И.Г.ЖУРКИН
Ю.М.НЕЙМАН

МЕТОДЫ ВЫЧИСЛЕНИЙ В ГЕОДЕЗИИ

И.Г.ЖУРКИН Ю.М.НЕЙМАН

И.Г.ЖУРКИН
Ю.М.НЕЙМАН

МЕТОДЫ ВЫЧИСЛЕНИЙ В ГЕОДЕЗИИ

Допущено Министерством высшего и среднего специального образования СССР в качестве учебного пособия для студентов геодезических специальностей вузов

МОСКВА "НЕДРА" 1988

ББК 26.1
Ж 91
УДК 528.063+519.6(075.8)

Рецензенты: проф., д-р техн. наук *Н. Д. Дроздов*, канд. физ.-мат. наук *Э. Б. Поляк*

Журкин И. Г., Нейман Ю. М.

Ж 91 **Методы** вычислений в геодезии: Учеб. пособие. — М.: Недра, 1988. — 304 с.: ил.

ISBN 5—247—00349—7

Рассмотрено применение численных методов при решении геодезических, фотограмметрических и картографических задач. Приведены элементы функционального анализа. Показаны источники ошибок, возникающих при работе с приближенными числами и связанных с выполнением арифметических операций на ЭВМ. Описаны прямые и итерационные методы решения систем линейных уравнений, устойчивость решения этих систем. Изложено вычисление собственных значений матриц. Освещены интерполяция и аппроксимация функций, а также методы численного интегрирования.

Для студентов геодезических специальностей вузов.

$\text{Ж}\frac{1802020000—405}{043(01)—88}$ 8—88 ББК 26.1

ISBN 5—247—00349—7

ПРЕДИСЛОВИЕ

Учебное пособие соответствует программе курса «Математические методы и модели расчета на ЭВМ» для студентов геодезических специальностей вузов. Особое внимание уделено различным вычислительным методам при решении задач геодезии, фотограмметрии и картографии. Иллюстрирующими примерами в основном являются геодезические задачи численного характера. Первым опытом подобной работы было выпущенное издательством «Недра» учебное пособие [4].

Для обоснования алгоритмов использован в основном традиционный математический аппарат, знакомый студентам всех инженерных специальностей. Необходимые дополнительные сведения собраны в гл. 1 и лишь в отдельных случаях приведены по всему тексту. Нетрадиционным для учебников является содержание гл. 2, посвященное оценкам ошибок округления при вычислениях на ЭВМ. Сравнительно подробно изложены наиболее хорошо зарекомендовавшие себя методы решения линейных алгебраических уравнений в главах 3 и 4. Уделено внимание оптимизации машинных алгоритмов при работе с разреженными матрицами коэффициентов, например, при уравнивании обширных геодезических сетей.

При изложении интерполирования и аппроксимации функций в главах 5, 6 предпринята попытка восполнить существующий в учебной литературе пробел относительно задачи коллокации: по заданным значениям разнородных функционалов на изучаемой функции требуется оценить значение нового функционала. Решение подобных задач составляет основу математического аппарата современной физической геодезии.

В гл. 7 рассмотрены основные методы численного интегрирования, которые обычно используются при решении геодезических задач.

Естественным продолжением гл. 7 являются методы численного интегрирования дифференциальных уравнений, широко применяемые в космической геодезии.

Главы 1, 2, 3, 4 написал И. Г. Журкин, главы 5, 6, 7 — Ю. М. Нейман.

Глава 1

ОСНОВНЫЕ СВЕДЕНИЯ ИЗ ТЕОРИИ КОНЕЧНОМЕРНЫХ ПРОСТРАНСТВ

§ 1.1. ЛИНЕЙНЫЕ (ВЕКТОРНЫЕ) ПРОСТРАНСТВА

Понятие линейного пространства является одним из основных в математике. Исключительно важная роль ему отводится и в численных методах линейной алгебры.

Рассмотрим множество L и поле P произвольной природы. Предположим, что для всех элементов из L определены операции сложения и умножения на числа из поля P. Будем называть элементы множества L *векторами* независимо от их конкретной природы.

Определение 1.1. Непустое множество L называется *линейным* или *векторным пространством*, если удовлетворяются следующие две группы условий.

I. В множестве L задана алгебраическая операция сложения, причем для $\forall x, y, z \in L$

1) $x+y=y+x$ (коммутативность),
2) $x+(y+z)=(x+y)+z$ (ассоциативность),
3) в L существует такой элемент «ø», что $x+\phi=x$,
4) для каждого x существует такой элемент $-x \in L$, что $x+(-x)=\phi$.

II. Для любого числа α из поля P и каждого $x \in L$ определен элемент $\alpha x \in L$, причем для $\forall x, y \in L$ и для $\forall \alpha, \beta \in P$

1) $\alpha(\beta x)=(\alpha\beta)x$,
2) $1 \cdot x = x$,
3) $(\alpha+\beta)x=\alpha x+\beta x$,
4) $\alpha(x+y)=\alpha x+\alpha y$.

Пространство L называется вещественным, если P — поле вещественных чисел, и комплексным, если P — поле комплексных чисел.

Приведем примеры линейных пространств.

