянутыми матричными нормами [18, 43 и др.]:

$$\frac{1}{n}\|\mathbf{A}\|_\infty \le \|\mathbf{A}\|_1 \le n\|\mathbf{A}\|_\infty,$$

$$\frac{1}{\sqrt{n}}\|\mathbf{A}\|_2 \le \|\mathbf{A}\|_\infty \le \sqrt{n}\|\mathbf{A}\|_2,$$

$$\frac{1}{\sqrt{n}}\|\mathbf{A}\|_F \le \|\mathbf{A}\|_2 \le \|\mathbf{A}\|_F,$$

$$\frac{1}{\sqrt{n}}M(\mathbf{A}) \le \|\mathbf{A}\|_F \le M(\mathbf{A}).$$

Введенные здесь понятия и определения легко распространяются на случай векторов и матриц с комплексными элементами (что делает оправданным присутствие модуля, например, в формулах (2) и (3)), а также на случай неквадратных матриц.

2. Об особенностях машинной арифметики. Пусть в основу запоминающего устройства вычислительной машины (компьютера) положены однотипные физические устройства (базисные элементы), имеющие r устойчивых состояний (как правило, $r := 2, 8, 16, \ldots$; чаще всего используется значение $r := 2$), причем каждому числу соответствует одинаковое количество k этих эле-

ментов и, кроме того, с помощью таких или более простых элементов фиксируется знак числа. Упорядоченные элементы образуют разрядную сетку машинного слова: в каждом разряде может быть записано только одно из базисных чисел $0, 1, \ldots, r-1$ (одна из r цифр r-ичной системы счисления) и в специальном разряде отображен знак «+» или «–».

При записи ***числа с фиксированной запятой***, кроме упомянутых параметров r (основания системы счисления) и k (количества разрядов, отводимых под запись цифр числа), указывается еще количество l разрядов, выделяемых под дробную часть числа. Таким образом, положительное вещественное число a, представляющее собой в r-ичной системе бесконечную, вообще говоря, непериодическую дробь, здесь будет отображено конечной последовательностью*

$$\alpha_1\alpha_2\ldots\alpha_{k-l}\alpha_{k-l+1}\ldots\alpha_{k-1}\alpha_k,$$

где $\alpha_i \in \{0, 1, \ldots, r-1\}$, т.е. реализуется приближенное равенство

$$a \approx \mathrm{fix}(a) := \alpha_1 r^{k-l-1} + \alpha_2 r^{k-l-2} + \cdots + \alpha_{k-l} r^0 + \\ + \alpha_{k-l+1} r^{-1} + \cdots + \alpha_{k-1} r^{-(l-1)} + \alpha_k r^{-l}.$$

Диапазон представляемых таким способом чисел определяется числами с наибольшими цифрами во всех разрядах, т.е. наименьшим числом $-(r-1)(r-1)\ldots(r-1)$ и наибольшим числом $(r-1)(r-1)\ldots(r-1)$, а ***абсолютная точность*** представления есть оценка величины $|a - \mathrm{fix}(a)|$, зависящая от способа округления. Она равна r^{-l} при простом отбрасывании «хвоста» $\alpha_{k+1} r^{-(l+1)} + \alpha_{k+2} r^{-(l+2)} + \ldots$ числа a и половине этой величины при ***правильном округлении*** (т.е. при увеличении α_k на единицу, если $\alpha_{k+1} > r/2$). Заметим, что абсолютная точность представления вещественных чисел с фиксированной запятой

---

* В записи не отражены разделители элементов последовательности.

