КАЗАНСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Р.З. ДАУТОВ, М.М. КАРЧЕВСКИЙ

ОСНОВЫ ЧИСЛЕННЫХ МЕТОДОВ ЛИНЕЙНОЙ АЛГЕБРЫ

Учебное пособие

КАЗАНЬ

2018

Оглавление

Предисловие

В данном пособии, отражающем опыт преподавания в институте вычислительной математики и информационных технологий Казанского федерального университета, рассматриваются численные методы решения разнообразных задач, которые традиционно относят к задачам линейной алгебры. Это — вопросы решения систем линейных алгебраических уравнений, обращения матриц, вычисление определителей, нахождения собственных чисел и собственных векторов матриц.

Предполагается, что читатель знаком с основными разделами линейной алгебры, например, в объеме книги [5]. Зачастую мы используем обозначения и результаты из [5] без дополнительных оговорок.

Для большинства вычислительных задач, встречающихся на практике, характерным является большой порядок матрицы. В связи с эти, там где это возможно, указываются оценки трудоемкости описываемых алгоритмов. Эти оценки имеют существенное значение для сравнительного анализа численных методов решения задач линейной алгебры.

Надо иметь в виду, что, как правило, исходные данные, например, матрица и правая часть системы линейных уравнений, оказывается известными лишь приближенно, с некоторой погрешностью. Погрешности округления, неизбежные при вычислениях, зачастую также удается интерпретировать как погрешности задания исходных данных. Поэтому важными оказывается исследование устойчивости задач линейной алгебры по отношению к возмущениям исходных данных. Этим вопросам в пособии посвящена отдельная глава.

Особое место в книге занимает заключительная глава. Она содержит набор практических, вычислительных, заданий, относящихся к большинству изучаемых в пособии методов. Выполняя эти задания, студенты более детально знакомятся с алгоритмами решения типичных задач линейной алгебры, а также получают навыки их программной реализации.

В процессе работы над книгой авторы пользовались неизменной поддержкой и консультациями сотрудников кафедр вычислительной и прикладной математики КФУ. Мы выражаем им нашу искреннюю благодарность.

Глава 1

Примеры задач, приводящих к системам линейных алгебраических уравнений

Многие задачи практики приводят к необходимости решать системы линейных уравнений. При конструировании инженерных сооружений, приборов, обработке результатов измерений, решении задач планирования производственного процесса и многих других задач техники, экономики, научного эксперимента приходится решать системы линейных уравнений.

Исследование ряда научно-технических и экономических проблем приводит к математическим моделям непосредственно в форме систем линейных алгебраических уравнений. Однако гораздо чаще системы линейных уравнений появляются в процессе математического моделирования как промежуточный этап при решении более сложной задачи, например, после дискретизации или линеаризации интегральных, дифференциальных, интегро-дифференциальных уравнений или систем уравнений такого сорта.

В данной главе приводится далеко неполный набор задач, при решении которых возникает необходимость в решении систем линейных алгебраических уравнений.

1. Системы нелинейных уравнений

Пусть требуется найти общий корень $x = (x_1, x_2, \ldots, x_n)$ заданных n функций $f_i(x_1, x_2, \ldots, x_n)$, т. е. решение следующей системы нелинейных алгебраических уравнений

$$\begin{gathered} f_1(x_1, x_2, \ldots, x_n) = 0, \\ f_2(x_1, x_2, \ldots, x_n) = 0, \\ \ldots \\ f_n(x_1, x_2, \ldots, x_n) = 0. \end{gathered}$$

Компактно эту систему запишем в виде

$$F(x) = 0, \tag{1.1}$$

где $x \in \mathbb{R}^n$ или $x \in \mathbb{C}^n$ в зависимости от вида F; F есть вектор функция, i-я компонента которого равна $f_i(x_1, x_2, \ldots, x_n)$.

Как правило, нельзя указать формулы, которые позволили бы найти x за конечное число арифметических операций. Поэтому для решения (1.1) обычно используются приближенные, итерационные, методы.

Немалое число итерационных методов определяется следующим образом: начиная с заданного начального приближения x^0 к решению, строится последовательность приближений x^k по формулам

$$A_k(x^k - x^{k-1}) + F(x^{k-1}) = 0, \quad k = 1, 2, \ldots, \tag{1.2}$$

где A_k — некоторая квадратная матрица порядка n. Способ ее задания определяет конкретный итерационный метод. Например, широко известный метод Ньютона определяется выбором $A_k = F'(x^{k-1})$, где $F'(x^{k-1})$ есть матрица Якоби отображения F в точке x^{k-1}, т. е. матрица $\{\partial f_i(x^{k-1})/\partial x_j\}_{i,j=1}^n$.

Положим $\triangle_k = x^k - x^{k-1}$, $b_k = -F(x^{k-1})$. Тогда для отыскания x^k согласно (1.2) необходимо выполнить следующие операции: a) вычислить A_k и b_k; b) решить систему линейных уравнений $A_k \triangle_k = b_k$; c) найти $x^k = x^{k-1} + \triangle_k$.

Таким образом, для реализации методов типа (1.2) надо уметь решать системы линейных алгебраических уравнений. Так обстоит дело с большинством известных итерационных методов решения нелинейных систем уравнений.

2. Приближение функций

Рассмотрим два метода приближения функций одной переменной.

1. Интерполяция функций. Пусть на отрезке $[a, b]$ задана вещественная функция f, значения которой известны в точках $x_0 < x_1 < \ldots < x_n$ этого отрезка. Требуется найти функцию y_n из некоторого заданного множества функций F_n, такую, что

$$y_n(x_i) = f_i, \quad i = 0, 1, \ldots, n, \tag{2.1}$$

где $f_i = f(x_i)$. Задача построения такой функции называется задачей интерполяции, а точки x_0, x_1, …, x_n — узлами интерполяции. Говорят, что задача интерполяции поставлена корректно, если при любых значениях f_i, $i = 0, 1, \ldots, n$, существует единственное решение задачи (2.1).

Линейный метод интерполяции заключается в том, что F_n определяется как множество всех линейных комбинаций заданных и до-

статочно простых для вычисления функций $\{\varphi_i\}_{i=0}^n$. В этом случае

$$y_n(x) = \sum_{j=0}^{n} c_j \varphi_j(x).$$

Функцию такого вида часто называют обобщенным полиномом степени n. Его неизвестные коэффициенты c_j находятся из условия интерполяции (2.1), которое принимает следующий вид:

$$\sum_{j=0}^{n} \varphi_j(x_i)\, c_j = f_i, \quad i = 0, 1, \ldots, n.$$

Эту систему линейных уравнений можно записать в матричной форме $A_n c = b_n$, где c есть вектор коэффициентов; квадратная матрица A_n имеет элементы $a_{ij} = \varphi_j(x_i)$, $b_n = (f_0, f_1, \ldots, f_n)^T$. Решая эту систему, находим вектор коэффициентов, а следовательно и интерполирующую функцию $y_n(x)$.

Известно, что линейный метод интерполяции при произвольном наборе узлов является корректным тогда и только тогда, когда любая функция $y_n \in F_n$, отличная от нулевой, имеет на $[a, b]$ не более, чем n различных корней (систему $\{\varphi_i(x)\}_{i=0}^n$ при этом называют системой Чебышева). При выполнении этого условия в F_n обязательно найдется система функций $\{l_i\}_{i=0}^n$, удовлетворяющая условиям

$$l_i(x_j) = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0 & i \neq j, \end{cases}, \quad i, j = 0, 1, \ldots, n.$$

Верно и обратное утверждение. Систему $\{l_i\}_{i=0}^n$ принято называть базисом Лагранжа в F_n, поскольку, как нетрудно видеть, любая функция $y_n \in F_n$ может быть однозначно представлена в виде

$$y_n(x) = \sum_{j=0}^{n} y_n(x_j) l_j(x). \tag{2.2}$$

Из (2.2) и (2.1) следует явная формула для интерполирующей $f(x)$ функции $y_n(x)$:

$$y_n(x) = \sum_{j=0}^{n} f(x_j) l_j(x).$$

Предполагается при этом, что мы можем построить базис Лагранжа в явном виде.

В общем случае для определения базиса Лагранжа приходится решать $n+1$ систему алгебраических уравнений с одной и той же матрицей A_n, но с разными правыми частями для определения коэффициентов разложения базисных функций по системе $\{\varphi_i(x)\}_{i=0}^n$. Легко видеть, что коэффициенты разложения $l_j(x)$ образуют j-й столбец матрицы A_n^{-1}.

Пример 1. Если элементами F_n являются алгебраические полиномы степени не выше n (в этом случае $\varphi_j(x) = x^j$), то говорят об алгебраической интерполяции, а функцию $y_n(x)$ называют интерполяционным полиномом. Легко проверяется, что базисные функции Лагранжа имеют вид

$$l_j(x) = \prod_{k=0,\,k\neq j}^{n} \frac{x - x_k}{x_j - x_k} = \frac{\omega_n(x)}{(x - x_j)\omega_n'(x_j)}, \quad j = 0, 1, \ldots, n. \tag{2.3}$$

Здесь $\omega_n(x) = \prod\limits_{k=0}^{n}(x - x_k)$. Таким образом для интерполяционного полинома верна формула

$$y_n(x) = \sum_{j=0}^{n} f(x_j)\, \frac{\omega_n(x)}{(x - x_j)\omega_n'(x_j)}\,, \tag{2.4}$$

известная как формула Лагранжа. Для практических вычислений более полезной является формула

$$y_n(x) = \Big(\sum_{i=0}^{n} \frac{\beta_i\, f(x_i)}{x - x_i}\Big) \Big/ \Big(\sum_{i=0}^{n} \frac{\beta_i}{x - x_i}\Big), \quad \beta_i = \frac{C}{\omega_n'(x_i)}\,, \tag{2.5}$$

где нормирующая постоянная C может быть выбрана произвольно. Формула (2.5) называется барицентрической.

Пример 2. Пусть F_n есть множество непрерывных функций на $[a, b]$, линейных на отрезках $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, n-1$. В этом случае говорят о кусочно-линейной интерполяции. Легко видеть, что интерполирующая функция $y_n(x)$ на $[x_i, x_{i+1}]$ определяется формулой

$$y_n(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i}\, f(x_i) + \frac{x - x_i}{x_{i+1} - x_i}\, f(x_{i+1}). \tag{2.6}$$

2. Метод наименьших квадратов. Пусть отрезке $[a, b]$ задана функция f, значения которой известны в точках $x_1 < x_2 < \ldots < x_N$

этого отрезка. Требуется найти коэффициенты обобщенного полинома $y_n(x) = \sum\limits_{j=0}^{n} c_j\varphi_j(x)$ степени $n \ll N$ так, чтобы минимизировать среднеквадратичное отклонение функции $y_n(x)$ от $f(x)$ на множестве узлов $\{x_i\}_{i=0}^{N}$. Под этим понимается, что коэффициенты $\{c_j\}_{j=0}^{n}$ находятся как решение задачи

$$\min_{c_0,c_1,\ldots,c_n} \Big[\frac{1}{N}\sum_{i=1}^{N}\Big(f(x_i) - \sum_{j=0}^{n} c_j\varphi_j(x_i)\Big)^2\Big]^{1/2}. \tag{2.7}$$

Такая задача часто встречается при обработке разнообразных экспериментальных данных.

Задача (2.7) есть задача на минимум функции $n+1$ переменной. Ее решение легко находится и совпадает с решением линейной системы алгебраических уравнений $A_n c = b_n$ размера $n+1$, где $A_n = \Phi_n^T\Phi_n$, $b_n = \Phi_n^T F_N$, а Φ_n есть прямоугольная матрица с элементами $\phi_{ij} = \varphi_j(x_i)$, $i = 1,\ldots,N$, $j = 0,\ldots,n$; вектор F_N имеет компоненты $f(x_i)$, $i = 1,\ldots,N$. Матрица A_n является симметричной. Нетрудно доказать, что она положительно определена, если $\{\varphi_i(x)\}_{i=0}^{n}$ является системой Чебышева на $[a,b]$. При этом следует учесть, что прямоугольная матрица Φ_n размера $N \times (n+1)$ имеет полный столбцевой ранг (равный $n+1$).

Выбирая те или иные системы функций $\{\varphi_i(x)\}_{i=0}^{n}$ (например, полагая $\varphi_i(x) = x^i$), получаем конкретный метод наименьших квадратов.

3. Задача Коши для дифференциальных уравнений

Рассмотрим для примера задачу Коши для системы линейных обыкновенных дифференциальных уравнений первого порядка.

Пусть задан конечный отрезок $[a,b]$ и вектор u_a длины $n \geqslant 1$. Пусть также для каждого $x \in [a,b]$ заданы квадратная матрица $A(x)$ и вектор $f(x)$ размера n. Элементы $A(x)$ и $f(x)$ обозначим через $a_{ij}(x)$ и $f_i(x)$ соответственно. Требуется найти n неизвестных функций $u_1(x),\ldots,u_n(x)$, удовлетворяющих для всех $x \in (a,b]$ уравнениям

$$u_i'(x) + \sum_{j=1}^{n} a_{ij}(x)u_j(x) = f_i(x), \quad i = 1,\ldots,n, \tag{3.1}$$

а также дополнительным условиям

$$u_i(a) = u_{a,i}, \quad i = 1,\ldots,n. \tag{3.2}$$

Такая задача называется задачей Коши.

Определим вектор функцию $u(x) = (u_1(x), \ldots, u_n(x))^T$. Тогда соотношениям (3.1), (3.2) можно придать более компактный вид:

$$u'(x) + A(x)u(x) = f(x), \quad u(a) = u_a. \tag{3.3}$$

Напомним, что по определению $u'(x) = (u_1'(x), \ldots, u_n'(x))^T$. Из теории обыкновенных дифференциальных уравнений известно, что задача (3.3) имеет единственное решение при произвольно заданном u_a, если все функции a_{ij} и f_i непрерывны на $[a, b]$. Эти условия в дальнейшем считаем выполненными.

Вообще говоря, для решения задач вида (3.3) используются приближенные методы, позволяющие найти решение с требуемой точностью. Опишем один такой метод, который относится к сеточным методам.

В сеточных методах неизвестные функции $u_i(x)$ определяются лишь на некотором дискретном множестве точек, называемом сеткой узлов или просто сеткой. Например, на равномерной сетке $\omega_h = \{x_i = a + ih,\ i = 0, 1, \ldots, N\}$. Здесь величина $h = (b - a)/N$ — шаг сетки, определяет расстояние между соседними узлами. Зная решение $u_i(x)$ в точках x_i, $i = 0, 1, \ldots, N$, решение в произвольной точке $x \in (a, b)$ можно вычислить, используя, например, кусочно-линейную интерполяцию.

При построени сеточного метода нам понадобятся формулы для приближенного вычисления производной функции. Пусть задана скалярная функция $f(x)$ и мы хотим приближенно вычислить $f'(x)$, используя при этом лишь значения функции $f(x)$. По определению

$$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}.$$

Следовательно, разностное отношение

$$f_x(x) = \frac{f(x + h) - f(x)}{h} \tag{3.4}$$

при малом h позволяет приближенно вычислить $f'(x)$. Сказанное остается справедливым и для разностного отношения

$$f_{\bar{x}}(x) = \frac{f(x) - f(x - h)}{h}. \tag{3.5}$$

Эти отношения являются простейшими формулами численного дифференцирования. Из формулы Тейлора $f(x \pm h) = f(x) \pm h\, f'(x) +$

$h^2/2f''(\xi(x))$, где $\xi(x) \in [a,b]$, следует, что погрешности этих формул имеют порядок малости $O(h)$, если $f(x)$ дважды дифференцируема на $[a,b]$. Аппроксимируя каждую компоненту $u_i'(x)$ вектор-функции $u'(x)$ одной из указанных выше формул, получим соответствующую формулу приближенного вычисления $u'(x)$.

Неявный метод Эйлера (также говорят неявная схема) для приближенного решения задачи Коши определяется следующим образом. Рассмотрим уравнение (3.3) в узле сетки x_i и заменим производную $u'(x_i)$ разностным отношением согласно (3.5). Вводя обозначение y^i для приближения к вектору $u(x_i)$, придем к формулам

$$\frac{y^i - y^{i-1}}{h} + A(x_i)y^i = f(x_i), \quad i = 1, 2, \ldots, N, \quad y^0 = u_a.$$

Определим матрицы $A^{(i)} = I + h\,A(x_i)$ и векторы $b^{(i)} = y^i + h\,f(x_i)$. Тогда формулировка неявного метода Эйлера будет выглядеть следующим образом: начиная с $y^0 = u_0$, для $i = 1, 2, \ldots, N$, найти y^i, решая систему алгебраических уравнений $A^{(i)}y^i = b^{(i)}$. Отметим, что в общем случае обратимости матриц $A^{(i)}$ можно добиться, если h выбрать достаточно малым. В этом случае все y^i определяются однозначно.

Вектор $e^i = u(x_i) - y^i$ есть погрешность приближенного решения в узле сетки x_i. Определим его среднеквадратичную норму

$$\|e^i\| = \Big(\frac{1}{n}\sum_{j=1}^{n} |e_j^i|^2\Big)^{1/2}.$$

Тогда величина $E = \max\limits_{i=1,\ldots,N} \|e^i\|$ — это максимальная погрешность приближенного решения. Для нее известна оценка $E \leqslant C\,h$, где постоянная C не зависит от h. Следовательно, уменьшая шаг сетки h (т. е. увеличивая число точек сетки N), мы можем добиться сколь угодно малой погрешности приближенного решения.

В зависимости от решаемой задачи целые n (число диффренциальных уравнений) и N (число точек сетки) могут быть большими числами. Это означает, что в этом случае решение задачи Коши неявным методом Эйлера (как и любым другим неявным методом), сводится решению большого числа линейных алгебраических уравнений большой размерности. Отметим также частный случай, когда матричная функция $A(x)$ не зависит от x, т. е. все $A(x)$ равны некоторой матрице A. В этом случае мы приходим к необходимости решения N уравнений $(E + h\,A)y^i = b_i$ с одной и той же матрицей, но с разными правыми частями.

Неявную схему можно использовать и для решения задачи Коши для нелинейных систем диффференциальных уравнений первого порядка вида

$$u'(x) = f(x, u(x)), \quad u(a) = u_a,$$

где f есть заданная вектор-функция. Расчетные формулы будут такими

$$\frac{y^i - y^{i-1}}{h} = f(x_i, y^i), \quad i = 1, 2, \ldots, N, \quad y^0 = u_a.$$

Следовательно, y^i является решением нелинейной системы уравнений вида $F^i(y) = 0$ при $F^i(y) = y - y^{i-1} - h\, f(x_i, y)$, которое можно найти, например, методом Ньютона; при этом вектор y^{i-1} является хорошим начальным приближением к решению этой системы.

4. Интегральные уравнения

Рассмотрим для примера линейное одномерное интегральное уравнение уравнение Фредгольма второго рода. Требуется найти функцию $u(x)$, определенную на отрезке $[a, b]$, и такую, что

$$u(x) - \lambda \int_a^b K(x, s)\, u(s)\, ds = f(x) \quad \forall\, x \in [a, b]. \tag{4.1}$$

Здесь число λ, а также непрерывные функции $K(x, s)$ и $f(x)$, считаются заданными. Функция $K(x, s)$ называется ядром интегрального уравнения.

Предполагая, что исходные данные таковы, что существует единственное решение этой задачи, рассмотрим один сеточный метод ее приближенного решения, известный как метод квадратур.

Пусть требуется вычислить определенный интеграл

$$I(f) = \int_a^b f(x) dx\,.$$

Квадратурной формулой (или просто квадратурой) называется формула для приближенного вычисления $I(f)$ вида

$$S_n(f) = \sum_{i=1}^{n} c_i f(x_i).$$

Числа c_i, как правило положительные, называются коэффициентами квадратуры, а x_i — узлами квадратуры.

Примеры. На $[a,b]$ введем равномерную сетку $x_i = a + (i-1)h$, $i = 1, \ldots, n$, с шагом $h = (b-a)/(n-1)$, а также непрерывную на $[a,b]$ функцию $y_n(x)$, имеющее представление (2.6) на каждом отрезке $[x_i, x_{i+1}]$. Приближение к $I(f)$ определим формулой $S_n(f) = I(y_n)$. Интеграл $I(y_n)$ легко вычислить и получить, что

$$S_n(f) = h(0.5f(x_1) + f(x_2) + \ldots + f(x_{n-1}) + 0.5f(x_n)).$$

Эта формула называется составной квадратурной формулой трапеций. Ясно, что ее коэффициенты задаются формулами $c_1 = c_n = h/2$, $c_2 = \ldots = c_{n-1} = h$.

Аналогично определяется составная квадратурная формула центральных прямоугольников. Она имеет вид

$$S_n(f) = h(f(x_{3/2}) + f(x_{5/2}) + \ldots + f(x_{n-1/2})),$$

где $x_{i+1/2} = (x_i + x_{i+1})/2$ — средние точки отрезка $[x_i, x_{i+1}]$ длины h.

Определим метод квадратур для решения уравнения (4). Пусть выбрана некоторая квадратура S_n. Рассмотрим уравнение (4) в узлах квадратурной формулы. Получим равенства

$$u(x_i) - \lambda \int_a^b K(x_i, s)\, u(s)\, ds = f(x_i), \quad i = 1, \ldots, n.$$

Для вычисления интеграла используем квадратурную формулу. Получим приближенные равенства

$$u(x_i) - \lambda \sum_{j=1}^{n} c_j K(x_i, x_j)\, u(x_j) \approx f(x_i), \quad i = 1, \ldots, n. \tag{4.2}$$

Здесь $u(x_i)$ есть значение точного решения задачи в точке сетки x_i. Для определения приближения y_i к $u(x_i)$ из (4.2) получим уравнения

$$y_i - \lambda \sum_{j=1}^{n} c_j K(x_i, x_j)\, y_j = f(x_i), \quad i = 1, \ldots, n. \tag{4.3}$$

Определим матрицу A_n с элементами $a_{ij} = \delta_{ij} - \lambda\, c_j K(x_i, x_j)$ и вектор $b_n = (f_1, f_2, \ldots, f_n)^T$. Тогда система уравнений (4.3) примет вид $A_n y = b_n$. Решая эту систему уравнений найдем вектор y, i-тая компонента y_i которого является приближением к $u(x_i)$. Отметим, что матрица A_n является симметричной, если ядро симметрично, т. е. $K(x,s) = K(s,x)$ для всех $x, s \in [a,b]$.

Применяя ту или иную квадратурную формулу, мы получим конкретный метод решения интегрального уравнения (4). Размер решаемой при этом системы $A_n y = b_n$ зависит от необходимой точности приближенного решения; чем точность выше, тем размер системы уравнений будет больше.

Метод квадратур применяется и для решения нелинейных интегральных уравнений

$$u(x) + \int_a^b K(x, s, u(s))\, ds = f(x) \quad \forall\, x \in [a, b].$$

В этом случае необходимо уметь решать нелинейные системы уравнений вида

$$y_i + \sum_{j=1}^{n} c_j K(x_i, x_j, y_j) - f(x_i) = 0, \quad i = 1, \ldots, n.$$

5. Краевые задачи для обыкновенных дифференциальных уравнений

Рассмотрим два сеточных метода для решения одномерной краевой задачи для линейного дифференциального уравнения второго порядка. Для заданных непрерывных функций $q(x)$, $f(x)$ и чисел u_a и u_b требуется найти решение задачи

$$-u''(x) + q(x)u(x) = f(x), \quad x \in (a, b), \tag{5.1}$$

$$u(a) = u_a, \quad u(b) = u_b. \tag{5.2}$$

В отличие от задачи Коши дополнительные условия заданы в двух граничных точках отрезка интегрирования, поэтому задача называется граничной или чаще — краевой. Условия $q(x) \geqslant 0$, $x \in [a, b]$, достаточно для существования ее единственного решения.

1. Конечно-разностная схема. Определим равномерную сетку $\omega_h = \{x_i = a + ih,\ i = 0, 1, \ldots, n\}$ на отрезке $[a, b]$ с шагом $h = (b-a)/n$ и рассмотрим уравнение (5.1) только во внутренних точках сетки. Получим

$$-u''(x_i) + q(x_i)u(x_i) = f(x_i), \quad i = 1, \ldots, n-1.$$

Определим формулу для приближенного вычисления $u''(x_i)$ как комбинацию разностных отношений (3.4) и (3.5):

$$u_{\bar{x}\,x}(x) = (u_{\bar{x}})_x(x) = \frac{u(x-h) - 2u(x) + u(x+h)}{h^2}\,. \tag{5.3}$$

Разложением в ряд Тейлора показывается, что $u''(x) - u_{\bar{x}\,x}(x) = O(h^2)$, если $u(x)$ достаточно гладкая функция. Используя формулу (5.3) для аппроксимации $u''(x_i)$, придем к приближенным равенствам

$$-u_{\bar{x}\,x}(x_i) + q(x_i)u(x_i) \approx f(x_i), \quad i = 1, \ldots, n-1.$$

Приближение y_i к $u(x_i)$ будем искать из равенств

$$-\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + q(x_i)y_i = f(x_i), \quad i = 1, \ldots, n-1.$$

Умножим обе части этих соотношений на h^2 и приведем подобные члены. Присоединяя к ним краевые условия $y_0 = u_a$, $y_n = u_b$, придем к дискретной задаче, которая называется конечно-разностной схемой: найти $y = (y_0, y_1, \ldots, y_n)^T$ из системы алгебраических уравнений

$$\begin{gathered} y_0 = u_a, \\ -y_{i-1} + \bigl(2 + h^2 q(x_i)\bigr) y_i - y_{i+1} = h^2 f(x_i), \quad i = 1, \ldots, n-1, \\ y_n = u_b. \end{gathered} \tag{5.4}$$

Матрица A_n этой системы имеет специальный вид: ее ненулевые элементы расположены лишь на трех диагоналях. Такие матрицы называются трехдиагональными. Известно, что A_n есть положительно определенная матрица, если $q(x) \geqslant 0$ на $[a, b]$. В этом случае система (5.4) однозначно разрешима.

Величина $E = \max_{i=0,\ldots,N} |u(x_i) - y_i|$ определяет максимальную погрешность приближенного решения задачи. Для достаточно гладких данных $q(x)$ и $f(x)$ для нее известна оценка $E \leqslant C\,h^2$, где постоянная C не зависит от h. Следовательно, уменьшая шаг сетки h (т.е. увеличивая число точек сетки n и размерность решаемой системы уравнений) мы можем добиться сколь угодно малой погрешности приближенного решения.

Зная приближенное решение $y(x)$ в точках x_i, $i = 0, 1, \ldots, N$, решение в произвольной точке $x \in (a, b)$ можно вычислить, используя кусочно-линейную интерполяцию.

2. Метод коллокаций. На отрезке $[a, b]$ введем неравномерную сетку узлов[1)]

$$x_i = \frac{a+b}{2} - \frac{b-a}{2} \cos\Bigl(\frac{i\pi}{n}\Bigr), \quad i = 0, \ldots, n. \tag{5.5}$$

[1)] Такой выбор узлов коллокаций позволяет, в частности, сохранить для аппроксимирующей системы уравнений основные свойства исходной задачи.

Отметим, что $x_0 = a$, $x_n = b$, а с ростом n узлы заметно сгущаются к этим граничным точкам. Отметим также, что узлы сетки являются экстремумами полинома Чебышева $T_n(z) = \cos(n \arccos(z))$, $z \in [-1, 1]$, сдвинутыми на отрезок $[a, b]$.

Будем искать приближенное решение исходной задачи (5.1), (5.2) в виде интерполяционного полинома $y_n(x)$. Положим $y_i = y_n(x_i)$, $i = 0, 1, \ldots, n$. Согласно формуле Лагранжа

$$y_n(x) = \sum_{j=0}^{n} y_j \, l_j(x) \,, \quad l_j(x) = \frac{\omega_n(x)}{(x - x_j)\omega_n'(x_j)} \,, \tag{5.6}$$

где $\omega_n(x) = \prod_{k=0}^{n}(x - x_k)$. Потребуем, чтобы $y_n(x)$ удовлетворял краевым условиям, а также дифференциальному уравнению, но не во всех точках (a, b) (это невозможно), а только во внутренних точках сетки. Таким образом придем к уравнениям

$$-\sum_{j=0}^{n} l_j''(x_i) \, y_j + q(x_i) y_i = f(x_i), \quad i = 1, \ldots, n-1, \tag{5.7}$$

$$y_0 = u_a, \quad y_n = u_b. \tag{5.8}$$

Укажем способ вычисления матрицы этой системы. Определим квадратные матрицы $D^{(k)} = \{l_j^{(k)}(x_i)\}_{i,j=0}^{n}$ размера $n + 1$ (матрицы дифференцирования). Из (5.6) следует, что

$$d_{ij}^{(1)} = l_j'(x_i) = \frac{\omega_n'(x_i)}{(x_i - x_j)\omega_n'(x_j)} = \frac{\beta_j/\beta_i}{(x_i - x_j)} \,, \quad i \neq j, \tag{5.9}$$

где $\beta_i = C/\omega_n'(x_i)$ — барицентрические веса. В силу специального выбора сетки оказывается, что C можно выбрать так, что $\beta_i = (-1)^i \gamma_i$, где $\gamma_0 = \gamma_n = 1/2$, $\gamma_i = 1$ при $i = 1, \ldots, n-1$.

Поскольку $\sum_{k=0}^{n} l_k(x) = 1$ для любого x, то дифференцированием находим диагональные элементы

$$d_{ii}^{(1)} = -\sum_{k=0,\, k\neq i}^{n} d_{ik}^{(1)}. \tag{5.10}$$

Элементы $D^{(2)}$ вычисляются по формулам

$$d_{ij}^{(2)} = \begin{cases} 2\, d_{ij}^{(1)} \big(d_{ii}^{(1)} - 1/(x_i - x_j) \big), & i \neq j \\ -\sum_{k=0,\, k\neq i}^{n} d_{ik}^{(2)}, & i = j. \end{cases} \tag{5.11}$$

Отметим, что из формулы $l_j'(x) = -\sum_{k=0}^{n} l_k'(x_j) l_k(x)$, $x \in [a,b]$, следует, что $D^{(2)} = \bigl(D^{(1)}\bigr)^2$.

Система (5.7), (5.8) в матричном виде принимает вид $A_n y = b_n$, где $y = (y_0, y_1, \ldots, y_n)^T$, $b_n = (u_a, f(x_1), \ldots, f(x_{n-1}), u_b)^T$, первая и последняя строки A_n равны $(1, 0, \ldots, 0)$ и $(0, \ldots, 0, 1)$ соответственно, а остальные элементы матрицы A_n имеют вид $-d_{ij}^{(2)} + q(x_i)\delta_{ij}$, при $i = 1, \ldots, n-1$, $j = 0, \ldots, n$.

Из системы уравнений $A_n y = b_n$ очевидным образом можно исключить y_0 и y_n и получить систему размера $n-1$ для определения $y_1, \ldots, y_{n-1}$. Оказывается, что матрица новой системы является симметричной и положительно определенной, если $q(x) \geqslant 0$. После решения системы приближенное решение $y_n(x)$ в любой точке $x \in (a,b)$ можно вычислить по барицентрической формуле

$$y_n(x) = \Bigl(\sum_{i=0}^{n} \frac{\beta_i\, y_i}{x - x_i}\Bigr) \Big/ \Bigl(\sum_{i=0}^{n} \frac{\beta_i}{x - x_i}\Bigr). \tag{5.12}$$

Для максимальной погрешности приближенного решения справедлива оценка $E = \max\limits_{x\in[a,b]} |u(x) - y_n(x)| \leqslant c\, n^{-s}$, где постоянная c не зависит от n, а $s = \min(n, m)$, где m — число непрерывных производных, которыми обладают функции q и f на отрезке $[a,b]$. Поэтому при гладких исходных данных этот метод имеет высокую точность уже при небольших значениях n.

Рассмотренные выше сеточные методы применимы и для решения нелинейной задачи вида

$$\begin{gathered} -u''(x) + q(x, u(x)) = f(x), \quad x \in (a,b), \\ u(a) = u_a, \quad u(b) = u_b. \end{gathered}$$

В этом случае необходимо уметь решать нелинейные системы алгебраических уравнений, получающихся из (5.4) или (5.7), заменой слагаемого $q(x_i)y_i$ на $q(x_i, y_i)$.

6. Краевые задачи для дифференциальных уравнений в частных производных

Рассмотрим метод конечных разностей для приближенного решения следующей задачи Дирихле для модельного уравнения эллиптического типа в прямоугольной области $\Omega = (0, L) \times (0, L)$:

$$-\frac{\partial^2 u(x)}{\partial x_1^2} - \frac{\partial^2 u(x)}{\partial x_2^2} + q(x)u(x) = f(x), \quad x \in \Omega, \tag{6.1}$$

$$u(x) = 0, \quad x \in \partial\Omega. \tag{6.2}$$

Здесь $x = (x_1, x_2)$; $\partial\Omega$ обозначает множество граничных точек Ω; $q(x)$, $f(x)$ заданные непрерывные функции. Условие $q(x) \geqslant 0$, $x \in \Omega$, является достаточным для существования и единственности решения задачи.

На области Ω зададим дискретное множество точек, в которых будем определять приближенное решение задачи (определим сетку). Для этого отрезки $[0, L]$ на осях x_1 и x_2 разобьем на n равных частей; пусть $h = L/n$. Через точки деления проведем прямые, параллельные соответствующим осям. В результате пересечения этих прямых получатся узлы, которые и образуют сетку. Те узлы (ih, jh), которые лежат внутри Ω, назовем внутренними. Их совокупность обозначим

$$\omega_h = \{(ih, jh) : \, i, j = 1, 2, \ldots, n-1\}.$$

Множество узлов сетки, принадлежащих $\partial\Omega$, назовем граничными и обозначим через γ_h.

Дискретный аналог задачи (6.1), (6.2) построим как и в одномерном случае: уравнения рассмотрим в точках сетки, и аппроксимируем вторые производные разностными отношениями (5.3). Придем к сеточным уравнениям

$$-\frac{y(x_1-h, x_2) - 2y(x_1, x_2) + y(x_1+h, x_2)}{h^2} - \frac{y(x_1, x_2-h) - 2y(x_1, x_2) + y(x_1, x_2+h)}{h^2} + q(x_1, x_2)y(x_1, x_2) = f(x_1, x_2), \quad (x_1, x_2) \in \omega_h, \tag{6.3}$$

$$y(x_1, x_2) = 0, \quad (x_1, x_2) \in \gamma_h. \tag{6.4}$$

Систему линейных алгебраических уравнений (6.3), (6.4) называют разностной схемой, а ее решение $y(x)$ — приближенным решением задачи (6.1), (6.2).

Запишем систему уравнений (6.3), (6.4) в матричном виде. Ясно, что неизвестными являются лишь значения $y(x)$ в точках ω_h; поскольку значения $y(x)$ в точках γ известны и равны нулю, то их нет необходимости включать в вектор неизвестных. Учитывая сказанное, уравнения (6.3) запишем в виде

$$-y(x_1, x_2-h) - y(x_1-h, x_2) + (4 + h^2 q(x_1, x_2))y(x_1, x_2) - y(x_1+h, x_2) - y(x_1, x_2+h) = h^2 f(x_1, x_2), \quad (x_1, x_2) \in \omega_h, \tag{6.5}$$

считая, что слагаемое вида $y(x_1, x_2 \pm h)$ или $y(x_1 \pm h, x_2)$ в этом равенстве опущено, если соответствующий ему узел сетки $(x_1, x_2 \pm h)$ или $(x_1 \pm h, x_2)$ принадлежит γ_h. Отметим, что такой коррекции требуют лишь уравнения, соответствующие приграничным узлам (т. е. узлам (ih, jh) при i или j равным 1 или $n-1$).

Чтобы определить вектор неизвестных, необходимо пронумеровать узлы сетки ω_h. Ясно, что это можно сделать многими способами. Выберем следующий способ: узлы ω_h пронумеруем слева-напрво и снизу-вверх, начиная с узла с координатой (h, h). А именно, примем, что узел (ih, jh) имеет номер l (т. е. $x_l = (ih, jh)$), если

$$l = (j-1)(n-1) + i, \quad i, j = 1, \ldots, n-1.$$

В такой нумерации уравнения (6.5) запишутся в виде

$$-y_{l-n+1} - y_{l-1} + d_l\, y_l - y_{l+1} - y_{l+n-1} = h^2 f_l, \quad l = 1, 2, \ldots, N, \quad (6.6)$$

где $N = (n-1)^2$, $d_l = 4 + h^2 q_l$, $q_l = q(x_l)$, $f_l = f(x_l)$. Уравнения (6.6) нужно скорректировать, опуская соответствующие слагаемые, если узел x_l является приграничным.

Уравнения (6.6) в матричном виде примут вид $A_N y = F_N$, где l-тая компонента y равна y_l, l-тая компонента F_N равна $h^2 f_l$, а матрица A_N размера N имеет следующий блочно-трехдиагональный вид:

$$A_N = \begin{bmatrix} T_1 & -I & & & \\ -I & T_2 & -I & & \\ & \cdots & \cdots & \cdots & \\ & & -I & T_{n-2} & -I \\ & & & -I & T_{n-1} \end{bmatrix},$$

где I есть единичная матрица размера $n-1$, T_k есть трехдиагональная матрица размера $n-1$ вида

$$T_k = \begin{bmatrix} d_{n_k+1} & -1 & & & \\ -1 & d_{n_k+2} & -1 & & \\ & \cdots & \cdots & \cdots & \\ & & -1 & d_{n_k+n-2} & -1 \\ & & & -1 & d_{n_k+n-1} \end{bmatrix},$$

$n_k = (k-1)(n-1)$.

Матрица A_N является симметричной и разреженной (подавляющее число ее элементов — нули, ненулевые элементы расположены лишь на пяти диагоналях). Если q есть неотрицательная функция, то доказывается, что она положительно определена. Отметим, что система уравнений может иметь большую размерность. Например, при $n \approx 100$ получаем $N \approx 10^4$.

Глава 2

Прямые методы решения систем линейных алгебраических уравнений

Метод решения системы линейных уравнений называется прямым (точным), если он позволяет найти ее точное решение за конечное число арифметических операций. Предполагается, что арифметические операции выполняются точно, а под одной арифметической операцией, кратко 1 flops (floating point operation), понимается любая из арифметических операций $+, -, *, /$. Количество требуемых для реализации метода арифметических операций называется *трудоемкостью* метода. Например, метод Гаусса, который мы опишем в дальнейшем, относится к прямым методам и имеет трудоемкость порядка $2n^3/3$ flops.

Все рассматриваемые в этой и в последующих главах векторы и матрицы предполагаются, вообще говоря, комплексными. Однако, при вещественных исходных данных излагаемые далее методы позволяют проводить все вычисления только с вещественными числами. В необходимых случаях даются соответствующие пояснения.

7. Трудоемкость базовых операций линейной алгебры

Рассмотрим предварительно трудоемкость простейших операций линейной алгебры.

1. Вычисление суммы векторов. Пусть требуется вычислить сумму z двух векторов x и y размера n. По определению компоненты вектора z вычисляются по формулам

$$z_i = x_i + y_i, \quad i = 1, \ldots, n.$$

Ясно, что трудоемкость метода составляет n flops.

2. Вычисление произведения матрицы и вектора. Пусть заданы матрица A размера n и вектор x. Рассмотрим задачу вычисления вектора $b = Ax$. По определению

$$b_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n,$$
$$b_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n,$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$
$$b_n = a_{n1}x_1 + a_{n2}x_2 + \ \dots\ + a_{nn}x_n,$$

или короче,

$$b_i = \sum_{j=1}^{n} a_{ij}\, x_j, \quad i = 1, 2, \dots, n. \tag{7.1}$$

Будем говорить, что эта формула определяет *метод* умножения матрицы A на заданный вектор. Заметим, что формула (7.1) ориентирована на работу со строками матрицы и определяет b_i как скалярное произведение i-той строки A на вектор-строку x.

Непосредственная реализация формул (7.1) в MATLAB приводит к следующей функции:

```
function b=Axrow(A,x)
n=numel(x);
b=zeros(size(x));
for i=1:n
  for j=1:n
    b(i)=b(i)+A(i,j)*x(j);
  end
end
```

В этой функции компоненты вектора b вычисляются последовательно друг за другом накоплением. Здесь и далее цикл по i означает цикл по строкам, а цикл по j — цикл по столбцам матрицы. Нетрудно видеть, что трудоемкость этой функции равна $2n^2$ flops.