1. Пусть P — произвольное поле. Рассмотрим множество всех строк (столбцов) $x=(x_1, x_2, \ldots, x_n)$; $y=(y_1, y_2, \ldots, y_n)$; ... фиксированной длины n, составленных из упорядоченной совокупности элементов поля P. Определим сложение строк и умножение строки на число из P:

$$\begin{gathered}(x_1, x_2, \ldots, x_n)+(y_1, y_2, \ldots, y_n)= \\ =(x_1+y_1, x_2+y_2, \ldots, x_n+y_n), \\ \alpha(x_1, x_2, \ldots, x_n)=(\alpha x_1, \alpha x_2, \ldots, \alpha x_n).\end{gathered} \tag{1.1}$$

Обе группы условий из определения 1.1 выполняются, значит, заданное множество строк с введенными операциями является линейным пространством. Если P — поле вещественных чисел, то приведенное в этом примере пространство называется *арифметическим действительным пространством* и обозначается R_n. Аналогично для поля комплексных чисел рассмотренное выше пространство называется *арифметическим комплексным пространством* и обозначается C_n.

2. Множество M_{nm} всех матриц $[n \times m]$ над полем P относительно сложения матриц и умножения матрицы на число из поля P является линейным пространством над этим полем. Нулевым вектором в данном пространстве служит нулевая матрица O_{nm}.

3. Множество всех непрерывных на отрезке $[a, b]$ вещественных функций вещественной переменной является линейным пространством над полем вещественных чисел с обычными операциями сложения функций и умножения их на числа. Это пространство принято обозначать символом $C[a, b]$.

Важным понятием в теории линейных пространств является понятие линейного подпространства.

Определение 1.2. Множество F линейного пространства L над полем P называется его *линейным подпространством*, если при тех же операциях, что и в пространствах L, оно само является линейным пространством над полем P.

Из определения непосредственно вытекает следующее утверждение: $F \subset L$ есть подпространство, если из $x \in F$, $y \in F$ следует, что $\alpha x + \beta y \in F$ при $\alpha, \beta \in P$. Справедливо и обратное утверждение, т. е. если $F \subset L$ — подпространство, то $\alpha x + \beta y \in F$ для всех $\alpha, \beta \in P$.

Во всяком линейном пространстве L имеется подпространство, состоящее из одного нулевого вектора, — нулевое подпространство, и подпространство, являющееся самим линейным пространством. Подпространство, отличное от L и содержащее хотя бы один ненулевой элемент, называется *собственным*.

Обратимся к примерам собственных подпространств.

1. Множество всех векторов $x = (x_1, x_2, x_3)$, лежащих в одной плоскости линейного арифметического пространства R_3, образует арифметическое подпространство R_2 пространства R_3.

2. Совокупность всех решений однородной системы линейных уравнений с n неизвестными представляет собой подпространство в линейном пространстве n-мерных столбцов.

3. Для произвольной системы векторов $\{a_i\}_{i=1 \div n}$ линейного пространства L над полем P множество всех линейных комбинаций векторов

$$\beta_1 a_1 + \beta_2 a_2 + \ldots + \beta_n a_n = \sum_{i=1}^{n} \beta_i a_i$$

с коэффициентами $\beta_1, \beta_2, \ldots, \beta_n$ из поля P, одновременно не

равными нулю, является линейным подпространством пространства L. Это подпространство получило название *линейной оболочки векторов* $\{a_i\}_{i=1\div n}$ и обозначается символом $L(a_1, a_2, \ldots, a_n)$. В частности, $L(a)$ — множество всех векторов вида $\lambda a, \lambda \in P$.

§ 1.2. НОРМИРОВАННЫЕ ПРОСТРАНСТВА

Определение 1.3. Линейное пространство L над числовым полем P называется *нормированным* пространством, если каждому вектору $x \in L$ поставлено в соответствие вещественное число $\|x\|$, называемое *нормой* вектора x и удовлетворяющее следующим условиям для любых векторов x, y и любого числа λ из поля P:

1) $\|x\| > 0$ для всех $x \neq 0$,

2) $\|\lambda x\| = |\lambda| \|x\|$ (аксиома абсолютной однородности),

3) $\|x+y\| \leqslant \|x\| + \|y\|$ (аксиома неравенства треугольника).

Вектор x называется *нормированным*, если выполняется равенство $\|x\| = 1$.

Из определения нормы вектора в линейном пространстве вытекают следующие ее свойства.

Свойство 1. $\|\varnothing\| = 0$, где $\varnothing$ — нулевой вектор в линейном пространстве L. Действительно, если $\lambda = 0$, то из условия 2 определения 1.3 следует, что $\|0 \cdot x\| = 0\|x\|$ и $\|\varnothing\| = 0$.

Свойство 2.

$$\| x - y \| \geqslant | \| x \| - \| y \| |.$$

Для доказательства этого свойства рассмотрим разность

$$\| x \| - \| y \| = \| (x - y) + y \| - \| y \|.$$

Использовав условие 3 из определения 1.3, имеем

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

После подстановки этого результата в выражение $\|x\| - \|y\|$ получим

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

Выполнив аналогичные преобразования и использовав условие 2 из определения 1.3, имеем

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

На основании двух полученных неравенств получаем окончательный вид свойства 2.

Свойство 3.

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

Доказательство этого свойства предлагаем провести само-

стоятельно. При этом рекомендуем воспользоваться свойством 2 и условием 3 из определения 1.3.

Свойство 4. Пусть $\|x\|$ — норма в линейном пространстве L. Тогда если β — положительное вещественное число, то функция $\varphi(x)=\beta\|x\|$ для всех $x\in L$ также является нормой.

Доказательство этого свойства следует в результате проверки трех условий, входящих в определение 1.3.