одинакова в любой части диапазона. В то же время ***относительная точность***, т.е. оценка величины $|a - \mathrm{fix}(a)|/|a|$ (или $|a - \mathrm{fix}(a)|/|\mathrm{fix}(a)|$), очевидно, может значительно различаться в зависимости от того, к чему близко число a: к нулю или к границе диапазона. Иными словами, *вещественные числа с фиксированной запятой имеют равномерную абсолютную плотность распределения на всем отрезке вещественной оси, определяемом границами диапазона, и неравномерную, возрастающую к границам отрезка, относительную плотность распределения.*

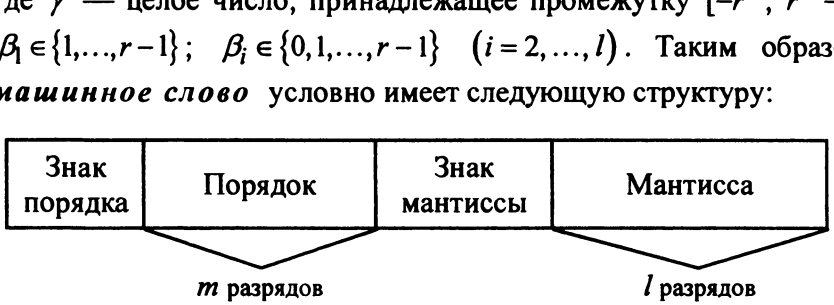
В основе значительно чаще употребляемого представления ***с плавающей запятой*** (или, что то же, ***с плавающей точкой***) лежит следующая экспоненциальная форма записи вещественного числа:

$$a := M \cdot r^p,$$

где r — ***основание***, p — ***порядок***, а M — ***мантисса***, такое число, что $r^{-1} \le |M| < 1 \left(= r^0\right)$. Если под мантиссу выделяется l r-ичных элементов, а под порядок m, то в системе записи с плавающей запятой вещественное число a представляется конечным числом $\mathrm{fl}(a)$ (от англ. *floating* — плавающий) вида

$$a \approx \mathrm{fl}(a) := \pm\left(\beta_1 r^{-1} + \beta_2 r^{-2} + \ldots + \beta_l r^{-l}\right) \cdot r^{\gamma},$$

где γ — целое число, принадлежащее промежутку $[-r^m, r^m - 1]$; $\beta_1 \in \{1, \ldots, r-1\}$; $\beta_i \in \{0, 1, \ldots, r-1\}$ $(i = 2, \ldots, l)$. Таким образом, ***машинное слово*** условно имеет следующую структуру:



Числа $\pm r^{r^m}$ определяют границы допустимого диапазона множества компьютерных чисел типа $\mathrm{fl}(a)$. Более содержательна

информация о диапазоне представимости положительных вещественных чисел, составляющем промежуток $[r^{-r^m}, r^{r^m-1}]$. Левую и правую границы этого отрезка называют соответственно ***машинным нулем*** и ***машинной бесконечностью***, так как числа из промежутка $[-r^{-r^m}, r^{-r^m}]$ машина должна заменить нулем, а числа, лежащие за пределами промежутка $[-r^{r^m-1}, r^{r^m-1}]$, она не воспринимает (без специальных ухищрений).

Важной характеристикой является число ε, называемое ***машинный эпсилон*** и обозначаемое обычно идентификатором **macheps**. Эта характеристика определяется как расстояние между единицей и ближайшим следующим за ней числом системы машинных чисел с плавающей запятой. Так как

$$1 = \left(1\cdot r^{-1}+0\cdot r^{-2}+\ldots+0\cdot r^{-l}+\ldots\right)\cdot r^{1},$$

а следующее за единицей машинное число есть

$$\left(1\cdot r^{-1}+0\cdot r^{-2}+\ldots+0\cdot r^{-(l-1)}+1\cdot r^{-l}\right)\cdot r^{1} = \mathrm{fl}\left(1+\varepsilon\right),$$

то за macheps можно принять величину

$$\varepsilon = 1\cdot r^{-l}\cdot r^{1} = r^{1-l}.$$

Это число напрямую связано с относительной погрешностью представления чисел в системе с плавающей запятой. Имеем:

$$\left|\frac{a-\mathrm{fl}(a)}{a}\right| = \frac{\beta_{l+1}r^{-(l+1)}+\beta_{l+2}r^{-(l+2)}+\ldots}{\beta_1 r^{-1}+\beta_2 r^{-2}+\ldots} \le \frac{1\cdot r^{-l}}{\beta_1\cdot r^{-1}} \le r^{1-l} = \varepsilon. \quad (4)$$