Алгоритм вычисления, реализованный в функции Axrow, принято называть строчно-ориентированным: в нем цикл по i предшествует циклу по j и в нем обрабатываются в цикле по j строки матрицы. Поменяв порядок циклов придем к другой реализации формул (7.1) (столбцово-ориентированной). В нем цикл по j предшествует циклу по i:

```
function b=Axcol(A,x)
n=numel(x);
b=zeros(size(x));
for j=1:n
  for i=1:n
    b(i)=b(i)+A(i,j)*x(j);
  end
end
```

В функции Axcol накоплением вычисляются вклады произведения Ax сразу во все компоненты вектора b и ее трудоемкость также равна $2n^2$ flops. В этой функции непосредственно реализован способ вычисления b, основанный на эквивалентной (7.1) формуле и ориен-

тированной на столбцы матрицы. Он имеет следующий вид:

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix} x_1 + \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{bmatrix} x_2 + \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{nn} \end{bmatrix} x_n.$$

Вместо языка MATLAB можно взять другой язык программирования (например, СИ, Паскаль, Фортран, ...) и написать аналоги функций Axrow и Axcol на этом языке. Практический интерес представляет ответ на следующий вопрос: *какая из полученных функций будет быстрее*, т. е. будет требовать меньшего времени для выполнения? На первый взгляд время работы функций не должно различаться. Однако это не так. *Ответ на этот важный с практической точки зрения вопрос зависит от языка программирования* и связан, главным образом, со способом хранения матриц (способом адресации элементов матриц). Из-за наличия в современных компьютерах многоуровнего кэша последовательное извлечение из оперативной памяти и сохранение чисел, расположенных в соседних ячейках памяти, производится намного быстрее, чем последовательное выполнение тех же операций с элементами, расположенными в памяти далеко друг от друга. В связи с этим, если элементы матрицы в памяти ЭВМ хранятся по строкам (как, например, в C, Паскаль, Python), то быстрее будет выполняться строчно-ориентированная функция Axrow. И наоборот, если элементы матрицы в памяти ЭВМ хранятся по столбцам (Fortran, MATLAB, OpenGL), то быстрее будет выполняться столбцово-ориентированная функция Axcol.

Будем говорить, что функции Axrow и Axcol реализуют *алгоритм* умножения матрицы A на заданный вектор. Эти функции демонстрируют разницу между методом и алгоритмом. В дальнейшем мы ограничимся указанием лишь метода решения задачи.

3. Вычисление произведения двух матриц. Пусть требуется вычислить произведения $C = AB$ двух заданных матриц размера n. По определению j-й столбец C есть произведение матрицы A и j-го столбца B, т. е.

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}, \quad i, j = 1, \ldots, n.$$

Вычисление c_{ij} накоплением требует $2n$ flops, поэтому трудоемкость вычисления C равна $2n^3$ flops.

8. Простые системы уравнений

Приведем примеры систем уравнений, которые легко решаются.

1. Системы с диагональной матрицей. Пусть D есть диагональная матрица с ненулевыми элементами d_i на диагонали, т. е. $D = \operatorname{diag}(d_1, d_2, \ldots, d_n)$). Тогда система уравнений $Dx = b$ элементарно решается за n flops, и вектор x находится по формулам $x_i = b_i/d_i$, $i = 1, \ldots, n$.

2. Системы с треугольной матрицей. Матрица A называется *нижней треугольной* (также левой треугольной), если $a_{ij} = 0$ при всех $j > i$. Аналогично, матрица A называется *верхней треугольной* (также правой треугольной), если $a_{ij} = 0$ при всех $i > j$. Как правило, нижние треугольные матрицы обозначаются буквой L (Lower, Left), а верхние треугольные — буквой U (Upper) или R (Right). Таким образом,

$$L = \begin{bmatrix} l_{11} & 0 & \ldots & 0 \\ l_{21} & l_{22} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ l_{n1} & l_{n2} & \ldots & l_{nn} \end{bmatrix}, \quad U = \begin{bmatrix} u_{11} & u_{12} & \ldots & u_{1n} \\ 0 & u_{22} & \ldots & u_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & u_{nn} \end{bmatrix}.$$

Матрица называется треугольной, если она является либо нижней треугольной, либо верхней треугольной.

Поскольку определитель L равен $|L| = l_{11}l_{22}\cdots l_{nn}$, и аналогично $|U| = u_{11}u_{22}\cdots u_{nn}$, то квадратные треугольные матрицы невырождены тогда и только тогда, когда все их диагональные элементы отличны от нуля.

Система уравнений $Lx = b$ в индексной форме имеет вид

$$\begin{aligned} l_{11}x_1 &= b_1, \\ l_{21}x_1 + l_{22}x_2 &= b_2, \\ \ldots\ldots\ldots\ldots\ldots & \quad \ldots \\ l_{n1}x_1 + l_{n2}x_2 + \ \ldots\ + l_{nn}x_n &= b_n. \end{aligned} \tag{8.1}$$

Решение этой системы находится последовательно: из первого уравнения определяется $x_1 = b_1/l_{11}$, из второго $x_2 = (b_2 - l_{21}x_1)/l_{22}$ и т. д. Таким образом,

$$x_i = \Big(b_i - \sum_{j=1}^{i-1} l_{ij}\, x_j\Big)/l_{ii}, \quad i = 1, 2, \ldots, n. \tag{8.2}$$

При $i = 1$ в (8.2) возникает сумма $\sum\limits_{j=1}^{0} \big(\dots \big)$. Подобные суммы здесь и далее считаются равными нулю.

Метод (8.2) решения системы $Lx = b$ называется *прямой подстановкой*. Определим трудоемкость этого алгоритма: при фиксированном i требуется $2(i-1)$ flops для вычисления суммы и 2 flops для вычисления x_i. Общее число операций равно

$$Q = 1 + 2\sum_{i=2}^{n} i = n^2 + n - 1 = n^2 + O(n) \text{ flops}\,.$$

Аналогично решается система $Ux = b$. Отличие в том, что сначала определяется $x_n = b_n/u_{nn}$, затем $x_{n-1} = (b_{n-1} - u_{n-1,n}x_n)/u_{n-1,n-1}$ и т. д. Таким образом,

$$x_i = \Big(b_i - \sum_{j=i+1}^{n} u_{ij}\, x_j\Big)/u_{ii}, \quad i = n, n-1, \dots, 1. \tag{8.3}$$

Метод (8.3) решения системы $Ux = b$ называется *обратной подстановкой*. Его трудоемкость также равна $n^2 + O(n)$ flops.

Обратим внимание, что суммарная трудоемкость прямого и обратного хода, т. е. трудоемкость последовательного решения двух треугольных систем $Ly = b$ и $Ux = y$ равна $2n^2 + O(n)$ flops, и при больших значениях n примерно равна трудоемкости вычисления $b = Ax$ при заданном x.

Отметим также замкнутость множества $\mathcal{L}$ всех обратимых нижних треугольных матриц (множества $\mathcal{U}$ всех обратимых верхних треугольных матриц). В самом деле, пусть $L, L_1, L_2 \in \mathcal{L}$. Тогда $L_1 + L_2 \in \mathcal{L}$ (что очевидно), $L_1 L_2 \in \mathcal{L}$ (непосредственно проверяется) и $L^{-1} \in \mathcal{L}$ (см. ниже упражнение 8.6). По определению нулевая матрица и единичная матрица являются элементами как $\mathcal{L}$, так и $\mathcal{U}$. Кроме того, $L_1 + L_2 = L_2 + L_1$, но, вообще говоря, $L_1 L_2 \neq L_2 L_1$, т. е. $\mathcal{L}$ (как и $\mathcal{U}$) есть коммутативная группа по сложению, и некоммутативная группа по умножению.

Квадратная матрица

$$L_k = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & l_{k,k} & 0 & \cdots & 0 \\ 0 & \cdots & l_{k+1,k} & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & l_{n,k} & 0 & \cdots & 1 \end{bmatrix} \tag{8.4}$$

называется *элементарной нижней треугольной*; она отличается от единичной матрицы лишь элементами k-го столбца. Важное свойство этой матрицы отмечено далее в упражнении 8.5.

3. Системы с унитарной матрицей. Матрица Q называется унитарной, если $Q^*Q = QQ^* = I$, где $Q^* = (\bar{Q})^T$; как обычно, черта означает комплексное сопряжение, T — транспонирование, I — единичная матрица. Равенство $Q^*Q = I$ ($QQ^* = I$) означает, что столбцы (строки) Q образуют ортонормированную систему из n векторов в пространстве $\mathbb{C}^n$. Вещественная унитарная матрица называется ортогональной.

Пусть требуется решить систему $Qx = b$. Умножая обе части этого равенства на Q^*, получим $x = Q^*b$. Трудоемкость такого метода решения есть $2n^2$ flops, если Q есть матрица общего вида.

Простейшим примером ортогональной матрицы является элементарная матрица перестановок (транспозиция). Матрица P_{ik} называется *элементарной матрицей перестановок*, если она получена из единичной матрицы перестановкой строк с номерами i и k. Например, матрицами перестановок третьего порядка являются матрицы

$$P_{12} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad P_{13} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad P_{23} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Упражнения.

8.1. Пусть P_{ik} — матрица перестановки.Показать, что вектор $P_{ik}x$ получается из вектора x перестановкой элементов с номерами i, k.

8.2. Как следствие показать, что матрица $P_{ik}A$ получается из матрицы A перестановкой строк с номерами i, k.

8.3. Пусть P_{ik} — матрица перестановки. Показать, что $P_{ik}^{-1} = P_{ik}^T = P_{ik}$.

8.4. Показать, что нижняя треугольная матрица L (с элементами l_{ij}) равна произведению элементарных нижних треугольных матриц L_k (см. (8.4)), т. е. $L = L_1L_2 \cdots L_{n-1}L_n$.

Указание. Проведите вычисления в соответствии со следующей расстановкой скобок: $L = L_1(L_2 \cdots (L_{n-2}(L_{n-1}L_n) \cdots)$, т. е. сначала перемножьте $L_{n-1}L_n$, результат умножьте слева на L_{n-2} и т. д.

8.5. Пусть L_k есть элементарная нижняя треугольная матрица и $l_{kk} \neq 0$. Показать, что

$$L_k^{-1} = \begin{bmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1/l_{k,k} & 0 & \dots & 0 \\ 0 & \dots & -l_{k+1,k}/l_{k,k} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & -l_{n,k}/l_{k,k} & 0 & \dots & 1 \end{bmatrix}.$$

8.6. Пусть L — нижняя треугольная матрица, у которой все элементы главной диагонали отличны от нуля. Показать, что матрица L^{-1} существует и является нижней треугольной матрицей. Показать, что аналогичное верно и для верхней треугольной матрицы.

9. Метод исключения Гаусса

В основе метода Гаусса, как, впрочем, и многих других методов решения систем линейных алгебраических уравнений

$$Ax = b, \tag{9.1}$$

лежит следующее утверждение. Пусть матрица B невырождена. Тогда система уравнений

$$BAx = Bb \tag{9.2}$$

эквивалентна системе (9.1), т. е. решение системы (9.2) — решение системы (9.1) и, наоборот, решение системы (9.1) — решение системы (9.2). Действительно, пусть x — решение системы (9.2). Тогда

$$B(Ax - b) = 0,$$

но матрица B невырождена, следовательно, $Ax - b = 0$. Обратное утверждение очевидно.

Матрица B выбирается так, чтобы матрица BA была проще матрицы A и решение системы (9.2) находилось легче, чем решение системы (9.1). В методе Гаусса матрица B конструируется при помощи элементарных нижних треугольных матриц так, чтобы матрица BA была верхней треугольной. Тогда решение системы (9.2) становится тривиальной задачей.

Приведем традиционное описание этого метода, ориентированное на операциях со строками A.

1. Расчетные формулы. Для удобства изложения положим $A^{(1)} = A$, $b^{(1)} = b$ и запишем исходную систему в индексной форме:

$$\begin{aligned}
&a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \ \ldots\ + a_{1n}^{(1)}x_n = b_1^{(1)},\\
&a_{21}^{(1)}x_1 + a_{22}^{(1)}x_2 + \ \ldots\ + a_{2n}^{(1)}x_n = b_2^{(1)},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_{n1}^{(1)}x_1 + a_{n2}^{(1)}x_2 + \ \ldots\ + a_{nn}^{(1)}x_n = b_n^{(1)}.
\end{aligned} \tag{9.3}$$

Предположим, что $a_{11}^{(1)} \neq 0$ и введем *множители*

$$l_{i1} = a_{i1}^{(1)}/a_{11}^{(1)},\ i = 2, \ldots, n. \tag{9.4}$$

Для каждого $i = 2, \ldots, n$, умножим обе части первого уравнения в (9.3) на l_{i1} и вычтем полученное равенство из i-го уравнения. Придем к новой (эквивалентной) системе $A^{(2)}x = b^{(2)}$ вида

$$a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \ \ldots\ + a_{1n}^{(1)}x_n = b_1^{(1)},$$

$$\begin{aligned} a_{22}^{(2)}x_2 + \ldots + a_{2n}^{(2)}x_n &= b_2^{(2)}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n2}^{(2)}x_2 + \ldots + a_{nn}^{(2)}x_n &= b_n^{(2)}. \end{aligned} \tag{9.5}$$

Согласно описанию, данному выше, новые элементы матрицы и правой части вычисляются по формулам

$$\begin{aligned} a_{ij}^{(2)} &= a_{ij}^{(1)} - l_{i1}a_{1j}^{(1)}, \quad i,j = 2,\ldots,n, \\ b_i^{(2)} &= b_i^{(1)} - l_{i1}b_1^{(1)}, \quad i = 2,\ldots,n. \end{aligned} \tag{9.6}$$

Говорят, что в системе (9.5) неизвестная x_1 исключена из уравнений со второго по n-е или, что матрица системы приведена к верхней треугольной форме в первом столбце. На этом заканчивается описание первого шага метода Гаусса.

На втором шаге проделаем аналогичные вычисления с подсистемой (9.5), включающей уравнения с номерами $2,3,\ldots,n$, и приведем матрицу системы к верхней треугольной форме во втором столбце. Это можно сделать, если $a_{22}^{(2)} \neq 0$. Повторяя вычисления, получим n эквивалентных систем

$$A^{(k)}x = b^{(k)}, \quad k = 1,2,\ldots,n, \tag{9.7}$$

с матрицами вида

$$A^{(k)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \ldots & a_{1k}^{(1)} & \ldots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \ldots & a_{2k}^{(2)} & \ldots & a_{2n}^{(2)} \\ \vdots & & \ddots & & & \vdots \\ 0 & \ldots & 0 & a_{kk}^{(k)} & \ldots & a_{kn}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \ldots & 0 & a_{nk}^{(k)} & \ldots & a_{nn}^{(k)} \end{bmatrix}, \quad k \geqslant 1. \tag{9.8}$$

Ясно, что при $k = n$ (после $n-1$ шага) получим систему $A^{(n)}x = b^{(n)}$ с верхней треугольной матрицей

$$\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \ldots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \ldots & a_{2n}^{(2)} \\ \vdots & & \ddots & \vdots \\ 0 & & & a_{nn}^{(n)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(n)} \end{bmatrix}, \tag{9.9}$$

решая которую получим решение первоначальной системы. Переход от исходной системы (9.3) к системе (9.9) называется *прямым ходом*

метода Гаусса. Решение системы (9.9) обратной подстановкой — *обратным ходом*. Элементы $a_{ii}^{(i)}$, $i = 1, \ldots, n$, называются *ведущими (главными) элементами метода Гаусса* и только на них производится деление в ходе вычислений. Для осуществимости метода *они должны быть отличны от нуля*.

Суммируя сказанное, приходим к следующим расчетным формулам. Для всех $k = 1, 2, \ldots, n-1$ сначала вычисляются множители

$$l_{ik} = a_{ik}^{(k)} / a_{kk}^{(k)}, \quad i = k+1, \ldots, n. \tag{9.10}$$

Затем вычисляются новые элементы матрицы $A^{(k+1)}$ и вектора $b^{(k+1)}$:

$$\begin{aligned} a_{ij}^{(k+1)} &= a_{ij}^{(k)} - l_{ik} a_{kj}^{(k)}, \quad i, j = k+1, \ldots, n, \\ b_i^{(k+1)} &= b_i^{(k)} - l_{ik} b_k^{(k)}, \quad i = k+1, \ldots, n. \end{aligned} \tag{9.11}$$

Можно заметить, что при программной реализации этих формул, элементы $a_{ij}^{(k+1)}$ можно хранить на месте элемента a_{ij} исходной матрицы, также как l_{ik} — на месте элемента a_{ik}, $b_i^{(k+1)}$ — на месте b_i.

2. Трудоемкость метода Гаусса. Ясно, что трудоемкость метода Гаусса вычисляется по формуле

$$Q = \sum_{k=1}^{n-1} (q_{mk} + q_{ak} + q_{bk}) + n^2 + n - 1,$$

где q_{mk}, q_{ak}, q_{bk} есть число операций, необходимых для вычисления множителей на шаге k, новых элементов матрицы $A^{(k+1)}$ и вектора $b^{(k+1)}$ по формулам (9.10), (9.11) соответственно, а $n^2 + n - 1$ есть трудоемкость обратной подстановки.

Используем хорошо известные формулы:

$$1 + 2 + \cdots + m = \frac{m(m+1)}{2},$$
$$1 + 2^2 + \cdots + m^2 = \frac{m(m+1)(2m+1)}{6}.$$

Ясно, что

$$q_m = \sum_{k=1}^{n-1} (n-k) = \sum_{k=1}^{n-1} k = (n-1)n/2.$$

Для вычисления $b^{(k+1)}$ требуется в два раза больше операций, т. е. $q_b = (n-1)n$. Наконец,

$$q_a = \sum_{k=1}^{n-1} 2(n-k)^2 = 2\sum_{k=1}^{n-1} k^2 = (n-1)n(2n-1)/3.$$

Суммарно, $Q = 2n^3/3 + 3n^2/2 - n/6 - 1 = 2n^3/3 + O(n^2)$ flops.

3. Матричная формулировка метода Гаусса. *LU* разложение матрицы. Для $k = 1, 2, \ldots, n-1$, определим элементарную треугольную матрицу L_k:

$$L_k = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & -l_{k+1,k} & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & -l_{n,k} & 0 & \cdots & 1 \end{bmatrix}. \tag{9.12}$$

Матрица L_k отличается от единичной только поддиагональными элементами k-го столбца.

Непосредственными вычислениями легко проверить (убедитесь!), что система уравнений после первого шага метода Гаусса равносильна системе $L_1 Ax = L_1 b$, т. е. $A^{(2)} = L_1 A$, $b^{(2)} = L_1 b$ (см. формулы (9.6)). Система уравнений после k-го шага равносильна системе $L_k A^{(k)} x = L_k b^{(k)}$, т.е. $A^{(k+1)} = L_k A^{(k)}$, $b^{(k+1)} = L_k b^{(k)}$ (см. формулы (9.11)). Обозначим $A^{(n)}$ через U. Тогда

$$U = L_{n-1} L_{n-2} \ldots L_1 A, \quad b^{(n)} = L_{n-1} L_{n-2} \ldots L_1 b.$$

Отсюда находим

$$A = LU, \tag{9.13}$$

где $L = L_1^{-1} L_2^{-1} \cdots L_{n-1}^{-1}$. Нетрудно видеть (см. упражнение 8.5), что

$$L_k^{-1} = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & l_{k+1,k} & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & l_{n,k} & 0 & \cdots & 1 \end{bmatrix},$$

а матрица L является нижней треугольной с единичной главной диагональю и поддиагональными элементами равными l_{ij} (см. упражнение 8.4). Если поддиагональные элементы матрицы L и элементы U

хранить на месте соответствующих элементов A, то приходим к следующему алгоритму LU разложения матрицы A (см. формулы (9.10), (9.11)):

```
for k = 1:n-1
  for i = k+1:n
    a(i,k) = a(i,k)/a(k,k);
    for j = k+1:n
      a(i,j) = a(i,j) - a(i,k)*a(k,j);
    end
  end
end
```

Этот алгоритм назовем kij – алгоритмом; kji – алгоритм получается перестановкой циклов по i и j. Он имеет вид

```
for k=1:n-1
  for i=k+1:n
     a(i,k)=a(i,k)/a(k,k);
  end
  for j=k+1:n,
    for i=k+1:n
      a(i,j)=a(i,j)-a(i,k)*a(k,j);
    end
  end
end
```

4. Условия применимости метода Гаусса. Описанный выше метод может быть реализован лишь в том случае, когда все ведущие элементы метода Гаусса отличны от нуля. Для этого невырожденности матрицы недостаточно. Следующий пример демонстрирует это:

$$A = A^{(1)} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad \det A = -1, \quad A^{(2)} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Выделим класс матриц, для которых метод Гаусса осуществим. Пусть

$$A_1 = a_{11}, \; A_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \; \ldots, A_n = \begin{vmatrix} a_{11} & a_{22} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix}$$

есть главные миноры матрицы A.

Теорема 1. *Для того, чтобы все ведущие элементы метода Гаусса были отличны от нуля необходимо и достаточно, чтобы все главные миноры матрицы A были ненулевыми.*

Доказательство. Напомним, что $a_{ij}^{(1)} = a_{ij}$, $i, j = 1, \ldots, n$. Пусть $A_i \neq 0$, $i = 1, \ldots, n$. Покажем по индукции, что тогда $a_{kk}^{(k)} \neq 0$

для всех $k = 1, \ldots, n$. Имеем, $a_{11}^{(1)} = a_{11} = A_1 \neq 0$. Пусть уже доказано, что $a_{11}^{(1)}, a_{22}^{(2)}, \ldots, a_{k-1,k-1}^{(k-1)}$ не равны нулю. Тогда, приводя минор A_k к треугольному виду при помощи преобразований прямого хода метода Гаусса, получим

$$A_k = \begin{vmatrix} a_{11}^{(1)} & a_{22}^{(1)} & \ldots & a_{1k}^{(1)} \\ 0 & a_{22}^{(2)} & \ldots & a_{2k}^{(2)} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & a_{kk}^{(k)} \end{vmatrix} = a_{11}^{(1)} a_{22}^{(2)} \ldots a_{kk}^{(k)}, \qquad (9.14)$$

следовательно, $a_{kk}^{(k)} \neq 0$, что завершает шаг индукции. Обратное утверждение теоремы есть следствие соотношения (9.14). $\square$[1)]

Следствием теоремы (1) и формулы (9.13) является

Теорема 2. *Пусть все главные миноры матрицы A отличны от нуля. Тогда справедливо единственное представление $A = LU$, где L нижняя треугольная матрица с единичной главной диагональю, U — верхняя треугольная матрица.*

Доказательство. Доказательства требует лишь единственность разложения. Предположим, что имеются два разложения $A = L_1U_1$ и $A = L_2U_2$, т. е. $L_1U_1 = L_2U_2$. Следовательно, $L_2^{-1}L_1 = U_2U_1^{-1}$, причем левая часть этого равенства представляет собой нижнюю треугольную матрицу с единичной диагональю, а правая часть — верхнюю треугольную матрицу. Это возможно только тогда, когда $L_2^{-1}L_1 = E$, $U_2U_1^{-1} = E$, т. е. при $L_1 = L_2$ и $U_1 = U_2$. $\square$

Приведем часто встречающиеся в приложениях примеры матриц, для которых метод Гаусса применим и, соответственно, справедлива теорема 2.

i) *Эрмитовы положительно определенные матрицы.* Напомним, что матрица $A = \{a_{ij}\}_{i,j=1}^{n}$ называется эрмитовой, если $A = A^*$. Матрица A называется положительно определенной, если

$$\sum_{i,j=1}^{n} a_{ij} x_j \bar{x}_i > 0 \quad \forall\, x \neq 0.$$

В соответствии с критерием Сильвестра эрмитова матрица положительно определена тогда, и только тогда, когда все ее главные миноры положительны.

[1)] Значком $\square$ отмечаем конец доказательства.

ii) *Матрицы с диагональным преобладанием.* Матрица, элементы которой удовлетворяют условию:

$$\sum_{j=1,\, j\neq i}^{n} |a_{ij}| < |a_{ii}|, \quad i = 1, 2, \ldots, n. \tag{9.15}$$

называется матрицей с диагональным преобладанием по строкам. Аналогично, матрицы, элементы которой удовлетворяют условию:

$$\sum_{i=1,\, i\neq j}^{n} |a_{ij}| < |a_{jj}|, \quad j = 1, 2, \ldots, n.$$

называются матрицей с диагональным преобладанием по столбцам. Говорят, что A есть матрица диагональным преобладанием, если она имеет диагональное преобладание либо по строкам, либо по столбцам.

Ясно, что если A есть матрица с диагональным преобладанием по строкам, то A^T — матрица с диагональным преобладанием по столбцам.

Теорема 3. *Пусть A есть матрица с диагональным преобладанием. Тогда все ее главные миноры отличны от нуля.*

Доказательство. Достаточно считать, что A имеет диагональное преобладание по строкам. Рассмотрим главный минор A_k, $k \geqslant 1$. Достаточно убедиться, что однородная система линейных уравнений

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + \ldots + a_{1k}x_k = 0,\\ &a_{21}x_1 + a_{22}x_2 + \ldots + a_{2k}x_k = 0,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kk}x_k = 0 \end{aligned} \tag{9.16}$$

с матрицей, составленной из элементов минора A_k, имеет только тривиальное решение. Предположим противное и пусть $\max\limits_{1\leqslant j\leqslant k} |x_j| = |x_i|$. Ясно, что $x_i \neq 0$. Поскольку $i \leqslant k$, то i-е уравнение системы (9.16) можно записать в виде

$$a_{ii}x_i = -\sum_{j=1,\, j\neq i}^{k} a_{ij}x_j.$$

Следовательно,

$$|a_{ii}||x_i| \leqslant \sum_{j=1,\, j\neq i}^{k} |a_{ij}|\,|x_j| \leqslant |x_i| \sum_{j=1,\, j\neq i}^{n} |a_{ij}|.$$

Отсюда следует, что $|a_{ii}| \leqslant \sum_{j=1,\ j\neq i}^{n} |a_{ij}|$, а это противоречит условию (9.15). $\square$

5. Компактная схема *LU* разложения матрицы. Посмотрим на разложение $A = LU$ как на уравнение $LU = A$ для определения элементов матриц L и U. Тогда получим n^2 уравнений

$$\sum_{k=1}^{n} l_{ik}u_{kj} = a_{ij}, \quad i, j = 1, \ldots, n. \tag{9.17}$$

Поскольку $l_{ii} = 1$, $l_{ik} = 0$, если $k > i$, и $u_{kj} = 0$, если $k > j$, то равенства (9.17) можно записать в виде

$$\sum_{k=1}^{\min(i,j)} l_{ik}u_{kj} = a_{ij}\,, \quad i, j = 1, \ldots, n.$$

Из этих равенств следуют соотношения

$$\sum_{k=1}^{j-1} l_{ik}u_{kj} + l_{ij}u_{jj} = a_{ij}\,, \quad i > j\,,$$

$$\sum_{k=1}^{i-1} l_{ik}u_{kj} + u_{ii} = a_{ij}\,, \quad i \leqslant j\,,$$

из которых вытекают формулы

$$l_{ij} = (a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj})/u_{jj}\,, \quad i > j\,, \tag{9.18}$$

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}\,, \quad i \leqslant j\,. \tag{9.19}$$

Из формул (9.18), (9.19) можно получить различные алгоритмы вычисления элементов L и U, если определить порядок их вычисления.

Например, следующий ijk-алгоритм позволяет вычислять элементы L и U построчно: для всех $i = 1, 2, \ldots, n$, сначала вычисляются l_{ij} по формулам (9.18) для всех $j = 1, \ldots, i-1$, а затем u_{ij} по формулам (9.19) для всех $j = i, \ldots, n$.

В jik-алгоритме элементы L и U вычисляются по столбцам: для всех $j = 1, 2, \ldots, n$, сначала вычисляются u_{ij} по формулам (9.19)

для всех $i = 1, \ldots, j$, а затем l_{ij} по формулам (9.18) для всех $i = j+1, \ldots, n$.

Легко проверить, что трудоемкость этих алгоритмов одна и та же и равна $(2/3)\,n^3 + O(n^2)$ flops. Отметим также, что как и в рассмотренном первоначально алгоритме LU разложения, элементы L и U в ходе вычисления можно располагать в соответствующих позициях матрицы A.

6. Метод Гаусса с выбором ведущего элемента по столбцу. Опишем модификацию изученного выше метода Гаусса, который применим для решения систем уравнений с любой невырожденной матрицей.

Выберем среди элементов первого столбца матрицы A максимальный по модулю. Пусть это есть элемент $a_{i_1,1}$. Он не может оказаться равным нулю, так как тогда все элементы первого столбца матрицы A — нули и, значит, $|A| = 0$, что противоречит условию $|A| \neq 0$.

Умножим обе части уравнения на матрицу перестановки $P_{i_1,1}$. В дальнейшем будем обозначать эту матрицу через P_1 (заметим, что она равна единичной, если максимальный по модулю элемент первого столбца матрицы A есть a_{11}). Получим

$$A^{(1)}x = b^{(1)}, \tag{9.20}$$

где $A^{(1)} = P_1 A$, $b^{(1)} = P_1 b$. Поясним, что матрица $A^{(1)}$ получается из матрицы A перестановкой первой и i_1-й строк, вектор-столбец $b^{(1)}$ получается из столбца b перестановкой первого и i_1-го элементов. Элементы матрицы $A^{(1)}$ обозначим через $a^{(1)}_{kl}$, элементы столбца $b^{(1)}$ — через $b^{(1)}_k$. По построению $a^{(1)}_{11} \neq 0$.

Теперь можем осуществить первый шаг рассмотренного ранее метода Гаусса и привести матрицу $A^{(1)}$ к верней треугольной форме в первом столбце. Это равносильно умножению обеих частей уравнения (9.20) на элементарную нижнюю треугольную матрицу L_1 вида (9.12), элементы которой определяются по формулам (9.4). В результате, придем к системе уравнений

$$A^{(2)}x = b^{(2)}, \tag{9.21}$$

где $A^{(2)} = L_1 A^{(1)} = L_1 P_1 A$, $b^{(2)} = L_1 b^{(1)} = L_1 P_1 b$. На этом заканчивается первый шаг исключения неизвестных.

На втором шаге среди элементов $a^{(2)}_{22}$, $a^{(2)}_{32}$, $\ldots$, $a^{(2)}_{n2}$ найдем максимальный по модулю. Пусть этот элемент есть $a^{(2)}_{i_2,2}$. Он не может равняться нулю. Действительно, если он равен нулю, то все числа

$a_{22}^{(2)}$, $a_{32}^{(2)}$, $\ldots$, $a_{n2}^{(2)}$ — нули и тогда, вычисляя $|A^{(2)}|$ разложением по первому столбцу, получим, что $|A^{(2)}| = 0$. С другой стороны, поскольку $|L_1| = 1$, а $|P_1| \neq 0$, то $|A^{(2)}| = |L_1|\,|P_1|\,|A| \neq 0$, что приводит к противоречию.

Умножим обе части уравнения (9.21) на матрицу $P_2 = P_{i_2,2}$, т. е. поменяем местами вторую и i_2-ю строки матрицы $A^{(2)}$. Получим

$$\tilde{A}^{(2)}x = P_2 L_1 P_1 b. \tag{9.22}$$

По определению элемент $\tilde{a}_{22}^{(2)} \neq 0$. Это позволяет осуществить второй шаг рассмотренного ранее метода Гаусса и привести матрицу $\tilde{A}^{(2)}$ к верней треугольной форме и во втором столбце. Это равносильно умножению обеих частей уравнения (9.23) на элементарную нижнюю треугольную матрицу L_2. В результате второго шага получим систему уравнений

$$A^{(3)}x = L_2 P_2 L_1 P_1 b, \tag{9.23}$$

где $A^{(3)} = L_2 P_2 L_1 P_1 A$.

Продолжая этот процесс, после $n-1$ шага получим систему уравнений с верхней треугольной матрицей $U = A^{(n)}$,

$$Ux = f \tag{9.24}$$

(очевидно, эквивалентную исходной), где

$$U = L_{n-1}P_{n-1}\cdots L_1 P_1 A, \tag{9.25}$$

$$f = L_{n-1}P_{n-1}\cdots L_1 P_1 b.$$

Решение системы (9.24) не вызывает затруднений.

Замечание 1. Выбор максимального по модулю элемента столбца при выполнении прямого хода метода Гаусса минимизирует влияние ошибок округления. Если не заботиться об ошибках округления, то на очередном шаге прямого хода метода Гаусса можно выбирать любой ненулевой элемент столбца.

Теорема 4. *Пусть* $|A| \neq 0$. *Тогда справедливо разложение* $PA = LU$, *где* L — *нижняя треугольная матрица с единичной главной диагональю,* U — *верхняя треугольная матрица,* $P = P_{i_{n-1},n-1}P_{i_{n-2},n-2}\cdots P_{i_1,1}$ — *матрица перестановок,* $i_k \geqslant k$, $k = 1,\ldots,n-1$.

Доказательство. Согласно формуле (9.25)

$$A = P_1 L_1^{-1}\cdots P_{n-2}L_{n-2}^{-1}P_{n-1}L_{n-1}^{-1}U. \tag{9.26}$$

Здесь мы учли, что произведение $P_k P_k$ есть единичная матрица. Это также позволяет эквивалентно преобразовать (9.26) к виду

$$P_{n-1}P_{n-2}\cdots P_1 A = \Big(P_{n-1}P_{n-2}\cdots P_2 L_1^{-1} P_2 P_3 \cdots P_{n-1}\Big) \Big(P_{n-1}\cdots P_3 L_2^{-1} P_3 \cdots P_{n-1}\Big)\cdots\Big(P_{n-1}L_{n-2}^{-1}P_{n-1}\Big)L_{n-1}^{-1}U = \\ = \big(\widetilde{L}_1^{-1}\widetilde{L}_2^{-1}\cdots\widetilde{L}_{n-2}^{-1}L_{n-1}^{-1}\big)U.$$

Отсюда следует утверждение теоремы. Действительно, каждая из матриц $\widetilde{L}_k^{-1}$ представляет собой элементарную нижнюю треугольную матрицу с единичной диагональю, отличающуюся от L_k^{-1} лишь перестановкой поддиагональных элементов в k-м столбце, а матрица $L = \widetilde{L}_1^{-1}\widetilde{L}_2^{-1}\ldots\widetilde{L}_{n-2}^{-1}L_{n-1}^{-1}$ есть нижняя треугольная с единичной диагональю. □

Программная реализации LU разложения матрицы методом Гаусса с выбором ведущего элемента по столбцу осуществляется также, как и описанное ранее LU разложение. Необходимо лишь внести изменения, связанные с перестановкой строк матрицы и запоминанием этих перестановок. Например, kij алгоритм примет вид:

```
function [A,p] = lukij(A)
n = size(A,1);
p = 1:n;
for k = 1:n-1
  [~, I] = max(abs(A(k:n,k)));
  row = I+k-1;
  a([k, row],:) = a([row, k],:);
  p([k, row])   = p([row, k]);
  for i = k+1:n
    a(i,k) = a(i,k)/a(k,k);
    for j = k+1:n
      a(i,j) = a(i,j) - a(i,k)*a(k,j);
    end
  end
end
```

В результате выполнения этой функции, матрицы L и U сохраняются на месте матрицы A.

Пусть $[LU, p] = lukij(A)$. Тогда команды $L = tril(LU, -1) + eye(n)$; $U = triu(LU)$ позволяют при необходимости получить L и U. Вектор перестановок p таков, что $A(p, :) = LU$.

Упражнение 9.1. Пусть диагональные элементы L и U равны единице. Получить формулы для элементов L^{-1} и U^{-1}. Оценить трудоемкость.

10. Метод Холесского.

Если матрица системы линейных уравнений эрмитова и положительно определена, можно добиться существенного сокращения числа операций и памяти, необходимых для разложения ее на треугольные множители. В основе соответствующего метода лежит

Теорема 1. *Пусть матрица A эрмитова и положительно определена. Тогда существует нижняя треугольная матрица L с положительными элементами на диагонали такая, что $A = LL^*$.*

Доказательство. Используем индукцию по порядку матрицы. Для матрицы первого порядка имеем тривиальное равенство $a_{11} = \sqrt{a_{11}}\sqrt{a_{11}}$. Пусть утверждение теоремы верно для матриц порядка $k > 1$. Покажем, что тогда оно верно и для матриц порядка $k+1$. Запишем матрицу A порядка $k+1$ как блочную:

$$A = \begin{bmatrix} A_k & a_k \\ a_k^* & a_{k+1,k+1} \end{bmatrix}.$$

Здесь A_k — матрица порядка k. Очевидно, она эрмитова и положительно определена. В силу предположения индукции $A_k = L_k L_k^*$, где L_k — нижняя треугольная матрица с положительными элементами на диагонали. Будем искать разложение матрицы A на треугольные множители в виде

$$A = LL^* = \begin{bmatrix} L_k & 0 \\ l_k^* & l_{k+1,k+1} \end{bmatrix} \begin{bmatrix} L_k^* & l_k \\ 0 & l_{k+1,k+1} \end{bmatrix}. \tag{10.1}$$

Выполняя умножение в правой части последнего равенства и сравнивая поблочно результат с матрицей A, получим систему линейных уравнений

$$L_k l_k = a_k \tag{10.2}$$

для определения вектора l_k и уравнение $l_k^* l_k + l_{k+1,k+1}^2 = a_{k+1,k+1}$ для элемента $l_{k+1,k+1}$. Можно считать, что $l_{k+1,k+1} > 0$, так как вследствие (10.1) имеем: $|A| = |A_k|\, l_{k+1,k+1}^2$, причем $|A_k|, |A| > 0$, так как по условию матрицы A_k, A положительно определены. Таким образом, для построения матрицы L нужно решить систему уравнений (10.2) с треугольной матрицей, а затем вычислить $l_{k+1,k+1}$ по формуле $l_{k+1,k+1} = \sqrt{a_{k+1,k+1} - l_k^* l_k}$. □

Доказательство теоремы 1, фактически, описывает алгоритм разложения на треугольные множители произвольной эрмитовой положительно определенной матрицы. Нетрудно видеть, что его реализация по затратам памяти и объему вычислений оказывается примерно

в два раза более экономичной, чем разложение на треугольные множители произвольной невырожденной матрицы.

После того, как матрица L построена, решение системы уравнений $Ax = b$ сводится к последовательному решению систем уравнений $Ly = b$, $L^*x = y$ с треугольными матрицами.

Упражнения.

10.1. Покажите, что трудоемкость метода Холесского равна $n^3/3 + O(n^2)$ flops..

10.2. Докажите, что при выполнении условий теоремы 1 нижняя треугольная матрица L в разложении $A = LL^*$ определяется однозначно.

11. Унитарная триангуляция матриц.

В этом параграфе будет доказана

Теорема 1. *Пусть A — произвольная квадратная матрица. Тогда существует унитарная матрица Q такая, что*

$$A = QR, \tag{11.1}$$

где R — верхняя треугольная матрица.

Если разложение (11.1) получено, то решение системы уравнений $Ax = b$ с невырожденной матрицей A сводится к вычислению вектора $f = Q^*b$ и решению системы уравнений $Rx = f$ с треугольной невырожденной матрицей.

При построении разложения (11.1) используются специальные унитарные матрицы, позволяющие решить следующую задачу.

Даны ненулевой вектор $a \in \mathbb{C}^n$ и вектор $i^1 = (1, 0, \ldots, 0)^T$. Требуется построить унитарную матрицу V такую, что $Va = \mu\, i^1$, где μ — число (ясно, что $|\mu| = |a|$, поскольку матрица V унитарна).

1. Матрицы вращения. Матрица

$$G_{kl} = \{g_{ij}\}_{i,j=1}^n, \quad 1 \leqslant k < l \leqslant n,$$

называется *матрицей вращения*, если $g_{ii} = 1$ при $i \neq k, l$, $g_{kk} = c$, $g_{ll} = \bar{c}$, $g_{kl} = -s$, $g_{lk} = \bar{s}$, все остальные элементы матрицы G_{kl} равны нулю, причем $|c|^2 + |s|^2 = 1$. Нетрудно видеть, что $G = G_{kl}$ — унитарная матрица.