Рассмотрим теперь арифметическое пространство над числовым полем P, в котором ξ_i при $i=1\div n$ — координаты вектора-столбца ξ. В этом пространстве наиболее употребительны следующие нормы:

1) $\|\xi\|_l=\sum\limits_{i=1}^{n}|\xi_i|$ (l-норма),

2) $\|\xi\|_E=(\sum\limits_{i=1}^{n}\xi_i^2)^{1/2}$ (*евклидова* норма в пространстве R_n)

или

$\|\xi\|_U=(\sum\limits_{i=1}^{n}|\xi_i|^2)^{1/2}$ (*унитарная* норма в пространстве C_n),

3) $\|\xi\|_c=\max\limits_i|\xi_i|$ (c — норма)

Для доказательства того, что приведенные выражения являются нормами, требуется выполнить проверку условий определения 1.3. Такая проверка для l-нормы и c-нормы не должна вызвать больших затруднений, и поэтому предлагаем выполнить ее самостоятельно. Проверка условий 1 и 3 для евклидовой нормы также несложна. Поэтому остановимся только на проверке условия 2 для евклидовой нормы, а именно покажем, что для любых векторов $\xi\in R_n$ и $\eta\in R_n$ выполняется условие

$$\sqrt{\sum_{i=1}^{n}(\xi_i+\eta_i)^2}\leqslant\sqrt{\sum_{i=1}^{n}\xi_i^2}+\sqrt{\sum_{i=1}^{n}\eta_i^2}. \tag{1.2}$$

После ввода обозначения $\gamma=\xi+\eta$ и символа нормы неравенство (1.2) примет вид

$$\|\gamma\|_E\leqslant\|\xi\|_E+\|\eta\|_E.$$

Выполнив ряд алгебраических операций, преобразуем правую часть этого неравенства

$$\|\xi\|_E+\|\eta\|_E=\sqrt{\|\gamma\|_E^2-2(\sum_{i=1}^{n}\xi_i\eta_i-\|\xi\|_E\|\eta\|_E)}. \tag{1.3}$$

Для случая, когда значение $\sum\limits_{i=1}^{n}\xi_i\eta_i\leqslant 0$, доказательство условия

(1.2) непосредственно следует из равенства (1.3). Рассмотрим теперь выражение вида

$$|\sum_{i=1}^{n}\xi_i\eta_i| - \|\xi\|_E\|\eta\|_E \leqslant \sum_{i=1}^{n}|\xi_i\eta_i| - \|\xi\|\|\eta\|.$$

Правую часть этого неравенства преобразуем следующим образом:

$$\sum_{i=1}^{n}|\xi_i\eta_i| - \|\xi\|_E\|\eta\|_E = \sum_{i=1}^{n}|\xi_i\eta_i| - \sqrt{(\sum_{i=1}^{n}\xi_i\eta_i)^2 + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}a_{ij}},$$

где

$$a_{ij} = (\xi_i\eta_j - \xi_j\eta_i)^2.$$

Анализ полученного выражения показывает, что

$$\sum_{i=1}^{n}|\xi_i\eta_i| - \|\xi\|_E\|\eta\|_E \leqslant 0$$

или

$$|\sum_{i=1}^{n}\xi_i\eta_i| \leqslant \|\xi\|_E\|\eta\|_E. \qquad (1.4)$$

Неравенство (1.4) получило название *неравенства Коши — Буняковского.* С учетом его из выражения (1.3) легко получить доказываемое неравенство (1.2).

В нормированном пространстве можно определить *расстояние* между векторами x и y как норму их разности, т. е. $\rho(x, y) = \|y - x\|$. Можно проверить, что число $\rho(x, y)$ при этом удовлетворяет следующим условиям:

1) $\rho(x, y) \geqslant 0$, причем для $\rho(x, y) = 0 \Leftrightarrow x = y$,
2) $\rho(x, y) = \rho(y, x)$, т. е. $\|y - x\| = \|x - y\|$,
3) $\rho(x, z) \leqslant \rho(x, y) + \rho(y, x)$ для $\forall\, x, y, z \in L$,
   т. е. $\|x - z\| \leqslant \|x - y\| + \|y - z\|$.

Определение 1.4. Если для любых элементов x, y, z множества M удовлетворяются три приведенных выше условия, то такое множество называется *метрическим пространством*, а число $\rho(x, y)$ — расстоянием между элементами $x, y \in M$.

С понятием расстояния в нормированных пространствах непосредственно связано такое понятие, как предел последовательности векторов в нормированном пространстве.

Определение 1.5. Последовательность вектора $\{x_n\}$, принадлежащая нормированному пространству, сходится к вектору x_0, если для каждого $\varepsilon > 0$ найдется такое $n_0(\varepsilon)$, что $\|x_n - x_0\| < \varepsilon$

для всех $n>n_0$. Вектор x_0 в этом случае называется *пределом* последовательности векторов $\{x_n\}$.

Приведем три свойства, связанных с пределом последовательности векторов.

Свойство 1. В нормированном пространстве из сходимости последовательности векторов $\{x_n\}$ к вектору x_0 следует сходимость числовой последовательности $\{\|x_n\|\}$ к норме $\|x_0\|$.

Из второго свойства для норм векторов следует, что

$$|\,\|x_n\| - \|x_0\|\,| \leqslant \|x_n - x_0\|.$$

Далее на основании определения 1.5 имеем $\|x_n - x_0\| < \varepsilon$ для всех $n>n_0$. Отсюда $|\,\|x_n\| - \|x_0\|\,| \leqslant \varepsilon$, а это означает не что иное, как условие сходимости числовой последовательности $\{\|x_n\|\}$, пределом которой является норма $\|x_0\|$.

Свойство 2. Если в нормированном пространстве последовательность векторов $\{x_n\}$ сходится, то предел этой последовательности единственный.

Предположим, что последовательность векторов $\{x_n\}$ имеет два предела x' и x''. Тогда на основании сформулированного выше второго свойства

$$\|x_n\| \longrightarrow \|x'\| \text{ и } \|x_n\| \longrightarrow \|x''\|.$$

Таким образом, получили одну числовую последовательность $\{\|x_n\|\}$, имеющую два предела. Но это, как известно из курса математического анализа, противоречит свойству единственности предела сходящейся числовой последовательности.