Таким образом, *машинный эпсилон служит мерой относительной точности представления вещественных чисел*, причем эта точность одинакова в любой части числового диапазона и зависит лишь от числа r-ичных разрядов, отводимых под мантиссу числа. В то же время оценка абсолютной погрешности

$$\left|a-\mathrm{fl}(a)\right| \le |a|\cdot r^{1-l}$$

показывает, что расстояние между вещественными числами и

конечными приближениями к ним в системе с плавающей запятой неодинаковы в разных частях числового диапазона: *абсолютная плотность машинных чисел больше вблизи нуля при одинаковой относительной плотности их распределения.*

Замечание 1. Если трактовать macheps как минимальное положительное действительное число, прибавление которого к 1 дает следующее за 1 число с плавающей запятой, то, очевидно, при правильном округлении значение macheps будет в два раза меньшим. Действительно, полагая $\varepsilon := \frac{1}{2} r^{1-l} = \frac{r}{2} \cdot r^{-(l+1)} \cdot r^{1}$, получаем

$$1+\varepsilon = \left(1 \cdot r^{-1} + 0 \cdot r^{-2} + \ldots + 0 \cdot r^{-l} + \frac{r}{2} \cdot r^{-(l+1)}\right) \cdot r^{1},$$

и значит, $\mathrm{fl}(1+\varepsilon) = 1 + r^{1-l} > 1$. «Мера дискретности» множества машинных чисел, как видим, остается той же: r^{1-l}.

Замечание 2. Величина macheps служит оценкой относительной точности представления вещественного числа a при условии, что $|a| > r^{-r^m}$. Если же $a \in [-r^{-r^m}, r^{-r^m}]$, то $\mathrm{fl}(a) \equiv 0$. Следовательно, относительная погрешность компьютерного числа $\mathrm{fl}(a)$ суть $|a - \mathrm{fl}(a)| / |a| \equiv 1$, т.е. является постоянной достаточно большой величиной, в то время как абсолютная погрешность не превосходит величины r^{-r^m}.

В большинстве персональных компьютеров, в частности, производимых на базе процессоров Intel в соответствии со стандартом IEEE 754, предусматривается наличие двух двоичных форматов с плавающей запятой: одинарной точности с выделением под мантиссу 24 разрядов (4 байта памяти) и двойной точности с 53-разрядной мантиссой (8 байт). Это соответствует 6...9 десятичным разрядам в первом и 15...17 десятичным разрядам во втором случаях [66]. Значения машинного нуля M_0 и машинной бесконечности M_∞ при этом регламентируются величинами $M_0 :\approx 1{,}2 \cdot 10^{-38}$, $M_\infty :\approx 3{,}4 \cdot 10^{38}$ при одинарной точности и $M_0 :\approx 2{,}2 \cdot 10^{-308}$, $M_\infty :\approx 1{,}8 \cdot 10^{308}$ при двойной точности.

В разных языках программирования машинным числам одинарной и двойной точности соответствуют свои типы представления данных: на Си — это `float` и `double`, на Паскале — `single` и `double`, на Фортране — `real` и `double precision`. Кроме двух указанных стандартов компьютерного отображения вещественных чисел, поддерживаемых аппаратно, языки программирования позволяют оперировать с машинными числами повышенной точности в расширенном диапазоне. Например, под числа типа `long double` в Си выделяется 10 байт памяти, что соответствует 18–20-разрядной мантиссе и границам диапазона положительных чисел $\approx 1{,}2 \cdot 10^{\mp 4932}$ [53]. На Паскале примерно такие же характеристики имеют числа типа `extended`. Языковые средства позволяют задавать числа и промежуточной точности между одинарной и двойной. Так, паскалевские шестибайтовые числа типа `real` имеют мантиссы, обеспечивающие точность в 11–12 десятичных разрядов.

С помощью несложных программ, написанных, например, по образцу из [28], можно самому конкретизировать параметры множеств машинных чисел разных типов представления в той или иной среде программирования в используемом компьютере.

Заметим, что стандартом IEEE допускается снижение границы машинного нуля до величины $\approx 1{,}4 \cdot 10^{-45}$ при одинарной точности и до величины $\approx 4{,}9 \cdot 10^{-324}$ при двойной точности (за счет укорачивания мантиссы).