Если числа c, s вещественны, то матрица G ортогональна. При этом порождаемое ей преобразование евклидова пространства $\mathbb{R}^n$ со стандартным скалярным произведением представляет собой поворот на угол $\varphi = \operatorname{arctg}(s/c)$ в двумерном подпространстве (плоскости), натянутом на векторы i^k, i^l естественного базиса пространства $\mathbb{R}^n$.

Матрица G^T, обратная к G, выполняет поворот в той же плоскости в обратном направлении.

Пусть a — произвольный вектор пространства $\mathbb{C}^n$. Ясно, что $(Ga)_i = a_i$ при $i \neq k, l$,

$$(Ga)_k = c\, a_k - s\, a_l,$$
$$(Ga)_l = \bar{s}\, a_k + \bar{c}\, a_l.$$

Положим $\rho = (|a_k|^2 + |a_l|^2)^{1/2}$. Пусть $c = 1, s = 0$, если $\rho = 0$, и $c = \bar{a}_k/\rho$, $s = -\bar{a}_l/\rho$, если $\rho > 0$. Тогда $(Ga)_k = \rho$, $(Ga)_l = 0$.

Теперь совершенно ясно, что если a — произвольный ненулевой вектор пространства $\mathbb{C}^n$, то выбирая последовательно числа $c_n, s_n, c_{n-1}, s_{n-1}, \ldots, c_2, s_2$, можно построить матрицы вращения $G_{1,n}$, $G_{1,n-1}, \ldots, G_{1,2}$ такие, что $Gx = |a|\, i^1$. Здесь $G = G_{1,2} \cdots G_{1,n-1} G_{1,n}$.

Таким образом, любой ненулевой вектор при помощи ортогональной матрицы можно преобразовать в вектор, совпадающий по направлению с вектором i^1 естественного базиса.

Замечание 1. Если вектор a принадлежит $\mathbb{R}^n$, то все матрицы $G_{1,n}, G_{1,n-1}, \ldots, G_{1,2}$, а следовательно, и матрица G — вещественные (ортогональные) матрицы.

Пусть теперь a, e — два произвольных ненулевых вектора пространства $\mathbb{C}^n$. Как только что было показано, существуют унитарные матрицы $G(a)$ и $G(e)$ такие, что $G(a)a = |a|\, i^1$, $G(e)e = |e|\, i^1$. Отсюда вытекает, что $Ga = \mu\, e$, где $\mu = |a|/|e|$, $G = G^*(e)G(a)$, т. е. для любой пары ненулевых векторов найдется унитарная матрица, преобразующая первый вектор в вектор, совпадающий по направлению со вторым.

2. Матрицы отражения. Пусть произвольно задан вектор $w = (w_1, w_2, \ldots, w_n\}^T \in \mathbb{C}^n$ единичной длины (матрица $n \times 1$). Матрица

$$H = H(w) = I - 2ww^* = \{\delta_{ij} - 2w_i \bar{w}_j\}_{i,j=1}^n \tag{11.2}$$

называется *матрицей отражения*. Отметим ряд ее свойств.

1. Матрица H эрмитова. Кроме того, она унитарна. Действительно,

$$H^*H = H^2 = I - 4ww^* + 4w(w^*w)w^* = I,$$

так как $w^*w = |w|^2 = 1$. Таким образом, $H = H^* = H^{-1}$.

2. Пусть $E_{n-1} = \{z \in \mathbb{C}^n : w^*z = (z, w) = 0\}$ — гиперплоскость размерности $n - 1$, нормальная к вектору w. Заметим, что

$$Hw = w - 2ww^*w = -w, \quad Hz = z - 2ww^*z = z, \quad z \in E_{n-1}. \tag{11.3}$$

Следовательно, H имеет однократное собственное значение равное -1, которому соответствует собственный вектор w, и собственное значение $+1$ кратности $n-1$, которому соответствует собственное подпространство E_{n-1}. Отсюда следует, что $|H| = -1$.

3. Пусть x — произвольный вектор, а z его проекция на гиперплоскость E_{n-1}. Ясно, что векторы x, z и w лежат в двумерной плоскости, нормальной к E_{n-1}, и x однозначно представим в виде $x = \alpha w + z$, где α некоторое число. Из равенств (11.3) вытекает, что $Rx = -\alpha w + z$ (сделайте рисунок!). Можно сказать, таким образом, что отображение, порождаемое матрицей H, выполняет отражение вектора x относительно гиперплоскости E_{n-1}, ортогональной вектору w. Это свойство матрицы H и позволяет называть ее матрицей отражения.

4. Пусть заданы векторы a, $e \in \mathbb{C}^n$, $|a| \neq 0$, $|e| = 1$ и φ есть аргумент (a, e), если $(a, e) \neq 0$[1)]. Рассмотрим задачу построения такой матрицы отражения $H = H(w)$, что $Ha = \mu e$, где $|\mu| = |a|$. Из геометрических соображений ясно, что эта задача имеет два решения[2)]. Положим

$$w = (a - \mu e)/\nu, \quad \nu = |a - \mu e|\,. \tag{11.4}$$

Имеем

$$\nu^2 = (a - \mu e, a - \mu e) = 2\,\mathrm{Re}\,(a, a - \mu e)\,, \tag{11.5}$$

$$H(w)a = a - \frac{2(a, a - \mu e)}{\nu^2}\,(a - \mu e). \tag{11.6}$$

Из формул (11.5), (11.6) следует, что $H(w)a = \mu e$, если $(a, a - \mu e)$ есть вещественное число, т. е. $\mathrm{Im}\,(a, \mu e) = \mathrm{Im}\,[\bar{\mu}\,(a, e)] = 0$. Это условие выполнено, если $(a, e) = 0$ и $\mu = \pm|a|$; в противном случчае, оно выполнено, если положить $\mu = \pm e^{i\varphi}\,|a|$.

Чаще всего в приложениях приходится рассматривать случай, когда $e = i^1 = (1, 0, \ldots, 0)^T$. В этом случае $(a, \mu e) = \bar{\mu} a_1$, поэтому следует положить $\mu = \pm|a| a_1/|a_1|$. Считается, конечно, что $a_1 \neq 0$. В противном случае полагаем $\mu = \pm|a|$. Следовательно, решение задачи $H(w)a = \mu i^1$ определяется формулой (11.2) при

$$v = (a_1 - \mu, a_2, \ldots, a_n)^T, \quad w = \frac{v}{|v|}\,, \quad \mu = \pm \begin{cases} |a|, & a_1 = 0, \\ \frac{|a|\,a_1}{|a_1|}, & a_1 \neq 0. \end{cases} \tag{11.7}$$

[1)] Напомним, что соотношение $z = e^{i\varphi}\,|z|$ задает тригонометрическое представление $z \in \mathbb{C}$, а φ называется аргументом z.

[2)] Проиллюстрируйте построение вектора w рисунком в двумерном вещественном случае.

Конкретное решение (т. е. знак μ) выбирается из дополнительных соображений, например, с целью получить более устойчивый к погрешностям округления алгоритм при вычислениях на ЭВМ.

5. Экономное вычисление w по формулам (11.7) требует $3n$ flops. Матрицу отражения в памяти ЭВМ можно не хранить; достаточно хранить только вектор w. Произведение $y = H(w)a$ при заданном a в этом случае вычисляется по формуле

$$y = (I - 2ww^*)a = a - \lambda\, w, \quad \lambda = 2\, w^* a, \tag{11.8}$$

а его трудоемкость равна $4n$ flops.

3. Доказательство теоремы 1. Доказательство является конструктивным и дает метод построения матриц Q, R, называемый *методом отражения*. Он состоит из $n-1$ шага и на k-м шаге матрица A преобразуется к матрице, имеющей верхнюю треугольную форму в k-м столбце. Обозначим через I_n единичную матрицу длины n.

Пусть a_j есть j-й столбец A. Если $a_1 = 0$, то перейдем ко второму шагу, полагая $H^{(1)} = I_n$, $A^{(1)} = A$. Иначе, выберем $H^{(1)} = H_1(w_1)$ как такую матрицу отражения, что $H^{(1)}a_1 = \mu_1\, i^1$ и вычислим $A^{(1)} = H^{(1)}A$. По определению

$$A^{(1)} = [H^{(1)}a_1,\, H^{(1)}a_2, \ldots, H^{(1)}a_n]\,.$$

На этом заканчивается первый шаг, после которого матрица $A^{(1)}$ имеет верхнюю треугольную форму в первом столбце и в блочном виде имеет представление

$$A^{(1)} = \begin{bmatrix} \mu_1 & c_1 \\ 0 & A_1 \end{bmatrix},$$

где $\mu_1 = \pm|a_1|a_{11}/|a_{11}|$, если $a_{11} \neq 0$, в противном случае $\mu_1 = \pm|a_1|$, A_1 — некоторая квадратная матрица размера $n-1$.

Подсчитаем трудоемкость этого шага. На вычисление w_1 требуется $3n$ flops. Вычисление произведений $H_1(w_1)a_2, \ldots, H_1(w_1)a_n$ требует $4n(n-1)$ flops. Таким образом, трудоемкость первого шага равна $4n^2 - n$ flops, если $a_1 \neq 0$.

Аналогично осуществляется второй шаг с той лишь разницей, что вычисления производятся с матрицей A_1. А именно, если первый столбец A_1 равен нулю, положим $H^{(2)} = I_n$, $A^{(2)} = A^{(1)}$. Иначе, определим $A^{(2)} = H^{(2)}A^{(1)}$, где матрица $H^{(2)}$ имеет вид

$$H^{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & H_2(w_2) \end{bmatrix}.$$

В этом случае

$$A^{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & H_2(w_2) \end{bmatrix} \begin{bmatrix} \mu_1 & c_1 \\ 0 & A_1 \end{bmatrix} = \begin{bmatrix} \mu_1 & c_1 \\ 0 & H_2(w_2)A_1 \end{bmatrix}.$$

Как и на первом шаге, выберем матрицу $H_2(w_2)$ как такую матрицу отражения, что $H_2(w_2)A_1$ имеет верхнюю треугольную форму в первом столбце. Размерность этой задачи на единицу меньше, чем на первом шаге, и равна $n-1$. Соответственно, трудоемкость второго шага равна $4(n-1)^2-(n-1)$ flops, если первый столбец A_1 отличен от нуля. Легко видеть, что матрица $H^{(2)}$ является унитарной.

Повторяя построения на k шаге определим унитарную матрицу

$$H^{(k)} = \begin{bmatrix} I_{k-1} & 0 \\ 0 & H_k(w_k) \end{bmatrix},$$

если матрица $A^{(k-1)}$ не имеет верхней треугольной формы в k-м столбце, в противном случае полагаем $H^{(k)} = I_n$, а также матрицу $A^{(k)} = H^{(k)}A^{(k-1)}$. Матрица $H_k(w_k)$ размера $n-k+1$ строится как соответствующая матрица отражения.

После $n-1$ шага получим унитарные матрицы $H^{(1)}$, $H^{(2)}$, $\ldots$, $H^{(n-1)}$ такие, что $H^{(n-1)}H^{(n-2)}\cdots H^{(1)}A = A^{(n-1)} = R$, где R — верхняя треугольная матрица. Следовательно, $A = QR$, где $Q = H^{(1)}H^{(2)}\cdots H^{(n-1)}$ суть унитарная матрица.

Трудоемкость метода равна

$$\sum_{k=2}^{n} \Big(4k^2 - k\Big) = \frac{4}{3}\,n^3 + O(n^2),$$

что при больших значениях n в два раз больше, чем требуется для разложения $PA = LU$ методом Гаусса. $\square$

Важным положительным качеством описанного метода является возможность его непосредственного применения для произвольной невырожденной матрицы без какой-либо перенумерации ее строк, а также его устойчивость к ошибкам округления. Последнее объясняется тем, что при унитарном преобразовании длина вектора не меняется.

Замечание 2. Без ограничения общности можно считать что все диагональные элементы матрицы R неотрицательны. В самом деле, если, например, на первом этапе описываемого в доказательстве теоремы 1 алгоритма получаем, что $\mu_1 = \pm\rho e^{i\varphi}$, то матрицу $H^{(1)}$

нужно заменить на матрицу $\pm e^{-i\varphi}H^{(1)}$, которая тоже, очевидно, унитарна. Аналогичное замечание относится и к последующим этапам построения матрицы R.

Замечание 3. Совершенно аналогично можно получить представление произвольной квадратной матрицы A в виде $A = QL$, где Q — унитарная матрица, а L — нижняя треугольная матрица, а также RQ и LQ разложения.

Упражнения.

11.1. Постройте алгоритм, аналогичный описанному при доказательстве теоремы 1 и основанный на использовании матриц вращения.

11.2. Докажите, что если матрица A невырождена, а диагональные элементы матрицы R считаются положительными, то матрицы Q, R в разложении (11.1) определяются однозначно.

11.3. Укажите метод построения разложений $A = QL$, $A = RQ$ и $A = LQ$.

12. Построение обратной матрицы

Задача построения обратной матрицы сводится к решению n систем линейных уравнений с одной и той же матрицей A и различными правыми частями. Действительно, обозначим матрицу A^{-1} через X. Тогда $AX = E$. Осталось записать это равенство подробнее:

$$Ax^k = i^k, \quad k = 1, 2, \ldots, n. \tag{12.1}$$

Здесь x^k — k-й столбец матрицы X,

$$i^k = (\underbrace{0, \ldots, 0}_{k-1}, 1, \underbrace{0, \ldots, 0}_{n-k})^T.$$

Рассмотрим два способа вычисления обратной матрицы.

1. Методом Гаусса с выбором ведущего элемента по столбцу вычислим матрицу перестановок P и треугольные матрицы L и R такие, что $PA = LU$. Это потребует $2/3n^3 + O(n^2)$ flops. Тогда X есть решение уравнения $LUX = P$, или $LUx^k = p^k$, где p^k есть k-й столбец P. Нахождение x^k требует решения систем $Ly = p^k$, $Ux^k = y$. Их суммарная трудоемкость равна $2n^2 + O(n)$ flops. Следовательно, матрица A^{-1} этим методом вычисляется за $(2 + 2/3)n^3 + O(n^2)$ flops.

2. Методом отражения найдем разложение $A = QR$, затратив $4/3n^3 + O(n^2)$ flops. Тогда $RX = Q^*$. Определение X из этого уравнения потребует $n^3 + O(n^2)$ flops. Суммарно, матрица A^{-1} этим методом вычисляется за $(1 + 4/3)n^3 + O(n^2)$ flops, что на $n^3/3$ flops меньше, чем в первом методе при больших n.

13. Метод прогонки для систем с трехдиагональными матрицами.

В приложениях довольно часто возникают системы уравнений с матрицами, большинство элементов которых — нули. Это так называемые разреженные матрицы. Процесс исключения неизвестных в таких системах (или разложение матриц на треугольные множители) во многих практически важных ситуациях удается организовать так, чтобы существенно сократить затраты памяти и объем необходимых вычислений.

Рассмотрим наиболее простой случай, а именно системы с матрицами, ненулевые элементы которых лежат лишь на главной и двух соседних с ней диагоналях. Системы такого вида часто возникают при приближенном решении задач математической физики. Соответствующие матрицы принято называть трехдиагональными.

Произвольную систему с трехдиагональной матрицей можно записать в следующем виде:

$$\begin{aligned} b_1x_1 + c_1x_2 &= f_1, \\ a_2x_1 + b_2x_2 + c_2x_3 &= f_2, \\ \dots\dots\dots\dots\dots\dots\dots\dots & \\ a_ix_{i-1} + b_ix_i + c_ix_{i+1} &= f_i, \\ \dots\dots\dots\dots\dots\dots\dots\dots & \\ a_nx_{n-1} + b_nx_n &= f_n. \end{aligned} \tag{13.1}$$

Разрешим первое уравнение системы относительно x_1. Получим:

$$x_1 = \alpha_2x_2 + \beta_2, \tag{13.2}$$

где

$$\alpha_2 = -\frac{c_1}{b_1}, \quad \beta_2 = \frac{f_1}{b_1}. \tag{13.3}$$

Используя соотношение (13.2) и второе уравнение системы (13.1), получим аналогичное выражение для x_2. Вообще, если $x_{i-1} = \alpha_ix_i + \beta_i$, то из i-го уравнения системы (13.1) получим

$$x_i = \alpha_{i+1}x_{i+1} + \beta_{i+1}, \quad i = 1, 2, \dots, n-1, \tag{13.4}$$

где

$$\alpha_{i+1} = -\frac{c_i}{b_i + a_i\alpha_i}, \quad \beta_{i+1} = \frac{f_i - a_i\beta_i}{b_i + a_i\alpha_i}. \tag{13.5}$$

Это означает, что формулы (13.4) справедливы для $i = 1, 2, \ldots, n-1$, формулы (13.5) — для $i = 2, 3, \ldots, n-1$. Используя (13.3) и (13.5), можно найти все $\alpha_i, \beta_i,\ i = 2, \ldots, n$.

Записывая теперь соотношение (13.4) при $i = n-1$ и последнее уравнение системы (13.1), получим

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

откуда находим, что $x_n = (f_n - a_n\beta_n)/(b_n + a_n\alpha_n)$, и, наконец, используя формулы (13.4) для $i = n-1, n-2, \ldots, 1$, найдем все остальные компоненты вектора x.

Описанный алгоритм носит название метода прогонки. Понятно, что — это метод Гаусса, записанный применительно к случаю трехдиагональной системы уравнений, причем процесс вычислений α_i, β_i (прямой ход метода прогонки) соответствует прямому ходу метода Гаусса, а вычисления по формулам (13.4) (обратный ход метода прогонки) соответствуют обратному ходу метода Гаусса.

Нетрудно подсчитать необходимые затраты: требуется примерно $8n$ flops и не более $6n$ ячеек памяти.

Метод может быть реализован, когда все знаменатели в формулах (13.3), (13.5) отличны от нуля. Учитывая связь метода прогонки с методом Гаусса, можно сказать, что данное условие выполнено, например, когда матрица системы (13.1) — матрица с диагональным преобладанием, т. е. $|c_1| < |b_1|$, $|a_n| < |b_n|$, $|a_i| + |c_i| < |b_i|$, $i = 2, \ldots, n-1$.

Глава 3

Вспомогательные сведения из теории операторов. Системы уравнений общего вида

14. Дефект и ранг линейного оператора.

Пусть $\mathcal{A}$ — линейный оператор, действующий из линейного пространства $\mathbf{X}$ в линейное пространство $\mathbf{Y}$.

Множество всех векторов y из пространства $\mathbf{Y}$ таких, что $y = \mathcal{A}x$ для некоторого $x \in \mathbf{X}$, называется *областью значений или образом* оператора и обозначается через $\mathrm{Im}(\mathcal{A})$.

Множество всех векторов $x \in \mathbf{X}$ таких, что $\mathcal{A}x = 0$, называется *ядром* оператора $\mathcal{A}$ и обозначается через $\mathrm{Ker}(\mathcal{A})$.

Теорема 1. *Множество* $\mathrm{Im}(\mathcal{A})$ *— линейное подпространство пространства* $\mathbf{Y}$.

Доказательство. Пусть $y^1, y^2 \in \mathrm{Im}(\mathcal{A})$. Тогда существуют $x^1, x^2 \in \mathbf{X}$ такие, что $y^1 = \mathcal{A}x^1$, $y^2 = \mathcal{A}x^2$. Для любых $\alpha, \beta \in \mathbb{C}$ отсюда получаем, что $\alpha y^1 + \beta y^2 = \alpha\mathcal{A}x^1 + \beta\mathcal{A}x^2$. Оператор $\mathcal{A}$ линеен, следовательно, $\alpha y^1 + \beta y^2 = \mathcal{A}(\alpha x^1 + \beta x^2)$, потому $\alpha y^1 + \beta y^2 \in \mathrm{Im}(\mathcal{A})$. □

Упражнение 14.1. Покажите, что $\mathrm{Ker}(\mathcal{A})$ — линейное подпространство пространства $\mathbf{X}$.

В дальнейшем полагаем, что пространства $\mathbf{X} = \mathbf{X}_n$, $\mathbf{Y} = \mathbf{Y}_m$ конечномерны (нижний индекс означает размерность). Размерность подпространства $\mathrm{Im}(\mathcal{A}) \subset \mathbf{Y}_m$ называется *рангом* оператора $\mathcal{A}$ и обозначается через $\mathrm{rank}(\mathcal{A})$.

Размерность ядра оператора $\mathcal{A}$ называется *дефектом* оператора $\mathcal{A}$ и обозначается через $\mathrm{def}(\mathcal{A})$.

Теорема 2. *Для любого линейного оператора* $\mathcal{A} : \mathbf{X}_n \to \mathbf{Y}_m$

$$\mathrm{rank}(\mathcal{A}) + \mathrm{def}(\mathcal{A}) = n. \tag{14.1}$$

Доказательство. Обозначим через M подпространство пространства $\mathbf{X}_n$ такое, что $\mathbf{X}_n = \mathrm{Ker}(\mathcal{A}) \dotplus M$. По теореме 1, с. 147, [5], имеем $n = \mathrm{def}(\mathcal{A}) + \dim(M)$. Теперь[1)] достаточно установить, что пространства M и $\mathrm{Im}(A)$ изоморфны. Для произвольного $x \in \mathbf{X}_n$

[1)] См. теорему 3, с. 159, [5],

имеем $x = x^0 + x^1$, где $x^0 \in \mathrm{Ker}(\mathcal{A})$, $x^1 \in M$, следовательно, $\mathcal{A}x = \mathcal{A}x^1$. Таким образом, всякий элемент из $\mathrm{Im}(\mathcal{A})$ — образ некоторого элемента из M. Осталось доказать, что если $\mathcal{A}x' = \mathcal{A}x''$ для $x', x'' \in M$, то $x' = x''$, т. е. оператор $\mathcal{A}$ осуществляет взаимноодно-значное отображение M на $\mathrm{Im}(\mathcal{A})$. Равенство $\mathcal{A}(x' - x'') = 0$ означает, что $x' - x'' \in \mathrm{Ker}(\mathcal{A})$. С другой стороны, M — подпространство, поэтому $x' - x'' \in M$. По теореме 7.2, с. 146, [5], отсюда получаем, что $x' - x'' = 0$. □

15. Ранг матрицы.

1. Пусть $A(m, n)$ — произвольная прямоугольная матрица. Будем трактовать ее столбцы как систему векторов пространства $\mathbb{C}^m$. Ранг этой системы векторов (см. §5 с. 121, [5]) назовем *рангом матрицы* $A(m, n)$. Ранг матрицы A будем обозначать через $\mathrm{rank}(A)$.

Теорема 1. *Пусть* $\mathcal{A} : \mathbf{X}_n \to \mathbf{Y}_m$, A_{eq} *— матрица оператора* $\mathcal{A}$ *относительно произвольным образом фиксированных базисов* $\{e_k\}_{k=1}^n \subset \mathbf{X}_n$, $\{q_k\}_{k=1}^m \subset \mathbf{Y}_m$. *Тогда* $\mathrm{rank}(A_{eq}) = \mathrm{rank}(\mathcal{A})$.

Доказательство. Пусть $x = \mathcal{E}_n\xi \in \mathbf{X}_n$. Тогда $\mathcal{A}x = \mathcal{Q}_m\eta$, где $\eta = A_{eq}\xi$ (см. п. 2, с. 162, [5]). Понятно, что вектор η принадлежит подпространству пространства $\mathbb{C}^m$, натянутому на столбцы матрицы A_{eq} и, следовательно, имеющему размерность, равную $\mathrm{rank}(A_{eq})$. Поскольку линейный оператор $\mathcal{Q}$ обратим, то, очевидно, указанное подпространство изоморфно $\mathrm{Im}\,\mathcal{A}$, следовательно, в силу теоремы 4, с. 160, [5], размерность $\mathrm{Im}(\mathcal{A})$ равна $\mathrm{rank}(A_{eq})$. □

Таким образом, ранг матрицы оператора инвариантен по отношению к выбору базисов, выбираемых при ее построении, и можно было бы дать эквивалентное определение ранга оператора как ранга его матрицы.

2. Матрицу $A(m, n)$ можно трактовать и как систему строк из пространства $\mathbb{C}^n$. Ранг этой системы строк обозначим через r_s.

Справедлива (см., например, [5]) следующая, на первый взгляд, неожиданная

Теорема 2. *Для любой матрицы* $A(m, n)$ *выполнено равенство* $r_s = \mathrm{rank}(A(m, n))$.

16. Системы линейных алгебраических уравнений. Условия разрешимости

1. Пусть $\mathcal{A}$ линейный оператор, действующий из конечномерного линейного пространства $\mathbf{X}_n$ в конечномерное линейное пространство $\mathbf{Y}_m$. Рассматривается уравнение

$$\mathcal{A}x = y, \tag{16.1}$$

где y — заданный элемент пространства $\mathbf{Y}_m$. При фактическом построении решений уравнения (16.1) вводят некоторые базисы $\mathcal{E}_n = \{e^k\}_{k=1}^n$, $\mathcal{Q}_m = \{q^k\}_{k=1}^m$ в пространствах $\mathbf{X}_n$, $\mathbf{Y}_m$ и переходят к системе линейных алгебраических уравнений относительно коэффициентов ξ разложения вектора x по базису $\mathcal{E}_n$, считая известными коэффициенты η разложения вектора y по базису $\mathcal{Q}_m$. В результате (см. п. 2, с. 162, [5]), получают

$$A_{eq}\xi = \eta, \tag{16.2}$$

где A_{eq} — матрица оператора $\mathcal{A}$.

Более подробная запись уравнения (16.2) дает

$$\sum_{j=1}^{n} a_{ij}^{(eq)}\xi_j = \eta_i, \quad i = 1, 2, \ldots, m. \tag{16.3}$$

Подчеркнем, что коэффициенты $a_{ij}^{(eq)}$ этой системы уравнений (элементы матрицы оператора $\mathcal{A}$) и столбец правой части $\eta_1, \eta_2, \ldots, \eta_m$ предполагаются известными, а числа $\xi_1, \xi_2, \ldots, \xi_n$ требуется найти.

В отличие от рассматривавшихся ранее систем линейных алгебраических уравнений у системы уравнений (16.3) количество уравнений и число неизвестных, вообще говоря, различны.

Задачи (16.1), (16.2) эквивалентны в том смысле, что если ξ — решение уравнения (16.2), то $x = \mathcal{E}_n\xi$ — решение уравнения (16.1) при $y = \mathcal{Q}_m\eta$, и наоборот, если x — решение уравнения (16.1), то коэффициенты разложения векторов x, y по соответствующим базисам связаны соотношением (16.2).

2. Получим необходимые и достаточные условия разрешимости системы линейных алгебраических уравнений

$$Ax = b, \tag{16.4}$$

где $A = A(m, n)$ — заданная прямоугольная матрица с комплексными, вообще говоря, элементами, b — заданный вектор из $\mathbb{C}^m$.

Обозначим через (A, b) матрицу размера $m \times (n+1)$, получающуюся присоединением к матрице A столбца b. Матрицу (A, b) принято называть *расширенной матрицей* системы (16.4).

Теорема 1 (Теорема Кронекера — Капелли[1]). *Для того, чтобы система уравнений* (16.4) *имела решение, необходимо и достаточно, чтобы ранги матриц A и (A, b) совпадали.*

Доказательство. Добавление столбца не уменьшает ранга матрицы, и, очевидно, что ранг сохраняется тогда и только тогда, когда b есть линейная комбинация столбцов матрицы A. Последнее эквивалентно тому, что существует вектор $x \in \mathbb{C}^n$, являющийся решением системы (16.4). □

Теорема 2 (матричная теорема Фредгольма[2]). *Для того, чтобы система линейных уравнений* (16.4) *имела решение, необходимо и достаточно, чтобы для любого решения однородной системы уравнений $zA = 0$ выполнялось равенство $zb = 0$.*

Поясним, что здесь b интерпретируется как вектор-столбец, а z — как вектор-строка.

Доказательство. Д о с т а т о ч н о с т ь. Пусть $r = \operatorname{rank}(A)$. Не ограничивая общности рассуждений, можно считать, что первые r строк матрицы A линейно независимы, Понятно, что тогда и первые r строк матрицы (A, b) линейно независимы. Если k-я строка матрицы A линейно выражается через ее первые r строк, то существует ненулевой вектор z такой, что $zA = 0$. Тогда по условию теоремы $zb = 0$, но это означает, что k-я строка матрицы (A, b) линейно выражается через ее первые r строк. Таким образом, ранги матриц A и (A, b) совпадают, и по теореме Кронекера — Капелли система (16.4) имеет решение. Н е о б х о д и м о с т ь. Пусть система уравнений (16.4) имеет решение, т. е. существует вектор $x \in \mathbb{C}^n$ такой, что $Ax = b$. Тогда для любого $z \in \mathbb{C}^m$ справедливо равенство $zAx = zb$. Очевидно, что если $zA = 0$, то $zb = 0$. □

17. Линейные уравнения в евклидовом пространстве

Теорема 1. *Пусть $\mathbf{X}_n, \mathbf{Y}_m$ — евклидовы пространства. Для любого линейного оператора $\mathcal{A} : \mathbf{X}_n \to \mathbf{Y}_m$ пространство $\mathbf{Y}_m$ допускает следующее ортогональное разложение:*

$$\mathbf{Y}_m = \operatorname{Ker}(\mathcal{A}^*) \oplus \operatorname{Im}(\mathcal{A}). \tag{17.1}$$

Здесь $\mathcal{A}^* : \mathbf{Y}_m \to \mathbf{X}_n$ — оператор сопряженный к $\mathcal{A}$.

[1] Альфредо Капелли (Alfredo Capelli; 1858 — 1916) — итальянский математик.
[2] Эрик Ивар Фредгольм (Erik Ivar Fredholm; 1866 — 1927) — шведский математик.

Доказательство. Пусть $y \in \mathrm{Im}(\mathcal{A})$, $y^1 \in \mathrm{Ker}(\mathcal{A}^*)$. Тогда существует $x \in \mathbf{X}_n$ такой, что $y = \mathcal{A}x$, следовательно,

$$(y, y^1) = (\mathcal{A}x, y^1) = (x, \mathcal{A}^* y^1) = 0,$$

т. е. y ортогонален $\mathrm{Ker}(\mathcal{A}^*)$. Если же вектор $y \in \mathbf{Y}_m$ ортогонален $\mathrm{Im}(\mathcal{A})$, то $(y, \mathcal{A}x) = 0$ для любого $x \in \mathbf{X}_n$, и тогда $(\mathcal{A}^* y, x) = 0$ для любого $x \in \mathbf{X}_n$, поэтому $\mathcal{A}^* y = 0$, т. е. $y \in \mathrm{Ker}(\mathcal{A}^*)$. Эти рассуждения показывают, что $\mathrm{Im}(\mathcal{A})$ — ортогональное дополнение $\mathrm{Ker}(\mathcal{A}^*)$, следовательно, по теореме 2, с. 153, [5], равенство (17.1) выполнено. □

Очевидно, что имеет место и следующее представление:

$$\mathbf{X}_n = \mathrm{Ker}(\mathcal{A}) \oplus \mathrm{Im}(\mathcal{A}^*). \tag{17.2}$$

Теорема 2. *Пусть оператор $\mathcal{A}$ действует из конечномерного евклидова пространства $\mathbf{X}_n$ в конечномерное евклидово пространство $\mathbf{Y}_m$. Тогда*

$$\mathrm{rank}(\mathcal{A}) = \mathrm{rank}(\mathcal{A}^*). \tag{17.3}$$

Доказательство. Оператор $\mathcal{A}$ осуществляет изоморфизм пространств $\mathrm{Im}(\mathcal{A}^*)$ и $\mathrm{Im}(\mathcal{A})$. Действительно, вследствие (17.2) для любого $x \in \mathbf{X}_n$ имеем $\mathcal{A}x = \mathcal{A}x^1$, где $x^1 \in \mathrm{Im}(\mathcal{A}^*)$, т. е. любой элемент $\mathrm{Im}(\mathcal{A})$ — образ некоторого элемента из $\mathrm{Im}(\mathcal{A}^*)$. Предполагая, что $\mathcal{A}x' = \mathcal{A}x''$ для несовпадающих x', x'' из $\mathrm{Im}(\mathcal{A}^*)$, получим, что $\mathcal{A}(x' - x'') = 0$, следовательно, $(x' - x'') \in \mathrm{Ker}(\mathcal{A})$. Поскольку $\mathrm{Im}(\mathcal{A}^*)$ — линейное подпространство, то $(x' - x'') \in \mathrm{Im}(\mathcal{A}^*)$. Вновь используя (17.2), получаем, что $x' - x'' = 0$. Таким образом, конечномерные пространства $\mathrm{Im}(\mathcal{A})$ и $\mathrm{Im}(\mathcal{A}^*)$ изоморфны, поэтому (см. теорему 3, с. 159, [5]) их размерности совпадают. □

Непосредственным следствием теоремы 1 является

Теорема 3 (Теорема Фредгольма). *Пусть $\mathbf{X}_n$, $\mathbf{Y}_m$ — евклидовы пространства, $\mathcal{A} : \mathbf{X}_n \to \mathbf{Y}_m$ — линейный оператор. Для того, чтобы уравнение*

$$\mathcal{A}x = y \tag{17.4}$$

имело решение, необходимо и достаточно, чтобы его правая часть была ортогональна любому решению однородного уравнения $\mathcal{A}^* z = 0$.

Упражнения.

17.1. Опираясь на теорему 2, докажите что $\mathrm{rank}(A) = \mathrm{rank}(A^T)$ для любой матрицы A.

17.2. Опираясь на представление (17.2), покажите, что если уравнение (17.4) разрешимо, то множество всех его решений содержит единственный элемент x_0 наименьшей длины. Этот элемент называется *нормальным решением* уравнения (17.4). Покажите, что $x_0 \in \mathrm{Im}(A^*)$.

18. Псевдорешение. Метод регуляризации Тихонова

1. Пусть оператор $\mathcal{A}$ действует из евклидова пространства $\mathbf{X}_n$ в евклидово пространство $\mathbf{Y}_m$, y — фиксированный вектор из $\mathbf{Y}_m$, x — произвольный вектор из $\mathbf{X}_n$. Вектор $\mathcal{A}x - y$ называется *невязкой*, соответствующей уравнению

$$\mathcal{A}x = y. \tag{18.1}$$

Вещественная функция

$$F(x) = |\mathcal{A}x - y|^2,$$

определенная на пространстве $\mathbf{X}_n$, называется *функцией* (*функционалом*) *невязки*. Если $\mathcal{A}x \neq y$, т. е. вектор x не является решением уравнения (18.1), то $F(x) > 0$. Естественно попытаться найти вектор x, который доставляет минимальное значение функции невязки.

Вектор $x \in \mathbf{X}_n$, минимизирующий функцию невязки, называют *псевдорешением* уравнения (18.1). Если уравнение (18.1) разрешимо, то любое его решение является псевдорешением.

2. Псевдорешение существует при любой правой части уравнения (18.1). В самом деле, в соответствии с разложением (17.1), с. 49, представим вектор y в виде $y = y^1 + y^0$, где $y^1 \in \mathrm{Im}(\mathcal{A})$, $y^0 \in \mathrm{Ker}(\mathcal{A}^*)$. Тогда для любого $x \in \mathbf{X}_n$ вектор $\mathcal{A}x - y^1$ принадлежит $\mathrm{Im}(\mathcal{A})$, и, следовательно,

$$F(x) = |\mathcal{A}x - y^1|^2 + |y^0|^2.$$

Очевидно, что минимальное значение функции F равно $|y^0|^2$ и достигается на векторе x, являющемся решением уравнения

$$\mathcal{A}x = y^1. \tag{18.2}$$

Поскольку $y^1 \in \mathrm{Im}(\mathcal{A})$, уравнение (18.2) разрешимо. Нормальное решение x^0 уравнения (18.2) называют *нормальным псевдорешением* уравнения (18.1).

3. Нетрудно убедиться, что $\mathcal{A}x_0 = \mathcal{P}y$, где $\mathcal{P}$ — оператор ортогонального проектирования $\mathbf{Y}_m$ на $\mathrm{Im}(\mathcal{A})$. В ходе локазательства теоремы 2 было показано, что оператор $\mathcal{A}$ осуществляет изоморфизм между $\mathrm{Im}(\mathcal{A}^*)$ и $\mathrm{Im}(\mathcal{A})$, поэтому существует линейный оператор $\mathcal{A}^+ : \mathbf{Y}_m \to \mathbf{X}_n$ такой, что для любого $y \in \mathbf{Y}_m$ справедливо равенство $x_0 = \mathcal{A}^+ y$, где x_0 — нормальное псевдорешение уравнения $\mathcal{A}x = y$. Оператор $\mathcal{A}^+$ называется псевдообратным по отношению к оператору $\mathcal{A}$. Нетрудно проверить, что если оператор $\mathcal{A}$ обратим, то $\mathcal{A}^+ = \mathcal{A}^{-1}$.

4. При любом $y \in \mathbf{Y}_m$ уравнение

$$\mathcal{A}^*\mathcal{A}x = \mathcal{A}^*y \tag{18.3}$$

разрешимо. Всякое его решение — псевдорешение уравнения (18.1). Действительно, так как $\mathcal{A}^*y^0 = 0$, то уравнение (18.3) эквивалентно уравнению

$$\mathcal{A}^*(\mathcal{A}x - y^1) = 0. \tag{18.4}$$

Уравнение (18.4) разрешимо, так как каждое решение уравнения (18.2) есть решение уравнения (18.4). Обратно, если x — решение уравнения (18.4), то вектор $\mathcal{A}x - y^1 \in \mathrm{Ker}(\mathcal{A}^*)$ и, следовательно (см. (17.1), с. 49), ортогонален $\mathrm{Im}(\mathcal{A})$, но, с другой стороны, $\mathcal{A}x - y^1 \in \mathrm{Im}(\mathcal{A})$, значит $\mathcal{A}x - y^1 = 0$, т. е. x — решение уравнения (18.2).

Уравнение (18.3) называется *трансформацией Гаусса* уравнения (18.1). Трансформация Гаусса любого линейного уравнения приводит к разрешимому уравнению.

5. При фактическом построении нормального псевдорешения можно использовать *метод регуляризации Тихонова*. Рассмотрим наряду с функционалом невязки так называемый *регуляризующий функционал* (*функционал Тихонова*)

$$F_\alpha(x) = F(x) + \alpha|x|^2 = |\mathcal{A}x - y|^2 + \alpha|x|^2. \tag{18.5}$$

Здесь α — положительное число, называемое *параметром регуляризации*.

Теорема 1. *При любом положительном α существует единственный вектор x_α, доставляющий минимальное значение функционалу F_α на пространстве $\mathbf{X}_n$, предел x_α при $\alpha \to 0$ существует и равен x^0.*

Доказательство. Введем в рассмотрение уравнение

$$\mathcal{A}^*\mathcal{A}x + \alpha x = \mathcal{A}^*y. \tag{18.6}$$

Это уравнение имеет единственное решение x_α при любом $y \in \mathbf{X}_n$. В самом деле, если x — решение соответствующего однородного уравнения, то умножая обе части этого уравнения скалярно на x, получим $|\mathcal{A}x|^2 + \alpha|x|^2 = 0$, откуда вследствие положительности α, получаем, что x=0. Выполняя теперь элементарные выкладки, с учетом равенства $\mathcal{A}^*y = \mathcal{A}^*\mathcal{A}x_\alpha + \alpha x_\alpha$ получим

$$F_\alpha(x) = (\mathcal{B}_\alpha(x - x_\alpha), x - x_\alpha) + (y, y) - (\mathcal{B}_\alpha x_\alpha, x_\alpha),$$

где $\mathcal{B}_\alpha = \mathcal{A}^*\mathcal{A} + \alpha I$. Поскольку $(\mathcal{B}_\alpha(x - x_\alpha), x - x_\alpha) > 0$ при любом $x \neq x_\alpha$, то x_α — единственная точка минимума функционала F_α. Таким образом,

$$F(x_\alpha) = |\mathcal{A}x_\alpha - y^1|^2 + |y^0|^2 + \alpha|x_\alpha|^2 \leqslant |\mathcal{A}x - y^1|^2 + |y^0|^2 + \alpha|x|^2 \quad \forall x \in \mathbf{X}_n.$$

Полагая здесь $x = x^0$, получим

$$|\mathcal{A}x_\alpha - y^1|^2 + \alpha|x_\alpha|^2 \leqslant \alpha|x^0|^2, \tag{18.7}$$

поэтому $|x_\alpha| \leqslant |x^0|$ и по теореме Больцано — Вейерштрасса можно указать такую последовательность $\alpha_k \to 0$ и такой вектор $x^* \in \mathbf{X}_n$ что $x_{\alpha_k} \to x^*$ при $\alpha_k \to 0$. Из (18.7) вытекает, что $Ax^* = y^1$, причем $|x^*| \leqslant |x^0|$. Поскольку нормальное псевдорешение единственно, то $x^* = x^0$. Вновь используя единственность нормального псевдорешения, получаем, что $x_\alpha \to x^0$ при любом способе стремления α к нулю. □

19. Сингулярное разложение оператора

1. Сингулярные базисы и сингулярные числа оператора. В этом параграфе будет показано, что для любого оператора $\mathcal{A}$, действующего из евклидова пространства $\mathbf{X}_n$ в евклидово пространство $\mathbf{Y}_m$, можно указать такие ортонормированные базисы $\{e^k\}_{k=1}^n \subset \mathbf{X}_n$ и $\{q^k\}_{k=1}^m \subset \mathbf{Y}_m$, что

$$\mathcal{A}e^k = \begin{cases} \rho_k q^k, & k \leqslant r, \\ 0 \quad , & k > r, \end{cases} \tag{19.1}$$

где $\rho_k > 0$, $k = 1, 2, \ldots, r$. Числа ρ_k называют *сингулярными числами* оператора $\mathcal{A}$[1)]. Базисы $\{e^k\}_{k=1}^n$, $\{q^k\}_{k=1}^m$, обеспечивающие выполнение соотношений (19.1), называются *сингулярными базисами* оператора $\mathcal{A}$.