Свойство 3. Алгебраические операции, введенные в нормированное пространство, непрерывны. Так, если $x_n \to x$ и $y_n \to y$, то $x_n + y_n \to x + y$, т. е. $\lim\limits_{n\to\infty} (x_n + y_n) = x + y$, а также если $\alpha_n \to \alpha$ и $x_n \to x$, то $\lim \alpha_n x_n = \alpha x$.

В качестве упражнения предоставляется доказать эти утверждения самостоятельно.

Целый ряд вопросов в вычислительных методах, в том числе и при оценке сходимости различных итерационных процессов, связан с понятием эквивалентности норм.

Определение 1.6. Нормы $\|x\|_1$ и $\|x\|_2$ назовем *эквивалентными*, если существуют такие положительные числа α_1 и α_2, что

$$\alpha_1 \|x\|_1 \leqslant \|x\|_2 \leqslant \alpha_2 \|x\|_1.$$

Теорема 1.1. Для сходимости последовательности векторов $\{x_k\}$ конечномерного линейного пространства по норме $\|*\|_1$ необходимо и достаточно, чтобы $\{x_k\}$ сходилась по любой эквивалентной ей норме.

Необходимость: пусть $x_k \to x_0$ по норме $\|*\|_1$. Тогда из определения 1.5 следует, что для ε/α_2, где $\varepsilon, \alpha_2 > 0$, найдется

такое $n_0(\varepsilon/\alpha_2)$, начиная с которого $\|x_k - x_0\|_1 < \varepsilon/\alpha_2$. В результате для эквивалентной нормы $\|*\|_2$ будем иметь

$$\|x_k - x_0\|_2 \leqslant \alpha_2 \|x_k - x_0\|_1 < \varepsilon,$$

а это и доказывает сходимость последовательности векторов $\{x_k\}$ к вектору x_0 по норме $\|*\|_2$, эквивалентной норме $\|*\|_1$.

Достаточность: пусть $x_k \to x_0$ по какой-то норме $\|*\|_2$, эквивалентной норме $\|*\|_1$. Тогда для $(\varepsilon\alpha_1)$, где $\varepsilon, \alpha_1 > 0$, найдется $m_0(\alpha_1\varepsilon)$, начиная с которого $\|x_k - x_0\| < \varepsilon\alpha_1$. Отсюда

$$\|x_k - x_0\|_1 \leqslant \|x_k - x_0\|_2 / \alpha_1 < \varepsilon,$$

что и доказывает достаточное условие теоремы.

Из приведенной теоремы следует, что l-норма, c-норма и евклидова норма в арифметическом пространстве попарно эквивалентны между собой. Поэтому если обеспечивается сходимость последовательности векторов по одной из перечисленных норм, то она обеспечивается и по любой другой норме, определенной в арифметическом пространстве.

В линейных пространствах определить сходимость последовательности векторов $\{x_k\}$ можно не только через норму, но и через сходимости последовательностей соответствующих координат векторов заданной последовательности $\{x_k\}$. Пусть система векторов $\{e_i\}_{i=1 \div n}$ — базис в n-мерном линейном пространстве X, тогда любой вектор x_m из $\{x_k\}$ можно представить как

$$x_m = \sum_{i=1}^{n} \xi_{m_i} e_i.$$

Если вектор $x_0 = \sum_{i=1}^{n} \xi_i^0 e_i$ и имеет место сходимость числовой последовательности $\{\xi_{k_i}\}$ к ξ_i^0 для всех $i = 1 \div n$, то будем говорить, что последовательность векторов $\{x_k\}$ сходится к вектору $\{x_0\}$. Назовем сходимость последовательности векторов $\{x_k\}$ к $\{x_0\}$ по описанному выше правилу *координатной сходимостью* (или сходимостью по координатам).

Теорема 1.2. В конечномерном нормированном пространстве определения сходимости по норме и координатам эквивалентны.

С доказательством этой теоремы можно познакомиться, например, в [6].

Приведем несколько символических записей для сходимости последовательности векторов $\{x_k\}$ к x_0. Так, из определения 1.5 следует:

$$\lim_{k \to \infty} x_k = x_0, \text{ если } \lim_{k \to \infty} \|x_k - x_0\| = 0,$$

или

$$x_k \underset{k\to\infty}{\longrightarrow} x_0, \text{ если } \| x_k - x_0 \| \underset{k\to\infty}{\longrightarrow} 0,$$

из определения координатной сходимости:

$$\lim_{k\to\infty} x_k = x_0, \text{ если } \lim_{k\to\infty} \xi_{k_i} = \xi_i^0 \text{ для всех } i = 1 \div n,$$

или

$$x_k \underset{k\to\infty}{\longrightarrow} x_0, \text{ если } \xi_{k_i} \underset{k\to\infty}{\longrightarrow} \xi_i^0 \text{ для всех } i = 1 \div n.$$

Из доказанной теоремы следует

$$\| x_k - x_0 \| \underset{k\to\infty}{\longrightarrow} 0 \Leftrightarrow \xi_{k_i} \longrightarrow \xi_i \text{ для всех } i = 1 \div n. \qquad (1.5)$$

§ 1.3. ЕВКЛИДОВО И УНИТАРНОЕ ПРОСТРАНСТВА

Евклидовы и унитарные пространства составляют наиболее распространенный для практики вычислений класс нормированных линейных пространств.