Обращаясь к арифметическим операциям над машинными числами, прежде всего, обратим внимание на то, что они утрачивают привычные свойства вещественных чисел. Особенно это касается свойств ассоциативности и дистрибутивности, нарушаемых при выполнении арифметических операций на любых ЭВМ (сохранение или несохранение свойства коммутативности связывают со способом округления чисел). Так, весьма утрированный пример сравнения выражения $\left(r^{D/2} \cdot r^{3D/4}\right) \cdot r^{-D/2}$ с выражением $r^{D/2} \cdot \left(r^{3D/4} \cdot r^{-D/2}\right)$, где число D таково, что r^D есть правая

граница числового диапазона, показывает существенность способа расстановок скобок при умножении: в первом случае машина выдаст сообщение о переполнении, поскольку $r^{D/2}\cdot r^{3D/4} > r^{D}$, а во втором будет получен правильный результат. Легко также представить ситуацию с тремя положительными числами a, b, c, когда расставляя по-разному скобки в выражении $a+b-c$, будем получать (или не получать вовсе) разные результаты.

Изучение погрешностей результатов арифметических операций над числами с плавающей запятой осуществляют с помощью представления

$$\mathrm{fl}(a) = a(1+\delta), \quad \text{где} \quad |\delta| \le \mathrm{macheps} \tag{5}$$

(чтобы убедиться в справедливости (5), достаточно ввести δ равенством $\delta := (\mathrm{fl}(a)-a)/a$, равносильным равенству, фигурирующему в (5), и воспользоваться доказанным в (4) неравенством $|\delta| \le \mathrm{macheps}$). Принимая во внимание, что операции над двумя машинными числами a и b осуществляются точно (здесь используется двойная длина машинного слова), после чего производится округление, результат любой арифметической операции $\otimes$ также может быть записан в виде

$$\mathrm{fl}(a \otimes b) = (a \otimes b)(1+\delta_1), \tag{6}$$

где $|\delta_1| \le \varepsilon$ (за исключением особых случаев).

Пусть складываются последовательно три положительных числа: a_1, a_2, a_3. Тогда, согласно (6), имеем:

$$\mathrm{fl}(a_1+a_2) = (a_1+a_2)(1+\delta_1), \quad \text{где} \quad |\delta_1| \le \varepsilon;$$

$$\mathrm{fl}\big((a_1+a_2)+a_3\big) = \big((a_1+a_2)(1+\delta_1)+a_3\big)(1+\delta_2) =$$

$$= (a_1+a_2)(1+\delta_1)(1+\delta_2) + a_3(1+\delta_2), \quad \text{где } |\delta_i| \le \varepsilon \quad (i=1,\,2).$$

Заменяя здесь δ_i бóльшим значением ε, получим следующую оценку абсолютной погрешности суммы трех слагаемых:

$$\big|\mathrm{fl}(a_1+a_2+a_3) - (a_1+a_2+a_3)\big| \le 2(a_1+a_2)\varepsilon + a_3\varepsilon + (a_1+a_2)\varepsilon^2.$$

Обращает на себя внимание неравноправность слагаемых в образовании погрешности суммы: меньшую роль в ней играет последнее слагаемое. Природа этого факта очевидна: первые слагаемые неявно (в просуммированном виде) участвуют в процессе каждого последующего сложения.

Если пренебречь степенями ε выше первой, то для суммы n положительных чисел a_i нетрудно получить [2] приближенную оценку абсолютной погрешности вида*

$$\left| \mathrm{fl}\left(\sum_{i=1}^{n} a_i \right) - \sum_{i=1}^{n} a_i \right| \lesssim \left| (n-1)(a_1 + a_2) + (n-2)a_3 + \ldots + 2a_{n-1} + a_n \right| \cdot \varepsilon$$

при последовательном суммировании, начинающемся с a_1. Очевидно, для того, чтобы эта погрешность была минимальной, *последовательность чисел нужно суммировать в порядке возрастания членов*. Только за счет этого можно добиться уменьшения погрешности, как показано в [17], в $n/\log_2 n$ раз.