[1)] Иногда в это множество удобно включать также $\min(m, n) - r$ нулей.

Как показывают соотношения (19.1), ненулевыми элементами матрицы A_{eq} оператора $\mathcal{A}$ относительно сингулярных базисов являются только числа $\rho_1, \rho_2, \ldots, \rho_r$, расположенные на диагонали главного (базисного) минора матрицы A_{eq}.

2. Построим сингулярные базисы оператора $\mathcal{A}$. Оператор $\mathcal{A}^*\mathcal{A}$ самосопряжен и неотрицателен (см. упражнение 2 на с. 223, [5]), следовательно (см. теорему 9, с. 226, и п. 3 с. 231, [5]), существуют ортонормированные собственные векторы $\{e^k\}_{k=1}^n$ оператора $\mathcal{A}^*\mathcal{A}$, все его собственные числа неотрицательны. Таким образом,

$$\mathcal{A}^*\mathcal{A}e^k = \rho_k^2 e^k, \quad k = 1, 2, \ldots, n. \tag{19.2}$$

Здесь $\rho_k^2 \geqslant 0$ — собственные числа оператора $\mathcal{A}^*\mathcal{A}$. Будем нумеровать их так, чтобы $\rho_1 \geqslant \rho_2 \geqslant \cdots \geqslant \rho_r > 0$, $\rho_{r+1} = \cdots = \rho_n = 0$. Положим $z^k = \mathcal{A}e^k$ для $k = 1, \ldots, r$ и заметим, что

$$(z^p, z^q) = (\mathcal{A}e^p, \mathcal{A}e^q) = (\mathcal{A}^*\mathcal{A}e^p, e^q) = \rho_p^2 (e^p, e^q).$$

Поэтому

$$(z^p, z^q) = \begin{cases} 0, & p \neq q, \\ \rho_p^2, & p = q, \end{cases} \tag{19.3}$$

следовательно, векторы

$$q^k = \rho_k^{-1}\mathcal{A}e^k, \quad k = 1, 2, \ldots, r, \tag{19.4}$$

образуют ортонормированную систему в пространстве $\mathbf{Y}_m$. Если окажется, что $r < m$, дополним ее произвольно векторами q^k, $k = r+1$, $r+2, \ldots, m$, до ортонормированного базиса пространства $\mathbf{Y}_m$. Из определения векторов $\{e^k\}_{k=1}^n$, $\{q^k\}_{k=1}^m$ сразу же вытекает справедливость (19.1).

3. Из (19.1) получаем, что векторы $\{q^k\}_{k=1}^r$ образуют базис в $\operatorname{Im}(\mathcal{A})$, но тогда из теоремы 1, с. 49, вытекает, что векторы $\{q^k\}_{k=r+1}^m$ образуют базис в $\operatorname{Ker}(\mathcal{A}^*)$, следовательно,

$$\mathcal{A}^* q^k = 0 \text{ для } k = r+1, \ldots, m. \tag{19.5}$$

Для $k = 1, 2, \ldots, r$ из (19.4), (19.2) получаем

$$\mathcal{A}^* q^k = \rho_k^{-1}\mathcal{A}^*\mathcal{A}e^k = \rho_k e^k. \tag{19.6}$$

4. Сопоставляя (19.6), (19.4), (19.5), будем иметь, что

$$\mathcal{A}\mathcal{A}^*q^k = \rho_k^2 q^k,\ k = 1, 2, \ldots, r, \quad \mathcal{A}\mathcal{A}^*q^k = 0,\ k = r+1, \ldots, m. \tag{19.7}$$

Из (19.2), (19.7) вытекает, что ненулевые собственные числа операторов $\mathcal{A}^*\mathcal{A}$ и $\mathcal{A}\mathcal{A}^*$ совпадают, т. е. спектры этих операторов могут отличаться лишь кратностью нулевого собственного числа[1)].

5. Из предыдущих рассуждений также следуют равенства

$$\operatorname{rank}(\mathcal{A}) = \operatorname{rank}(\mathcal{A}^*\mathcal{A}) = \operatorname{rank}(\mathcal{A}\mathcal{A}^*),$$

$$\operatorname{def}(\mathcal{A}^*\mathcal{A}) = n - \operatorname{rank}(\mathcal{A}), \quad \operatorname{def}(\mathcal{A}\mathcal{A}^*) = m - \operatorname{rank}(\mathcal{A}).$$

6. Понятно, что ранг r оператора $\mathcal{A}$ равен количеству ненулевых сингулярных чисел оператора $\mathcal{A}$. Это наблюдение открывает реальную возможность вычисления ранга оператора $\mathcal{A}$: нужно решить задачу на собственные значения для самосопряженного неотрицательного оператора $\mathcal{A}^*\mathcal{A}$ и определить количество ненулевых собственных чисел. Именно таким способом обычно пользуются в вычислительной практике. Ясно также, что собственные векторы $\{e^i\}_{i=r+1}^n$ оператора $\mathcal{A}^*\mathcal{A}$ образуют ортонормированный базис ядра оператора $\mathcal{A}$.

7. Если сингулярные числа и сингулярные базисы оператора $\mathcal{A}$ найдены, то построение псевдорешения (см. п. 18, с. 51) уравнения

$$\mathcal{A}x = y \tag{19.8}$$

не вызывает затруднений. В самом деле, как было показано в п. 4, с. 52, любое решение уравнения

$$\mathcal{A}^*\mathcal{A}x = \mathcal{A}^*y \tag{19.9}$$

есть псевдорешение уравнения (19.8). Представляя векторы x и y в виде разложений по сингулярным базисам, $x = \sum_{k=1}^{n} \xi_k e^k$, $y = \sum_{k=1}^{m} \eta_k q^k$, и используя затем соотношения (19.2), (19.5), (19.6), получим как следствие уравнения (19.9), что

$$\sum_{k=1}^{r} (\rho_k^2 \xi_k - \rho_k \eta_k) e^k = 0, \tag{19.10}$$

[1)] См. соответствующие определения в [5].

откуда вытекает, что $\xi_k = \eta_k/\rho_k$ для $k = 1, 2, \ldots, r$. Таким образом, любой вектор

$$x = \sum_{k=1}^{r} (\eta_k/\rho_k) e^k + \sum_{k=r+1}^{n} \xi_k e^k, \tag{19.11}$$

где $\xi_{r+1}, \ldots, \xi_n$ — произвольные числа, есть псевдорешение уравнения (19.8).

Если $y \in \mathrm{Im}(\mathcal{A})$, т. е. уравнение (19.8) разрешимо, то формула (19.11) дает общее решение (см. § 1, гл. 10, [5]) уравнения (19.8). Действительно, в этом случае вектор $x^0 = \sum\limits_{k=1}^{r} (\eta_k/\rho_k) e^k$ есть частное решение уравнения (19.8), а $\sum\limits_{k=r+1}^{n} \xi_k e^k$ — общее решение соответствующего однородного уравнения.

8. Для любого псевдорешения x уравнения (19.8) имеем

$$|x|^2 = \sum_{k=1}^{r} (\eta_k/\rho_k)^2 + \sum_{k=r+1}^{n} \xi_k^2.$$

Полагая $\xi_{r+1}, \ldots, \xi_n = 0$, получим псевдорешение с минимальной длиной, т. е. нормальное псевдорешение. Оно ортогонально ядру оператора $\mathcal{A}$.

Упражнения.

19.1. Покажите, что модуль определителя любого оператора, действующего в конечномерном пространстве, равен произведению всех сингулярных чисел этого оператора.

19.2. Пусть $A \in M_{m,n}$ — матрица ранга r. Покажите, что существуют унитарные матрицы $U \in M_m$, $V \in M_n$ такие, что

$$A = UDV, \tag{19.12}$$

где

$$D = \begin{pmatrix} R & O_{1,2} \\ O_{2,1} & O_{2,2} \end{pmatrix} \in M_{m,n}$$

есть блочная 2×2 матрица, $R = \mathrm{diag}(\rho_1, \rho_2, \ldots, \rho_r)$, все элементы диагонали R положительны, все элементы матриц $O_{1,2}$, $O_{2,1}$, $O_{2,2}$ — нули.

Формула (19.12) определяет так называемое *сингулярное* разложение прямоугольной матрицы. Числа $\rho_1, \rho_2, \ldots, \rho_r$ — сингулярные числа матрицы A.

19.3. Покажите, что сингулярные числа матриц A и UAV, где U, V — произвольные унитарные матрицы соответствующих размеров, совпадают (говорят поэтому, что сингулярные числа матрицы инвариантны по отношению к унитарным преобразованиям).

19.4. Пусть $A \in M_{m,n}$ — произвольная матрица, $\rho_1, \rho_2, \ldots, \rho_r$ — ее сингулярные числа. Докажите, что

$$\max_{1 \leqslant k \leqslant r} \rho_k \leqslant \left(\sum_{i,j=1}^{m,n} |a_{ij}|^2 \right)^{1/2}. \tag{19.13}$$

9. Сингулярные числа оператора характеризуют чувствительность решения линейного уравнения по отношению к изменению его правой части. Пусть $\mathcal{A}$ — невырожденный оператор, действующий в евклидовом пространстве $\mathbf{X}_n$. Рассмотрим наряду с уравнением

$$\mathcal{A}x = y \tag{19.14}$$

уравнение

$$\mathcal{A}x = \tilde{y}. \tag{19.15}$$

Поскольку оператор $\mathcal{A}$ невырожден, оба уравнения однозначно разрешимы. Пусть x — решение уравнения (19.14), $\tilde{x}$ — решение уравнения (19.15). Величину $\delta_x = |x - \tilde{x}|/|x|$ называют величиной *относительного изменения решения* при изменении правой части. Выясним, как она зависит от $\delta_y = |y - \tilde{y}|/|y|$ — величины *относительного изменения правой части.* Представим векторы y, $\tilde{y}$ в виде разложений по сингулярному базису: $y = \sum\limits_{k=1}^{n} \eta_k q^k$, $\tilde{y} = \sum\limits_{k=1}^{n} \tilde{\eta}_k q^k$. Тогда вследствие (19.1) получим $x = \mathcal{A}^{-1}y = \sum\limits_{k=1}^{n} \dfrac{\eta_k}{\rho_k} e^k$, $\tilde{x} = \mathcal{A}^{-1}\tilde{y} = \sum\limits_{k=1}^{n} \dfrac{\tilde{\eta}_k}{\rho_k} e^k$, поэтому, используя неравенства $\rho_1 \geqslant \rho_2 \geqslant \cdots \geqslant \rho_n > 0$, будем иметь, что

$$\delta_x^2 = \frac{\sum\limits_{k=1}^{n} \dfrac{|\eta_k - \tilde{\eta}_k|^2}{\rho_k^2}}{\sum\limits_{k=1}^{n} \dfrac{|\eta_k|^2}{\rho_k^2}} \leqslant \frac{\rho_1^2}{\rho_n^2} \frac{\sum\limits_{k=1}^{n} |\eta_k - \tilde{\eta}_k|^2}{\sum\limits_{k=1}^{n} |\eta_k|^2} = \frac{\rho_1^2}{\rho_n^2}\delta_y^2. \tag{19.16}$$

Таким образом,

$$\delta_x \leqslant \frac{\rho_1}{\rho_n}\delta_y. \tag{19.17}$$

Величина ρ_1/ρ_n, характеризующая устойчивость решения уравнения (19.14) по отношению к изменению его правой части, называется *числом обусловленности* оператора $\mathcal{A}$ и обозначается через $\operatorname{cond}(\mathcal{A})$. Очевидно, $\operatorname{cond}(\mathcal{A}) \geqslant 1$ для любого оператора $\mathcal{A}$.

Упражнения.

19.5. Покажите, что при определенном выборе y и $\tilde{y}$ неравенство (19.17) превращается в равенство, и в этом смысле оценка (19.17) неулучшаема.

19.6. Приведите примеры операторов, для которых число обусловленности равно единице.

Глава 4

Нормы векторов и матриц

20. Основные неравенства

Вещественная функция f вещественной переменной называется *выпуклой* на интервале (a, b), если для любых точек x_1, x_2 из этого интервала и для любого $t \in [0, 1]$ выполнено неравенство

$$f(tx_1 + (1-t)x_2) \leqslant tf(x_1) + (1-t)f(x_2). \tag{20.1}$$

Геометрически это означает, что любая точка графика функции f на отрезке $[x_1, x_2]$ лежит ниже хорды, стягивающей точки $(x_1, f(x_1))$, $(x_2, f(x_2))$, или на этой же хорде.

Теорема 1. *Пусть функция f дифференцируема на интервале (a, b), и ее производная не убывает на интервале (a, b). Тогда f — выпуклая функция на интервале (a, b).*

Доказательство. Достаточно установить, что при любых $x_1, x_2 \in (a, b)$, $x_1 < x_2$, функция φ вещественной переменной t, задаваемая равенством

$$\varphi(t) = f((1-t)x_1 + tx_2) - (1-t)f(x_1) - tf(x_2)$$

неположительна для всех t из отрезка $[0, 1]$. Нетрудно видеть, что $\varphi(0) = 0$, $\varphi(1) = 0$, а $\varphi'(t)$ не убывает на отрезке $[0, 1]$. Пусть t — произвольным образом фиксированная точка из интервала $(0, 1)$. Используя формулу конечных приращений Лагранжа, получим, что $\varphi(t) = \varphi(t) - \varphi(0) = t\varphi'(t_1)$, где t_1 — некоторая точка из интервала $(0, t)$. Аналогично получаем, что $\varphi(t) = (t-1)\varphi'(t_2)$, где t_2 — точка из интервала $(t, 1)$. Отсюда следует, что

$$\varphi(t) = t(t-1)(\varphi'(t_2) - \varphi'(t_1)) \leqslant 0. \ \square$$

При помощи теоремы 1 легко доказывается, что функция $-\ln(x)$ выпукла на интервале $(0, \infty)$. Поэтому для любых положительных чисел a, b и любых $p, q > 1$ и таких, что $1/p + 1/q = 1$,

$$\ln(a^p/p + b^q/q) \geqslant \ln(a^p)/p + \ln(b^q)/q = \ln(ab),$$

следовательно, $ab \leqslant a^p/p + b^q/q$. Очевидно, что последнее неравенство верно и при $ab = 0$. Далее, поскольку $|ab| \leqslant |a||b|$, то

$$|ab| \leqslant |a|^p/p + |b|^q/q \tag{20.2}$$

для любых, вообще говоря, комплексных чисел a, b. Неравенство (20.2) называют *неравенством Юнга*[1)].

Теорема 2 (неравенство Гёльдера). *Пусть* $x, y \in \mathbb{C}^n$, $p > 1$, $1/p + 1/q = 1$. *Тогда*

$$\left|\sum_{k=1}^{n} x_k y_k\right| \leqslant \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p} \left(\sum_{k=1}^{n} |y_k|^q\right)^{1/q}. \tag{20.3}$$

Доказательство. Доказываемое неравенство выполнено, если хотя бы один из векторов x, y равен нулю. Для ненулевых x, y, используя неравенство Юнга, получим при $l = 1, 2, \ldots, n$

$$\frac{|x_l|}{\left(\sum\limits_{k=1}^{n} |x_k|^p\right)^{1/p}} \frac{|y_l|}{\left(\sum\limits_{k=1}^{n} |y_k|^q\right)^{1/q}} \leqslant \frac{|x_l|^p}{p\sum\limits_{k=1}^{n} |x_k|^p} + \frac{|y_l|^q}{q\sum\limits_{k=1}^{n} |y_k|^q}.$$

Суммируя все эти неравенства, будем иметь

$$\sum_{k=1}^{n} |x_k||y_k| \leqslant \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p} \left(\sum_{k=1}^{n} |y_k|^q\right)^{1/q},$$

откуда, очевидно, следует (20.3). □

При $p = 2$ неравенство (20.3) называют *неравенством Коши*.

Теорема 3 (неравенство Минковского). *Пусть* $x, y \in \mathbb{C}^n$, $p > 1$. *Тогда*

$$\left(\sum_{k=1}^{n} |x_k + y_k|^p\right)^{1/p} \leqslant \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p} + \left(\sum_{k=1}^{n} |y_k|^p\right)^{1/p}. \tag{20.4}$$

Доказательство. Будем считать x, y такими, что левая часть неравенства (20.4) положительна, так как в противном случае неравенство (20.4) выполняется очевидным образом. Ясно, что

[1)] Уильям Генри Юнг (William Henry Young; 1863 — 1942) — английский математик.

$$\sum_{k=1}^{n}|x_k+y_k|^p=\sum_{k=1}^{n}|x_k+y_k|^{p-1}|x_k+y_k|\leqslant$$
$$\leqslant\sum_{k=1}^{n}|x_k+y_k|^{p-1}|x_k|+\sum_{k=1}^{n}|x_k+y_k|^{p-1}|y_k|. \quad (20.5)$$

Оценим суммы в правой части последнего неравенства, используя неравенство Гёльдера:

$$\sum_{k=1}^{n}|x_k+y_k|^{p-1}|x_k|\leqslant\left(\sum_{k=1}^{n}|x_k+y_k|^{(p-1)q}\right)^{1/q}\left(\sum_{k=1}^{n}|x_k|^p\right)^{1/p}, \quad (20.6)$$

$$\sum_{k=1}^{n}|x_k+y_k|^{p-1}|y_k|\leqslant\left(\sum_{k=1}^{n}|x_k+y_k|^{(p-1)q}\right)^{1/q}\left(\sum_{k=1}^{n}|y_k|^p\right)^{1/p}, \quad (20.7)$$

где $1/p+1/q=1$ и, следовательно, $(p-1)q=p$. Поэтому из (20.5)–(20.7) вытекает, что

$$\sum_{k=1}^{n}|x_k+y_k|^p\leqslant$$
$$\leqslant\left(\sum_{k=1}^{n}|x_k+y_k|^p\right)^{1/q}\left(\left(\sum_{k=1}^{n}|x_k|^p\right)^{1/p}+\left(\sum_{k=1}^{n}|y_k|^p\right)^{1/p}\right),$$

откуда, учитывая равенство $1-1/q=1/p$, получим (20.4). □

21. Нормы на пространстве $\mathbb{C}^n$

1. Наряду с введенным выше понятием длины (или модуля) вектора $x\in\mathbb{C}^n$ во многих случаях оказывается удобным использовать более общее понятие, а именно, понятие нормы вектора.

Будем говорить, что на пространстве $\mathbb{C}^n$ введена *норма*, если каждому вектору $x\in\mathbb{C}^n$ однозначно поставлено в соответствие вещественное число $\|x\|$ (читается: норма x). При этом должны быть выполнены следующие условия (*аксиомы нормы*):

1) $\|x\|\geqslant 0$ для любого $x\in\mathbb{C}^n$, равенства $\|x\|=0$ и $x=0$ эквивалентны;

2) $\|\alpha x\|=|\alpha|\|x\|$ для любых $x\in\mathbb{C}^n$, $\alpha\in\mathbb{C}$;

3) $\|x+y\|\leqslant\|x\|+\|y\|$ для любых $x,y\in\mathbb{C}^n$.

Условие 3) обычно называют *неравенством треугольника.*

Отметим неравенство

4) $\big|\|x\| - \|y\|\big| \leqslant \|x - y\| \quad \forall\, x, y \in \mathbb{C}^n,$

которое вытекает из аксиомы 3). В самом деле,

$$\|x\| = \|x - y + y\| \leqslant \|x - y\| + \|y\|.$$

Аналогично,

$$\|y\| \leqslant \|x - y\| + \|x\|.$$

Неравенство 4) есть просто более краткая запись этих неравенств.

2. Приведем примеры норм на пространстве $\mathbb{C}^n$.

1) Пусть $p \geqslant 1$. Равенство $\|x\|_p = \Big(\sum\limits_{k=1}^{n} |x_k|^p\Big)^{1/p}$ определяет норму. Действительно, аксиомы 1), 2) выполнены очевидным образом, а неравенство 3) при $p = 1$ непосредственно вытекает из свойств модуля, а при $p > 1$ совпадает с неравенством Минковского (20.4).Отметим, что $\|x\|_2^2 = |x|^2 = (x, x)$, для любого $x \in \mathbb{C}^n$, здесь и далее в этой главе $(\cdot, \cdot)$ — стандартное скалярное произведение на пространстве $\mathbb{C}^n$.

2) Положим $\|x\|_\infty = \max\limits_{1 \leqslant k \leqslant n} |x_k|$. Элементарно проверяется, что это равенство определяет норму.

3) Пусть A — эрмитова положительно определенная матрица. Функция $\|x\|_A = (Ax, x)^{1/2}$ есть норма на пространстве $\mathbb{C}^n$. Для обоснования этого факта достаточно вспомнить, что соотношение $(x, y)_A = (Ax, y)$ определяет скалярное произведение на пространстве $\mathbb{C}^n$ (см. упражнение 1, с. 223, а также п. 2, с. 128, [5]).

3. Любая норма непрерывна на всем пространства $\mathbb{C}^n$. В самом деле, пусть x, y — произвольные точки $\mathbb{C}^n$. Представим их в виде разложений по естественному базису пространства $\mathbb{C}^n$: $x = \sum\limits_{k=1}^{n} x_k i^k$, $y = \sum\limits_{k=1}^{n} y_k i^k$. Используя теперь неравенство треугольника, получим $\|x - y\| \leqslant \sum\limits_{k=1}^{n} \|i^k\| |x_k - y_k|$, откуда, очевидно, вытекает, что если x стремится к y, то $\|x - y\|$ стремится к нулю.

4. Будем говорить, что последовательность $\{x^k\} \subset \mathbb{C}^n$ *сходится* к вектору $x \in \mathbb{C}^n$ *по норме*, если $\lim\limits_{k\to\infty} \|x - x^k\| = 0$. В п. 3, фактически, показано, что если последовательность векторов сходится *покомпонентно*, то она сходится и по любой норме, введенной на пространстве $\mathbb{C}^n$. Ниже будет установлено, что справедливо и обратное утверждение.

5. Говорят, что нормы $\|\cdot\|_{(1)}$ и $\|\cdot\|_{(2)}$ *эквивалентны* если существуют положительные постоянные c_1 и c_2 такие, что

$$c_1\|x\|_{(1)} \leqslant \|x\|_{(2)} \leqslant c_2\|x\|_{(1)} \quad \forall\, x \in \mathbb{C}^n. \tag{21.1}$$

Теорема 1. *Любые две нормы на пространстве $\mathbb{C}^n$ эквивалентны.*

Доказательство. Отношение эквивалентности норм, очевидно, транзитивно. Поэтому достаточно показать, что любая норма $\|\cdot\|$ эквивалентна норме $\|\cdot\|_2 = |\cdot|$, т. е. показать, что существуют положительные постоянные c_1, c_2 такие, что

$$c_1|x| \leqslant \|x\| \leqslant c_2|x| \quad \forall\, x \in \mathbb{C}^n. \tag{21.2}$$

Пусть $S_1(0)$ — множество всех векторов из пространства $\mathbb{C}^n$, удовлетворяющих условию $|x| = 1$ ($S_1(0)$ — сфера единичного радиуса с центром в нуле). Это множество ограничено и замкнуто в пространстве $\mathbb{C}^n$. Функция $\varphi(x_1, x_2 \ldots, x_n) = \|x\|$, как показано в п. 3, непрерывна на $\mathbb{C}^n$. Поэтому по теореме Вейерштрасса (см. курс математического анализа) существуют точки x^1, x^2, принадлежащие $S_1(0)$, и такие, что $\|x^1\| = \min\limits_{x\in S_1(0)} \|x\|$, $\|x^2\| = \max\limits_{x\in S_1(0)} \|x\|$. Положим $c_1 = \|x^1\|$, $c_2 = \|x^2\|$. Ясно, что $0 \leqslant c_1 \leqslant c_2$. Причем c_1 не может равняться нулю, так как в противном случае $x^1 = 0$, но, с другой стороны, $x^1 \in S_1(0)$, поэтому $|x^1| = 1$, и, стало быть, $x^1 \neq 0$. Итак, для любого $x \in S_1(0)$ выполнены неравенства $0 < c_1 \leqslant \|x\| \leqslant c_2$. Пусть теперь x — произвольный вектор из $\mathbb{C}^n$, не равный нулю. Тогда, очевидно, вектор $(1/|x|)x$ принадлежит $S_1(0)$, следовательно, $c_1 \leqslant \|(1/|x|)x\| \leqslant c_2$, откуда вытекает, что для вектора x выполнены неравенства (21.2). Если $x = 0$, то неравенства (21.2) выполняются очевидным образом. □

6. Из теоремы 1 вытекает, что всякая норма на пространстве $\mathbb{C}^n$ эквивалентна норме $\|\cdot\|_2$, поэтому из сходимости последовательности векторов по любой норме вытекает ее покомпонентная сходимость. Важно иметь в виду, что постоянные c_1, c_2, вообще говоря, зависят

от n, т. е. от размерности пространства $\mathbb{C}^n$. Приведем, например, следующие оценки:

$$\|x\|_\infty \leqslant \|x\|_p \quad \forall\, x \in \mathbb{C}^n \text{ при любом } p \geqslant 1; \tag{21.3}$$

$$\|x\|_p \leqslant \|x\|_q \quad \forall\, x \in \mathbb{C}^n, \text{ если } p \geqslant q \geqslant 1; \tag{21.4}$$

$$\|x\|_p \leqslant n^{1/p-1/q}\|x\|_q \quad \forall\, x \in \mathbb{C}^n, \text{если } q > p \geqslant 1; \tag{21.5}$$

$$\|x\|_p \leqslant n^{1/p}\|x\|_\infty \quad \forall x \in \mathbb{C}^n \text{ при любом } p \geqslant 1. \tag{21.6}$$

Прежде чем доказывать эти неравенства заметим, что они являются точными, т. е. для каждого из них можно указать такой ненулевой вектор x, на котором неравенство превращается в равенство. Именно, первые два неравенства обращаются в равенства, например, при $x = (1, 0, \ldots, 0)$, а последние два — при $x = (1, 1, \ldots, 1)$.

Приведем теперь соответствующие доказательства.

1) Пусть $\|x\|_\infty \equiv \max\limits_{1\leqslant k\leqslant n} |x_k| = |x_i|$. Очевидно, что

$$|x_i| = (|x_i|^p)^{1/p} \leqslant \left(\sum_{k=1}^n |x_k|^p\right)^{1/p} = \|x\|_p.$$

2) Выполнив очевидные выкладки, получим

$$\|x\|_p = \left(\sum_{k=1}^n |x_k|^q |x_k|^{p-q}\right)^{1/p} \leqslant \|x\|_\infty^{(p-q)/p}\|x\|_q^{q/p},$$

откуда, используя (21.3), приходим к (21.4).

3) Представим $|x_k|^p$ в виде $|x_k|^p \cdot 1$ и используем для оценки $\|x\|_p$ неравенство Гёльдера с показателями $t = q/p > 1$ и $r = t/(t-1) = q/(q-p)$. Получим, что $\|x\|_p =$

$$= \left(\sum_{k=1}^n |x_k|^p\right)^{1/p} \leqslant \left(\sum_{k=1}^n |x_k|^q\right)^{1/q} \left(\sum_{k=1}^n 1\right)^{(q-p)/(pq)} = n^{1/p-1/q}\|x\|_q.$$

Доказательство неравенства (21.6) читатель легко выполнит самостоятельно.

Упражнение 21.1. Показать, что для любого $x \in \mathbb{C}^n$ выполнено предельное соотношение $\|x\|_\infty = \lim\limits_{p\to\infty} \|x\|_p$.

7. Норма вектора называется *абсолютной*, если она зависит только от модулей компонент вектора. Например, норма $\|\cdot\|_p$ при $p \geqslant 1$ абсолютна, норма на пространстве $\mathbb{C}^2$, определяемая равенством

$$\|x\| = (|x_1|^2 + |x_2|^2 - \operatorname{Re}(x_1 x_2))^{1/2},$$

не абсолютна.

Пусть $D = \operatorname{diag}(d_1, d_2, \ldots, d_n)$, $0 \leqslant d_i \leqslant 1$, $i = 1, 2, \ldots, n$, $x \in \mathbb{C}^n$. Тогда для любой абсолютной нормы $\|Dx\| \leqslant \|x\|$. Очевидно, достаточно убедиться в этом, когда $D = \operatorname{diag}(1, \ldots, 1, d_k, 1, \ldots, 1)$, $d_k \in [0, 1]$. Имеем

$$Dx = \frac{1}{2}(1 - d_k)(x_1, x_2, \ldots, -x_k, \ldots, x_n) + \frac{1}{2}(1 - d_k)x + d_k x,$$

следовательно, $\|Dx\| \leqslant \frac{1}{2}(1 - d_k)\|x\| + \frac{1}{2}(1 - d_k)\|x\| + d_k\|x\| = \|x\|$.

Норма на пространстве $\mathbb{C}^n$ называется *монотонной*, если из неравенств $|x_k| \leqslant |y_k|$, $k = 1, 2, \ldots, n$, следует, что $\|x\| \leqslant \|y\|$. Всякая монотонная норма является абсолютной. Действительно, если норма монотонна, то для любого вектора x выполнены неравенства

$$\|(|x_1|, |x_2|, \ldots, |x_n|)\| \leqslant \|(x_1, x_2, \ldots, x_n)\| \leqslant \|(|x_1|, |x_2|, \ldots, |x_n|)\|.$$

Обратно, всякая абсолютная норма монотонна. В самом деле, если для векторов x, y имеем, что $|x_k| \leqslant |y_k|$, $k = 1, 2, \ldots, n$, то существует матрица $D = \operatorname{diag}(d_1 e^{i\varphi_1}, d_2 e^{i\varphi_2}, \ldots, d_n e^{i\varphi_n})$[1)], $0 \leqslant d_k \leqslant 1$, $k = 1, 2, \ldots, n$, такая, что $x = Dy$. Используя теперь определение абсолютной нормы и неравенство, установленное в п. 7, нетрудно убедиться, что $\|x\| \leqslant \|y\|$.

22. Теорема Хана — Банаха. Дуальные нормы

1. Будем говорить, что на пространстве $\mathbb{C}^n$ задан *вещественный линейный* функционал f, если каждому $x \in \mathbb{C}^n$ поставлено в соответствие однозначно вещественное число $f(x)$ и

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y) \quad \forall\, x, y \in \mathbb{C}^n,\ \alpha, \beta \in \mathbb{R}. \tag{22.1}$$

Будем говорить, что на пространстве $\mathbb{C}^n$ задан *линейный* функционал f, если каждому $x \in \mathbb{C}^n$ поставлено в соответствие однозначно комплексное число $f(x)$ и это соответствие линейно, т. е.

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y) \quad \forall\, x, y \in \mathbb{C}^n,\ \alpha, \beta \in \mathbb{C}. \tag{22.2}$$

[1)] Напомним, что по определению $e^{i\varphi} = \cos\varphi + i\sin\varphi$.

2. Если на пространстве $\mathbb{C}^n$ определена некоторая норма $\|\cdot\|$, то каждому линейному функционалу f (вещественному или комплексному) можно поставить в соответствие его *норму* $\|f\|$, полагая

$$\|f\| = \sup_{x\in\mathbb{C}^n,\ x\neq 0} \frac{|f(x)|}{\|x\|} = \sup_{x\in\mathbb{C}^n,\ \|x\|=1} |f(x)|. \tag{22.3}$$

Для каждого линейного функционала

$$\|f\| < \infty. \tag{22.4}$$

Докажем неравенство (22.4) применительно к вещественному случаю. Для комплексного случая рассуждения аналогичны и несколько проще. Пусть $z = (z_1, z_2, \ldots, z_n) \in \mathbb{C}^n$, $\|z\| = 1$. Будем считать, что $z_k = x_k + iy_k$, $x_k, y_k \in \mathbb{R}$, $k = 1, 2, \ldots, n$. Имеем

$$f(z) = f(\sum_{k=1}^{n}(x_k + iy_k)i_k) = \sum_{k=1}^{n}(x_k f(i_k) + y_k f(ii_k)),$$

следовательно, $|f(z)| \leqslant \max(\max\limits_{1\leqslant k\leqslant n} |f(i_k)|, \max\limits_{1\leqslant k\leqslant n} |f(ii_k)|) \sum\limits_{k=1}^{n} |z_k|$. Поскольку все нормы на пространстве $\mathbb{C}^n$ эквивалентны, отсюда вытекает, что $|f(z)| \leqslant c\|z\| = c$, где c — постоянная, зависящая только от n, а это и означает справедливость (22.4).

Теорема 1 (Хан — Банах). *Пусть L — подпространство пространства $\mathbb{C}^n$, f — линейный функционал, определенный на L,*

$$\|f\| = \sup_{x\in L,\, \|x\|=1} |f(x)|. \tag{22.5}$$

Существует линейный функционал F, определенный на $\mathbb{C}^n$ такой, что $F(x) = f(x)$ для всех $x \in L$ и

$$\|F\| = \sup_{x\in\mathbb{C}^n,\, \|x\|=1} |F(x)| = \|f\|^{1)}. \tag{22.6}$$

Доказательство. Предположим сначала, что f — вещественный линейный функционал. Естественно, мы считаем, что f — не нуль тождественный, поэтому без ограничения общности рассуждений можно положить, что $\|f\| = 1$. Исключим из рассмотрения тривиальный случай, когда $L = \mathbb{C}^n$, и пусть $u \notin L$, а $L_1 \supset L$ — множество

[1)] Говорят, что F есть продолжение функционала f на все пространство $\mathbb{C}^n$ с сохранением нормы.

векторов вида $x + tu$, где $x \in L$, $t \in \mathbb{R}$. Вследствие неравенства треугольника для любых $x, y \in L$ имеем

$$f(x) - f(y) \leqslant \|x - y\| \leqslant \|x + u\| + \|y + u\|,$$

поэтому $f(x) - \|x + u\| \leqslant f(y) + \|y + u\|$, и, значит существует число a такое. что

$$\sup_{x \in L}(f(x) - \|x + u\|) \leqslant a \leqslant \inf_{x \in L}(f(x) + \|x + u\|). \tag{22.7}$$

Определим функционал f_1 на L_1, полагая $f_1(x + tu) = f(x) - at$ (проверьте, что f_1 вещественный линейный функционал!). Из (22.7) следует, что $|f(x) - a| \leqslant \|x + u\|$ $\forall x \in L$, значит,

$$|f_1(x + u)| \leqslant \|x + u\| \quad \forall x \in L.$$

При $t \neq 0$ получаем $f_1(x + tu) = tf_1(t^{-1}x + u)$, поэтому

$$|f_1(x + tu)| = |t||f_1(t^{-1}x + u)| \leqslant |t|\|t^{-1}x + u)\| = \|x + tu\|,$$

или $|f_1(x)| \leqslant \|x\|$ $\forall x \in L_1$. Рассуждая точно так же, построим вещественный линейный функционал f_2, определенный на множестве векторов $L_2 \supset L_1$ вида $x + t(iu)$, где $x \in L_1$, $t \in \mathbb{R}$, такой, что

$$|f_2(x)| \leqslant \|x\| \ \forall x \in L_2.$$

Нетрудно видеть, что множество L_2 совпадает, с подпространством пространства $\mathbb{C}^n$, натянутым на базис подпространства L и вектор u. Таким образом, построено продолжение вещественного линейного функционала f, заданного на L, на более широкое подпространство. Последовательно увеличивая размерность подпространств, мы построим вещественный линейный функционал F, определенный на всем пространстве $\mathbb{C}^n$, такой, что $F(x) = f(x)$ $\forall x \in L$, и

$$|F(x)| \leqslant \|x\| \ \forall\, x \in \mathbb{C}^n.$$

Из последней оценки и определения (22.5) вытекает, что $\|F\| = \|f\|$.

Пусть теперь f — линейный (комплексный) функционал, определенный на L. Представим его в виде $f(x) = g(x) + ih(x)$ $\forall\, x \in L$, где g, h — линейные вещественные функционалы, определенные на L. Вследствие линейности f получаем

$$f(ix) = g(ix) + ih(ix) = if(x) = ig(x) - h(x),$$

откуда $h(x) = -g(ix)$, поэтому $f(x) = g(x) - ig(ix)$. По условию $\|f\| = 1$, следовательно, $\|g\| \leqslant 1$. Используя конструкцию, описанную в предыдущей части доказательства, построим линейный вещественный функционал $G(x)$, определенный на всем пространстве $\mathbb{C}^n$, такой, что $G(x) = g(x)\ \forall\, x \in L$, $|G(x)| \leqslant \|x\|\ \forall\, x \in \mathbb{C}^n$. Пусть далее $F(x) = G(x) - iG(ix)\ \forall\, x \in \mathbb{C}^n$. Ясно, что $F(x) = f(x)\ \forall\, x \in L$. Покажем, что функционал F линеен. Для этого (в дополнение к предыдущему) достаточно установить, что $F(ix) = iF(x)\ \forall\, x \in \mathbb{C}^n$, а это непосредственно следует из определения. Действительно,

$$F(ix) = G(ix) + iG(x) = i(G(x) - iG(ix)).$$

Осталось убедится в справедливости равенства (22.6). Фиксируем произвольно $x \in \mathbb{C}^n$. Выберем вещественное число θ так, чтобы $F(x)e^{i\theta}$ было неотрицательно. Тогда

$$|F(x)| = F(e^{i\theta}x) = G(e^{i\theta}x) \leqslant \|e^{i\theta}x\| = \|x\|.$$

Вместе с (22.5) это неравенство гарантирует выполнение (22.6). □

Следствие 1. *Пусть $x_0 \in \mathbb{C}^n$. Существует линейный функционал F, определенный на $\mathbb{C}^n$, такой, что $F(x_0) = \|x_0\|$, $\|F\| = 1$.*

Доказательство. Ведем в рассмотрение подпространство L пространства $\mathbb{C}^n$ векторов вида $\alpha x_0, \alpha \in \mathbb{C}$. Определим на этом подпространстве линейный функционал f, полагая $f(\alpha x_0) = \alpha\|x_0\|$. Тогда, очевидно, $f(x_0) = \|x_0\|$, $\|f\| = 1$. Осталось, пользуясь теоремой Хана — Банаха, продолжить функционал f на все пространство $\mathbb{C}^n$ с сохранением нормы. □

3. Пространство $\mathbb{C}^n$ можно рассматривать как евклидово, определив на нем скалярное произведение (например, стандартное). По теореме Рисса (см. с. 213, [5]) всякому линейному функционалу f на $\mathbb{C}^n$ можно поставить в соответствие один и только один вектор $y \in \mathbb{C}^n$ такой что $f(x) = (x, y)\ \forall\, x \in \mathbb{C}^n$, и, наоборот, всякий вектор $y \in \mathbb{C}^n$ порождает линейный функционал: $f(x) = (x, y)\ \forall\, x \in \mathbb{C}^n$. Пусть $\|\cdot\|$ — некоторая норма на пространстве $\mathbb{C}^n$. Для каждого $y \in \mathbb{C}^n$ положим

$$\|y\|_* = \|f\| = \sup_{x \in \mathbb{C}^n,\, x \neq 0} \frac{|(x, y)|}{\|x\|} = \sup_{x \in \mathbb{C}^n,\, \|x\| = 1} |(x, y)|. \tag{22.8}$$

Элементарно проверяется что соотношение (22.8) определяет норму на пространстве $\mathbb{C}^n$. Эта норма называется *дуальной* по отношению к исходной норме. Следующая теорема показывает, что понятие дуальности норм взаимно.