Определение 1.7. Вещественное линейное пространство E называется *евклидовым*, если каждой паре векторов x, y из E поставлено в соответствие вещественное число (x, y), называемое *скалярным* (или *внутренним*) произведением, для которого выполняются следующие условия:

1) $(x, y) = (y, x)$,
2) $(\alpha x, y) = \alpha(x, y)$, где α — вещественное число,
3) $(x+y, z) = (x, z) + (y, z)$ для $\forall\, x, y, z \in E$,
4) $(x, x) \geqslant 0$, причем $(x, x) = 0 \Leftrightarrow x = 0$.

Аналогично евклидову пространству определяется унитарное пространство, основным полем в котором служит поле комплексных чисел.

Определение 1.8. Комплексное линейное пространство U называется *унитарным*, если каждой паре векторов x, y из U поставлено в соответствие комплексное число (x, y), называемое *скалярным* (или *внутренним*) произведением, для которого выполняются следующие условия:

1) $(x, y) = \overline{(y, x)}$,
2) $(\alpha x, y) = \alpha(x, y)$, где α — комплексное число,
3) $[x+y, z) = (x, z) + (y, z)$ для $\forall\, x, y, z \in U$,
4) $(x, x) \geqslant 0$, причем $(x, x) = 0 \Leftrightarrow x = 0$.

Черта в первом условии означает комплексное сопряжение.

Остановимся на свойствах скалярного произведения, введенного в линейном пространстве.

Свойство 1. В комплексном линейном пространстве

$$(x, \alpha y) = \bar{\alpha}(x, y).$$

Покажем это. Из условия 1 определения 1.8 следует, что

$$(x, \alpha y) = \overline{(\alpha y, x)}.$$

Воспользовавшись теперь условием 2, получим

$$(x, \alpha y) = \bar{\alpha}\overline{(y, x)},$$

и после применения условия 1 последнее равенство преобразуется так, что

$$(x, \alpha y) = \bar{\alpha}\overline{(x, y)} = \bar{\alpha}(x, y).$$

Для действительного линейного пространства это свойство принимает вид

$$(x, \alpha y) = \alpha(x, y).$$

Свойство 2. В действительном (или комплексном) линейном пространстве выполняется равенство

$$(z, x+y) = (z, x) + (z, y).$$

Доказательство этого свойства в порядке упражнения предоставляем читателю.

Приведем несколько примеров.

1. В арифметическом пространстве R_n, в котором ξ_i при $i = 1 \div n$ — координаты вектора-столбца (строки) ξ, определим умножение столбцов ξ и η формулой

$$(\xi_1 \xi_2 \dots \xi_n)^{\tau}(\eta_1 \eta_2 \dots \eta_n)^{\tau} = \sum_{i=1}^{n} \xi_i \eta_i. \tag{1.6}$$

Покажем, что введенная операция является операцией скалярного умножения между векторами в пространстве R_n. Для этого проверим для введенной операции выполнение всех условий, указанных в определении 1.7. Проверка 1-го условия вытекает из числового равенства

$$\sum_{i=1}^{n} \xi_i \eta_i = \sum_{i=1}^{n} \eta_i \xi_i.$$

Пусть $\xi + \eta = \gamma$, где $\gamma \in R_n$, тогда исходя из формулы (1.6), умножение между столбцами γ и $q \in R_n$ примет вид

$$\sum_{i=1}^{n} \gamma_i q_i = \sum_{i=1}^{n} (\xi_i + \eta_i) q_i = \sum_{i=1}^{n} (\xi_i q_i + \eta_i q_i) =$$

$$= \sum_{i=1}^{n} \xi_i q_i + \sum_{i=1}^{n} \eta_i q_i.$$

Отсюда следует выполнение 3-го условия. По аналогии можно выполнить проверку и оставшихся условий, которую оставляем для читателя.

Таким образом, пространство R_n с операцией между его векторами, определяемой по формуле (1.6), является евклидовым пространством, а сама операция — скалярным произведением в R_n.

2. В пространстве $C[a, b]$ определим скалярное произведение функций формулой

$$(f(x), g(x)) = \int_a^b f(x)g(x)dx. \tag{1.7}$$

Условия 1, 2 и 3 в определении 1.7 выполняются из следующих свойств определенного интеграла:

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

$$\int_a^b \alpha\gamma(x)dx = \alpha \int_a^b \gamma(x)dx,$$

$$\int_a^b [f(x)+g(x)]dx = \int_a^b f(x)dx + \int_a^b g(x)dx.$$

Для проверки условия 4 этого определения воспользуемся такими свойствами:

1) определенный интеграл от функции $f(x) \geqslant 0$ является величиной положительной;

2) неотрицательная функция, интеграл от которой на отрезке $[a, b]$ равен нулю, тождественно равна нулю на этом отрезке.

В евклидовом пространстве можно ввести норму с помощью формулы

$$\| x \|_E = \sqrt{(x, x)} = \sqrt{\sum_{i=1}^{n} x_i^2}, \tag{1.8}$$

где $x \in E$,

а в унитарном пространстве

$$\| x \|_U = \sqrt{(x, x)} = \sqrt{\sum_{i=1}^{n} | x_i |^2}, \tag{1.9}$$

где $x \in U$.

Доказательство того, что выражение (1.9) является нормой в пространстве R_n, было дано при рассмотрении евклидовой нормы $\|\xi\|_E$. В том же примере доказывается и неравенство Коши — Буняковского для евклидовых пространств, которое в обозначениях, принятых для скалярного произведения, имеет вид

$$| (x, y) |^2 \leqslant (x, x)(y, y). \tag{1.10}$$

Нетрудно проверить, что равенство в формуле (1.10) достигается только тогда, когда векторы x, y коллинеарны, т. е. $x=\alpha y$, где α — число из поля P.