На основе изучения погрешности произведения нескольких чисел строят алгоритмы оптимального умножения.

Пусть требуется перемножить n чисел a_i, таких, что $|a_1| \le |a_2| \le \ldots \le |a_n|$. Погрешность произведения будет минимальной, если находить его по следующей схеме: умножать a_1 последовательно на a_n, a_{n-1}, ... до тех пор, пока модуль частичного произведения не станет большим единицы, затем это частичное произведение умножать на a_2, a_3, ... до тех пор, пока новое частичное произведение не станет по модулю меньшим единицы, и так далее до исчерпывания всех сомножителей [17].

Естественно, что расплатой за выигрыш в точности при реализации таких алгоритмов будет проигрыш в скорости счета.

---

* Знак «$\lesssim$» используется для обозначения неравенства в смысле главных (линейных) частей.

СПИСОК ЛИТЕРАТУРЫ

1. *Амосов А.А., Дубинский Ю.А., Копченова Н.В.* Вычислительные методы для инженеров. — М.: Высш. шк., 1994.

2. *Арушанян О.Б., Залеткин С.В.* Численное решение обыкновенных дифференциальных уравнений на Фортране. — М.: Изд-во МГУ, 1990.

3. *Бахвалов Н.С.* Численные методы. — М.: Наука, 1973.

4. *Бахвалов Н.С., Жидков Н.П., Кобельков Г.М.* Численные методы. — М.: Лаборатория Базовых Знаний, 2001.

5. *Бахвалов Н.С., Лапин А.В., Чижонков Е.В.* Численные методы в задачах и упражнениях. — М.: Высш. шк., 2000.

6. *Беланов А.А.* Решение алгебраических уравнений методом Лобачевского. — М.: Наука, 1989.

7. *Березин И.С., Жидков Н.П.* Методы вычислений. Т.2. — М.: Физматгиз, 1962.

8. *Бут Э.Д.* Численные методы. — М.: Физматгиз, 1959.

9. *Валях Е.* Последовательно-параллельные вычисления. — М.: Мир, 1985.

10. *Васин В.В., Агеев А.Л.* Некорректные задачи с априорной информацией. — Екатеринбург: УИФ «Наука», 1993.

11. *Вержбицкий В.М.* Обращение матриц и решение нелинейных систем. — Ижевск: Изд. ИМИ, 1980.

12. *Вержбицкий В.М.* Численные методы (линейная алгебра и нелинейные уравнения). — М.: Высш. шк., 2000.

13. *Вержбицкий В.М.* Основы численных методов. — 2-е изд. — М.: Высш. шк., 2005.

14. *Вержбицкий В.М., Качурина Т.И.* Ортогональное и сингулярное разложения матриц. — Ижевск: Изд. ИжГТУ, 2005.

15. *Верлань А.Ф., Сизиков В.С.* Интегральные уравнения: методы, алгоритмы, программы. — Киев: Наукова думка, 1986.

16. *Воеводин В.В.* Численные методы алгебры (теория и алгорифмы). — М.: Наука, 1966.

17. *Воеводин В.В.* Вычислительные основы линейной алгебры. — М.: Наука, 1977.

18. *Воеводин В.В., Кузнецов Ю.А.* Матрицы и вычисления. — М.: Наука, 1984.

19. *Волков Б.А.* Численные методы. — М.: Наука, 1979.

20. *Гантмахер Ф.Р.* Теория матриц. — 4-е изд. — М.: Наука, 1988.

21. *Годунов С.К.* Решение систем линейных уравнений. — Новосибирск: Наука, 1980.

22. *Годунов С.К., Рябенький В.С.* Введение в теорию разностных схем. — М.: Физматгиз, 1962.

23. *Голуб Дж., Ван Лоун Ч.* Матричные вычисления. — М.: Мир, 1999.

24. *Двайт Г.Б.* Таблица интегралов и другие математические формулы. — М.: Наука, 1966.

25. *Демидович Б.П., Марон И.А.* Основы вычислительной математики. — М.: Наука, 1970.

26. *Деммель Дж.* Вычислительная линейная алгебра. Теория и приложения. — М.: Мир, 2001.