Теорема 2. *Пусть $\|\cdot\|$ — произвольная норма пространстве $\mathbb{C}^n$, $\|\cdot\|_*$ — дуальная по отношению к ней норма. Тогда*

$$\|x\| = \sup_{y\in\mathbb{C}^n,\, \|y\|_*=1} |(x,y)|. \tag{22.9}$$

Доказательство. Непосредственно из определения дуальной нормы вытекает, что для любого не равного нулю $y \in \mathbb{C}^n$ справедливо неравенство $\|x\| \geqslant |(x,y)|/\|y\|_*$, причем в силу следствия 1 можно указать такой вектор y, для которого $\|x\| = |(x,y)|/\|y\|_*$. Эти рассуждения показывают, что равенство (22.9) выполнено. □

В ходе доказательства теоремы 2 мы установили, что справедливо

Следствие 2. *Для любых $x, y \in \mathbb{C}^n$ выполнено неравенство*

$$|(x,y)| \leqslant \|x\|\|y\|_*. \tag{22.10}$$

Неравенство (22.10) называют *обобщенным* неравенством Коши — Буняковского.

Пример. Нормы $\|\cdot\|_p$, $\|\cdot\|_q$ при $p > 1$, $1/p+1/q = 1$ дуальны, если под скалярным произведением на $\mathbb{C}^n$ понимать стандартное скалярное произведение. В самом деле, для любых $x, y \in \mathbb{C}^n$ по неравенству Гёльдера (см. (20.3)) имеем $|(x,y)| \leqslant \|x\|_p\|y\|_q$. Пусть $x_k = \rho_k e^{i\varphi_k}$, $k = 1, 2, \ldots, n$. Положим $y_k = \rho_k^{p-1} e^{i\varphi_k}$, $k = 1, 2, \ldots, n$. Элементарные вычисления показывают, что $|(x,y)| = \|x\|_p\|y\|_q$. Следовательно, $\|x\|_p = \sup\limits_{y\in\mathbb{C}^n,\, y\neq 0} |(x,y)|/\|y\|_q$.

Упражнение 22.1. Докажите, что нормы $\|\cdot\|_1$ и $\|\cdot\|_\infty$ дуальны относительно стандартного скалярного произведения на $\mathbb{C}^n$.

23. Нормы на пространстве матриц

1. Как и ранее, через $M_{m,n}$ будем обозначать множество всех прямоугольных матриц с m строками и n столбцами с комплексными, вообще говоря, элементами. При $m = n$ будем писать M_n. Определив на множестве $M_{m,n}$ обычным образом операции сложения двух матриц и умножения матрицы на число, мы превратим его в комплексное линейное пространство размерности mn. На этом линейном пространстве введем норму, т. е. поставим в соответствие каждой матрице $A \in M_{m,n}$ число $\|A\|$ так, что:

1) $\|A\| \geqslant 0$ для любой матрицы $A \in M_{m,n}$, равенства $\|A\| = 0$ и $A = 0$ эквивалентны;

2) $\|\alpha A\| = |\alpha|\|A\|$ для любой матрицы $A \in M_{m,n}$ и любого $\alpha \in \mathbb{C}$;

3) $\|A + B\| \leqslant \|A\| + \|B\|$ для любых матриц $A, B \in M_{m,n}$.

Говорят в этом случае, что на пространстве матриц $M_{m,n}$ введена *векторная норма*. Понятно, что она обладает всеми свойствами, которые были изучены в предыдущем параграфе применительно к норме векторов.

Часто используют так называемые *согласованные нормы* на пространстве матриц. При этом дополнительно к 1)–3) должна выполняться аксиома

4) $\|AB\|_{mp} \leqslant \|A\|_{mn}\|B\|_{np}$ для любых матриц $A \in M_{mn}, B \in M_{np}$.

Здесь нижними индексами помечены нормы на соответствующих пространствах матриц.

Не всякие векторные нормы на пространстве матриц является согласованными. Пусть, например,

$$\|A\| = \max_{1\leqslant i,j\leqslant n} |a_{ij}| \tag{23.1}$$

для $A \in M_n$. Очевидно, это — векторная норма, но она не является согласованной на M_n. Действительно, если

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \text{ то } AA = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix},$$

причем $\|A\| = 1$, $\|AA\| = 2$, и неравенство $\|AA\| \leqslant \|A\|\|A\|$ не выполнено.

Упражнение 23.1. Пусть $\|\cdot\|$ — согласованная норма на M_n, $S \in M_n$ — произвольная невырожденная матрица. Покажите, что формула $\|A\|_{(s)} = \|SAS^{-1}\|$ $\forall A \in M_n$ определяет согласованная норму на M_n.

2. Приведем важные примеры согласованных матричных норм.

1) Положим $\|A\|_{l_1} = \sum\limits_{i,j=1}^{n} |a_{ij}|$ для $A \in M_n$. Очевидно, три первых аксиомы нормы выполнены. Проверим аксиому 4). По определению для $A, B \in M_n$ имеем

$$\|AB\|_{l_1} = \sum_{i,j=1}^{n} \left| \sum_{k=1}^{n} a_{ik}b_{kj} \right|,$$

следовательно,

$$\|AB\|_{l_1} \leqslant \sum_{i,j,k=1}^{n} |a_{ik}||b_{kj}|.$$

Добавляя к сумме в правой части последнего неравенства неотрицательные слагаемые, усилим неравенство:

$$\|AB\|_{l_1} \leqslant \sum_{i,j,k,m=1}^{n} |a_{ik}||b_{mj}|.$$

Осталось заметить, что

$$\sum_{i,j,k,m=1}^{n} |a_{ik}||b_{mj}| = \sum_{i,k}^{n} |a_{ik}| \sum_{j,m=1}^{n} |b_{mj}| = \|A\|_{l_1}\|B\|_{l_1}.$$

2) Положим $\|A\|_E = \left(\sum\limits_{i,j=1}^{m,n} |a_{ij}|^2 \right)^{1/2}$ для $A \in M_{m,n}$. Эта норма порождается естественным скалярным произведением на пространстве $\mathbb{C}^{mn}$, поэтому три первых аксиомы для нее выполняются. Норму $\|A\|_E$ обычно называют *евклидовой* нормой или нормой *Фробениуса*[1)]. Докажем справедливость четвертой аксиомы для этой нормы, опираясь на неравенство Коши (см. с. 59). Пусть $A \in M_{m,n}$, $B \in M_{n,p}$. Тогда

$$\|AB\|_E^2 = \sum_{i,j=1}^{m,p} \left| \sum_{k=1}^{n} a_{ik}b_{kj} \right|^2 \leqslant \sum_{i,j=1}^{m,p} \sum_{k=1}^{n} |a_{ik}|^2 \sum_{k=1}^{n} |b_{kj}|^2 =$$

$$= \sum_{i,k=1}^{m,n} |a_{ik}|^2 \sum_{k,j=1}^{n,p} |b_{kj}|^2 = \|A\|_E^2 \|B\|_E^2.$$

Упражнение 23.2. Доказать, что норма $\|A\| = n \max\limits_{1 \leqslant i,j \leqslant n} |a_{ij}|$ является согласованной на пространстве M_n.

3. Пусть $A \in M_{m,n}$, $\|\cdot\|_{(m)}$, $\|\cdot\|_{(n)}$ — некоторые нормы на пространствах $\mathbb{C}^m$, $\mathbb{C}^n$, соответственно. Тогда существует неотрицательное число N_A такое, что

$$\|Ax\|_{(m)} \leqslant N_A \|x\|_{(n)} \quad \forall\, x \in \mathbb{C}^n. \tag{23.2}$$

В самом деле, поскольку всякая норма $\|\cdot\|$ на $\mathbb{C}^n$ эквивалентна норме $\|\cdot\|_\infty$, то $c_1\|x\|_\infty \leqslant \|x\|_{(n)}$ $\forall\, x \in \mathbb{C}^n$, $\|x\|_{(m)} \leqslant c_2\|x\|_\infty$ $\forall\, x \in \mathbb{C}^m$, где c_1, c_2 — положительные не зависящие от x постоянные. Поэтому справедлива следующая цепочка неравенств:

[1)] Фердинанд Георг Фробениус (Ferdinand Georg Frobenius; 1849 — 1917) — немецкий математик.

$$\|Ax\|_{(m)} \leqslant c_2\|Ax\|_\infty = c_2 \max_{1\leqslant i\leqslant m}\left|\sum_{j=1}^{n} a_{ij}x_j\right| \leqslant c_2\|x\|_\infty \max_{1\leqslant i\leqslant m}\sum_{j=1}^{n}|a_{ij}| \leqslant$$
$$\leqslant (c_2/c_1)\max_{1\leqslant i\leqslant m}\sum_{j=1}^{n}|a_{ij}|\|x\|_{(n)}.$$

Обозначим через $\nu(A)$ точную нижнюю грань всех чисел N_A, для которых выполнено (23.2). Очевидно, что можно дать и другое, эквивалентное, определение функции ν на пространстве $M_{m,n}$:

$$\nu(A) = \sup_{x\in\mathbb{C}^n,\, x\neq 0}\frac{\|Ax\|_m}{\|x\|_n} = \sup_{x\in\mathbb{C}^n,\, \|x\|_n=1}\|Ax\|_m. \tag{23.3}$$

Понятно, что

$$\|Ax\|_m \leqslant \nu(A)\|x\|_n \quad \forall x \in \mathbb{C}^n.$$

Упражнение 23.3. Докажите, что для функции ν выполнены все аксиомы согласованной матричной нормы.

Матричную норму, сконструированную указанным способом, называют *подчиненной* норме векторов или *операторной* нормой.

Упражнение 23.4. Докажите, что при любом способе определения норм на пространствах $\mathbb{C}^m$, $\mathbb{C}^n$ существует вектор $x^0 \in \mathbb{C}^n$ такой, что $\|x^0\|_n = 1$ и

$$\|Ax^0\|_m = \sup_{x\in\mathbb{C}^n,\, \|x\|_n=1}\|Ax\|_m,$$

т. е. в определении (23.3) символ точной верхней грани можно заменить на символ максимума.

Нетрудно убедиться, что при любом способе задания нормы на $\mathbb{C}^n$ подчиненная норма единичной матрицы (порядка n) равна единице.

Не всякая норма, определенная на M_n, подчинена какой либо норме векторов. Например, норма Фробениуса не подчинена никакой норме векторов, так как $\|I\|_E = \sqrt{n}$. Норма (23.1) также не является операторной, так как она не согласованная норма на M_n

4. Приведем примеры вычисления подчиненных матричных норм.

1) Пусть норма на пространстве $\mathbb{C}^n$ определена, как в п. 2, с. 61, равенством $\|x\|_1 = \sum\limits_{k=1}^{n}|x_k|$. Тогда подчиненная норма матрицы есть

$$\|A\|_1 = \max_{x\in\mathbb{C}^n,\, \|x\|_1=1}\|Ax\|_1.$$

Нетрудно видеть, что для любого вектора $x \in \mathbb{C}^n$, $\|x\|_1 = 1$,

$$\|Ax\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leqslant \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}||x_j| = \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}| \leqslant$$

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

Предположим, что $\max\limits_{1 \leqslant j \leqslant n} \sum\limits_{i=1}^{n} |a_{ij}| = \sum\limits_{i=1}^{n} |a_{ik}|$, и положим, что $\tilde{x}$ есть вектор естественного базиса пространства $\mathbb{C}^n$ такой, что $\tilde{x}_k = 1$, а все остальные координаты вектора $\tilde{x}$ равны нулю. Ясно, что $\|\tilde{x}\|_1 = 1$, а $\|A\tilde{x}\|_1 = \sum\limits_{i=1}^{n} |a_{ik}|$. Таким образом, доказано, что

$$\|A\|_1 = \max_{x \in \mathbb{C}^n,\, \|x\|_1 = 1} \|Ax\|_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} |a_{ij}|.$$

Поэтому норму $\|A\|_1$ часто называют *столбцовой* нормой матрицы A.

2) Определим норму на $\mathbb{C}^n$ равенством $\|x\|_\infty = \max\limits_{1 \leqslant k \leqslant n} |x_k|$. Тогда для любого $x \in \mathbb{C}^n$ такого, что $\|x\|_\infty = 1$

$$\|Ax\|_\infty = \max_{1 \leqslant i \leqslant n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leqslant \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |a_{ij}||x_j| \leqslant$$

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

Положим, что $\max\limits_{1 \leqslant i \leqslant 1} \sum\limits_{j=1}^{n} |a_{ij}| = \sum\limits_{j=1}^{n} |a_{kj}|$ и определим вектор $\tilde{x} \in \mathbb{C}^n$ при помощи соотношений

$$\tilde{x}_j = \begin{cases} \bar{a}_{kj}/|a_{kj}|, & a_{kj} \neq 0, \\ 1, & a_{kj} = 0, \end{cases}$$

где $j = 1, 2, \ldots, n$, черта, как обычно, есть знак комплексного сопряжения. Ясно, что $\|\tilde{x}\|_\infty = 1$, причем элементарные выкладки показывают, что для любого $i = 1, 2, \ldots, n$ выполнено неравенство

$$\left| \sum_{j=1}^{n} a_{ij} \tilde{x}_j \right| \leqslant \sum_{j=1}^{n} |a_{ij}| \leqslant \sum_{j=1}^{n} |a_{kj}|,$$

а для $i = k$

$$\left|\sum_{j=1}^{n} a_{ij}\tilde{x}_j\right| = \sum_{j=1}^{n} |a_{kj}|,$$

т. е. $\|A\tilde{x}\|_\infty = \max\limits_{1\leqslant i\leqslant 1} \sum\limits_{j=1}^{n} |a_{ij}|$. Таким образом,

$$\|A\|_\infty = \max_{x\in\mathbb{C}^n,\, \|x\|_\infty=1} \|Ax\|_\infty = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|.$$

Норму $\|A\|_\infty$ часто называют *строчной* нормой матрицы A.

3) Введем на пространствах $\mathbb{C}^m$, $\mathbb{C}^n$ норму, согласованную со стандартным скалярным произведением, т. е. положим $\|x\|_2 = |x|$. Для любого $x \in \mathbb{C}^n$ имеем $\|Ax\|_2^2 = (Ax, Ax) = (A^*Ax, x)$. Матрица A^*A эрмитова и неотрицательна. Поэтому существует ортонормированный базис $\{e^k\}_{k=1}^n$ такой, что $A^*Ae^k = \rho_k^2 e^k$, $\rho_k = \rho_k(A)$ — неотрицательные числа, сингулярные числа матрицы A, $k = 1, 2, \ldots, n$ (см. по этому поводу п. 2, с. 54, и приводимые там ссылки). Представим вектор x в виде разложения по базису $x = \sum\limits_{k=1}^{n} \xi_k e^k$ и предположим, что $\|x\|_2 = 1$. Тогда $\sum\limits_{k=1}^{n} |\xi_k|^2 = 1$, $\|Ax\|_2^2 = \sum\limits_{k=1}^{n} \rho_k^2 |\xi_k|^2 \leqslant \max\limits_{1\leqslant k\leqslant n} \rho_k^2$. Пусть $\rho_j = \max\limits_{1\leqslant k\leqslant n} \rho_k$. Полагая $\tilde{x} = e^j$, получим $\|A\tilde{x}\|^2 = \rho_j^2$. Таким образом, доказано, что $\max\limits_{x\in\mathbb{C}^n,\, \|x\|_2=1} \|Ax\|_2 = \max\limits_{1\leqslant k\leqslant n} \rho_k$, т. е.

$$\|A\|_2 = \max_{1\leqslant k\leqslant n} \rho_k(A). \tag{23.4}$$

Отметим следующий интересный для многих приложений частный случай. Будем считать, что матрица $A \in M_n$ эрмитова, т. е. $A = A^*$. Тогда, очевидно $\rho_k(A) = |\lambda_k(A)|$, $k = 1, 2, \ldots, n$, где через $\lambda_k(A)$ обозначены собственные числа матрицы A. Таким образом, для любой эрмитовой матрицы

$$\|A\|_2 = \max_{1\leqslant k\leqslant n} |\lambda_k(A)| = \max_{x\in\mathbb{C}^n,\, x\neq 0} \frac{|(Ax, x)|}{(x, x)} = \rho(A), \tag{23.5}$$

где $\rho(A)$ — спектральный радиус матрицы A (см. с. 212, [5]). Норму $\|A\|_2$ в связи с этим часто называют *спектральной*.

Упражнение 23.5. Докажите, что если матрица A обратима, то

$$\operatorname{cond}(A) = \|A\|_2 \|A^{-1}\|_2$$

(см. с. 57).

Замечание 1. Часто применяют обозначение

$$\operatorname{cond}(A) = \operatorname{cond}_2(A).$$

5. Вычисление сингулярных чисел матрицы, вообще говоря, — довольно сложная задача. Поэтому полезно получить некоторую оценку величины $\|A\|_2$, просто выражаемую через элементы матрицы A. Докажем, что для любой матрицы $A \in M_{mn}$ справедливо неравенство $\|A\|_2 \leqslant \|A\|_E$. С этой целью заметим, что элементарные выкладки приводят к равенству[1)] $\operatorname{tr}(A^*A) = \sum\limits_{i,j=1}^{m,n} |a_{ij}|^2$. С другой стороны, $\operatorname{tr}(A^*A) = \sum\limits_{k=1}^{n} \rho_k^2(A) \geqslant \max\limits_{1\leqslant k\leqslant n} \rho_k^2(A)$, следовательно,

$$\|A\|_2 = \max_{1\leqslant k\leqslant n} \rho_k(A) \leqslant \left(\sum_{i,j=1}^{m,n} |a_{ij}|^2 \right)^{1/2} = \|A\|_E. \tag{23.6}$$

Упражнение 23.6. Докажите, что для любой матрицы A: 1) нормы $\|A\|_2$ и $\|A\|_E$ не меняются при умножении A (слева или справа) на любую унитарную матрицу; 2) $\|A\|_2 = \|A^*\|_2$.

6. Знание согласованной нормы матрицы оказывается, в частности, полезным при оценке ее спектрального радиуса, а именно, для любой квадратной матрицы A справедливо неравенство

$$\rho(A) \leqslant \|A\|, \tag{23.7}$$

где $\|A\|$ — любая согласованая норма матрицы A. В самом деле, пусть λ, x — собственная пара матрицы A, а X — квадратная матрица, столбцами (одинаковыми) которой служит вектор x. Тогда, очевидно, $AX = \lambda X$ и

$$|\lambda|\|X\| = \|AX\| \leqslant \|A\|\|X\|$$

для любой согласованной матричной нормы, причем $\|X\| \neq 0$, так как вектор x по определению собственного вектора не равен нулю. Таким образом, для любого собственного числа λ матрицы A верно неравенство $|\lambda| \leqslant \|A\|$, а это эквивалентно (23.7).

Из оценки (23.7) очевидным образом вытекает

Следствие 1. *Если некоторая согласованная норма матрицы $A \in M_n$ меньше единицы, то A — сходящаяся матрица.*

[1)] Здесь след матрицы вычисляется как сумма элементов ее главной диагонали.

Глава 5

Элементы теории возмущений

24. Задача на собственные значения для эрмитовой матрицы

1. Пусть A, B — эрмитовы матрицы порядка n, $\lambda_k(A)$, $\lambda_k(B)$, $k = 1, 2, \ldots, n$, — их собственные числа. Записав очевидное равенство $A = B + (A - B)$ и воспользовавшись неравенствами (11.1), с. 231, [5], а затем неравенством (23.7), с. 74, получим, что[1)]

$$\max_{1\leqslant k\leqslant n} |\lambda_k(A) - \lambda_k(B)| \leqslant \max_{1\leqslant k\leqslant n} |\lambda_k(A - B)| \tag{24.1}$$

$$\max_{1\leqslant k\leqslant n} |\lambda_k(A) - \lambda_k(B)| \leqslant \|A - B\|, \tag{24.2}$$

где $\|\cdot\|$ — любая согласованная матричная норма. Выбирая в качестве нормы матрицы норму Фробениуса (см. с. 70), получим, что

$$\max_{1\leqslant k\leqslant n} |\lambda_k(A) - \lambda_k(B)| \leqslant \left(\sum_{i,j=1}^{n} |a_{ij} - b_{ij}|^2\right)^{1/2}. \tag{24.3}$$

Неравенства (24.1)–(24.3) обычно называют *неравенствами Вейля.*

Полагая, что $|a_{ij} - b_{ij}| \leqslant \varepsilon$, будем иметь, что

$$\max_{1\leqslant k\leqslant n} |\lambda_k(A) - \lambda_k(B)| \leqslant n\varepsilon. \tag{24.4}$$

Нетрудно убедиться, что если $A = I$, а все элементы матрицы E равны $\varepsilon > 0$, то $\max\limits_{1\leqslant k\leqslant n} |\lambda_k(A) - \lambda_k(A + E)| = n\varepsilon$, т. е. оценка (24.3) неулучшаема на множестве всех эрмитовых матриц.

2. В следующей теореме рассматриваются специальные возмущения эрмитовой матрицы.

[1)] Собственные числа эрмитовой матрицы будем всегда считать упорядоченными по невозрастанию, т. е. $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$.

Теорема 1 («относительная» теорема Вейля). *Пусть λ_k, $k = 1, 2, \ldots, n$ — собственные числа эрмитовой матрицы A, порядка n, $\tilde{\lambda}_k$, $k = 1, 2, \ldots, n$ — собственные числа матрицы X^*AX, где X — произвольная невырожденная матрица. Тогда*

$$|\tilde{\lambda}_i - \lambda_i| \leqslant \lambda_i \|I - X^*X\|, \quad i = 1, 2, \ldots, n, \tag{24.5}$$

где $\|\cdot\|$ — любая согласованная матричная норма.

Доказательство. Фиксируем целое $i \in [1, n]$ и запишем очевидное равенство $X^*(A - \lambda_i I)X = H + F$, где $H = X^*AX - \lambda_i I$, $F = \lambda_i(I - X^*X)$. Легко проверяется, что i-м собственным числом матрицы $A - \lambda_i I$ будет нуль. Используя теорему 1, с. 261, [5][1)], нетрудно убедиться, что i-м собственным числом матрицы $X^*(A - \lambda_i I)X$ также будет нуль. Матрица H в качестве i-го собственного числа имеет $\tilde{\lambda}_i - \lambda_i$, поэтому, применяя неравенство (24.2), получим (24.5). □

Теорема 1 показывает, что при замене матрицы A на X^*AX с невырожденной матрицей X нулевые собственные числа сохраняются, а для ненулевых гарантируется оценка относительной погрешности.

$$\frac{|\tilde{\lambda}_i - \lambda_i|}{|\lambda_i|} \leqslant \|I - X^*X\|, \quad i = 1, 2, \ldots, n.$$

25. Собственные числа произвольной матрицы

1. Пусть $A = \{a_{ij}\}_{i,j=1}^n$ — произвольная квадратная матрица Положим

$$R_i(A) = \sum_{1 \leqslant j \leqslant n,\ j \neq i} |a_{ij}|, \quad i = 1, 2, \ldots, n,$$

$$C_j(A) = \sum_{1 \leqslant i \leqslant n,\ i \neq j} |a_{ij}|, \ j = 1, 2, \ldots, n.$$

Теорема 1 (Гершгорин[2)]). *Все собственные числа произвольной квадратной матрицы A порядка n лежат в объединении кругов*

$$G_i^R = \{z \in \mathbb{C} \colon |z - a_{ii}| \leqslant R_i(A)\}, \quad i = 1, 2, \ldots, n. \tag{25.1}$$

[1)] В [5] упомянутая теорема сформулирована и доказана применительно к симметричным вещественным матрицам, случай эрмитовых матриц рассматривается точно так же.

[2)] Семён Аронович Гершгорин (1901–1933) — советский математик.

Доказательство. Пусть λ, x — собственная пара матрицы A и пусть x_i максимальная по модулю компонента вектора x. Очевидно, $x_i \neq 0$. Из определения собственной пары вытекает равенство

$$(a_{ii} - \lambda)x_i = \sum_{1 \leqslant j \leqslant n,\ j \neq i} a_{ij}x_j,$$

следовательно, $|a_{ii} - \lambda||x_i| \leqslant R_i(A)|x_i|$, и $|a_{ii} - \lambda| \leqslant R_i(A)$. Таким образом, каждое собственные число матрицы A принадлежит одному из кругов G_i, $i = 1, 2, \ldots, n$. $\square$

Поскольку все собственные числа матриц A, A^T совпадают, то все они лежат также в объединении кругов

$$G_i^C = \{z\colon |z - a_{ii}| \leqslant C_i(A)\}, \quad i = 1, 2, \ldots, n. \tag{25.2}$$

Это есть так называемый столбцовый вариант теоремы Гершгорина.

Теорему 1 можно трактовать как теорему о возмущениях диагональной матрицы $D = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$. Она показывает, что если недиагональные элементы матрицы A малы, то ее собственные числа мало отличаются от собственных чисел матрицы D.

Следующие две теоремы, называемые теоремами Бауэра — Файка распространяют теорему Гершгорина на более общий класс матриц, а именно, на матрицы, подобные диагональным, иначе говоря, на матрицы простой структуры (см. § 6, с. 189, [5]).

Теорема 2. *Пусть для квадратной матрицы $A = \{a_{ij}\}_{i,j=1}^n$ существует невырожденная матрица V такая, что*

$$V^{-1}AV = \Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n), \tag{25.3}$$

$B = \{b_{ij}\}_{i,j=1}^n$ — произвольная квадратная матрица. Тогда все собственные числа матрицы $A + B$ лежат в объединении кругов

$$G_i = \{z\colon |z - \lambda_i| \leqslant \|B\|\|V\|\|V^{-1}\|\}, \quad i = 1, 2, \ldots, n. \tag{25.4}$$

Под нормой матрицы здесь может пониматься любая норма, подчиненная абсолютной норме векторов.

Доказательство. Пусть λ, x есть собственная пара матрицы $A + B$, т. е. $(A + B)x = \lambda x$. Тогда $(\lambda I - \Lambda)V^{-1}x = V^{-1}BVV^{-1}x$, откуда (см. п. 7, с. 64) получаем $\min\limits_{1 \leqslant i \leqslant n} |\lambda - \lambda_i|\|V^{-1}x\| \leqslant \|B\|\|V^{-1}\|\|V\|\|V^{-1}x\|$, но $V^{-1}x \neq 0$, следовательно, $\min\limits_{1 \leqslant i \leqslant n} |\lambda - \lambda_i| \leqslant \|B\|\|V^{-1}\|\|V\|$, поэтому $\lambda \in \bigcup\limits_{i=1}^n G_i$. $\square$

Теорема 3. *Пусть выполнены условия теоремы* 2. *Тогда все собственные числа матрицы* $A+B$ *лежат в объединении кругов*

$$G_i = \{z\colon |z-\lambda_i| \leqslant n s_i \|B\|_2\}, \quad i = 1, 2, \ldots, n, \tag{25.5}$$

где $s_i = \|u^i\|_2 \|v^i\|_2 / |(u^i, v^i)|$, v^i — *i-й столбец матрицы* V, u_i — *i-й столбец матрицы* $U = (V^{-1})^*$.

Замечание 1. Ясно, что v_i, λ_i, $i = 1, 2, \ldots, n$, — собственные пары матрицы A, $u_i, \bar{\lambda}_i$, $i = 1, 2, \ldots, n$, — собственные пары матрицы A^*. Каждое из чисел s_i, $i = 1, 2, \ldots, n$, не меньше единицы. Их называют *коэффициентами перекоса* соответствующих собственных векторов матрицы A. Если λ — алгебраически простое собственное число матрицы A, то, очевидно, $\bar{\lambda}$ — алгебраически простое собственное число матрицы A^*. Отвечающие им собственные подпространства одномерны и, следовательно, соответствующий коэффициент перекоса определяется однозначно.

Доказательство теоремы 3. Собственное числа матриц $A+B$ и $\Lambda + V^{-1}BV = \Lambda + \widetilde{B}$, где $\widetilde{B} = U^*BV$, совпадают. Используя столбцовую теорему Гершгорина, получим, что все собственные числа матрицы $\Lambda + \widetilde{B}$ лежат в объединении кругов

$$G'_i = \{z\colon |z-\lambda_i-\tilde{b}_{ii}| \leqslant C_i(\widetilde{B})\}, \quad i = 1, 2, \ldots, n.$$

Заметим теперь, что $|z-\lambda_i-\tilde{b}_{ii}| \geqslant |z-\lambda_i| - |\tilde{b}_{ii}|$, $C_i(\widetilde{B}) + |\tilde{b}_{ii}| = \|\tilde{b}^i\|_1$, где, как обычно, $\tilde{b}^i$ — i-й столбец матрицы $\widetilde{B}$. Отсюда вытекает, что все собственные числа матрицы $A+B$ лежат в объединении кругов

$$G''_k = \{z\colon |z-\lambda_k| \leqslant \|\tilde{b}^k\|_1\}, \quad k = 1, 2, \ldots, n.$$

Оценим $\|\tilde{b}^k\|_1$. Введем в рассмотрение векторы $t^k \in \mathbb{C}^n$ с компонентами

$$t^k_j = \begin{cases} \tilde{b}^k_j / |\tilde{b}^k_j|, & \tilde{b}^k_j \neq 0, \\ 0, & \tilde{b}^k_j = 0. \end{cases}$$

Элементарно проверяется равенство $\|\tilde{b}^k\|_1 = (\widetilde{B} i^k, t^k)$, где i^k — столбец единичной матрицы. Отсюда, используя неравенство Коши — Буняковского, получаем

$$\|\tilde{b}^k\|_1 = (BVi^k, Ut^k) \leqslant \|B\|_2 \|U\|_2 \|v^k\|_2 \|t^k\|_2. \tag{25.6}$$

Нетрудно убедиться, что $\|t^k\|_2 \leqslant \sqrt{n}$. Далее, вследствие (23.6), с. 74, имеем $\|U\|_2 \leqslant \left(\sum\limits_{k=1}^{n} \|u^k\|_2^2 \right)^{1/2}$. Столбцы матрицы U определяются,

очевидно, с точностью до постоянных ненулевых множителей. Нормируем их так, чтобы $\|u^k\|_2 = 1$ для всех $k = 1, 2, \ldots, n$. Очевидно, при этом столбцы матрицы V должны быть нормированы так, чтобы $(v^k, u^k) = 1$ для всех $k = 1, 2, \ldots, n$. При этом будем иметь $\|v^k\|_2 = \|v^k\|_2\|u^k\|_2/|(u^k, v^k)| = s_k$. Таким образом, из (25.6) получаем, что $\|\tilde{b}^k\|_1 \leqslant ns_k\|B\|_2$. $\square$

26. Возмущения и обратимость матрицы

1. Пусть $A \in M_n$ — обратимая матрица, т. е. $|A| \neq 0$. Пусть, далее, $B \in M_n$. Возникает вопрос, при каких условиях на B матрица $A + B$ будет также обратимой? Поскольку $A + B = A(I + A^{-1}B)$, то для существования матрицы, обратной к $A+B$, очевидно, необходимо и достаточно, чтобы спектр матрицы $A^{-1}B$ не содержал -1. Отсюда вытекают следующие практически важные достаточные условия обратимости матрицы $A + B$:

1) матрица $A + B$ обратима, если $\rho(A^{-1}B) < 1$;
2) матрица $A + B$ обратима, если $\|A^{-1}B\| < 1$;
3) матрица $A + B$ обратима, если $\|A^{-1}\|\|B\| < 1$.

Здесь и далее в этом пункте под нормой матрицы понимается согласованная матричная норма. Третье условие часто записывают так:

$$\operatorname{cond}(A)(\|B\|/\|A\|) < 1, \tag{26.1}$$

где $\operatorname{cond}(A) = \|A^{-1}\|\|A\|$. Это число называют числом обусловленност матрицы A (ср. с п. 9, с. 57). Условие (26.1) можно интерпретировать следующим образом: матрица $A + B$ обратима, если относительное возмущение матрицы A, т. е. $\|B\|/\|A\|$, мало по сравнению с ее числом обусловленности.

2. Пример. Пусть $A = \{a_{ij}\}_{i,j=1}^n$ — произвольная квадратная матрица Напомним, что A — матрица с *диагональным преобладанием по строкам*, если[1)]

$$|a_{ii}| > R_i(A) \quad \forall\, i = 1, 2, \ldots, n, \tag{26.2}$$

и A — матрица с *диагональным преобладанием по столбцам*, если

$$|a_{ii}| > C_i(A) \quad \forall\, i = 1, 2, \ldots, n. \tag{26.3}$$

Покажем, что если A — матрица с диагональным преобладанием по строкам, то она невырождена. Пусть $D = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$.

[1)] См. обозначения в § 25, с. 76.

Вследствие условия (26.2) матрица D невырождена. Запишем матрицу A в виде $A = D + (A - D)$. Вновь используя условие (26.2), получим, что $\|D^{-1}(A - D)\|_\infty < 1$, значит выполнено условие 2, и матрица A невырождена. Поскольку определители матриц A и A^T совпадают, то матрица с диагональным преобладанием по столбцам также невырождена.

Упражнение 26.1. Покажите, что если выполнено условие (26.2), или (26.3), то все главные миноры матрицы A отличны от нуля.

Теорема 1. *Пусть матрицы A и $\tilde{A} = A + B$ обратимы. Тогда*

$$\frac{\|A^{-1} - \tilde{A}^{-1}\|}{\|\tilde{A}^{-1}\|} \leqslant \|A^{-1}B\|. \tag{26.4}$$

Если $\|A^{-1}B\| < 1$, то

$$\|\tilde{A}^{-1}\| \leqslant \frac{\|A^{-1}\|}{1 - \|A^{-1}B\|}, \tag{26.5}$$

$$\frac{\|A^{-1} - \tilde{A}^{-1}\|}{\|A^{-1}\|} \leqslant \frac{\|A^{-1}B\|}{1 - \|A^{-1}B\|}. \tag{26.6}$$

Доказательство. По условию теоремы $I = (A + B)\tilde{A}^{-1}$, следовательно, $A^{-1} = (I + A^{-1}B)\tilde{A}^{-1}$, поэтому $A^{-1} - \tilde{A}^{-1} = A^{-1}B\tilde{A}^{-1}$. Отсюда, очевидно, следует (26.7). Далее, $\tilde{A}^{-1} = A^{-1} - A^{-1}B\tilde{A}^{-1}$, значит, $\|\tilde{A}^{-1}\| \leqslant \|A^{-1}\| + \|A^{-1}B\|\|\tilde{A}^{-1}\|$, откуда вытекает (26.5). Наконец, (26.6) — очевидное следствие (26.7), (26.5). □

Из теоремы 1 непосредственно вытекает

Следствие 1. *Пусть матрицы A и $\tilde{A} = A+B$ обратимы. Тогда*

$$\frac{\|A^{-1} - \tilde{A}^{-1}\|}{\|\tilde{A}^{-1}\|} \leqslant \operatorname{cond}(A)(\|B\|/\|A\|). \tag{26.7}$$

Если $\operatorname{cond}(A)(\|B\|/\|A\|) < 1$, то

$$\|\tilde{A}^{-1}\| \leqslant \frac{\|A^{-1}\|}{1 - \operatorname{cond}(A)(\|B\|/\|A\|)}, \tag{26.8}$$

$$\frac{\|A^{-1} - \tilde{A}^{-1}\|}{\|A^{-1}\|} \leqslant \frac{\operatorname{cond}(A)(\|B\|/\|A\|)}{1 - \operatorname{cond}(A)(\|B\|/\|A\|)|}. \tag{26.9}$$

3. Следующая теорема показывает, что «расстояние» от невырожденной матрицы A до ближайшей вырожденной матрицы характеризуется величиной $1/\operatorname{cond}(A)$.

Теорема 2. *Пусть матрица A обратима, матрица $A+B$ вырождена, тогда*

$$\|B\|/\|A\| \geqslant 1/\operatorname{cond}(A). \tag{26.10}$$

Если при этом под нормой матрицы понимать норму, подчиненную некоторой норме векторов, то найдется такая матрица B, что

$$\|B\|/\|A\| = 1/\operatorname{cond}(A), \tag{26.11}$$

а матрица $A+B$ вырождена.

Доказательство. Как было указано выше, если матрица A а обратима, а матрица $A+B$ вырождена, то спектр матрицы $A^{-1}B$ содержит число -1, значит $\rho(A^{-1}B) \geqslant 1$, но

$$\rho(A^{-1}B) \leqslant \|A^{-1}B\| \leqslant \|A^{-1}\|\|B\|,$$

т. е. $\|B\| \geqslant 1/\|A^{-1}\|$. Последнее неравенство эквивалентно (26.10). Переходим к доказательству второй части теоремы. Из определения подчиненной нормы матрицы следует, что существует вектор x такой, что $\|x\| = 1$, $\|A^{-1}x\| = \|A^{-1}\|$. Положим $y = \|A^{-1}\|^{-1}A^{-1}x$. Тогда $\|y\| = 1$, $Ay = \|A^{-1}\|^{-1}x$. По следствию 1, с. 67, на пространстве $\mathbb{C}^n$ существует линейный функционал f такой, что $f(y) = \|y\| = 1$, и $\|f\| = \sup\limits_{v\in\mathbb{C}^n,\,\|v\|=1} |f(v)| = 1$. Определим матрицу B действием ее на векторы при помощи соотношения

$$Bv = -(f(v)/\|A^{-1}\|)x \quad \forall\, v \in \mathbb{C}^n.$$

Ясно, что $By = -\|A^{-1}\|^{-1}x$, поэтому $(A+B)y = 0$, значит,

$$\det(A+B) = 0.$$

Кроме того,

$$\|B\| = \sup_{v\in\mathbb{C}^n,\,\|v\|=1} \|Bv\| = \|A^{-1}\|^{-1} \sup_{v\in\mathbb{C}^n,\,\|v\|=1} |f(v)| = \|A^{-1}\|^{-1}.$$

Полученное равенство эквивалентно (26.11). □

27. Устойчивость систем линейных уравнений

1. В этом параграфе норма матриц считается согласованной с нормой векторов. Следующая теорема устанавливает связь относительного возмущениям матрицы системы и ее правой части с относительным возмущением решения. Главную роль в получаемых здесь оценках играет число обусловленности матрицы системы уравнений.

Теорема 1. *Пусть матрица A обратима, матрица B такова, что $\|A^{-1}B\| < 1$, вектор x — решение системы уравнений*

$$Ax = y, \tag{27.1}$$

вектор $\tilde{x}$ — решение системы уравнений

$$\tilde{A}\tilde{x} = y + b, \quad \tilde{A} = A + B. \tag{27.2}$$

Тогда

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leqslant \frac{\operatorname{cond}(A)}{1 - \|A^{-1}B\|} \left(\frac{\|b\|}{\|y\|} + \frac{\|B\|}{\|A\|} \right). \tag{27.3}$$

Если дополнительно потребовать, чтобы выполнялось условие

$$\|A^{-1}\|\|B\| < 1,$$

то

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leqslant \frac{\operatorname{cond}(A)}{1 - \operatorname{cond}(A)(\|B\|/\|A\|)} \left(\frac{\|b\|}{\|y\|} + \frac{\|B\|}{\|A\|} \right). \tag{27.4}$$

Доказательство. По условию теоремы матрицы A^{-1} и $\tilde{A}^{-1}$ существуют, поэтому $x = A^{-1}y$, $\tilde{x} = \tilde{A}^{-1}(y + b)$, следовательно, $\tilde{x} - x = \tilde{A}^{-1}b + (\tilde{A}^{-1} - A^{-1})y$, и

$$\|x - \tilde{x}\| \leqslant \|\tilde{A}^{-1}\|\|b\| + \|\tilde{A}^{-1} - A^{-1}\|\|y\|,$$

откуда, используя (26.5), (26.6) и неравенство $\|y\| \leqslant \|A\|\|x\|$, после элементарных преобразований получим (27.3). Оценка (27.4) есть очевидное следствие (27.3). □

2. Пусть некоторым способом найден вектор $\tilde{x}$, который мы считаем приближением к решению уравнения (27.1). Наша цель — оценить погрешность $\|x - \tilde{x}\|$ через норму невязки $\|A\tilde{x} - y\|$. Введем используемую в дальнейшем вспомогательную величину. Пусть матрица A обратима, $x \neq 0$, $Ax = y$. Положим $\eta = \|A\|\|x\|/\|y\|$. Очевидно, что $\eta \geqslant 1$, и поскольку $\|x\| \leqslant \|A^{-1}\|\|y\|$, то $\eta \leqslant \|A\|\|A^{-1}\| = \operatorname{cond}(A)$. Для $\tilde{x} \in \mathbb{C}^n$ положим $r = A\tilde{x} - y$. Тогда $x - \tilde{x} = A^{-1}r$,

$$\|x - \tilde{x}\| \leqslant \|A^{-1}\|\|r\|.$$

Поэтому

$$\|x - \tilde{x}\|/\|x\| \leqslant (\operatorname{cond}(A)/\eta)\|r\|/\|y\|, \tag{27.5}$$

и как следствие

$$\|x - \tilde{x}\|/\|x\| \leqslant \operatorname{cond}(A)\|r\|/\|y\|. \tag{27.6}$$

Оценка (27.5) показывает, что чем ближе величина η к величине $\operatorname{cond}(A)$, тем лучше относительная погрешность оценивается относительной невязкой приближенного решения.