Покажем, что в евклидовом (унитарном) пространстве скалярное произведение непрерывно относительно сходимости по норме. Действительно, если $x_n \to x$ и $y_n \to y$, то на основании свойства 1 определения 1.5 следует, что числовые последовательности $\|x_n\|$ и $\|y_n\|$ ограничены сверху, в связи с чем найдется такое число M, что $\|x_n\|_E \leqslant M$ и $\|y_n\|_E \leqslant M$. Отсюда

$$\begin{gathered}|(x_n, y_n)-(x, y)|=|(x_n, y_n)-(x, y_n)+ \\ +(x, y_n)-(x, y)|=|((x_n-x), y_n)+(x, (y_n-y))| \leqslant \\ \leqslant \|x_n-x\|_E\|y_n\|_E+\|x\|_E\|y_n-y\|_E \leqslant \|x_n-x\|_E M+ \\ +\|x\|_E\|y_n-y\|_E \longrightarrow 0 \text{ при } n \longrightarrow \infty.\end{gathered}$$

Задание в линейном пространстве L скалярного произведения позволяет ввести в этом пространстве не только норму вектора, но и угол между векторами.

Определение 1.9. Угол φ между ненулевыми векторами x и y определяется равенством

$$\cos\varphi=(x, y)/(\|x\|_E\|y\|_E). \tag{1.11}$$

Из неравенства Коши — Буняковского непосредственно следует, что в любом евклидовом (унитарном) пространстве можно определить угол по формуле (1.11), при этом косинус угла между векторами по модулю не превосходит единицы.

Векторы x и y называются *ортогональными*, если скалярное произведение между ними равно нулю.

Определение 1.10. Система ненулевых векторов евклидова (унитарного) пространства называется *ортогональной*, если она состоит из одного вектора или ее векторы попарно ортогональны. Нормированная ортогональная система векторов называется *ортонормированной*.

Убедимся в том, что система ортогональных векторов *линейно независима*. Предположим, что для системы ортогональных векторов $x_1, x_2, \ldots, x_i, \ldots, x_m$ пространства E_n выполняется равенство

$$\alpha_1x_1+\alpha_2x_2+\ldots+\alpha_ix_i+\ldots+\alpha_mx_m=0.$$

Умножая обе части этого равенства скалярно на x_i, получим

$$\alpha_1(x_1, x_i)+\alpha_2(x_2, x_i)+\ldots+\alpha_i(x_i, x_i)+\ldots+\alpha_m(x_m, x_i)=0.$$

Поскольку (x_i, x_k) при $i\neq k$ и $(x_i, x_i)\neq 0$, то имеем $\alpha_i=0$ при всех $i=1\div m$, а это и указывает на независимость системы векторов $\{x_i\}_{i=1\div m}$. Естественно, что доказанное выше утверждение распространяется и на систему ортонормированных векторов.

Напомним теперь определение базиса линейного (векторного) пространства, в соответствии с которым любая совокуп-

ность n линейно независимых векторов n-мерного линейного пространства называется его *базисом*. На основании данного определения и условия линейной независимости системы ортогональных векторов следует вывод о том, что в n-мерном линейном пространстве любая система ортогональных векторов $x_1, x_2, \ldots, x_i, \ldots, x_n$ образует базис в этом пространстве. Кроме того, если система векторов $e_1, e_2, \ldots, e_n$ — базис n-мерного евклидова пространства E_n и в этом пространстве имеется вектор $x=x_1e_1+x_2e_2+\ldots+x_ne_n$, для которого выполняется равенство $(x,x)=x_1^2+x_2^2+\ldots+x_n^2$, то базис $\{e_i\}_{i=1\div n}$ является ортонормированным. Поэтому евклидова норма любого вектора $y\in E_n$ будет определяться в этом базисе как $\|y\|_E=(\sum\limits_{i=1}^{n} y_i^2)^{1/2}$, где y_i — координаты вектора y в базисе $\{e_i\}_{i=1\div n}$. Доказать сформулированные утверждения предлагаем читателю.

Теорема 1.3. В любом конечномерном евклидовом (унитарном) пространстве существует ортонормированный базис.

Доказательство. Пусть система векторов $x_1, x_2, \ldots, x_n$ образует базис в n-мерном линейном пространстве L_n. Используя линейные операции сложения векторов и умножения вектора на скаляр, построим в этом пространстве новую систему векторов $\{y_i\}_{i=1\div n}$ по правилу

$$\begin{aligned} &y_1=x_1,\ y_2=x_2-\alpha_{21}y_1,\ \ldots,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &y_i=x_i-\alpha_{i1}y_1-\alpha_{i2}y_2-\ldots-\alpha_{i,i-1}y_{i-1},\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &y_n=x_n-\alpha_{n1}y_1-\alpha_{n2}y_2-\ldots-\alpha_{n,n-1}y_{n-1}. \end{aligned} \tag{1.12}$$

Элементы α_{ij} будем вычислять по формулам:

$$\alpha_{ij}=\begin{cases}\alpha_{ij}=0 \text{ при } i<j,\\ \alpha_{ij}=(x_i, y_i)/(y_i, y_i) \text{ при } i>j.\end{cases} \tag{1.13}$$

Из формул (1.13) следует, что преобразования (1.12) возможны только при условии, что $y_i\neq 0$ для всех $i=1\div n$. Для $i=1$ и $i=2$ выполнение этого условия непосредственно вытекает из первой строчки преобразований (1.12). Допустим, что при $i=k$ вектор $y_k\neq 0$. Тогда базисный вектор x_k можно представить в виде линейной комбинации системы $\{y_i\}_{i=1\div k}$, т. е.

$$x_k=y_k+\alpha_{k1}y_1+\alpha_{k2}y_2+\ \ldots\ +\alpha_{k,k-1}y_{k-1},$$

которая является независимой и образует базис в k-мерном подпространстве исходного линейного пространства L_n. Отсюда вытекает, что система векторов $y_1, y_2, \ldots, y_k, x_{k+1}$ независима, а следовательно, вектор

$$y_{k+1}=x_{k+1}-\alpha_{k+1,1}y_1-\alpha_{k+1,2}y_2-\ldots-\alpha_{k+1,k}y_k$$

не равен нулю. Тем самым нами установлено, что в системе векторов $\{y_i\}_{i=1\div n}$ нет нулевых векторов.