27. *Джордж А., Лю Дж.* Численное решение больших разреженных систем уравнений. — М.: Мир, 1984.

28. *Дэннис Дж., Шнабель Р.* Численные методы безусловной оптимизации и решения нелинейных уравнений. — М.: Мир, 1988.

29. *Журкин И.Г., Нейман Ю.М.* Методы вычислений в геодезии. — М.: Недра, 1988.

30. *Икрамов Х.Д.* Численные методы линейной алгебры (решение линейных уравнений) // Математика, кибернетика. — №4. — М.: Знание, 1987.

31. *Икрамов Х.Д.* Численное решение матричных уравнений. — М.: Наука, 1984.

32. *Икрамов Х.Д.* Несимметричная проблема собственных значений. — М.: Наука, 1991.

33. *Ильин В.П., Кузнецов Ю.А.* Трехдиагональные матрицы и их приложения. — М.: Наука, 1985.

34. *Калиткин Н.Н.* Численные методы. — М.: Наука, 1978.

35. *Канторович Л.В., Акилов Г.П.* Функциональный анализ. — М.: Наука, 1977.

36. *Каханер Д., Моулер К., Нэш С.* Численные методы и программное обеспечение. — М.: Мир, 1998.

37. *Киреев В.И., Пантелеев А.В.* Численные методы в примерах и задачах. — М.: Высш.шк., 2006.

38. *Коллатц Л.* Задачи на собственные значения с техническими приложениями. — М. : Наука, 1968.

39. *Коллатц Л.* Функциональный анализ и вычислительная математика. — М.: Мир, 1969.

40. *Коллатц Л., Альбрехт Ю.* Задачи по прикладной математике. — М.: Мир, 1978.

41. *Косарев В.И.* 12 лекций по вычислительной математике. — М.: Изд-во МФТИ, 1995.

42. *Крылов В.И., Бобков В.В., Монастырный П.И.* Вычислительные методы. Т.1. — М.: Наука, 1976.

43. *Крылов В.И., Бобков В.В., Монастырный П.И.* Начала теории вычислительных методов. Линейная алгебра и нелинейные уравнения. — Минск: Наука и техника, 1985.

44. *Локуциевский О.В., Гавриков М.Б.* Начала численного анализа. — М.: ТОО «Янус», 1995.

45. *Лоусон Ч., Хенсон Р.* Численное решение задач метода наименьших квадратов. — М.: Наука, 1986.

46. *Маркус М., Минк Х.* Обзор по теории матриц и матричных неравенств. — М.: Наука, 1972.

47. Математика и САПР. В 2-х кн. Кн. 1 / *П.Шенен, М.Коснар, И.Гардан и др.* — М.: Мир, 1988.

48. *Молчанов И.Н.* Машинные методы решения прикладных задач. Дифференциальные уравнения. — Киев: Наукова думка, 1988.

49. *Морозов В.А.* Регулярные методы решения некорректно поставленных задач. — М.: Наука, 1987.

50. *Ортега Дж.* Введение в параллельные и векторные методы решения линейных систем. — М.: Мир, 1991.

51. *Ортега Дж., Пул У.* Введение в численные методы решения дифференциальных уравнений. — М.: Наука, 1986.

52. *Ортега Дж., Рейнболдт В.* Итерационные методы решения нелинейных систем уравнений со многими неизвестными. — М.: Мир, 1975.

53. *Паппас К., Мюррей У.* Эффективная работа: Visual C++.NET. — СПб.: Питер, 2002.

54. *Парлетт Б.* Симметричная проблема собственных значений. — М.: Мир, 1983.

55. *Плис А.И., Сливина Н.А.* Лабораторный практикум по высшей математике. — М.: Высш. шк., 1994.

56. *Прэтт У.* Цифровая обработка изображений. — М.: Мир, 1982.

57. *Райс Дж.* Матричные вычисления и математическое обеспечение. — М.: Мир, 1984.

58. *Рыжиков Ю.И.* Решение научно-технических задач на персональном компьютере. — СПб.: КОРОНА принт, 2000.