Глава 6

Итерационные методы решения систем линейных уравнений

При реализации прямых методов важно, чтобы все данные располагались в оперативной (быстрой) памяти компьютера. Если порядок системы настолько велик, что оперативной памяти для реализации метода недостаточно, то время, затрачиваемое на решение системы, существенно увеличивается. Для таких систем предпочтительнее оказываются итерационные методы. Основная идея этих методов состоит в построении последовательности векторов x^k, $k = 1, 2, \ldots$, сходящейся к решению x системы $Ax = b$. За приближенное решение принимается вектор x^k при достаточно большом k. В качестве критерия окончания итерационного процесса обычно принимают либо достаточную близость двух соседних приближений x^k и x^{k+1}, либо достаточную малость невязки $Ax^k - b$.

28. Простейшие итерационные методы

Всюду в дальнейшем через z^k будем обозначать вектор $x^k - x$, где x — решение системы

$$Ax = b, \tag{28.1}$$

т. е. *погрешность приближения* с номером x^k.

1. Метод Якоби[1)]. Будем считать, что все диагональные элементы матрицы A отличны от нуля. Перепишем систему (28.1), разрешая каждое уравнение относительно переменной, стоящей на диагонали:

$$x_i = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j + \frac{b_i}{a_{ii}}, \quad i = 1, 2, \ldots, n. \tag{28.2}$$

Выберем некоторое начальное приближение $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)$ и построим последовательность векторов $x^1, x^2, \ldots$, определяя вектор x^{k+1} по уже найденному вектору x^k при помощи соотношений:

$$x_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^k - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^k + \frac{b_i}{a_{ii}}, \quad i = 1, 2, \ldots, n. \tag{28.3}$$

[1)] Карл Густав Якоб Яко́би (Carl Gustav Jacob Jacobi; 1804 — 1851) — немецкий математик.

Формулы (28.3) определяют итерационный метод решения системы (28.1), называемый *методом Якоби*.

Укажем легко проверяемое достаточное условие сходимости этого метода. Напомним, что для матрицы A выполнено условие диагонального преобладания по строкам, если

$$q = \max_{1\leqslant i\leqslant n} \sum_{j=1,\ j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1. \tag{28.4}$$

Теорема 1. *Пусть матрица A системы (28.1) — матрица с диагональным преобладанием по строкам. Тогда итерационный метод Якоби сходится при любом начальном приближении x^0; справедлива следующая оценка скорости сходимости:*

$$\|z^k\|_\infty \leqslant q^k \|z^0\|_\infty. \tag{28.5}$$

Доказательство. Пусть x — решение системы уравнений (28.1). Вычитая почленно из равенства (28.3) равенство (28.2), получим

$$z_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} z_j^k - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} z_j^k, \quad i = 1, 2, \ldots, n,$$

следовательно,

$$|z_i^{k+1}| \leqslant \sum_{j=1}^{i-1} \frac{|a_{ij}|}{|a_{ii}|} |z_j^k| + \sum_{j=i+1}^{n} \frac{|a_{ij}|}{|a_{ii}|} |z_j^k| \leqslant \left(\sum_{j=1}^{i-1} \frac{|a_{ij}|}{|a_{ii}|} + \sum_{j=i+1}^{n} \frac{|a_{ij}|}{|a_{ii}|} \right) \max_{1\leqslant j\leqslant n} |z_j^k| =$$

$$= q \max_{1\leqslant j\leqslant n} |z_j^k|, \quad i = 1, 2, \ldots, n,$$

откуда вытекает, что

$$\|z^{k+1}\|_\infty \leqslant q \|z_j^k\|_\infty$$

для любого $k = 0, 1, \ldots$, поэтому

$$\|z^k\|_\infty \leqslant q^k \|z_j^0\|_\infty \to 0$$

при $k \to \infty$, поскольку $0 < q < 1$, а это и означает, что $x^k \to x$. □

Оценка (28.5) показывает, что, чем меньше q, т. е. чем выше диагональное преобладание матрицы A, тем быстрее сходится метод Якоби.

2. Метод Зейделя. Формулы (28.3) допускают естественную модификацию. Именно, при вычислении x_i^{k+1} будем использовать уже найденные компоненты вектора x^{k+1}, т. е. $x_1^{k+1}, x_2^{k+1}, \dots x_{i-1}^{k+1}$. В результате приходим к итерационному *методу Зейделя*[1)]:

$$x_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{k+1} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^k + \frac{b_i}{a_{ii}}, \quad i = 1, 2, \dots, n, \ k = 0, 1, \dots \tag{28.6}$$

Метод Зейделя позволяет более экономно расходовать память компьютера, поскольку в данном случае вновь получаемые компоненты вектора x^{k+1} можно размещать на месте соответствующих компонент вектора x^k, в то время как при реализации метода Якоби все компоненты векторов x^k, x^{k+1} должны одновременно находиться в памяти компьютера.

Достаточное условие сходимости и оценку скорости сходимости метода Зейделя дает

Теорема 2. *Пусть матрица A — матрица с диагональным преобладанием по строкам. Тогда метод Зейделя сходится при любом начальном приближении x^0; справедлива оценка скорости сходимости:*

$$\|z_j^k\|_\infty \leqslant q^k \|z^0\|_\infty, \tag{28.7}$$

где q определяется (28.4).

Доказательство. Вычитая почленно из равенства (28.6) равенство (28.2), получим

$$z_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} z_j^{k+1} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} z_j^k, \quad i = 1, 2, \dots, n. \tag{28.8}$$

Пусть $\max\limits_{1\leqslant j\leqslant n} |z_j^{k+1}| = |z_l^{k+1}|$. Из l-того уравнения системы (28.8) вытекает, что

$$|z_l^{k+1}| \leqslant \alpha_l \max_{1\leqslant j\leqslant n} |z_j^{k+1}| + \beta_l \max_{1\leqslant j\leqslant n} |z_j^k|,$$

где

$$\alpha_l = \sum_{j=1}^{l-1} \frac{|a_{lj}|}{|a_{ll}|}, \quad \beta_l = \sum_{j=l+1}^{n} \frac{|a_{lj}|}{|a_{ll}|},$$

[1)] Филипп Людвиг Зейдель (Philipp Ludwig von Seidel; 1821 — 1896) — немецкий математик и астроном.

следовательно,

$$\|z^{k+1}\|_\infty \leqslant \frac{\beta_l}{1-\alpha_l}\|z^k\|_\infty.$$

Из условия (28.4) получаем, что $\alpha_l+\beta_l \leqslant q < 1$, но тогда и $q\alpha_l+\beta_l \leqslant q$, таким образом, $\beta_l/(1-\alpha_l) \leqslant q$, поэтому $\|z^{k+1}\|_\infty \leqslant q \max \|z^k\|_\infty$ для любого $k \geqslant 0$. Дальнейшие рассуждения совпадают с соответствующими рассуждениями из доказательства предыдущей теоремы. □

3. Метод редаксации. Зачастую существенного ускорения сходимости можно добиться за счет введения в расчетные формулы числового параметра. В качестве примера приведем итерационный процесс

$$x_i^{k+1} = (1-\omega)x_i^k + \omega\left(-\sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ii}}x_j^{k+1} - \sum_{j=i+1}^{n}\frac{a_{ij}}{a_{ii}}x_j^k + \frac{b_i}{a_{ii}}\right), \qquad (28.9)$$

$i = 1, 2, \ldots, n$, $k = 0, 1, \ldots$ Этот метод называется *методом релаксации*, число ω — *релаксационным параметром*. При $\omega = 1$ метод переходит в метод Зейделя.

Ясно, что по затратам памяти и объему вычислений на каждом шаге итераций метод релаксации не отличается от метода Зейделя.

29. Элементы общей теории итерационных методов

1. Далее наряду со стандартным скалярным произведением будем использовать так называемое *энергетическое скалярное произведение* и соответствующую ему норму на пространстве $\mathbb{C}^n$. Именно, если $D \in M_n$ — эрмитова положительно определенная матрица, то по определению $(x, y)_D = (Dx, y)$, будем полагать также, что $\|x\|_D = (Dx, x)^{1/2}$ [1)].

2. Придадим итерационным методам, рассмотренным в предыдущих пунктах, матричные формулировки. Начнем с метода Якоби. Нетрудно видеть, что равенства (28.3) можно записать в матричном виде

$$D(x^{k+1} - x^k) + Ax^k = b, \qquad (29.1)$$

где $D = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$. Для того, чтобы придать матричную форму записи методам Зейделя и релаксации, обозначим через L нижнюю треугольную матрицу, поддиагональные элементы которой

[1)] См. по этому поводу п. 2, с. 61.

совпадают с соответствующими элементами матрицы A, а все диагональные элементы равны нулю. Через R обозначим верхнюю треугольную матрицу такую, что $A = L + D + R$. Равенства (28.9) могут быть переписаны тогда в следующем виде:

$$\frac{1}{\omega}(D + \omega L)(x^{k+1} - x^k) + Ax^k = b. \tag{29.2}$$

3. Будем рассматривать общий класс итерационных методов, определяемых соотношениями

$$\frac{1}{\tau}B(x^{k+1} - x^k) + Ax^k = b, \quad k = 0, 1, \ldots, \tag{29.3}$$

вектор x^0 считается заданным. Здесь B — невырожденная матрица, $\tau > 0$ — число, называемое *итерационным параметром*. Для того, чтобы найти вектор x^{k+1} по уже известному вектору x^k, решим систему линейных уравнений

$$Bw^k = r^k, \tag{29.4}$$

где $r^k = Ax^k - b$, и положим $x^{k+1} = x^k - \tau w^k$.

Очевидно, при построении итерационного метода (29.3) матрица B должна выбираться так, чтобы решение системы уравнений вида (29.4) выполнялось намного быстрее, чем решение исходной системы уравнений (28.1).

Итерационные методы Якоби, Зейделя и релаксации являются частными случаями метода (29.3). Например, в случае метода Якоби $B = D$, $\tau = 1$.

4. Наша ближайшая цель — получить условия на матрицу B и параметр τ, обеспечивающие сходимость метода (29.3).

Если x — решение системы (28.1), то, очевидно,

$$\frac{1}{\tau}B(x - x) + Ax = b. \tag{29.5}$$

Вычитая почленно равенства (29.3), (29.5), получим

$$\frac{1}{\tau}B(z^{k+1} - z^k) + Az^k = 0, \tag{29.6}$$

откуда

$$z^{k+1} = Sz^k, \tag{29.7}$$

где

$$S = I - \tau B^{-1}A. \tag{29.8}$$

и, следовательно,

$$z^k = S^k z^0, \tag{29.9}$$

Понятно, что сходимость итерационного метода (29.3) определяется свойствами матрицы S, которую обычно называют матрицей шага итерационного метода (29.3).

Теорема 1. *Для того, чтобы итерационный метод* (29.3) *сходился при любом начальном приближении x^0, необходимо и достаточно, чтобы спектральный радиус $\rho(S)$ матрицы S был меньше единицы.*

Доказательство. Необходимость. Пусть λ — собственное число матрицы S такое, что $|\lambda| \geqslant 1$, e — соответствующий этому собственному числу нормированный собственный вектор матрицы S. Выберем в качестве начального приближения в итерационном методе (29.3) вектор $x^0 = x + e$, где x — решение уравнения (28.1). Тогда в соответствии с (29.9) имеем $z^k = \lambda^k e$, следовательно, $|z^k| = |\lambda|^k$. Очевидно, либо $|z^k| \to \infty$ при $k \to \infty$, либо $|z^k| = 1$ для всех $k = 1, 2, \ldots$, т. е. метод (29.3) не сходится. Достаточность. Если спектральный радиус матрицы S меньше единицы, то она является сходящейся матрицей (см. с. 212, [5]), т. е. $S^k \to 0$ при $k \to \infty$, и тогда из (29.9) вытекает, что $z^k \to 0$ при $k \to \infty$. $\square$

Из теоремы 1 и оценки (23.7), с. 74, сразу же вытекает

Следствие 1. *Для сходимости итерационного метода* (29.3) *достаточно, чтобы для какой-либо согласованной нормы выполнялось условие $\|S\| < 1$.*

Например, при $\tau = 1$ итерационный метод (29.3) сходится, если матрицы A и B достаточно близки, т. е. $\|B^{-1}\|\|B - A\| < 1$. Используя оценку (26.8), с. 80, получим отсюда, что итерационный метод (29.3) сходится, если $\tau = 1$ и

$$\frac{\|B - A\|}{\|A\|} \operatorname{cond}(A) < 1/2.$$

Опираясь на теорему 1, получим часто используемое для систем уравнений с эрмитовой положительно определенной матрицей условие сходимости итерационного процесса (29.3).

Теорема 2 (Самарский[1]). *Пусть матрица A положительно определена и пусть для любого не равного нулю вектора x из $\mathbb{C}^n$*

[1] Александр Андреевич Самарский (1919 — 2008) — советский, российский математик.

выполнено неравенство

$$(B_1x, x) > (\tau/2)(Ax, x), \tag{29.10}$$

где $B_1 = (1/2)(B + B^*)$. *Тогда матрица* B *невырождена, и итерационный процесс* (29.3) *сходится при любом начальном приближении* x^0. *Обратно, если матрица* A *положительно определена и итерационный процесс* (29.3) *сходится при любом начальном приближении* x^0, *то выполнено условие* (29.10).

Доказательство. Невырожденность матрицы B сразу же следует из условия (29.10) и положительной определенности матрицы A (см. упражнение 3 на с. 223, [5]). Покажем, что если выполнено условие (29.10), то $\rho(S) < 1$, где S — матрица, определенная равенством (29.8). Вследствие теоремы 1 отсюда будет вытекать сходимость итерационного метода (29.3). Пусть λ, x — собственная пара матрицы S. Тогда $Bx - \tau Ax = \lambda Bx$, поэтому

$$\lambda = \frac{(Bx, x) - \tau(Ax, x)}{(Bx, x)}.$$

Используя формулу (6.3), с. 221, [5], представим матрицу B в виде

$$B = B_1 + iB_2, \tag{29.11}$$

где $B_1 = (1/2)(B + B^*)$, B_2 — эрмитовы матрицы. Тогда

$$\lambda = \frac{(B_1x, x) - \tau(Ax, x) + i(B_2x, x)}{(B_1x, x) + i(B_2x, x)},$$

следовательно,

$$|\lambda|^2 = \frac{((B_1x, x) - \tau(Ax, x))^2 + (B_2x, x)^2}{(B_1x, x)^2 + (B_2x, x)^2}.$$

Запишем последнее равенство в виде

$$|\lambda|^2 = \frac{(1 - a)^2 + b^2}{1 + b^2}, \tag{29.12}$$

где $a = \tau(Ax, x)/(B_1x, x)$, $b = (B_2x, x)/(B_1x, x)$. Из условия (29.10) получаем, что $0 < a < 2$, поэтому $|1 - a| < 1$, откуда, очевидно, вытекает, что $|\lambda| < 1$. Для доказательства второй части теоремы достаточно заметить, что если итерационный процесс (29.3) сходится при любом начальном приближении, то по теореме 1 все собственные числа матрицы S по модулю строго меньше единицы, и тогда

из представления (29.12) получаем, что $0 < a < 2$, следовательно, условие (29.10) выполнено. $\square$

Теорема 3. *Пусть выполнены условия теоремы* 2. *Тогда для погрешностей итерационного процесса* (29.3) *при любом* $k \geqslant 0$ *выполнено неравенство*

$$(Az^{k+1}, z^{k+1}) < (Az^k, z^k), \tag{29.13}$$

если $z^k \neq 0$.

Доказательство. Используя тривиальное тождество

$$z^k = (1/2)(z^{k+1} + z^k) - (1/2)(z^{k+1} - z^k),$$

преобразуем уравнение (29.6) к виду

$$\frac{1}{\tau}(B - (\tau/2)A)(z^{k+1} - z^k) + (1/2)A(z^{k+1} + z^k) = 0.$$

Умножая теперь скалярно обе части последнего равенства на вектор $2(z^{k+1} - z^k)$ и используя представление (29.11), после элементарных преобразований получим

$$\begin{aligned}\frac{2}{\tau}((B_1 - (\tau/2)A)(z^{k+1} - z^k), z^{k+1} - z^k) + \\ + i\frac{2}{\tau}(B_2(z^{k+1} - z^k), z^{k+1} - z^k) + \\ + (Az^{k+1}, z^{k+1}) - (Az^k, z^k) + i\,\mathrm{Im}(Az^k, z^{k+1}) = 0,\end{aligned}$$

поэтому

$$\begin{aligned}\frac{2}{\tau}((B_1 - (\tau/2)A)(z^{k+1} - z^k), z^{k+1} - z^k) + \\ + (Az^{k+1}, z^{k+1}) - (Az^k, z^k) = 0.\end{aligned} \tag{29.14}$$

Если $z^k \neq 0$, то вследствие невырожденности оператора B из (29.6) вытекает, что $z^{k+1} - z^k \neq 0$. Тогда на основании условия (29.10) из равенства (29.14) получаем, что $(Az^{k+1}, z^{k+1}) - (Az^k, z^k) < 0$. $\square$

5. Если матрица A положительно определена, то уравнение

$$Ax = b \tag{29.15}$$

эквивалентно задаче минимизации функции (функционала)

$$F(x) = (Ax, x) - 2\,\mathrm{Re}(x, b)\ ^{1)}. \tag{29.16}$$

[1)] Функционал F часто называют *энергетическим*. Это связано с задачами физики, в которых возникают уравнения с положительно определенными матрицами.

Действительно, пусть $\hat{x}$ — решение уравнения (29.15). Тогда

$$F(x) = (Ax, x) - 2\,\mathrm{Re}(x, A\hat{x}) = \\ = (Ax, x) - 2\,\mathrm{Re}(x, A\hat{x}) + (A\hat{x}, \hat{x}) - (A\hat{x}, \hat{x}) = \\ = (A(x - \hat{x}), x - \hat{x}) - (A\hat{x}, \hat{x}), \quad (29.17)$$

следовательно, функции $F(x)$ и $F_0(x) = (A(x - \hat{x}), x - \hat{x})$ отличаются на постоянное слагаемое. Поскольку матрица A положительно определена, то единственной точкой минимума функции F_0, а стало быть, и функции F является $\hat{x}$. Вследствие (29.17) неравенство (29.13) можно записать в виде

$$F(x^{k+1}) < F(x^k). \quad (29.18)$$

Таким образом, можно сказать, что при выполнении условий теоремы 2 итерационный процесс (29.3) является *релаксационным*[1]: каждое последующее приближение уменьшает значение функционала F. Используя полученные в предыдущем пункте общие результаты, исследуем сходимость метода релаксации (29.2).

Теорема 4. *Пусть матрица A положительно определена,*

$$0 < \omega < 2. \quad (29.19)$$

Тогда итерационный метод релаксации (29.2) *сходится при любом начальном приближении x^0.*

Доказательство. Будем опираться на теорему 2. В рассматриваемом случае $B = D + \omega L$, $\tau = \omega$, $B_1 = D + (\omega/2)(L + L^*)$, $A = D + L + L^*$, и условие (29.10) принимает вид $(Dx, x) > (\omega/2)(Dx, x)$ для любого $x \neq 0$. Все диагональные элементы положительно определенной матрицы положительны[2], поэтому матрица D положительно определена, и условие (29.10) выполнено, если выполнено условие (29.19). □

Теорема 5. *Условие* (29.19) *необходимо для сходимости итерационного процесса* (28.9).

Доказательство. Запишем равенство (29.8) в виде

$$(D + \omega L)S = (D + \omega L) - \omega A = (1 - \omega)D - \omega R. \quad (29.20)$$

Поскольку L и R — строго треугольные матрицы, а D — диагональная матрица, все диагональные элементы которой отличны от нуля,

[1] Релаксация (лат. relaxatio) — уменьшение напряжения, ослабление.

[2] Действительно, $a_{kk} = (Ai^k, i^k) > 0$, $k = 1, 2, \ldots, n$.

то, вычисляя определители левой и правой частей равенства (29.20), получим, что $\det(S) = (1-\omega)^n$, следовательно (см. (7.7), с. 193, [5]),

$$\prod_{k=1}^{n} |\lambda_k| = |1-\omega|^n, \tag{29.21}$$

где λ_1, λ_2, $\ldots$,λ_n — собственные числа матрицы S. Если условие (29.19) нарушено, то $|1-\omega| > 1$, и среди собственных чисел λ_k матрицы S есть хотя бы одно, модуль которого больше единицы, но тогда по теореме 1 найдется начальное приближение x^0, при котором итерационный процесс (28.9) не сходится. □

6. Оптимизация итерационного параметра. Из доказательства теоремы 1 видно, что итерационный процесс (29.3) сходится тем быстрее, чем меньше спектральный радиус матрицы $S = I - \tau B^{-1}A$. В связи с этим возникает задача отыскания такого (*оптимального*) значения итерационного параметра τ, при котором величина $\rho(S)$ принимает минимальное значение.

Наиболее просто эта задача решается в случае, когда матрицы A, B положительно определены. Поскольку в рассматриваемом случае $B = B^*$, т. е. в представлении (29.11) матрица B_2 равна нулю, то из (29.12) получаем, что для любой собственной пары λ, x матрицы S справедливо равенство

$$|\lambda| = \left|1 - \tau\frac{(Ax,x)}{(Bx,x)}\right|. \tag{29.22}$$

Нетрудно видеть, что если x — собственный вектор матрицы S, то x — собственный вектор матрицы $B^{-1}A$ и, следовательно, x — собственный вектор задачи

$$Ax = \lambda Bx \tag{29.23}$$

(см. подробнее § 13, с. 237, [5]). Очевидно, справедливо и обратное: любой собственный вектор задачи (29.23) есть собственный вектор матрицы S.

Для любой собственной пары x, λ задачи (29.23) справедливо равенство $(Ax,x) = \lambda(Bx,x)$. Поэтому все собственные числа задачи (29.23) положительны. Пусть m — минимальное, а M максимальное из этих чисел. Тогда для любого собственного вектора x матрицы S справедливы неравенства

$$m \leqslant \frac{(Ax,x)}{(Bx,x)} \leqslant M. \tag{29.24}$$

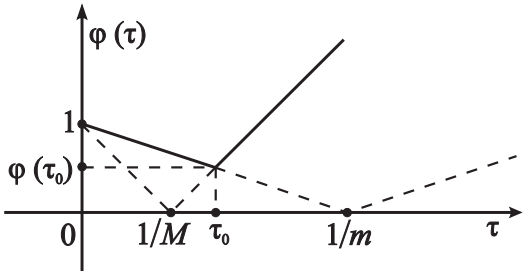

Рис. 1. К выбору оптимального итерационного параметра

Полученные оценки являются точными, поскольку соответствующие неравенства (29.24) превращаются в равенства, если в качестве x взять собственный вектор, отвечающий m или M.

Нетрудно видеть, что функция $g(\mu) = |1 - \tau\mu|$ вещественного переменного μ на любом ограниченном отрезке вещественной оси достигает максимального значения на одном из концов этого отрезка. Поэтому, используя соотношения (29.22), (29.24), получаем, что

$$\rho(S) = \varphi(\tau) = \max\{|1 - \tau m|, |1 - \tau M|\}. \tag{29.25}$$

График функции $\varphi(\tau)$ при $\tau \geqslant 0$ изображен на рис. 1. Нетрудно убедиться, что

$$\min_{\tau \geqslant 0} \varphi(\tau) = \varphi(\tau_0) = \rho_0 = (M - m)/(M + m), \tag{29.26}$$

где $\tau_0 = 2/(M + m)$.

Таким образом, итерационный процесс (29.3) при оптимальном значении итерационного параметра $\tau = \tau_0$ сходится тем быстрее, чем больше m/M, т. е. чем меньше разброс собственных чисел задачи (29.23).

Упражнение 29.1. Покажите, что если матрицы A, B эрмитовы и положительно определены, то итерационный метод (29.3) сходится при любом $\tau \in (0, 2/M)$.

7. При сделанных в предыдущем пункте предположениях о матрицах A, B удается получить оценки скорости сходимости сходимости итерационного метода (29.3).

Нам потребуются далее следующие вспомогательные построения[1)]. Пусть A — эрмитова неотрицательная матрица порядка n, $\{e^k\}_{k=1}^n$ — ортонормированная система ее собственных векторов: $Ae^k = \lambda_k e^k$. Очевидно, что $\lambda_k \geqslant 0$ при $k = 1, 2, \ldots, n$. Определим матрицу $A^{1/2}$ ее действием на векторы базиса $A^{1/2}e^k = \lambda_k^{1/2} e^k$, $k = 1, 2, \ldots, n$. Нетрудно убедиться, что матрица $A^{1/2}$ эрмитова

[1)] Более подробное изложение см. в [5], с. 236.

и неотрицательна, причем, если матрица A положительно определена, то и матрица $A^{1/2}$ положительно определена. Очевидно, что $(A^{1/2})^2 = A$. Полезно отметить также, что матрицы $A^{1/2}$ и A перестановочны, т. е. $A^{1/2}A = AA^{1/2}$. Матрицу $A^{1/2}$ называют *корнем квадратным* из матрицы A. Если матрица A положительно определена, то $(A^{1/2})^{-1} = (A^{-1})^{1/2}$. Будем использовать обозначение $(A^{-1})^{1/2} = A^{-1/2}$.

Теорема 6. *Пусть матрицы A, B эрмитовы и положительно определены. Тогда для приближений, построенных по итерационному методу* (29.3) *при $\tau = \tau_0$, справедливы следующие оценки:*

$$\|x^k - \hat{x}\|_A \leqslant \rho_0^k \|x^0 - \hat{x}\|_A, \quad k = 1, 2, \ldots \tag{29.27}$$

Доказательство. Используя (29.7), нетрудно убедиться, что

$$A^{1/2} z^{k+1} = (I - \tau_0 A^{1/2} B^{-1} A^{1/2}) A^{1/2} z^k,$$

следовательно,

$$\|z^{k+1}\|_A \leqslant \|(I - \tau_0 A^{1/2} B^{-1} A^{1/2})\|_2 \|z^k\|_A.$$

Матрица $I - \tau_0 A^{1/2} B^{-1} A^{1/2}$, очевидно, эрмитова. Поэтому

$$\|(I - \tau_0 A^{1/2} B^{-1} A^{1/2})\|_2 = \rho(I - \tau_0 A^{1/2} B^{-1} A^{1/2}).$$

Пусть y, λ — собственная пара матрицы $A^{1/2} B^{-1} A^{1/2}$, т. е.

$$A^{1/2} B^{-1} A^{1/2} y = \lambda y. \tag{29.28}$$

Матрица $B^{-1} A^{1/2}$ обратима. Полагая $y = A^{-1/2} B x$, получим, что собственные значения задачи (29.28) совпадают с собственными значениями задачи (29.23). Поэтому, проводя рассуждения полностью совпадающие с выполненными в п. 6, получим, что $\|z^{k+1}\|_A \leqslant \rho_0 \|z^k\|_A$, откуда, очевидно, следует (29.27). □

Упражнение 29.2. Покажите, что если выполнены условия теоремы 6, то справедливы оценки

$$\|x^k - \hat{x}\|_B \leqslant \rho_0^k \|x^0 - \hat{x}\|_B, \quad k = 1, 2, \ldots \tag{29.29}$$

30. Итерационные методы вариационного типа

В этом параграфе рассматривается задача о решении системы линейных алгебраических уравнений

$$Ax = b \tag{30.1}$$

с эрмитовой положительно определенной матрицей $A \in M_n$. Как было показано в предыдущем параграфе, для решения таких уравнений можно применять итерационные методы с оптимальным выбором параметров. Следует однако иметь в виду, что вычисление оптимальных значений параметров требует знания границ спектра некоторой вспомогательной задачи на собственные значения. Итерационные методы вариационного типа свободны от этого недостатка. Входящие в них параметры меняются от шага к шагу и вычисляются в ходе итерационного процесса. При этом методы автоматически настраиваются на оптимальную скорость сходимости.

1. Метод наискорейшего спуска. Пусть x^0 — некоторое начальное приближение к решению уравнения (30.1). Все последующие приближения будем вычислять по формуле

$$x^{k+1} = x^k - \tau_{k+1} r^k, \quad k = 0, 1, \ldots \tag{30.2}$$

Здесь и всюду далее

$$r^k = Ax^k - b. \tag{30.3}$$

Параметр $\tau_{k+1} \geqslant 0$ на каждом шаге итерационного метода будем выбирать так чтобы минимизировать энергетическую норму погрешности

$$\|x^{k+1} - \hat{x}\|_A = (A(x^{k+1} - \hat{x}), x^{k+1} - \hat{x})^{1/2}.$$

Нетрудно проверить, что $\|x^{k+1} - \hat{x}\|_A = \|r^{k+1}\|_{A^{-1}}$. Из (30.2) очевидным образом следует, что $r^{k+1} = r^k - \tau_{k+1} A r^k, \quad k = 0, 1, \ldots$ Таким образом, параметр τ_{k+1} должен быть выбран так, чтобы минимизировать величину $\|r^k - \tau_{k+1} A r^k\|_{A^{-1}}$. Элементарные выкладки дают:

$$\|r^k - \tau_{k+1} A r^k\|^2_{A^{-1}} = (A^{-1} r^k, r^k) - 2\tau_{k+1}(r^k, r^k) + \tau^2_{k+1}(Ar^k, r^k).$$

Отсюда получаем, что

$$\tau_{k+1} = \frac{(r^k, r^k)}{(Ar^k, r^k)}. \tag{30.4}$$

Формулы (30.2)–(30.4) полностью определяют метод наискорейшего спуска.

2. Исследуем сходимость метода наискорейшего спуска

Теорема 1. *Если A — эрмитова положительно определенная матрица, то метод* (30.2)–(30.4) *сходится при любом начальном приближении x^0. Имеет место следующая оценка скорости сходимости*

$$\|x^k - \hat{x}\|_A \leqslant \rho_0^k \|x^0 - \hat{x}\|_A, \quad k = 1, 2, \ldots, \tag{30.5}$$

где $\rho_0 = (M-m)/(M+m) = (\mathrm{cond}(A)-1)/(\mathrm{cond}(A)+1)$, M, m *— максимальное и минимальное собственные числа матрицы* A *соответственно.*

Доказательство. Из соотношения (30.2) и способа определения параметра τ_{k+1} вытекает, что

$$\|z^{k+1}\|_A = \|z^k - \tau_{k+1}Az^k\|_A \leqslant \|z^k - \tau_0 Az^k\|_A,$$

где $\tau_0 = 2/(M+m)$. Далее, как при доказательстве теоремы 6, с. 94 (полагая $B = I$), получим, что $\|z^k - \tau_0 Az^k\|_A \leqslant \rho_0\|z^k\|_A$. □

3. В случае, когда матрица A и вектор b вещественны, метод наискорейшего спуска допускает простую интерпретацию, оправдывающую его название. Пусть $x \in \mathbb{R}^n$,

$$\nabla F(x) = \left(\frac{\partial F(x)}{\partial x_1}, \frac{\partial F(x)}{\partial x_2}, \ldots, \frac{\partial F(x)}{\partial x_n}\right)$$

есть градиент дифференцируемой функции F в точке x. Элементарные вычисления дают, что для функции F, определенной равенством (29.16), с. 90, $\nabla F(x) = 2(Ax - b)$. Отсюда следует (см. (30.2)), что при любом $\tau_{k+1} > 0$ точка x^{k+1} лежит на луче, проходящем через точку x^k в направлении наибыстрейшего убывания функции F (см. курс математического анализа). Описанный выше способ определения параметра τ_{k+1} означает, что точка x^{k+1} есть точка минимума функции F на указанном луче (см. (29.17), с. 91).

4. Как показывает оценка (30.5), сходимость метода наискорейшего спуска существенно замедляется с ухудшением обусловленности матрицы A. Однако на практике довольно часто удается исправить положение, переходя к решению эквивалентной системы, матрица которой имеет меньшее число обусловленности.

Пусть B — эрмитова положительно определенная матрица. Преобразуем систему (30.1) к следующему виду:

$$B^{-1/2}AB^{-1/2}B^{1/2}x = B^{-1/2}b.$$

Полагая

$$C = B^{-1/2}AB^{-1/2}, \quad y = B^{1/2}x, \quad f = B^{-1/2}b, \tag{30.6}$$

получим

$$Cy = f. \tag{30.7}$$

Матрица C эрмитова и положительно определена (докажите!).

Запишем формулы метода наискорейшего спуска и соответствующую оценку погрешности применительно к уравнению (30.7):

$$y^{k+1} = y^k - \tau_{k+1}(Cy^k - f), \quad k = 0, 1, \ldots, \tag{30.8}$$

$$\tau_{k+1} = \frac{(r_C^k, r_C^k)}{(Cr_C^k, r_C^k)}, \tag{30.9}$$

где

$$r_C^k = Cy^k - f, \tag{30.10}$$

$$\|y^k - \hat{y}\|_C \leqslant \rho_0^k(C)\|y^0 - \hat{y}\|_C, \quad k = 1, 2, \ldots, \tag{30.11}$$

$\rho_0(C) = (M_C - m_C)/(M_C + m_C) = (\mathrm{cond}(C) - 1)/(\mathrm{cond}(C) + 1)$, M_C, m_C — максимальное и минимальное собственные числа матрицы C соответственно.

На первый взгляд, полученные формулы кажутся практически бесполезными, так как матрица C и вектор f не определены конструктивно. Тем не менее, попробуем эти формулы преобразовать.

Умножим обе части уравнения (30.8) на $B^{-1/2}$ и воспользуемся затем равенствами (30.6). В результате получим

$$x^{k+1} = x^k - \tau_{k+1}w^k, \quad k = 0, 1, \ldots, \tag{30.12}$$

где $w^j = B^{-1}r^j$, $r^j = Ax^j - b$, $x^j = B^{-1/2}y^j$, $j = 0, 1, \ldots$ Далее, преобразуем формулу (30.13) с использованием (30.6), (30.10). Получим

$$\tau_{k+1} = \frac{(w^k, r^k)}{(Aw^k, w^k)}, \quad k = 0, 1, \ldots \tag{30.13}$$

Наконец, аналогичные преобразования приводят оценку (30.11) к виду

$$\|x^k - \hat{x}\|_A \leqslant \rho_0^k(C)\|x^0 - \hat{x}\|_A, \quad k = 1, 2, \ldots \tag{30.14}$$

Вычисления по формулам (30.12), (30.13) обычно проводятся следующим образом. Сначала по известному приближению x^k вычисляется невязка $r^k = Ax^k - b$, затем путем решения уравнения

$$Bw^k = r^k \tag{30.15}$$

находится так называемая поправка w^k, при помощи формулы (30.13) вычисляется итерационный параметр τ_{k+1}, и, наконец, по формуле (30.12) находится следующее приближение x^{k+1}. Важно подчеркнуть, что каждый шаг полученного итерационного метода требует решения системы линейных алгебраических уравнений вида (30.15).

Как показывает оценка (30.14), скорость сходимости метода определяется собственными числами задачи $Cy = \lambda y$. Более подробная ее запись с использованием (30.6) дает $B^{-1/2}AB^{-1/2}y = \lambda y$. Матрица $B^{-1/2}$ обратима, поэтому, полагая $B^{-1/2}y = x$, приходим к эквивалентной задаче $Ax = \lambda Bx$, уже рассматривавшейся нами ранее (см. (29.23), с. 92).

Выбор матрицы B должен быть подчинен двум противоречивым требованиям. С одной стороны, матрица B в определенном смысле должна быть близка к матрице A, так как отношение M_C/m_C должно быть, как можно, ближе к единице. С другой стороны, матрица B должна быть существенно проще матрицы A, чтобы решение системы уравнений вида (30.15) было намного менее трудоемким, чем решение системы уравнений с матрицей A.

Сравнение оценок (29.27) и (30.14) показывает, что метод наискорейшего спуска сходится не медленнее, чем метод (29.3) при оптимальном выборе итерационного параметра.

Метод, описанный в настоящем пункте, часто называют *предобусловленным* методом наискорейшего спуска.

5. Метод сопряженных градиентов. Этот метод решения уравнения (30.1) можно рассматривать как непосредственное обобщение метода наискорейшего спуска: по заданному начальному приближению $x^0 \in \mathbb{C}^n$ строятся векторы $x^1, x^2, \ldots$ при помощи соотношений

$$x^k = x^0 - \sum_{j=1}^{k} \alpha_j^{(k)} A^{j-1}(Ax^0 - b), \quad k = 1, 2, \ldots; \tag{30.16}$$

при каждом k числа $\alpha_j^{(k)}$, $j = 1, 2, \ldots, k$, определяются так, чтобы норма погрешности $\|x^k - \hat{x}\|_A$ принимала минимальное значение.

Теорема 2. *При каждом $k = 1, 2, \ldots$ приближение x^k, определенное указанным выше способом, существует, и, более того, оно единственно.*

Доказательство. Применяя те же обозначения, что и в предыдущих пунктах, из (30.16) получим, что

$$r^k = r^0 - \sum_{j=1}^{k} \alpha_j^{(k)} A^j r^0, \quad k = 1, 2, \ldots \tag{30.17}$$

Отсюда следует, что

$$\|x^k - \hat{x}\|_A = \|r^k\|_{A^{-1}} = \|r^0 - \sum_{j=1}^{k} \alpha_j^{(k)} A^j r^0\|_{A^{-1}}, \quad k = 1, 2, \ldots \quad (30.18)$$

Таким образом, можно считать, что параметры $\alpha_j^{(k)}$, $j = 1, 2, \ldots, k$, выбираются из условия минимума нормы, записанной в правой части равенства (30.18). Иными словами, разыскивается элемент наилучшего приближения к r^0 в смысле нормы $\|\cdot\|_{A^{-1}}$ в подпространстве, натянутом на векторы Ar^0, A^2r^0, $\ldots$, A^kr^0. Как известно (см. § 3, с. 148, [5]), такой элемент существует и определяется однозначно. Поэтому и вектор r^k определяется при помощи соотношения (30.17) однозначно. Зная r^k, вектор x^k определим однозначно при помощи (30.3). □

6. Теорема 2 гарантирует существование приближений по методу сопряженных градиентов, но не указывает эффективного способа их вычисления. Построение соответствующих расчетных формулы основано на устанавливаемых ниже свойствах последовательности невязок r^k, $k = 0, 1, \ldots$

Лемма 1. *При любом $k = 1, 2, \ldots$ выполнены равенства*

$$(r^k, A^j r^0) = 0, \quad j = 0, 1, \ldots, k-1. \quad (30.19)$$

Доказательство. Как отмечалось в ходе доказательства теоремы 2, вектор $\sum_{j=1}^{k} \alpha_j^{(k)} A^j r^0$ есть ортогональная проекция вектора r^0 в смысле скалярного произведения $(\cdot, \cdot)_{A^{-1}}$ на подпространство, натянутое на векторы $A^j r^0$, $j = 1, 2, \ldots, k$. Поэтому (см. § 3, с. 148, [5]) для вектора r^k, определяемого формулой (30.17), выполнены равенства

$$(r^k, A^j r^0)_{A^{-1}} = 0, \quad j = 1, 2, \ldots, k,$$

эквивалентные (30.19). □

Следствие 1. *При любом $k = 1, 2, \ldots$ выполнены равенства*

$$(r^k, r^j) = 0, \quad j = 0, 1, \ldots, k-1, \quad (30.20)$$

$$(Ar^k, r^j) = 0, \quad j = 0, 1, \ldots, k-2. \quad (30.21)$$

Доказательство. Для того, чтобы получить (30.20), запишем r^j по формуле (30.17), а затем воспользуемся (30.19). Вследствие эрмитовости матрицы A и равенства (30.17) будем иметь

$$(Ar^k, r^j) = (r^k, Ar^j) = (r^k, Ar^0 - \sum_{l=1}^{j} \alpha_l^{(j)} A^{l+1} r^0).$$

На основании (30.19) отсюда получаем (30.21). □

Замечание 1. Говорят, что ненулевые векторы $x, y \in \mathbb{C}^n$ сопряженны относительно эрмитовой матрицы A, если $(Ax, y) = 0$. Вектор r^j пропорционален градиенту функционала F в точке $x^j \in \mathbb{R}^n$, если матрица A и вектор b вещественны (см. п. 3, с. 96). Равенства (30.21) оправдывают, таким образом, название изучаемого здесь метода.