Рассмотрим теперь скалярные произведения (y_i, y_j) при условии, что $i>j$ и $i=1\div n$. Тогда для $i=2$ и $j=1$ имеем

$$(y_2, y_1)=(x_2, x_1)-\alpha_{21}(y_1, x_1),$$

где

$$\alpha_{21}=(x_2, y_1)/(y_1, y_1).$$

После несложных преобразований данного выражения получаем $(y_2, y_1)=0$. Если предположить попарную ортогональность векторов $y_1, y_2, y_s, \ldots, y_i$, то имеем систему равенств:

$$\begin{aligned}
&(y_{i+1}, y_1)=(x_{i+1}, y_1)-\alpha_{i+1,1}(y_1, y_1),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&(y_{i+1}, y_s)=(x_{i+1}, y_s)-\alpha_{i+1,s}(y_s, y_s),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&(y_{i+1}, y_i)=(x_{i+1}, y_i)-\alpha_{i+1,i}(y_i, y_i).
\end{aligned}$$

Здесь численные значения $\alpha_{i+1,\,s}$ вычисляются по формулам (1.13). Выполнив ряд вычислений, убедимся, что стоящие справа в приведенных равенствах скалярные произведения равны нулю при всех $s=1\div i$. Таким образом, получили, что система векторов $\{y_s\}_{s=1\div(i+1)}$ состоит из попарно ортогональных векторов, а следовательно, попарно ортогональна и вся система векторов $\{y_i\}_{i=1\div n}$, которая образует базис в пространстве L_n.

Если теперь каждый из векторов y_i поделить на число, равное евклидовой норме этого вектора, то получится ортонормированный базис, образованный векторами

$$y_1^0=\frac{y_1}{\|y_1\|_E},\quad y_2^0=\frac{y_2}{\|y_2\|_E},\ \ldots,\ y_i^0=\frac{y_i}{\|y_i\|_E},\ \ldots,\ y_n^0=\frac{y_n}{\|y_n\|_E}. \qquad (1.14)$$

Процесс перехода от базиса $\{x_i\}_{i=1\div n}$ к ортонормированному базису $\{y_i^0\}_{i=1\div n}$ по формулам (1.12), (1.13) и (1.14) получил название *метода ортогонализации Грамма — Шмидта.*

§ 1.4. АЛГОРИТМЫ ОРТОГОНАЛЬНЫХ РАЗЛОЖЕНИЙ МАТРИЦ. ОПРЕДЕЛИТЕЛЬ ГРАМА

Ортогонализация Грама — Шмидта может быть представлена и в матричной форме записи. Для этого перепишем равенства (1.12) в виде

$$\begin{aligned}
&x_1=y_1,\\
&x_2=\alpha_{21}y_1+y_2,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&x_i=\alpha_{i1}y_1+\alpha_{i2}y_2+\ldots+\alpha_{i,i-1}y_{i-1}+y_i,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&x_n=\alpha_{n1}y_1+\alpha_{n2}y_2+\ldots+\alpha_{n,n-1}y_{n-1}+y_n.
\end{aligned} \qquad (1.15)$$

Если теперь определить A как матрицу, у которой j-столбцом является вектор x_j из (1.15), Q — как матрицу с j-столбцом в

виде вектора y_i и R — как верхнюю треугольную матрицу с единичными диагональными и наддиагональными элементами, задаваемыми $\alpha_{i,j}$ и вычисляемыми по формулам (1.13), то выражение (1.15) можно представить в виде

$$A=QR, \tag{1.16}$$

где $A=(x_1, x_2, \ldots, x_j, \ldots, x_n)$, $Q=(y_1, y_2, \ldots, y_j, \ldots, y_n)$ — ортогональная матрица, и

$$R=\begin{pmatrix} 1 & \alpha_{21} & \ldots & \alpha_{i1} & \ldots & \alpha_{n-1,1} & \alpha_{n1} \\ 0 & 1 & \ldots & \alpha_{i2} & \ldots & \alpha_{n-1,2} & \alpha_{n2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 01 & \ldots & \alpha_{n-1,i} & \alpha_{ni} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & \ldots & 01 & \alpha_{n,n-1} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \ldots & \ldots & 0 & 1 \end{pmatrix} \tag{1.17}$$

Для окончательного описания процесса перехода в пространстве L_n от произвольного базиса $(x_1, x_2, \ldots, x_n)$ к ортонормированному $(y_1^0, y_2^0, \ldots, y_n^0)$ остается отразить в матричной форме преобразования (1.14). Они представляются как

$$(y_1 y_2 \ldots y_n)=(y_1^0 y_2^0 \ldots y_n^0)D, \tag{1.18}$$

где D — диагональная матрица с диагональю $\operatorname{diag}(\|y_1\| \|y_2\| \ldots \ldots \|y_n\|)$. После подстановки (1.18) в (1.16) получаем ортогонализацию Грама — Шмидта в матричной форме записи

$$A=NDR, \tag{1.19}$$

в которой $N=(y_1^0 y_2^0 \ldots y_n^0)$ — ортонормированная матрица.