59. *Рябенький В.С.* Введение в вычислительную математику. — М.: Наука, 1994.

60. *Самарский А.А., Вабищевич П.Н., Самарская Е.А.* Задачи и упражнения по численным методам. Изд.3. — М.: КомКнига, 2007.

61. *Самарский А.А., Гулин А.В.* Численные методы. — М.: Наука, 1989.

62. *Самарский А.А., Николаев Е.С.* Методы решения сеточных уравнений. — М.: Наука, 1978.

63. Сборник задач по методам вычислений / Под ред. П.И. Мо-

настырного. — М.: Наука, 1994.

64. *Стренг Г.* Линейная алгебра и ее применения. — М.: Мир, 1980.

65. *Треногин В.А.* Функциональный анализ. — М.: Наука, 1980.

66. *Турчак Л.И., Плотников П.В.* Основы численных методов. — 2-е изд., перераб. и доп. — М.: ФИЗМАТЛИТ, 2002.

67. *Уилкинсон Дж.Х.* Алгебраическая проблема собственных значений. — М.: Наука, 1970.

68. *Уилкинсон Дж.Х., Райнш К.* Справочник алгоритмов на языке АЛГОЛ. Линейная алгебра. — М.: Машиностроение, 1976.

69. *Фаддеев Д.К., Фаддеева В.Н.* Вычислительные методы линейной алгебры. — М.: Физматгиз, 1960.

70. *Форсайт Дж., Малькольм М., Моулер К.* Машинные методы математических вычислений. — М.: Мир, 1980.

71. *Форсайт Дж., Молер К.* Численное решение систем линейных алгебраических уравнений. — М.: Мир, 1969.

72. *Хемминг Р.В.* Численные методы. — М.: Наука, 1968.

73. *Хорн Р., Джонсон Ч.* Матричный анализ. — М.: Мир, 1989.

74. *Шолохович Ф.А., Васин В.В.* Основы высшей математики. — Екатеринбург: Изд-во Урал. ун-та, 1999.

75. *Эстербю О., Златев З.* Прямые методы для разреженных матриц. — М.: Мир, 1987.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

УКАЗАТЕЛЬ ОБОЗНАЧЕНИЙ И СОКРАЩЕНИЙ

$:=$	положить по определению; присвоить 7
$:\approx$	принять приближенно 245
$\lesssim$	неравенство в смысле главных (линейных) частей 339
$O(\cdot)$, $o(\cdot)$	символы Ландау 159
δ_{ij}	символ Кронекера 60
$\mathrm{fix}(a)$	машинное число с фиксированной запятой 333
$\mathrm{fl}(a)$	машинное число с плавающей запятой 302, 334
masheps	машинный эпсилон 335
$\mathrm{sgn}_+(x)$	функция «сигнум-плюс» 26
$\delta_{\mathbf{x}}$, $\delta_{\mathbf{A}}$	оценка относительной погрешности вектора $\mathbf{x}$, матрицы $\mathbf{A}$ 309
$\mathbf{e}_i$	единичный вектор (орт) 25, 56
m_-, m_+	индексы отрицательности и положительности матрицы 150
$\mathbf{x}^+$	псевдорешение СЛАУ $\mathbf{Ax} = \mathbf{b}$ 259, 281
$\mathbf{x}_0$	нормальное решение СЛАУ $\mathbf{Ax} = \mathbf{b}$ 260
$\mathbf{x}_\alpha^*$	α - регуляризованное решение СЛАУ $\mathbf{Ax} = \mathbf{b}$ 322
$\mathbf{E}$	единичная матрица 7, 55
$\mathbf{H}$	матрица Хаусхолдера (отражения) 24
$\mathbf{H}_n$	$n \times n$ -матрица Гильберта 310
$\mathbf{L}$	левая (нижняя) треугольная матрица 9, 11
$\mathbf{R}$, $\mathbf{U}$	правая (верхняя) треугольная матрица 9, 11
$\mathbf{T}_{ij}$	матрица плоских вращений 36
$\mathbf{A}^{-1}$	матрица, обратная к $\mathbf{A}$ 16,48, 56
$\mathbf{A}^+$	псевдообратная матрица 261
$\mathbf{A}^*$	матрица, сопряженная с $\mathbf{A}$ 19, 115
$\mathbf{A}^T$, $\mathbf{x}^T$	транспонированные матрица, вектор 8, 24
$\det \mathbf{A}$	определитель матрицы $\mathbf{A}$ 7, 47, 54
$\mathbf{D}$, $\mathrm{diag}(\cdot;\ldots;\cdot)$	диагональная матрица 20, 31, 254