Лемма 2. *Пусть*

$$x = e^1 + e^2 + \cdots + e^p, \tag{30.22}$$

где $e^1, e^2, \ldots, e^p$ собственные векторы эрмитовой матрицы A, отвечающие всем попарно различным собственным числам $\lambda_1, \lambda_2, \ldots, \lambda_p$ этой матрицы[1)]*. Тогда векторы x, Ax, $\ldots$, $A^{p-1}x$ линейно независимы.*

Доказательство. Пусть существуют $c_0, c_1, \ldots, c_{p-1}$ такие, что $c_0 x + c_1 Ax + \cdots + c_{p-1} A^{p-1} x = 0$. Используя (30.22), после элементарных преобразований отсюда получим, что $\sum\limits_{k=1}^{p} \sum\limits_{j=0}^{p-1} c_j \lambda_k^j e^k = 0$. По теореме 2, с. 186, [5] векторы $e^1, e^2, \ldots, e^p$ линейно независимы, следовательно,

$$\sum_{j=0}^{p-1} c_j \lambda_k^j = 0, \quad k = 1, 2, \ldots, p. \tag{30.23}$$

Равенства (30.23) можно рассматривать как систему линейных уравнений относительно $c_0, c_1, \ldots, c_{p-1}$. Определитель системы (30.23) есть определитель Вандермонда. Он отличен от нуля, поскольку по условию теоремы все числа $\lambda_1, \lambda_2, \ldots, \lambda_p$ попарно различны. Таким образом, система (30.23) может иметь только тривиальное решение. □

Лемма 3. *Пусть*

$$r^0 = e^1 + e^2 + \cdots + e^p, \tag{30.24}$$

[1)] Согласно (9.1), с. 227, [5] любой вектор из $\mathbb{C}^n$ может быть представлен в указанном виде.

где $e^1, e^2, \ldots, e^p$ *— собственные векторы матрицы* A*, отвечающие ее попарно различным собственным числам. Тогда* $r^p = 0$*, и* $r^k \neq 0$ *при* $k = 1, 2, \ldots, p-1$.

Доказательство. По лемме 2 векторы

$$r^0, Ar^0, \ldots, A^{p-1}r^0 \tag{30.25}$$

линейно независимы, поэтому векторы r^k, определяемые соотношениями (30.17), при $k < p$ не могут равняться нулю. Пусть теперь $k = p$. Обозначим через L_p подпространство пространства $\mathbb{C}^n$, натянутое на векторы $e^1, e^2, \cdots, e^p$. Очевидно, что $\dim L_p = p$. Векторы (30.25) принадлежат L_p и образуют его базис. Вследствие обратимости матрицы A векторы $Ar^0, A^2r^0, \ldots, A^pr^0$, также образуют базис в L_p. По построению вектор $\sum\limits_{j=1}^{p} \alpha_j^{(p)} A^j r^0$ есть элемент наилучшего приближения к вектору $r^0 \in L_p$, поэтому $r^p = r^0 - \sum\limits_{j=1}^{p} \alpha_j^{(p)} A^j r^0 = 0$. □

Следствие 2. *При любом начальном приближении* $x^0 \in \mathbb{C}^n$ *метод сопряженных градиентов дает точное решение системы* (30.1) *не больше чем через* n *итераций* .

Справедливость этого утверждения сразу же вытекает из того факта, что любой вектор r^0 пространства $\mathbb{C}^n$ представим в виде (30.24) при некотором $p \leqslant n$ (см. сноску на предыдущей странице).

Замечание 2. Метод сопряженных градиентов может трактоваться, таким образом, и как прямой метод решения систем линейных алгебраических уравнений. Однако этот вывод справедлив лишь при отсутствии ошибок округления, что при реальных вычислениях недостижимо. На практике метод сопряженных градиентов используется исключительно как итерационный.

Лемма 4. *Пусть выполнены условия леммы* 3*. Тогда* $\alpha_k^{(k)} \neq 0$ *при любом* $k = 1, 2, \ldots, p$.

Доказательство. Предположим, что $\alpha_p^{(p)} = 0$. Тогда из (30.17) при $k = p$ получаем, что

$$r^0 - \sum_{j=1}^{p-1} \alpha_j^{(p)} A^j r^0 = 0,$$

а это противоречит линейной независимости системы векторов (30.25).

Если $\alpha_k^{(k)} = 0$ при некотором $k < p$, то из (30.17) вытекает, что

$$r^k = r^0 - \sum_{j=1}^{k-1} \alpha_j^{(k)} A^j r^0. \tag{30.26}$$

По лемме 3 имеем, что $r^k \neq 0$, по лемме 1 для r^k выполнены соотношения (30.19). Получили, что ненулевой вектор r^k одновременно является линейной комбинацией некоторого набора векторов и ортогонален к каждому из векторов этого набора, чего быть не может. □

Лемма 5. *Пусть выполнены условия леммы* 3. *Тогда при любом* $k \leqslant p-1$ *векторы*

$$r^0, r^1, \ldots, r^{k-1}, Ar^{k-1} \tag{30.27}$$

образуют базис в подпространстве S_{k+1}, *натянутом на векторы* r^0, $Ar^0, \ldots, A^k r^0$.

Доказательство. Принадлежность векторов системы (30.27) подпространству S_{k+1} сразу же вытекает из равенств (30.17). Осталось доказать их линейную независимость. По следствию 1 векторы $r^0, r^1, \ldots, r^{k-1}$ линейно независимы. В силу (30.17) они принадлежат S_k. Вектор Ar^{k-1} можно записать в виде

$$Ar^{k-1} = Ar^0 - \sum_{j=1}^{k-2} \alpha_j^{(k-1)} A^{j+1} r^0 + \alpha_{k-1}^{(k-1)} A^k r^0,$$

причем по лемме 4 величина $\alpha_{k-1}^{(k-1)}$ отлична от нуля. Отсюда вытекает, что вектор Ar^{k-1} не принадлежит S_k. Таким образом, векторы (30.27) линейно независимы. □

Из леммы 5 непосредственно вытекает

Следствие 3. *При любом* $k \leqslant p-1$ *существуют и однозначно определены числа* $\gamma_0^{(k)}, \gamma_1^{(k)}, \ldots, \gamma_k^{(k)}$ *такие, что*

$$r^k = \sum_{j=0}^{k-1} \gamma_j^{(k)} r^j + \gamma_k^{(k)} Ar^{k-1}. \tag{30.28}$$

7. Покажем теперь, как можно вычислить коэффициенты в разложении (30.28).

Пусть $k = 1$. Тогда $r^1 = \gamma_0^{(1)} r^0 + \gamma_1^{(1)} A r^0$. С другой стороны, по формуле (30.17) получаем, что

$$r^1 = r^0 - \alpha_1^{(1)} A r^0. \tag{30.29}$$

Векторы r^0, Ar^0 линейно независимы, следовательно, $\gamma_0^{(1)} = 1$. Используя теперь равенство $(r^1, r^0) = 0$ (см. (30.20)), получим, что

$$\gamma_1^{(1)} = -\frac{(r^0, r^0)}{(Ar^0, r^0)}. \tag{30.30}$$

Из (30.3), (30.29), (30.30), очевидно, вытекает, что

$$x^1 = x^0 - \frac{(r^0, r^0)}{(Ar^0, r^0)} r^0. \tag{30.31}$$

т. е. первое приближение по методу сопряженных градиентов, как и следовало ожидать, совпадает с первым приближением по методу наискорейшего спуска.

Далее, пусть $k > 1$. При любом $l < k$ вследствие (30.28), (30.20) получаем, что

$$0 = \sum_{j=0}^{k-1} \gamma_j^{(k)} (r^j, r^l) + \gamma_k^{(k)} (Ar^{k-1}, r^l). \tag{30.32}$$

Из (30.32), используя (30.20), (30.21), найдем, что

$$\gamma_0^{(k)}, \gamma_1^{(k)}, \ldots, \gamma_{k-3}^{(k)} = 0.$$

Поэтому (см. (30.28))

$$r^k = \gamma_{k-2}^{(k)} r^{k-2} + \gamma_{k-1}^{(k)} r^{k-1} + \gamma_k^{(k)} A r^{k-1}. \tag{30.33}$$

Запишем r^{k-2}, r^{k-1} по формуле (30.17) и подставим в (30.33). Получим, что

$$r^k = (\gamma_{k-2}^{(k)} + \gamma_{k-1}^{(k)}) r^0 + \sum_{j=1}^{k} \delta_j^{(k)} A^j r^0 \tag{30.34}$$

с некоторыми коэффициентами $\delta_j^{(k)}$. Сравнивая (30.34) с (30.17) и используя линейную независимость системы векторов (30.25), будем иметь, что $\gamma_{k-2}^{(k)} + \gamma_{k-1}^{(k)} = 1$. Таким образом, соотношению (30.33) можно придать следующий вид:

$$r^k = \alpha_k r^{k-1} + (1 - \alpha_k) r^{k-2} + \beta_k A r^{k-1}. \tag{30.35}$$

Осталось найти числа α_k, β_k. Умножая равенство (30.35) почленно сначала на r^{k-1}, а затем на r^{k-2}, получим систему линейных уравнений для их определения:

$$\begin{aligned} \alpha_k(r^{k-1}, r^{k-1}) + \beta_k(Ar^{k-1}, r^{k-1}) &= 0, \\ -\alpha_k(r^{k-2}, r^{k-2}) + \beta_k(Ar^{k-1}, r^{k-2}) &= -(r^{k-2}, r^{k-2}). \end{aligned} \tag{30.36}$$

Построенные нами формулы позволяют последовательно вычислить все приближения по методу сопряженных градиентов. В самом деле, используя соотношения (30.3) и умножая обе части равенства (30.35) на A^{-1}, в дополнение к (30.31), (30.36) будем иметь

$$x^k = \alpha_k x^{k-1} + (1 - \alpha_k)x^{k-2} + \beta_k r^{k-1} \tag{30.37}$$

при $k = 2, 3, \ldots$

Важно подчеркнуть, что в отличие от всех рассмотренных нами ранее итерационных методов метод сопряженных градиентов требует при вычислении каждого последующего приближения x^k, $k \geqslant 2$, знания не одного, а двух предыдущих приближений, т. е. x^{k-1}, x^{k-2}. Это предъявляет дополнительные требования к памяти компьютера.

8. Известны и другие варианты расчетных формул для построения последовательности x^2, x^3, $\ldots$, различающиеся объемом необходимых вычислений и требуемой памятью компьютера. При отсутствии ошибок округлений все они в силу теоремы 2 эквивалентны, но при решении той или иной конкретной системы уравнений результаты их работы могут различаться (иногда значительно) именно из-за неизбежных ошибок округления.

Приведем пример расчетных формул метода сопряженных градиентов, отличных от (30.36), (30.37):

$$\alpha_1 = 1, \quad \tau_1 = (r^0, r^0)/(Ar^0, r^0),$$

$$\tau_k = \frac{(r^{k-1}, r^{k-1})}{(Ar^{k-1}, r^{k-1})}, \quad \alpha_k = \left(1 - \frac{\tau_k}{\tau_{k-1}} \frac{(r^{k-1}, r^{k-1})}{(r^{k-2}, r^{k-2})} \frac{1}{\alpha_{k-1}}\right)^{-1}, \tag{30.38}$$

$$x^k = \alpha_k(x^{k-1} - \tau_k r^{k-1}) + (1 - \alpha_k)x^{k-2}, k = 2, 3, \ldots. \tag{30.39}$$

Упражнение 30.1. Получите формулы (30.38), (30.39) и интерпретируйте их аналогично п. 3, с. 96.

Указания. Положите $\beta_k = -\alpha_k\tau_k$. Исключите (Ar^{k-1}, r^{k-2}) из второго уравнения (30.36), используя равенство (30.35), записанное для r^{k-1}. Сопоставьте (30.38), (30.39) с (30.2)–(30.4).

Рекуррентные формулы (30.38) экономичнее формул (30.36), так как не требуют вычисления скалярного произведения (Ar^{k-1}, r^{k-2}).

Чаще всего метод сопряженных градиентов реализуют в виде следующего алгоритма По заданному вектору x^0 находят

$$r^0 = b - Ax^0,\ \rho_0 = (r^0, r^0), \tag{30.40}$$

$$p^1 = r^0,\ q^1 = Ap^1,\ \alpha_1 = \rho_0/(p^1, q^1), \tag{30.41}$$

$$x^1 = x^0 + \alpha_1 p^1, \quad r^1 = r^0 - \alpha_1 q^1. \tag{30.42}$$

Затем для $i = 2, 3, \ldots$ последовательно вычисляют

$$\begin{aligned} &\rho_{i-1} = (r^{i-1}, r^{i-1}),\\ &\beta_{i-1} = \rho_{i-1}/\rho_{i-2},\\ &p^i = r^{i-1} + \beta_{i-1} p^{i-1},\\ &q^i = Ap^i,\\ &\alpha_i = \rho_{i-1}/(p^i, q^i),\\ &x^i = x^{i-1} + \alpha_i p^i,\\ &r^i = r^{i-1} - \alpha_i q^i. \end{aligned} \tag{30.43}$$

Как уже отмечалось, все описанные выше варианты реализации метода сопряженных градиентов при отсутствии ошибок округления приводят к одной и той же последовательности приближений x^0, x^1, x^2, ... Однако в реальных вычислениях они различаются по степени устойчивости к ошибкам округления. Опыт показывает, что предпочтительнее использовать алгоритм (30.40)–(30.43).

9. Переходим к оценке скорости сходимости метода сопряженных градиентов.

Теорема 3. *При любом* $k \geqslant 1$

$$\|x^k - \hat{x}\|_A \leqslant q_k \|x^0 - \hat{x}\|_A. \tag{30.44}$$

Здесь

$$q_k = \frac{2\rho_1^k}{1 + \rho_1^{2k}}, \quad \rho_1 = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}},$$

M, m — максимальное и минимальное собственные числа матрицы A соответственно.

Доказательство. Представим равенства (30.16) в виде

$$z^k = P_k(A) z^0, \quad k = 1, 2, \ldots, \tag{30.45}$$

где $P_k(A) = I - \sum\limits_{j=1}^{k} \alpha_j^{(k)} A^j$. Отсюда следует, что[1)]

$$A^{1/2} z^k = P_k(A) A^{1/2} z^0, \quad k = 1, 2, \ldots \tag{30.46}$$

Числа $\alpha_j^{(k)}$, $k = 1, 2, \ldots$, $j = 1, 2, \ldots, k$, определяющие приближения по методу сопряженных градиентов, таковы, что

$$\|z^k\|_A = \|A^{1/2} z^k\|_2 = \|P_k(A) A^{1/2} z^0\|_2 \leqslant \|Q_k(A) A^{1/2} z^0\|_2, \tag{30.47}$$

где $Q_k(A) = I - \sum\limits_{j=1}^{k} \beta_j^{(k)} A^j$, а $\beta_j^{(k)}$, $j = 1, 2, \ldots, k$, — какие угодно числа. Из (30.47) получаем, что

$$\|z^k\|_A \leqslant \|Q_k(A)\|_2 \|z^0\|_A. \tag{30.48}$$

Будем считать, что все $\beta_j^{(k)}$, $j = 1, 2, \ldots, k$, вещественны. Тогда матрица $Q_k(A)$ эрмитова, следовательно,

$$\|Q_k(A)\|_2 = \max_{1 \leqslant i \leqslant n} |Q_k(\lambda_i(A))| \leqslant \max_{m \leqslant \mu \leqslant M} |Q_k(\mu)| \tag{30.49}$$

(см. (23.5), с. 73). Полином Q_k имеет степень k, причем $Q_k(0) = 1$. Понятно, что числа $\beta_j^{(k)}$, $j = 1, 2, \ldots, k$, можно выбрать так, что

$$Q_k(\mu) = \frac{T_k\left(\frac{\mu\tau_0 - 1}{\rho_0}\right)}{T_k\left(-\frac{1}{\rho_0}\right)} \qquad \forall\, \mu \in \mathbb{R}, \tag{30.50}$$

где T_k — полином Чебышева порядка k (см. п. 2.2, с. 141, [5]). а величины ρ_0, τ_0 определены так же, как в теореме 1, с. 95. Нетрудно проверить, что $|(\mu\tau_0 - 1)/\rho_0| \leqslant 1$ при $\mu \in [m, M]$. Поэтому (см. формулу (10.10), с. 142, [5])

$$\max_{m \leqslant \mu \leqslant M} |Q_k(\mu)| = \frac{1}{\left|T_k\left(-\frac{1}{\rho_0}\right)\right|}. \tag{30.51}$$

Поскольку $\rho_0 < 1$, для вычисления $T_k(-1/\rho_0)$ нужно использовать формулу предшествующую (10.10), с. 142, [5]. В результате, будем иметь, что

$$\max_{m \leqslant \mu \leqslant M} |Q_k(\mu)| = q_k. \tag{30.52}$$

Из (30.48), (30.49), (30.52) следует (30.44). □

[1)] Мы использовали тот факт, что корень из эрмитовой неотрицательной матрицы A перестановочен с матрицей A (см. с. 93).

10. Оценка (30.44) показывает, что метод сопряженных градиентов сходится существенно быстрее метода наискорейшего спуска. Преимущество становится особенно заметным с увеличением отношения M/m, т. е. с ухудшением обусловленности матрицы A.

11. Оценка (30.44) не может быть улучшена, в том смысле, что не существует полинома Q_k степени k, равного единице в нуле, и такого, что $\max\limits_{m\leqslant\mu\leqslant M}|Q_k(\mu)|<q_k$. Предполагая противное, мы должны написать, что

$$\max_{m\leqslant\mu\leqslant M}|Q_k(\mu)|<\max_{m\leqslant\mu\leqslant M}\frac{\left|T_k\left(\frac{\mu\tau_0-1}{\rho_0}\right)\right|}{\left|T_k\left(-\frac{1}{\rho_0}\right)\right|} \tag{30.53}$$

для некоторого полинома Q_k степени k такого, что $Q_k(0)=1$. Выполним в неравенстве (30.53) замену переменной, полагая

$$t=\frac{\mu\tau_0-1}{\rho_0}.$$

В результате, как нетрудно убедиться, получим, что

$$\max_{-1\leqslant t\leqslant 1}|\tilde{Q}_k(t)|<\max_{-1\leqslant t\leqslant 1}|\tilde{T}_k(t)|, \tag{30.54}$$

где $\tilde{Q}_k(t)=Q_k((\rho_0 t+1)/\tau_0)$, $\tilde{T}_k(t)=T_k(t)/T_k\left(-1/\rho_0\right)$. Пусть

$$t_j=\cos\frac{\pi j}{k},\quad j=0,1,\ldots,k.$$

Используя формулу (10,11), с. 127, [1], будем иметь, что

$$|\tilde{T}_k(t_j)|=\max_{-1\leqslant t\leqslant 1}|\tilde{T}_k(t)|,\quad j=0,1,\ldots,k,$$

причем при любых $j=0,1,\ldots,k-1$ знаки величин $\tilde{T}_k(t_j)$ и $\tilde{T}_k(t_{j+1})$ противоположны. Введем в рассмотрение полином R_k степени не выше k, полагая $R_k(t)=\tilde{Q}_k(t)-\tilde{T}_k(t)$. Нетрудно сообразить, что при любых $j=0,1,\ldots,k-1$ знаки величин $R_k(t_j)$ и $R_k(t_{j+1})$ также противоположны. Это означает, что полином R_k имеет на интервале $(-1,1)$ не менее k корней. Кроме того, $R_k(-1/\rho_0)=0$, а $-1/\rho_0<-1$. Таким образом, полином R_k имеет не менее $k+1$ корней, что невозможно.

Задача. Аналогично п. 4, с. 96, опишите и исследуйте предобусловленный вариант метода сопряженных градиентов.

Глава 7

Алгебраическая проблема собственных значений

Под алгебраической проблемой собственных значений понимают задачу отыскания собственных чисел и собственных векторов матрицы. Различают полную проблему собственных значений, т. е. нахождение всех собственных чисел и собственных векторов, и частичную проблему собственных значений, т. е. отыскание лишь некоторых собственных чисел и соответствующих им собственных векторов.

Понятно, что методы решения частичной проблемы собственных значений должны быть более простыми. Мы рассмотрим примеры методов обоих классов.

31. Методы прямой и обратной итераций

1. Метод прямой итерации. Этот метод предназначен для отыскания максимального по модулю собственного числа матрицы и соответствующего ему собственного вектора.

Выберем некоторое нормированное начальное приближение y^0 и образуем последовательность нормированных векторов $y^1, y^2, \ldots$ Именно, для $k = 0, 1, \ldots$ вычисляем $x^{k+1} = Ay^k$, $y^{k+1} = x^{k+1}/|x^{k+1}|$. Строим также последовательность чисел $\lambda^{(k)} = (Ay^k, y^k)$, $k = 1, 2, \ldots$

На протяжении этого параграфа предполагаем, что матрица A эрмитова. Собственные числа с матрицы A будем нумеровать в порядке неубывания их модулей: $|\lambda_1| \leqslant |\lambda_2| \leqslant \cdots \leqslant |\lambda_{n-1}| \leqslant |\lambda_n|$. Через $e^1, e^2, \ldots, e^n$ будем обозначать соответствующие ортонормированные собственные векторы. По теореме об ортогональном разложении евклидова пространства при любом $k = 0, 1, 2, \ldots$

$$y^k = c_k e^n + s_k u^k, \tag{31.1}$$

где $(e^n, u^k) = 0$, $|u^k| = 1$, $c_k = (y^k, e^n)$, $s_k = (y^k, u^k)$, $|c_k|^2 + |s_k|^2 = 1$. Далее будем использовать обозначение $t_k = s_k/c_k$. Чем меньше $|t_k|$, тем ближе y^k по направлению к собственному вектору матрицы A, отвечающему λ_n.

Полезно отметить, что случае вещественной матрицы A числа c_k, s_k вещественны, поэтому $c_k = \cos\varphi_k$, $s_k = \sin\varphi_k$, $t_k = \operatorname{tg}\varphi_k$, где φ_k

можно интерпретировать как угол, образованный векторами e_n, y_k (сделайте рисунок!).

Теорема 1. *Пусть* $|\lambda_{n-1}| < |\lambda_n|$, $c_0 \neq 0$. *Тогда* $t_k \to 0$, $\lambda^{(k)} \to \lambda_n$ *при* $k \to \infty$. *Справедливы следующие оценки скорости сходимости:*

$$|t_k| \leqslant \left|\frac{\lambda_{n-1}}{\lambda_n}\right|^k |t_0|, \quad |\lambda^{(k)} - \lambda_n| \leqslant (|\lambda_{n-1}| + |\lambda_n|) \left|\frac{\lambda_{n-1}}{\lambda_n}\right|^{2k}. \qquad (31.2)$$

Доказательство. В соответствии с изучаемым алгоритмом

$$y^{k+1} = \alpha(c_k A e^n + s_k A u^k) = \alpha(c_k \lambda_n e^n + s_k A u^k).$$

Здесь α — число, выбираемое так, чтобы вектор y^{k+1} имел единичную длину. Перепишем последнее равенство в виде

$$y^{k+1} = \alpha(c_k\, \lambda_n e^n + |Au^k| s_k\, Au^k/|Au^k|\,) = \\ = c_{k+1} e^n + s_{k+1} Au^k/|Au^k|. \quad (31.3)$$

Мы учли здесь, что вектор Au^k ортогонален e^n, поскольку

$$(Au^k, e^n) = (u^k, Ae^n) = \lambda_n (u^k, e^n) = 0.$$

Из (31.3) вытекает, что

$$t_{k+1} = \frac{|Au^k|}{\lambda_n} t_k.$$

Записывая разложение u^k по базису собственных векторов матрицы A и учитывая ортогональность u^k и e^n, получим, что $Au^k = \sum\limits_{i=1}^{n-1} \alpha_i \lambda_i e^i$, следовательно,

$$|Au^k|^2 = \sum_{i=1}^{n-1} \alpha_i^2 \lambda_i^2 \leqslant \lambda_{n-1}^2 \sum_{i=1}^{n-1} c_i^2 = \lambda_{n-1}^2 |u^k|^2 = \lambda_{n-1}^2. \qquad (31.4)$$

Таким образом,

$$|t_{k+1}| \leqslant \frac{|\lambda_{n-1}|}{|\lambda_n|} |t_k|,$$

и первая оценка (31.2) доказана.

Вновь используя представление (31.1) и то, что $(Au^k, e^n) = 0$, получим:

$$\lambda_n - \lambda^{(k)} = \lambda_n - (A(c_k\, e^n + s_k\, u^k), c_k\, e^n + s_k\, u^k) =$$
$$= \lambda_n - (c_k\, \lambda_n e^n + s_k\, Au^k, c_k\, e^n + s_k\, u^k) =$$
$$= \lambda_n - (\lambda_n |c_k|^2 + (Au^k, u^k)|s_k|^2) = (\lambda_n - (Au^k, u^k))|s_k|^2.$$

Отсюда вследствие (31.4) вытекает, что

$$|\lambda_n - \lambda^{(k)}| \leqslant |\lambda_n + \lambda_{n-1}||s_k|^2 \leqslant |\lambda_n + \lambda_{n-1}||t_k|^2,$$

Вместе с уже полученной первой оценкой (31.2) это завершает доказательство теоремы. □

Условие $c_0 \neq 0$ на практике не слишком обременительно. Если оно нарушается, то при проведении итераций за счет ошибок округления приближения обязательно выйдут из гиперплоскости, ортогональной e^n.

2. Метод обратной итерации. Метод предназначен для отыскания минимального по модулю собственного числа и соответствующего ему собственного вектора и состоит в следующем: выбираем нормированное начальное приближение y^0 и строим последовательность векторов $y^1, y^2, \ldots$ по формулам: $x^{k+1} = A^{-1}y^k$, $y^{k+1} = x^{k+1}/|x^{k+1}|$, а также числа $\lambda^{(k)} = (Ay^k, y^k)$, $k = 0, 1, 2, \ldots$

При реализации метода выгоднее не строить и хранить матрицу A^{-1}, а решать на каждой итерации систему линейных уравнений $Ax^{k+1} = y^k$. Предварительно целесообразно представить матрицу A в виде LU или QR разложения (см. гл. 2).

Относительно сходимости метода справедлива теорема, полностью аналогичная теореме 1, но на этот раз скорость сходимости характеризуется отношением $|\lambda_1|/|\lambda_2| < 1$.

2.1. Метод обратной итерации со сдвигом. Рассмотрим обобщение предыдущего метода, а именно переход от вектора y^k к y^{k+1} будем выполнять по формулам: $(A-\sigma I)x^{k+1} = y^k$, $y^{k+1} = x^{k+1}/|x^{k+1}|$. Здесь σ — параметр, называемый сдвигом. Последовательность $\lambda^{(k)}$, $k = 1, 2, \ldots$, по-прежнему, определяется формулой $\lambda^{(k)} = (Ay^k, y^k)$. Сходимость этого метода исследуется по той же схеме, что и в теореме 1. При этом оказывается, что $\lambda^{(k)} \to \lambda_j$, где номер j характеризуется условием $|\lambda_j - \sigma| < |\lambda_i - \sigma| \quad \forall\, i \neq j$, а последовательность y^k сходится к соответствующему собственному вектору e^j. Таким образом, метод позволяет находить собственное число матрицы A, ближайшее к заданному числу σ.

32. Метод Якоби решения задач на собственные значения

1. В этом параграфе излагается *метод Якоби*, который можно применять для приближенного отыскания собственных чисел и собственных векторов эрмитовых матриц. Как и все методы, используемые в настоящее время для приближенного решения задач на собственные значения, метод Якоби является итерационным. В самых общих чертах, идея его состоит в следующем. Пусть A — диагональная матрица. Тогда собственные числа матрицы A есть ее диагональные элементы. Метод Якоби для любой эрмитовой матрицы A дает способ построения последовательности матриц $A_1, A_2, \ldots, A_k, \ldots$ таких, что каждая из матриц этой последовательности эрмитова, подобна матрице A и с увеличением номера k становится все более близкой к диагональной. В качестве приближенных значений собственных чисел матрицы A берутся диагональные элементы матрицы A_k, как только все ее внедиагональные элементы оказываются достаточно малыми.

Итак, пусть A — эрмитова матрица порядка n, $Q = \{q_{ij}\}_{i,j=1}^{n}$ — матрица, отличающаяся от единичной лишь четырьмя элементами:

$$q_{k,k} = \cos\varphi,\ q_{ll} = \cos\varphi,\ q_{kl} = -q\sin\varphi,\ q_{lk} = \bar{q}\sin\varphi, \tag{32.1}$$

где $1 \leqslant k < l \leqslant n$, φ — вещественное число, q — вообще говоря, комплексное число, $|q| = 1$. Очевидно, Q — унитарная матрица[1)].

Образуем по матрице A матрицу $\widehat{A} = Q^T A Q$ и попытаемся выбрать параметры матрицы Q, т. е. числа $k,\ l,\ \varphi,\ q$, так, чтобы матрица $\widehat{A}$ была максимально близка к диагональной.

Нетрудно убедиться, что матрица $\tilde{A} = Q^T A$ отличается от матрицы A лишь элементами строк с номерами k, l, причем

$$\begin{aligned} \tilde{a}_{k,j} &= \ \ a_{kj}\cos\varphi + a_{lj}\bar{q}\sin\varphi, \\ \tilde{a}_{l,j} &= -a_{kj}q\sin\varphi + a_{lj}\cos\varphi, \quad j = 1, \ldots, n. \end{aligned} \tag{32.2}$$

Аналогично, матрица $\hat{A} = \tilde{A}Q$ отличается от матрицы $\tilde{A}$ лишь элементами столбцов с номерами k, l, причем

$$\begin{aligned} \hat{a}_{j,k} &= \ \ \tilde{a}_{jk}\cos\varphi + \tilde{a}_{jl}q\sin\varphi, \\ \hat{a}_{j,l} &= -\tilde{a}_{jk}\bar{q}\sin\varphi + \tilde{a}_{jl}\cos\varphi, \quad j = 1, \ldots, n. \end{aligned} \tag{32.3}$$

Из (32.2), (32.3) сразу же следует, что

[1)] Матрица Q есть частный случай матрицы вращения, описанной в п. 1, с. 38.

$$|\tilde{a}_{k,j}|^2 + |\tilde{a}_{l,j}|^2 = |a_{k,j}|^2 + |a_{l,j}|^2, \quad |\hat{a}_{j,k}|^2 + |\hat{a}_{j,l}|^2 = |a_{j,k}|^2 + |a_{j,l}|^2, \quad j = 1, \ldots, n, \tag{32.4}$$

$$\hat{a}_{kl} = \bar{q}(a_{ll} - a_{kk}) \cos\varphi \sin\varphi + a_{kl} \cos^2\varphi - \bar{q}^2 a_{lk} \sin^2\varphi. \tag{32.5}$$

Вычислим сумму квадратов модулей внедиагональных элементов матрицы $\hat{A}$. Используя соотношения (32.2) – (32.4), нетрудно получить, что

$$\sum_{i \neq j} |\widehat{a}_{ij}|^2 = \sum_{i \neq j} |a_{ij}|^2 - 2|a_{kl}|^2 + |\hat{a}_{kl}|^2. \tag{32.6}$$

Определим теперь числа k, l из условия

$$|a_{kl}| = \max_{i \neq j} |a_{ij}|. \tag{32.7}$$

Поскольку A — эрмитова матрица, то $a_{lk} = \bar{a}_{kl}$, и из (32.5) с учетом того, что $1/\bar{q} = q$, будем иметь, что

$$\hat{a}_{kl} = \bar{q} \left(\frac{a_{ll} - a_{kk}}{2} \sin 2\varphi + q a_{kl} \cos^2\varphi - \bar{q}\bar{a}_{kl} \sin^2\varphi \right).$$

Будем считать, что $a_{kl} \neq 0$. В противном случае матрица диагональна, и ее собственные числа определяются тривиальным образом. Положим

$$q = |a_{kl}|/a_{kl}. \tag{32.8}$$

Тогда

$$\hat{a}_{kl} = \bar{q} \left(\frac{a_{ll} - a_{kk}}{2} \sin 2\varphi + |a_{kl}| \cos 2\varphi \right). \tag{32.9}$$

Выберем затем угол φ так, чтобы

$$|a_{kl}| \cos 2\varphi + \frac{1}{2}(a_{ll} - a_{kk}) \sin 2\varphi = 0,$$

или

$$\operatorname{tg} 2\varphi = \frac{2|a_{kl}|}{a_{kk} - a_{ll}}. \tag{32.10}$$

При указанном выборе параметров, определяющих матрицу Q, сумма квадратов модулей внедиагональных элементов матрицы $\widehat{A}$ принимает наименьшее значение.

Теперь можно описать метод Якоби. Пусть $A_0 = A$. Образуем последовательность матриц $A_1, A_2, \ldots$ при помощи рекуррентной формулы

$$A_{p+1} = Q_p^T A_p Q_p, \quad p = 0, 1, \ldots, \tag{32.11}$$

где параметры матрицы Q_p определяются так, чтобы сделать сумму квадратов внедиагональных элементов матрицы A_{p+1} минимально возможной, т. е. по формулам вида (32.7), (32.8), (32.10).

Вычисления проводят до тех пор, пока все внедиагональные элементы матрицы A_p не станут достаточно малыми. Тогда в качестве приближений к собственным числам матрицы A принимают диагональные элементы матрицы A_p, а столбцы матрицы $Q_0Q_1\cdots Q_p$ считают приближениями к собственным векторам матрицы A.

2. При исследовании сходимости метода Якоби существенно используется

Теорема 1. *Пусть параметры матрицы Q определяются согласно формулам* (32.7), (32.8), (32.10). *Тогда*

$$\sum_{i\neq j}|\widehat{a}_{ij}|^2 \leqslant \rho\sum_{i\neq j}|a_{ij}|^2, \tag{32.12}$$

где

$$0<\rho=1-\frac{2}{n(n-1)}<1$$

при $n>2$.

Доказательство. Вследствие (32.10) из (32.6) получаем

$$\sum_{i\neq j}|\widehat{a}_{ij}|^2=\sum_{i\neq j}|a_{ij}|^2-2|a_{kl}|^2, \tag{32.13}$$

а на основании (32.7)

$$\sum_{i\neq j}|a_{ij}|^2\leqslant|a_{kl}|^2n(n-1). \tag{32.14}$$

Здесь учтено, что матрица порядка n имеет n^2-n внедиагональных элементов. Из (32.13), (32.14) очевидным образом следует (32.12). □

3. Докажем сходимость метода Якоби. Пусть $A_p=\{a_{ij}^{(p)}\}_{i,j=1}^n$. Из рекуррентной формулы (32.11) и леммы 1 вытекает, что

$$\sum_{i\neq j}|a_{ij}^{(p)}|^2\leqslant\rho\sum_{i\neq j}|a_{ij}^{(p-1)}|^2\leqslant\cdots\leqslant\rho^p\sum_{i\neq j}|a_{ij}|^2\to 0 \text{ при } p\to\infty.$$

Это означает, что по любому заданному $\varepsilon>0$ можно указать целое положительное число p такое, что

$$|a_{ij}^{(p)}|\leqslant\varepsilon/n,\quad i\neq j,\quad i,j=1,2,\ldots,n. \tag{32.15}$$

Обозначим через Λ_p диагональную матрицу, на диагонали которой расположены диагональные элементы матрицы A_p. В соответствии с оценками (32.15), а также (24.4), с. 75, можем написать:

$$|\lambda_k(A_p) - \lambda_k^{(p)}| \leqslant \varepsilon, \quad k = 1, 2, \ldots, n,$$

где $\lambda_k^{(p)}$, $k = 1, \ldots, n$, — диагональные элементы матрицы Λ_p, упорядоченные по неубыванию, $\lambda_k(A_p)$ — так же упорядоченные собственные числа матрицы A_p. Вследствие (32.11) имеем $A_p = T_p^T A T_p$, где $T_p = Q_0 Q_1 \ldots Q_p$, т. е. матрицы A_p и A подобны, а значит, их собственные числа совпадают, поэтому

$$|\lambda_k(A) - \lambda_k^{(p)}| \leqslant \varepsilon, \quad k = 1, 2, \ldots, n. \tag{32.16}$$

Таким образом, выполнив определенное количество итераций, мы получим приближенные значения собственных чисел матрицы A с любой наперед заданной точностью.

4. Применяя метод Якоби для приближенного отыскания собственных чисел и собственных векторов симметричной вещественной матрицы, в формулах (32.1) параметр q следует положить равным единице. Соответственно в формуле (32.10) нужно заменить $|a_{kl}|$ на a_{kl}. Все выше полученные оценки при этом сохраняются.

33. QR-алгоритм

Этот алгоритм является одним из наиболее эффективных методов отыскания всех собственных чисел матрицы невысокого порядка. Формально метод чрезвычайно прост. Пусть $A \in M_n$. Положим $A_0 = A$ и образуем последовательность матриц A_0, A_1, $\ldots$ по следующему правилу.

Если матрица A_k известна, то:

1) представляем матрицу в виде (см. §11, гл. 2)

$$A_k = Q_k R_k, \tag{33.1}$$

где Q_k — унитарная, R_k — верхняя треугольная матрицы (такое представление может быть получено, например, методом отражений, см. гл. 2, с. 38),

2) строим матрицу

$$A_{k+1} = R_k Q_k. \tag{33.2}$$

Нетрудно видеть, что матрицы A_k, A_{k+1} унитарно подобны. В самом деле, из (33.1), (33.2), очевидно вытекает, что

$$A_{k+1} = Q_k^* A_k Q_k. \tag{33.3}$$

Из (33.3) получаем, что

$$A_{k+1} = S_k^* A S_k, \tag{33.4}$$

где,

$$S_k = Q_1 Q_2 \cdots Q_k, \tag{33.5}$$

т. е. каждая их матриц построенной последовательности унитарно подобна матрице A.

Исследование сходимости QR-алгоритма проведем в предположении, что A — нормальная матрица, т. е. $AA^* = A^*A$. Подробнее о нормальных матрицах см., например, в [5].

Теорема 1. *Пусть A — нормальная матрица. Тогда последовательность треугольных матриц R_k, $k = 0, 1, \ldots$, построенных при помощи алгоритма* (33.1), (33.2), *сходится к диагональной матрице.*

Для доказательства теоремы 1 нам потребуются некоторые вспомогательные результаты. Кроме того, будем придерживаться следующих соглашений. Если $X \in M_n$, то столбцы этой матрицы будем обозначать через x^i, а строки через x_i, $i = 1, 2, \ldots, n$. Под нормой векторов будем понимать норму $\|\cdot\|_2$. В разложении (33.1) диагональные элементы матриц R_k предполагаются неотрицательными (см. замечание 2 на с. 42).