$\mathbb{N}$	множество натуральных чисел 303
$\mathbb{N}_0$	множество целых неотрицательных чисел 87
$\mathbb{R}, \mathbb{R}_n$	множества вещественных чисел, n -мерных векторов с вещественными координатами 326
$\mathbb{C}, \mathbb{C}_n$	множества комплексных чисел, n -мерных векторов с комплексными координатами 115
$\mathrm{Sp}\,\mathbf{A}$	след матрицы $\mathbf{A}$ 164
$\mathrm{rank}\,\mathbf{A}$	ранг матрицы $\mathbf{A}$ 254
$\lambda_{\mathbf{A}}, \lambda, \lambda_i$	собственное число матрицы $\mathbf{A}$ 142
$\{\lambda, \mathbf{x}\}$	собственная пара матрицы 142
$\rho_{\mathbf{A}}$	спектральный радиус матрицы $\mathbf{A}$ 91, 311, 330
$\rho(\mathbf{x})$	отношение Рэлея 146
$\mathrm{cond}\,\mathbf{A}, \nu(\mathbf{A})$	число (мера) обусловленности матрицы $\mathbf{A}$ 255, 308, 311
σ_i	сингулярное число матрицы 228
$\mathbf{\Sigma}$	диагональная матрица из сингулярных чисел 9, 230, 254
$(\mathbf{x}, \mathbf{y})$	скалярное произведение векторов $\mathbf{x}$ и $\mathbf{y}$ 327
$\|\cdot\|$	норма вектора, матрицы 326, 328
$\sum_{i=k}^{n} a_i$	сумма чисел a_i от a_k до a_n 12
$\prod_{i=k}^{n} a_i$	произведение чисел a_i от a_k до a_n 71
$\langle a_i \rangle$	среднее арифметическое чисел a_i 169
$\{a, b, ..., c\}$	множество элементов $a, b, ..., c$ 14, 52
inf, sup	точные нижняя и верхняя грани 311

ЛЗНК	линейная задача о наименьших квадратах 5, 274, 280
МНК	метод наименьших квадратов 271
МПИ	метод простых итераций 90
МСГ	метод сопряженных градиентов 126
ПВР	метод последовательной верхней релаксации 119
СЛАУ	система линейных алгебраических уравнений 3, 45
ЭВМ	электронно-вычислительная машина (компьютер) 45, 337

Учебное издание

Вержбицкий Валентин Михайлович

ВЫЧИСЛИТЕЛЬНАЯ ЛИНЕЙНАЯ АЛГЕБРА

Редактор *Т.А. Садчикова*
Внешнее оформление *К.И. Мандель*
Художественный редактор *А.Ю. Войткевич*
Технический редактор *Л.А. Маркова*
Корректор *Г.Н. Петрова*
Компьютерная верстка *В.М. Вержбицкого*

Изд. № РЕНТ-553. Подп. в печать 20.01.09. Формат $60 \times 88^{1}/_{16}$.
Бум. офсетная. Гарнитура «Таймс». Печать офсетная.
Объем 21,56 усл. печ. л., 22,30 усл. кр.-отт.
Тираж 3000 экз. Заказ № 10122.

ОАО «Издательство «Высшая школа»,
127994, Москва, Неглинная ул., д. 29/14, стр. 1.
Тел.: (495) 694-04-56. http://www.vshkola.ru. E-mail: info_vshkola@mail.ru

Отдел реализации: (495) 694-07-69, 694-31-47; факс: (495) 694-34-86.
E-mail: sales_vshkola@mail.ru

Отпечатано в ОАО «Тульская типография».
300600, г. Тула, пр. Ленина, 109.