Лемма 1. *Пусть матрицы A_k, Q_k, R_k, $k = 0, 1, \ldots$, построены по матрице A при помощи алгоритма* (33.1), (33.2). *Тогда*[1)]

$$\|a^{(k),i}\| = \|a_i^{(k)}\|, \quad i = 1, 2, \ldots \tag{33.6}$$

$$\|r^{(k),i}\| = \|a^{(k),i}\|, \ \|r_i^{(k)}\| = \|a_i^{(k+1)}\|, \quad i = 1, 2, \ldots, \tag{33.7}$$

$$\sum_{i=1}^{m} \sum_{j=m+1}^{n} |r_{ij}^{(k)}|^2 = \sum_{i=1}^{m} \|r_i^{(k)}\|^2 - \sum_{i=1}^{m} \|r^{(k),i}\|^2 = \\ = \sum_{i=1}^{m} \|a_i^{(k+1)}\|^2 - \sum_{i=1}^{m} \|a_i^{(k)}\|^2 \tag{33.8}$$

для $m = 1, 2, \ldots, n-1$,

$$|r_{mm}^{(k)}|^2 = \|a_m^{(k+1)}\|^2 - \sum_{j=i+1}^{n} |r_{mj}^{(k)}|^2 \tag{33.9}$$

[1)] Поясним, что индекс (k) обозначает принадлежность соответствующих строк и столбцов матрице с номером k.

для $m = 1, 2, \ldots, n$.

Доказательство. Равенства (33.6) непосредственно следуют из того, что если матрица A нормальна, то вследствие (33.4) и все матрицы A_k, $k = 1, 2, \ldots$, также — нормальные матрицы. Выполнение (33.7) обеспечивается равенствами (33.1), (33.2) и тем, что матрицы Q_k унитарны и потому не меняют длин векторов. Первое равенство (33.8) легко проверяется непосредственными вычислениями, второе следует из (33.7), (33.6). Для обоснования (33.9) достаточно учесть второе равенство (33.7). □

Доказательство теоремы 1. Очевидно, достаточно установить, что последовательности

$$\Delta_m^{(k)} = \sum_{i=1}^{m} \sum_{j=m+1}^{n} |r_{ij}^{(k)}|^2, \quad m = 1, 2, \ldots, n-1,$$

стремятся к нулю, а последовательности $r_{mm}^{(k)}$, $m = 1, 2, \ldots, n$, сходятся при $k \to \infty$. Пусть $\sigma_m^{(k)} = \sum\limits_{i=1}^{m} \|a_i^{(k)}\|^2$. Из (33.8) вытекает, что каждая из последовательностей $\sigma_m^{(k)}$ не убывает. Из равенств (33.3) (см. также упражнение на с. 74) получаем, что $\sigma_m^{(k)} \leqslant \|A\|_E$ для всех $m = 1, 2, \ldots, n-1$, $k = 1, 2, \ldots$ Таким образом, все последовательности $\sigma_m^{(k)}$ являются сходящимися. Но тогда из (33.8) вытекает, что $\Delta_m^{(k)} = \sigma_m^{(k+1)} - \sigma_m^{(k)} \to 0$ при $k \to \infty$ для всех $m = 1, 2, \ldots, n-1$. Заметим теперь, что $\|a_1^{(k+1)}\|^2 = \sigma_1^{(k+1)}$, $\|a_m^{(k+1)}\|^2 = \sigma_m^{(k+1)} - \sigma_{m-1}^{(k+1)}$ для $m = 2, 3, \ldots, n-1$. Поэтому из равенств (33.9) вытекает, что все последовательности $|r_{mm}^{(k)}|$, $m = 1, 2, \ldots, n$, являются сходящимися. Осталось напомнить, что по принятому соглашению $|r_{mm}^{(k)}| = r_{mm}^{(k)}$ для всех $m = 1, 2, \ldots, n$, $k = 1, 2, \ldots$ □

Теорема 2. *Пусть* $A \in M_n$ *— нормальная матрица, как и выше,* R_k, $k = 0, 1, \ldots$, *— последовательность треугольных матриц, построенных при помощи алгоритма* (33.1), (33.2),

$$r_{11}^{(k)} \geqslant r_{22}^{(k)} \geqslant \cdots \geqslant r_{nn}^{(k)} \geqslant 0$$

есть диагональные элементы матрицы R_k, *упорядоченные по невозрастанию,*

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$$

есть модули собственных чисел матрицы A. *Тогда* $r_{ii}^{(k)} \to |\lambda_i|$ *для* $i = 1, 2, \ldots, n$ *при* $k \to \infty$.

Доказательство. Вследствие (33.4) имеем $A_{k+1}A_{k+1}^* = S_k^* A A^* S_k$, т. е. для любого $k = 1, 2, \ldots$ спектры матриц $A_{k+1}A_{k+1}^*$ и AA^* совпадают. Из (33.2) получаем, что $A_{k+1}A_{k+1}^* = R_k R_k^*$. По теореме 1 последовательность матриц $R_k R_k^*$ стремится к диагональной матрице. Поэтому, рассуждая, как при обосновании метода Якоби (см. п. 3, с. 113), получаем, что $(r_{ii}^{(k)})^2 \to \rho_i^2$ при $k \to \infty$ для $i = 1, 2, \ldots, n$. Здесь $\rho_1^2, \rho_2^2, \ldots, \rho_n^2$ — собственные числа матрицы AA^*, упорядоченные по неубыванию. Напомним, что $\rho_1, \rho_2, \ldots, \rho_n \geqslant 0$ — сингулярные числа матрицы A, а для нормальной матрицы сингулярные числа и модули собственных чисел совпадают, что непосредственно вытекает из теоремы 4, с. 225, [5]. □

Отметим очевидное, но полезное

Следствие 1. *Если матрица A самосопряжена и неотрицательно определена, то $(r_{ii}^{(k)}) \to \lambda_i$, $k \to \infty$, $i = 1, 2, \ldots, n$.*

Наиболее трудоемким при реализации QR-алгоритма является вычисление на каждом шаге разложения матрицы на треугольный и унитарный сомножители. Это требует $4n^3/3 + O(n^2)$ flops (см. с. 42). Поэтому, обычно перед проведением итераций матрицу A подобным преобразованием приводят к такой форме, для которой QR-разложение требует существенно меньших затрат, и которая сохраняется на всех шагах QR-алгоритма. Особенно эффективно этот подход может быть реализован для эрмитовых матриц.

Теорема 3. *Пусть $A \in M_n$ — эрмитова матрица. Тогда существует унитарная матрица U такая, что матрица*

$$UAU^* \tag{33.10}$$

есть трехдиагональная (эрмитова) матрица.

Доказательство. Представим матрицу A в блочном виде

$$A = \begin{bmatrix} \alpha_1 & a_1^* \\ a_1 & M_1 \end{bmatrix}.$$

Здесь α_1 — число, a_1 — столбец. Пусть

$$U_1 = \begin{bmatrix} 1 & 0^T \\ 0 & V_1 \end{bmatrix},$$

где V_1 — унитарная матрица порядка $n - 1$. Тогда

$$U_1 A U_1^* = \begin{bmatrix} \alpha_1 & (V_1 a_1)^* \\ V_1 a_1 & V_1 M_1 V_1^* \end{bmatrix}.$$

Рассуждая, как при доказательстве теоремы 1, с. 41, мы можем построить матрицу V_1 так, чтобы все элементы столбца $V_1 a_1$, начиная со второго были равны нулю. Аналогичные рассуждения можно провести по отношению к матрице $V_1 M_1 V_1^*$ и так далее. $\square$

Упражнения.

33.1. Покажите, что алгоритм, описанный в доказательстве теоремы 3, требует порядка $4n^3/3$ flops[1)].

33.2. Предполагая, что матрица Q в QR алгоритме строится с использованием матриц отражения, докажите, что если матрица A — эрмитова трехдиагональная матрица, то и все матрицы A_k, $k = 1, 2, \ldots$, также эрмитовы и трехдиагональны. Покажите, что QR разложение эрмитовой трехдиагональной матрицы требует $O(n)$ flops.

33.3. Выясните, какую структуру будет иметь матрица UAU^*, построенная в ходе доказательства теоремы 3, если отказаться от предположения об эрмитовости матрицы A. Матрицы такой структуры называются матрицами Хессенберга.

[1)] В [7] указана модификация этого алгоритма, позволяющая в вещественном случае уменьшить затраты вдвое.

Глава 8

Практикум по численным методам

С целью закрепления теоретических знаний по численным методам линейной алгебры и навыков программирования, студентам могут быть предложены практические (лабораторные) задания, если это предусмотрено учебным планом. Ниже приводятся типовые задания. Далее мы будем предполагать, что задания выполняются в среде программирования MATLAB. Все встречающиеся ниже векторы и матрицы вещественны.

34. Варианты систем линейных уравнений

Для выполнения заданий потребуются как тестовые, так и содержательные примеры матриц и соответствующих им систем уравнений. Приведем ряд таких примеров.

Тестовые матрицы и системы. Тестовые матрицы и системы уравнений необходимы для отладки написанных студентом функций. При этом могут потребоваться матрицы со специальными свойствами (такими, как симметричность, положительная определенность и т. д.). Следующие рекомендации могут помочь при создании тестов.

1. Квадратная матрица A общего вида заданного размера n может задана с использованием генератора случайных чисел. Полезно элементы A генерировать равномерно распределенными на отрезке $[0, 1]$. Такие матрицы с вероятностью, практически равной единице, имеют ненулевой определитель достаточно хорошо обусловлены. Соответствующая функция MATLAB имеет вид $A = rand(n, n)$.

2. Команды $A = rand(n, n);\ A = A + A^T$; генерируют симметричную матрицу A общего вида с ненулевым определителем. Здесь A^T, как обычно, есть транспонированная к A матрица.

3. Если A — квадратная матрица из теста 1 или 2, то достаточно увеличить диагональные элементы A на n, чтобы получить матрицу с диагональным преобладанием (как по строкам, так и столбцам). В MATLAB это достигается командой $A = A + n * eye(n, n)$ (функция $E = eye(n, n)$ генерирует единичную матрицу размера n).

4. Матрица A с элементами $a_{ij} = \min\{i, j\}$ является симметричной и положительно определенной, а обратная к ней является трехдиагональной с целыми элементами; она задается командой $A = gallery('minij', n)$.

5. Пусть A матрица из предыдущего теста, E — матрица, все элементы которой равны единицы. Тогда матрица $B = 2A - E$ (т. е. $B = 2 * A - ones(size(A))$) является симметричной, а ее собственные числа равны $\lambda_k(B) = 0.5 \sec((2k-1)\pi/(4n))^2$, $k = 1 : n$.

6. Команда $A = gallery('tridiag', a, b, c)$ генерирует трехдиагональную матрицу размера n в разреженном формате (sparse). Вектор b (длины n) определяет диагональные элементы A, а a и c — соответсвенно, под- и наддиагональные элементы (векторы длины $n-1$).

7. Команда $A = gallery('randjorth', n, n, cond, 1, 1)$ генерирует квадратную матрицу размера $2n$ с заданным числом обусловленности равным $cond$.

8. При отладке функций решения систем уравнений $Ax = b$ с тестовой матрицей A вектор правой части b можно определить следующим образом. Выберем случайным образом решение системы x. Например, положим $x = rand(n, 1)$ и вычислим $b = Ax$. Таким образом мы получили тестовую систему уравнений с известным решением x. Если теперь мы решим систему $Ax = b$ и найдем ее решение (обозначим его через y; из-за ошибок округления при вычислениях вектор y, вообще говоря, отличается от x), то вектор $x - y$ определяет погрешность решения. Команда $e = norm(x - y, inf)$ позволяет вычислить максимальную погрешность решения. А именно, норму $\|x - y\|_\infty = \max\limits_{i=1,\dots,n} |x_i - y_i|$.

Вариант 1. Матрица A и вектор b получаются в результате дискретизации тем или иным методом квадратур (см. § 4, гл. 1) интегрального уравнения Фредгольма второго рода

$$u(x) - \lambda \int_a^b K(x, s)\, u(s)\, ds = f(x) \quad \forall\, x \in [a, b]. \tag{34.1}$$

Система уравнений имеет вид

$$y_i - \lambda \sum_{j=1}^{n} c_j K(x_i, x_j)\, y_j = f(x_i), \quad i = 1, \dots, n,$$

где x_i, c_i, $i = 1, \dots, n$, есть узлы и коэффициенты квадратурной формулы. Матрица A определяется как матрица этой системы, а вектор

b — как вектор правой части. Таким образом,

$$A = \{\delta_{ij} - \lambda\, c_j K(x_i, x_j)\}_{i,j=1}^{n}, \quad b = (f_1, f_2, \ldots, f_n)^T.$$

Решение $y = (y_1, \ldots, y_n)^T$ системы $Ay = b$ дает приближенное решение интегрального уравнения: y_i является приближением к $u(x_i)$.

Конкретные матрица A и вектор b получаются, если определить интервал $[a, b]$, число λ, ядро $K(x, s)$, функцию $f(x)$, а также квадратурную формулу. Ниже мы укажем дополнительно точное решение уравнения (34.1), чтобы графически можно было бы сравнить точность найденного приближенного решения и его зависимость от числа узлов сетки n. Таким образом, помимо тестирования метода решения системы алгебраических уравнений, предлагается попутно протестировать метод решения интегрального уравнения Фредгольма второго рода.

Отметим, что варианты 1a и 1b приводят к несимметричной матрице A, остальные — к симметричной.

Вариант 1a. Квадратура — составная формула центральных прямоугольников,

$$[a, b] = [0, 1], \quad \lambda = 0.5, \quad K(x, s) = x\, e^s, \quad f(x) = e^{-x}, \quad u(x) = x + e^{-x}.$$

Вариант 1b. Квадратура — составная формула трапеций,

$$[a, b] = [0, 1], \quad \lambda = 0.5, \quad K(x, s) = (x + 1)\, e^{-xs},$$
$$f(x) = e^{-x} - 0.5 + 0.5e^{-(x+1)}, \quad u(x) = e^{-x}.$$

Вариант 1c. Квадратура — составная формула центральных прямоугольников,

$$[a, b] = [-\pi, \pi], \quad \lambda = 0.3/\pi, \quad K(x, s) = 1/\big(0.64 \cos^2\big((x + s)/2\big) - 1\big),$$
$$f(x) = 25 - 16 \sin^2(x), \quad u(x) = 17/2 + (128/17)\cos(2x).$$

Вариант 1d. Квадратура — составная формула трапеций,

$$[a, b] = [-1, 1], \quad \lambda = 1, \quad K(x, s) = \operatorname{sh}(x + s),$$
$$f(x) = x^2, \quad u(x) = x^2 + \alpha \operatorname{sh}(x) + \beta \operatorname{ch}(x),$$
$$\alpha = (6 \operatorname{sh}(1) - 4 \operatorname{ch}(1))/\big(2 - \operatorname{sh}^2(2)/4\big), \quad \beta = \alpha(\operatorname{sh}(2)/2 - 1).$$

Вариант 1e. Квадратура — составная формула центральных прямоугольников,

$$[a, b] = [0, 3\pi], \quad \lambda = 1, \quad K(x, s) = \cos(x + s),$$

$$f(x) = (1 - 3\pi/2)\cos(x), \quad u(x) = \cos(x).$$

Вариант 1f. Квадратура — составная формула трапеций,

$$[a,b] = [0,1], \quad \lambda = -3, \quad K(x,s) = (xs)^2 - 4xs + 1,$$
$$f(x) = 2\pi^2\cos(2\pi\, x), \quad u(x) = 2\pi^2\cos(2\pi\, x) + 5(2x^2 - 1)/3.$$

Вариант 2. Матрица A и вектор b получаются в результате дискретизации методом коллокаций краевой задачи (см. § 5, гл. 1)

$$-u''(x) + q(x)u(x) = f(x), \quad x \in (a,b), \tag{34.2}$$
$$u(a) = u_a, \quad u(b) = u_b.$$

Для формирования матрицы A необходимо выполнить следующие вычисления:

1) вычислить сетку узлов $\{x_i\}$ по формуле (5.5), гл. 1;
2) Вычислить матрицы $D^{(1)}$ и $D^{(2)}$ размера $n+1$ по формулам (5.9), (5.10) и (5.11) гл. 1. Сформировать матрицу $D = -D^2 + \mathrm{diag}(q(x_0), q(x_1), \ldots, q(x_n))$;
3) вычислить вектор столбец $F = (f(x_1), f(x_2), \ldots, f(x_{n-1}))^T$;
4) вычислить $b = F - D_{2:n,1}u_a - D_{2:n,n+1}u_b$, где $D_{2:n,k}$ — элементы k-го столбца D со второго по n-й;
5) A получается из D вычеркиванием строк и столбцов с номерами 1 и $n+1$.

После решения системы $Az = b$ находится приближенное решение задачи (34.2) в виде $y = (u_a, z_1, \ldots, z_{n-1}, u_b)^T$.

Конкретные матрица A и вектор b получаются, если определить интервал $[a,b]$, функции $q(x)$ и $f(x)$. Ниже мы укажем дополнительно точное решение уравнения (34.2), чтобы можно было графически продемонстрировать точность найденного приближенного решения и ее зависимость от числа узлов сетки n. Для построения графиков $u(x)$ и $y_n(x)$ достаточно использовать равномерную сетку узлов $t = \{t_i\}_{i=1}^N$ с шагом $h = (b-a)/(N-1)$ при $100 \leqslant N \leqslant 200$. Для вычисления $y_n(x)$ в этих узлах можно использовать формулу (5.12). Таким образом, помимо тестирования метода решения системы алгебраических уравнений предлагается попутно протестировать метод решения краевой задачи (34.2).

Отметим, что матрица A является симметричной и положительно определенной, если $q(x) \geqslant 0$. Поскольку $u(x)$ известно, то числа u_a и u_b определяются равенствами $u_a = u(a)$, $u_b = u(b)$.

Вариант 2a.

$$[a,b]=[0,1],\quad q(x)=1/\varepsilon,\ \varepsilon=0.05,\quad f(x)=0,$$
$$u=\big(\exp(-x/\varepsilon^{1/2})-\exp((x-2)/\varepsilon^{1/2})\big)/\big(1-\exp(-2/\varepsilon^{1/2})\big).$$

Вариант 2b.

$$[a,b]=[-1,1],\quad q(x)=1/\varepsilon,\ \varepsilon=0.05,\quad f(x)=(1/\varepsilon+\pi^2)\cos(\pi x),$$
$$u=\cos(\pi x)+\exp((x-1)/\varepsilon^{1/2})+\exp(-(x+1)/\varepsilon^{1/2}).$$

Вариант 2c.

$$[a,b]=[0,\pi],\ \ q(x)=\sin(x),\ \ f(x)=\big(9+\sin(x)\big)\sin(3x),\ \ u=\sin(3x).$$

Вариант 2d.

$$[a,b]=[0,2],\ \ q(x)=x^2,\ \ f(x)=(4+x^2)\cos(2x),\ \ u=\cos(2x).$$

Вариант 2e.

$$[a,b]=[0,3],\quad q(x)=(1+x)^2,\quad f(x)=1-6/(1+x)^4,$$
$$u=1/(1+x)^2.$$

Вариант 2f.

$$[a,b]=[-2,2],\quad q(x)=4\cos^2(2x),\quad f(x)=\sin^2(4x)-16\cos(4x),$$
$$u=\sin^2(2x).$$

Вариант 3. Матрица A и вектор b получаются в результате дискретизации методом конечных разностей краевой задачи (см.§ 5, гл. 1)

$$-u''(x)+q(x)u(x)=f(x),\quad x\in(a,b),\qquad(34.3)$$
$$u(a)=u_a,\quad u(b)=u_b.$$

Система алгебраических уравнений имеет вид

$$y_0=u_a,$$
$$-y_{i-1}+\big(2+h^2q(x_i)\big)y_i-y_{i+1}=h^2f(x_i),\quad i=1,\dots,n-1,$$
$$y_n=u_b.$$

Поскольку y_0 и y_n известны, то их можно исключить из системы и получить новую систему для определения неизвестных $y_1, y_2, \ldots, y_{n-1}$:

$$\begin{aligned} \big(2 + h^2 q(x_1)\big) y_1 - y_2 &= h^2 f(x_1) + u_a, \\ -y_{i-1} + \big(2 + h^2 q(x_i)\big) y_i - y_{i+1} &= h^2 f(x_i), \quad i = 2, \ldots, n-2, \\ -y_{n-2} + \big(2 + h^2 q(x_{n-1})\big) y_{n-1} &= h^2 f(x_{n-1}) + u_b. \end{aligned} \tag{34.4}$$

Матрица A размера $N = n - 1$ этой новой системы является симметричной, трехдиагональной и с диагональным преобладанием, если $q(x) \geqslant 0$. Методы решения таких систем как правило не требуют хранения A в памяти ЭВМ: достаточно трех векторов для хранения элементов, расположенных на ненулевых диагоналях.

Конкретная система уравнений получается, если определить интервал $[a, b]$, функции $q(x)$ и $f(x)$. Ниже мы укажем дополнительно точное решение уравнения (34.3), чтобы можно было графически продемонстрировать точность найденного приближенного решения и ее зависимость от числа узлов сетки n. Таким образом, помимо тестирования метода решения системы алгебраических уравнений, предлагается попутно протестировать конечно-разностный метод решения краевой задачи (34.3).

Отметим, что поскольку $u(x)$ известно, то числа u_a и u_b определяются равенствами $u_a = u(a)$, $u_b = u(b)$.

Вариант 3a.

$$[a, b] = [0, 1], \quad q(x) = 1/\varepsilon, \ \varepsilon = 0.05, \quad f(x) = 0,$$
$$u = \big(\exp(-x/\varepsilon^{1/2}) - \exp((x-2)/\varepsilon^{1/2})\big) / \big(1 - \exp(-2/\varepsilon^{1/2})\big).$$

Вариант 3b.

$$[a, b] = [-1, 1], \quad q(x) = 1/\varepsilon, \ \varepsilon = 0.05, \quad f(x) = (1/\varepsilon + \pi^2)\cos(\pi x),$$
$$u = \cos(\pi x) + \exp((x-1)/\varepsilon^{1/2}) + \exp(-(x+1)/\varepsilon^{1/2}).$$

Вариант 3c.

$$[a, b] = [0, \pi], \ \ q(x) = \sin(x), \ \ f(x) = \big(9 + \sin(x)\big)\sin(3x), \ \ u = \sin(3x).$$

Вариант 3d.

$$[a, b] = [0, 2], \ \ q(x) = x^2, \ \ f(x) = (4 + x^2)\cos(2x), \ \ u = \cos(2x).$$

Вариант 3e.

$$[a, b] = [0, 3], \quad q(x) = (1 + x)^2, \quad f(x) = 1 - 6/(1 + x)^4,$$

$$u = 1/(1+x)^2.$$

Вариант 3f.

$$[a,b] = [-2,2], \quad q(x) = 4\cos^2(2x), \quad f(x) = \sin^2(4x) - 16\cos(4x),$$
$$u = \sin^2(2x).$$

35. Задание 1. Решение трехдиагональных систем уравнений

1. Целью задания является закрепление теоретических знаний и приобретение практических навыков при решении систем линейных алгебраических уравнений с трехдиагональными матрицами методом прогонки и итерационными методами, а также ознакомление с конечно-разностным методом решения краевых задач для обыкновенных дифференциальных уравнений.

2. Для анализа методов каждый студент получает одну из систем уравнений варианта 3.

3. Считая, что система уравнений имеет общий вид

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = f_i, \quad i = 1, \ldots, N,$$

при $a_1 = 0$, $c_N = 0$, требуется реализовать и отладить на тестовом примере следующие методы.

a) Метод прогонки. Метод должен быть реализован в виде отдельной функции с входными параметрами (a, b, c, f) и выходным x. Здесь a, b, c — векторы коэффициентов системы, f — вектор правой части;

b) Итерационный метод Якоби. Метод должен быть реализован в виде отдельной функции с входными параметрами $(a, b, c, f, tol, x0, maxiter)$ и выходными параметрами $[x, niter, r]$. Здесь дополнительно: tol — критерий точности, $x0$ — начальное приближение к x, $maxiter$ — максимальное число итераций (итерации выполняются пока не выполнено условие $\|x^{k+1} - x^k\|_\infty \leqslant tol$ близости соседних итераций), $niter$ — число итераций, потребовавшихся для достижения критерия точности, r — вектор норм невязок на итерациях, т. е. $r(k) = \|r^k\|_\infty$. Предусмотреть, что входные параметры $tol, x0, maxiter$ (или часть из них) могут быть не указаны при вызове функции; в этом случае принять по умолчанию $tol = 10^{-6}$, $x0 = 0$, $maxiter = 10n$. Также предусмотреть, что параметр $niter$ и (или) r

при вызове функции могут быть не указаны, а также сообщение на экран, если достигнуто максимальное число итераций;

c) Итерационный метод Зейделя. Метод должен быть реализован в виде отдельной функции (с теми же входными и выходными параметрами, что и для метода Якоби);

d) Итерационный метод релаксации. Метод должен быть реализован в виде отдельной функции с входными параметрами $(a, b, c, f, omega, tol, x0, maxiter)$ и выходными — $[x, niter, r]$. Здесь $omega$ — параметр метода;

4. Для значений $n = 10, 20, 50, 100$ решить заданную систему уравнений методом прогонки. При каждом n требуется определить максимальную погрешность решения разностной схемы, т. е. величину $e_n = \|u - y\|_\infty = \max\limits_{i=1:n} |u(x_i) - y_i|$ и в одних осях построить графики u и y. Кроме того, требуется составить таблицу из трех строк, откладывая в первой строке значения n, во второй — значения e_n, в третьей — $e_n\, n^2$. Анализируя графики и таблицу, студент должен сделать выводы о точности разностной схемы.

5. При $n = 100$ решить заданную систему методами прогонки, Якоби, Зейделя и релаксации при $\omega = 1.7$, $\omega = 1.8$, $\omega = 1.9$, полагая $tol = 10^{-6}$, $maxiter = 10000$, $x0$ равным случайному вектору. Составить таблицу с пятью столбцами (по числу итерационных методов) и тремя строками. В строках указываются соответствующие итерационному методу значения $niter$, $r(niter)$, а также истинная погрешность $e_n = \|y_p - y_{it}\|_\infty$, где y_p — решение, полученное методом прогонки, y_{it} — решение, полученное итерационным методом при достижении критерия точности. Кроме того, в одних осях необходимо построить графики норм невязок на итерациях (r) всех методов.

Анализируя графики и таблицу, студент должен сделать выводы о поведении норм невязок методов на итерациях, скорости сходимости методов, влиянии параметра ω на скорость сходимости метода релаксации.

6. Для значений $n = 50$, 100, 200, 400, 800 решить заданную систему методами прогонки, Якоби, Зейделя и релаксации при одном выбранном студентом параметре ω. Выбрать $tol = 10^{-4}$, $maxiter = 10000$, $x0 = 0$. Требуется составить три таблицы с тремя строками (по числу итерационных методов) и пятью столбцами (по числу значений n). В первой таблице в строках указываются соответствующие итерационному методу значения $niter$; во второй таблице — $r(niter)$, в третьей — истинная погрешность $e_n = \|y_p - y_{it}\|_\infty$, где y_p — решение,

полученное методом прогонки, y_{it} — решение, полученное итерационным методом при достижении критерия точности.

Анализируя графики и таблицы, студент должен сделать выводы о скорости сходимости методов, а также обоснованно выбрать наилучший итерационный метод.

36. Задание 2. Метод Гаусса

1. Целью задания является закрепление теоретических знаний и приобретение практических навыков решения систем линейных алгебраических уравнений, а также ознакомление студента с одним из приближенных методов решения краевых задач для обыкновенных дифференциальных уравнений.

2. Для анализа методов каждый студент получает одну из систем уравнений варианта 2.

3. Требуется реализовать и отладить на тестовом примере следующие алгоритмы метода Гаусса без выбора ведущего элемента для решения системы уравнений $Ax = b$.

a) kij-алгоритм. Метод должен быть реализован в виде отдельной функции с входными параметрами (A, b) и выходным x.

b) kji-алгоритм. Метод должен быть реализован в виде отдельной функции с входными параметрами (A, b) и выходным x.

c) ijk-алгоритм LU разложения. Метод должен быть реализован в виде отдельной функции с входным параметром A и выходными $[L, U]$;

d) jik-алгоритм LU разложения. Метод должен быть реализован в виде отдельной функции с входным параметром A и выходными $[L, U]$;

e) алгоритм решения системы $LUx = b$ в виде отдельной функции с входными параметрами $[L, U, b]$ и выходным x;

Во всех функциях a)–d) необходимо предусмотреть сообщение на экран, если разложение матрицы невозможно осуществить.

5. Для значений $n = 100, 500, 1000, 1500, 2000$ решить тестовую систему, используя функции a)–e). Требуется составить три таблицы, две из которых имеют шесть строк и пять столбцов, третья — две строки и пять столбцов. В этих таблицах в столбцах указываются значения n. В первой таблице в строках указывается время работы

соответствующей функции; во второй таблице — погрешность найденного решения; в третьей — число обусловленности матрицы A, вычисленное при помощи MATLAB функции *cond*.

Анализируя таблицы, студент должен сделать выводы о накоплении погрешности при реализации метода Гаусса и соответствии времени работы теоретическим ожиданиям.

6. Написать функцию формирования заданной системы уравнений из варианта 2 с входными параметрами (q, f, ua, ub, m) и выходными $[A, b]$, где q, f — указатели на функции дифференциального уравнения, ua, ub данные краевых условий u_a и u_b, m — размер матрицы A.

7. Для значений $n = 5, 10, 20, 50$ решить заданную систему уравнений kij – алгоритмом. При каждом n требуется определить максимальную погрешность решения приближенного метода, т. е. величину $e_n = \|u - y\|_\infty = \max_{i=1:N} |u(t_i) - y_n(t_i)|$ и в одних осях построить графики u и y_n. Кроме того, требуется составить таблицу из двух строк, откладывая в первой строке значения n, во второй — значения e_n. Анализируя графики и таблицу, студент должен сделать выводы о точности дискретной схемы.

37. Задание 3. Метод Гаусса с выбором главного элемента

1. Целью задания является закрепление теоретических знаний и приобретение практических навыков решения систем линейных алгебраических уравнений, а также ознакомление студента с одним из приближенных методов решения интегральных уравнений.

2. Для анализа методов каждый студент получает одну из систем уравнений варианта 1.

3. Требуется реализовать и отладить на тестовом примере следующие алгоритмы метода Гаусса с выбором ведущего элемента по столбцу для решения системы уравнений $Ax = b$.

a) kij-алгоритм. Метод должен быть реализован в виде отдельной функции с входными параметрами (A, b) и выходным x.

b) kji-алгоритм. Метод должен быть реализован в виде отдельной функции с входными параметрами (A, b) и выходным x.

c) ijk-алгоритм LU разложения. Метод должен быть реализован в виде отдельной функции с входным параметром A и выходными $[L, U, p]$, где p — вектор перестановок такой, что $A(p, :) = L * U$.

Соответствующая p матрица перестановок удовлетворяет равенству $PA = LU$;

d) jik-алгоритм LU разложения. Метод должен быть реализован в виде отдельной функции с входным параметром A и выходными $[L, U, p\,]$;

e) алгоритм решения системы $LUx = Pb$ в виде отдельной функции с входными параметрами $[L, U, p, b]$ и выходным x;

Во всех функциях a)–d) необходимо предусмотреть сообщение на экран, если разложение матрицы невозможно осуществить.

5. Для значений $n = 100, 500, 1000, 1500, 2000$ решить тестовую систему, используя функции a)–e). Требуется составить три таблицы, две из которых имеют шесть строк и пять столбцов, третья — две строки и пять столбцов. В этих таблицах в столбцах указываются значения n. В первой таблице в строках указывается время работы соответствующей функции; во второй таблице — погрешность найденного решения; в третьей — число обусловленности матрицы A, вычисленное при помощи MATLAB функции *cond*.

Анализируя таблицы, студент должен сделать выводы о накоплении погрешности при реализации метода Гаусса и соответствии времени работы теоретическим ожиданиям.

6. Написать функцию формирования заданной системы уравнений из варианта 1 с входными параметрами $(K, f, lambda, a, b, n)$ и выходными $[A, b]$, где K, f — указатели на функции ядра и правой части интегрального уравнения, a, b отрезок интегрирования, n — размер матрицы A.

7. Для значений $n = 5, 10, 20, 50$ решить заданную систему уравнений kij-алгоритмом. При каждом n требуется определить максимальную погрешность решения приближенного метода, т.е. величину $e_n = \|u - y\|_\infty = \max\limits_{1 \leqslant i \leqslant n} |u(x_i) - y_i|$ и в одних осях построить графики u и y. Кроме того, требуется составить таблицу из двух строк, откладывая в первой строке значения n, во второй — значения e_n. Анализируя графики и таблицу, студент должен сделать выводы о точности рассматриваемого метода квадратур.

38. Задание 4. Итерационные методы вариационного типа

1. Целью задания является закрепление теоретических знаний и приобретение практических навыков решения систем линейных алгебраических уравнений итерационными методами.

2. Для анализа методов каждый студент должен создать две тестовые системы построения разреженных систем алгебраических уравнений с симметричной и положительно определенной матрицей порядка n, реализовав их в виде отдельных функций с входным параметром n и выходными $[A, b, x]$, где n — размер матрицы A системы, b — вектор-столбец правой части, x — точное решение системы.

3. Требуется запрограммировать следующие итерационные методы решения системы уравнений $Ax = b$.

a) Метод наискорейшего спуска. Он должен быть реализован в виде отдельной функции с входными параметрами $(A, b, x0, tol, maxiter)$ и выходными параметрами $[x, niter, r]$. Здесь $x0$ — начальное приближение к x, tol — критерий точности, $maxiter$ — максимальное число итераций; $niter$ — число итераций, потребовавшихся для достижения критерия точности, r — вектор норм невязок на итерациях, т.е. $r(k) = \|r^k\|_\infty$. Итерации заканчиваются, если выполнено одно из условий: либо $\|r^k\|_\infty \leqslant tol\,\|b\|_\infty$, либо $k > maxiter$. Здесь k — номер итерации.

Предусмотреть, что входные параметры $x0, tol, maxiter$ (или часть из них) могут быть не указаны при вызове функции; в этом случае принять по умолчанию $tol = 10^{-6}$, $x0 = 0$, $maxiter = 10n$. Также предусмотреть, что параметр $niter$ и (или) r при вызове функции могут быть не указаны, а также сообщение на экран, если достигнуто максимальное число итераций;

b) Метод сопряженных градиентов. Он должен быть реализован аналогично методу наискорейшего спуска. Предполагается использование расчетных формул (30.40)–(30.43), с. 105.

4. Для значений $n = 500, 2500, 10000, 25000, 50000$ решить тестовые системы, используя реализованные методы, и параметры $tol = 10^{-6}$, $x0 = 0$, $maxiter = 10n$. Требуется в одних осях построить графики норм невязок на итерациях (r) обоих методов. Кроме того, необходимо составить четыре таблицы, три из которых имеют три строки и пять столбцов, четвертая — две строки и пять столбцов. В этих таблицах в столбцах указываются значения n. В первой таблице в строках указывается время работы соответствующей функции; во второй таблице — истинная погрешность найденного решения, т. е. величина $\|x - x_{it}\|_\infty$, где x — решение системы, x_{it} — решение, полученное итерационным методом при достижении критерия точности; в третьей — число итераций, в четвертой — число обусловленности матрицы A, вычисленное при помощи MATLAB функции $rcond$.

Анализируя графики и таблицы студент должен сделать выводы о

поведении норм невязок методов на итерациях и обоснованно выбрать наилучший метод.

39. Задание 5. Метод Якоби решения задачи на собственные значения

1. Целью задания является закрепление теоретических знаний и приобретение практических навыков решения полной проблемы на собственные значения матриц.

2. Каждый студент должен построить две симметричные матрицы A порядка n с известным набором собственных чисел и векторов и реализовать их в виде отдельных функций с входным параметром n и выходными параметрами $[A, lamda, U]$, где n — размер матрицы A, $lamda$ — вектор собственных чисел, U — матрица, столбцы которой есть собственные векторы A, соответствующие $lamda$.

3. Требуется реализовать метод Якоби определения собственных чисел и векторов симметричной матрицы A. Он должен быть реализован в виде отдельной функции с входными параметрами $(A, tol, maxiter)$ и выходными параметрами $[lamda, U, niter]$. Здесь tol — критерий точности, $maxiter$ — максимальное число итераций; $niter$ — число итераций, потребовавшихся для достижения критерия точности.

Предусмотреть, что входные параметры $tol, maxiter$ могут быть не указаны при вызове функции; в этом случае принять по умолчанию $tol = 10^{-3}$, $maxiter = 10n$. Также предусмотреть, что параметр $niter$ и (или) U при вызове функции могут быть не указаны. Выдать сообщение на экран, если достигнуто максимальное число итераций.

4. Для значений $n = 10, 50, 250, 500, 1000$ найти собственные числа и векторы тестовых матриц при $maxiter = 10n$ и $tol = 10^{-3}$, $tol = 10^{-4}$, $tol = 10^{-5}$.

Для каждого варианта вычислений требуется составить таблицу с четырьмя строками и пятью столбцами. В первой строке таблицы указываются значения n; во второй — время работы функции; в третьей — погрешность определения собственных чисел; в четвертой — погрешность определения собственных векторов, т. е. величина $\max\limits_{i=1,\dots,n} \|u_i - v_i\|_\infty$, где u_i — точный собственный вектор, v_i — найденный методом Якоби.

Анализируя таблицы, студент должен сделать выводы об эффективности метода Якоби a) при определении собственных чисел; b) при определении собственных векторов.

Основные обозначения

$A = A(m,n) = \{a_{ij}\}_{i,j=1}^{m,n}$ — матрица из m строк и n столбцов, вообще говоря, комплексная.

$A = A(n) = \{a_{ij}\}_{i,j=1}^{n}$ — квадратная матрица порядка n.

A^{-1} — матрица, обратная к матрице A.

A^T — транспонированная матрица.

$A^* = (\bar{A})^T$ — сопряженная матрица.

I — единичная матрица.

E — матрица, все элементы которой равны единице.

$\mathbb{R}^n$ — линейное пространство всех упорядоченных наборов (векторов) $(x_1, x_2, \ldots, x_n)$ вещественных чисел со стандартным скалярным произведением $(x,y) = \sum\limits_{i=1}^{n} x_i y_i$.

$\mathbb{C}^n$ — линейное пространство всех упорядоченных наборов (векторов) $(x_1, x_2, \ldots, x_n)$ комплексных чисел со стандартным скалярным произведением $(x,y) = \sum\limits_{i=1}^{n} x_i \bar{y}_i$.

$i^k = (\underbrace{0, \ldots, 0}_{k-1}, 1, \underbrace{0, \ldots, 0}_{n-k})$, $k = 1, 2, \ldots, n$, — векторы естественного базиса пространства $\mathbb{R}^n$ ($\mathbb{C}^n$).

$M_{m,n}$ — множество всех прямоугольных матриц с m строками и n столбцами.

M_n — множество всех квадратных матриц порядка n.

$\mathbf{X}_n$ — n-мерное линейное (евклидово) пространство, как правило, над полем комплексных чисел.

$\mathcal{A}$ — линейный оператор, действующий из $\mathbf{X}_n$ в $\mathbf{Y}_m$.

$\mathcal{A}^*$ — сопряженный к $\mathcal{A}$ оператор.

flops — floating point operation — арифметическая операция с плавающей точкой.

Литература

1. **Бахвалов Н.С., Жидков Н.П., Кобельков Г.М.** Численные методы. — М.: БИНОМ. Лаб. знаний, 2015.
2. **Воеводин В.В.** Линейная алгебра. — Изд-во «Лань», 2009.
3. **Голуб Дж., Ван Лоун Ч.** Матричные вычисления. — М.: Мир, 1999.
4. **Деммель Дж.** Вычислительная линейная алгебра. Теория и приложения . — М.: Мир, 2001.
5. **Карчевский Е.М., Карчевский М.М.** Лекции по линейной алгебре и аналитической геометрим. — Казань: Казан. ун-т, 2014.
6. **Квасов Б.И.** Численные методы анализа и линейной алгебры. Использование Matlab и Scilab (Электронный ресурс) — СПб. : Лань, 2016.
7. **Парлетт Б.** Симметричная проблема собственных значений. — М.: Мир, 1983.
8. **Самарский А.А., Гулин А.В.** Численные методы. — М.: Наука, 1989.
9. **Хорн Р., Джонсон Ч.** Матричный анализ. — М.: Мир, 1989.
10. **Higham N.J.** Accuracy and Stability of Numerical Algorithms. — SIAM, 1996.
11. **Stewart G.W., Sun J.** Matrix perturbation theory. — Academic Press, 1990.
12. **Zhang F.** Matrix theory: basic results and techniques. — Springer, 1999.