Среди ортонормированных матриц большое применение в методах вычислений получили матрицы перестановок.

Определение 1.11. Квадратная матрица P размера $[n \times n]$ называется матрицей перестановки, если ее элементы являются результатом перестановки $(P_1, P_2, \ldots, P_n)$ последовательности символов 1, 2, ..., n согласно условию

$$P_{ij}=\begin{cases} 1 \text{ при } j=P_i, \\ 0 \text{ при остальных значениях } j. \end{cases} \tag{1.20}$$

Нетрудно проверить, что при умножении матрицы перестановки на вектор $x=(x_1 x_2 \ldots x_n)^{\mathrm{T}}$ получается вектор $\hat{x}=(x_{p_1} x_{p_2} \ldots \ldots x_{p_n})^{\mathrm{T}}$. Если теперь матрицу A размера $[n \times m]$ представить в виде $A=(a_1 a_2 \ldots a_j \ldots a_m)$, где a_j — вектор-столбец матрицы A, равный $(a_{1j} a_{2j} \ldots a_{nj})^{\mathrm{T}}$, то результатом умножения PA будет матрица $\hat{A}$ размера $[n \times m]$:

$$\hat{A}=(\hat{a}_1 \hat{a}_2 \ldots \hat{a}_j \ldots \hat{a}_m), \tag{1.21}$$

в которой вектор-столбец $\hat{a}_j=(a_{p_1j}a_{q_2j}\dots a_{p_nj})^{\mathrm{T}}$ для всех $j=1\div n$. Таким образом, в матрице $\hat{A}$ по сравнению с матрицей A в соответствии с заданной перестановкой $(P_1P_2\dots P_n)$ переставлены только строки. Для того чтобы получить из матрицы B размера $[m\times n]$ матрицу $\hat{B}$ с аналогичной перестановкой $(P_1P_2\dots P_n)$, но уже столбцов, требуется выполнить умножение матрицы B справа на матрицу P^{T}, т. е.

$$\hat{B}=BP^{\mathrm{T}}. \tag{1.22}$$

Будем называть перестановку двух строк (столбцов) матрицы или двух компонент вектора *элементарной перестановкой. Матрицу элементарной перестановки* k и l номеров строк или столбцов матрицы, или компонент вектора будем обозначать P_{kl}. Матрица P_{kl} отличается от единичной матрицы такого же размера только четырьмя элементами: $P_{ii}=P_{jj}=0$; $P_{ij}=P_{ji}=1$. Так как $P_{kl}=P^{\mathrm{T}}{}_{kl}$, то использование матрицы элементарной перестановки для транспозиции строк или столбцов в исходной матрице A зависит только от ее умножения слева или справа на матрицу P_{kl}. Можно показать, что любая матрица перестановок P получается путем последовательного перемножения матриц элементарных перестановок, т. е.

$$P=P_{k_tl_t}P_{k_{t-1}l_{t-1}}\cdot\ \cdot\ \cdot P_{k_2l_2}P_{k_1l_1}, \tag{1.23}$$

где число элементарных перестановок t равно числу несовпадений в перестановках (1, 2, ..., n) и $(P_1P_2\dots P_n)$ минус 1. В порядке упражнения вывод формулы (1.23) рекомендуем выполнить самостоятельно.

В качестве примера рассмотрим матрицу

$$A=\begin{pmatrix}0 & 0 & a_{13} & a_{14}\\ 0 & a_{22} & a_{23} & a_{24}\\ 0 & 0 & 0 & a_{34}\\ a_{41} & a_{42} & a_{43} & a_{44}\end{pmatrix},$$

которую только за счет перестановки строк можно привести к треугольному виду. Для решения этой задачи требуется выполнить следующие преобразования: 1-ю строку матрицы A переставить на место 3-й строки, 2-ю строку — на место 2-й строки, т. е. оставить на месте, 3-ю строку — на место 4-й строки и 4-ю строку — на место 1-й строки. Эти преобразования описываются перестановкой вида (3, 2, 4, 1). На основании формулы (1.20) напишем матрицу перестановки P:

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

Эту матрицу можно получить и по формуле (1.23) как произведение двух матриц элементарных преобразований, т. е.

$$P=P_{k_2i_2}P_{k_1i_1}=\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&0&1\\0&0&1&0\end{pmatrix}\begin{pmatrix}0&0&0&1\\0&1&0&0\\0&0&1&0\\1&0&0&0\end{pmatrix}.$$

Выполнив умножение PA, получим требуемый результат

$$\hat{A}=PA=\begin{pmatrix}a_{41}&a_{42}&a_{43}&a_{44}\\0&a_{22}&a_{23}&a_{24}\\0&0&a_{13}&a_{14}\\0&0&0&a_{34}\end{pmatrix}.$$

Пусть в n-мерном линейном пространстве задана система векторов $\{x_i\}_{i=1\div m}$, из которых $(m-r)$ векторов линейно зависимы. Задача состоит в ортогонализации r — линейно независимых векторов. Отметим, что между величинами m, n, r существует шесть вариантов в постановке данной задачи. Это когда

1) $m=n$, $r=n$; 2) $m=n$, $r<m$; 3) $m<n$, $r=m$;
4) $m<n$, $r<m$; 5) $m>n$, $r=n$; 6) $m>n$, $r<n$.

Первый из перечисленных вариантов решается методом ортогонализации Грама — Шмидта, рассмотренным в § 1.3. Из остальных вариантов наиболее общий характер носят варианты 4 и 6, которые с учетом замены индекса m на n, а n на m полностью совпадают. Поэтому для решения поставленной выше задачи достаточно рассмотреть вариант 4.

Теорема 1.4. Матрица A размера $[n\times m]$ при $m<n$ и $r=\mathrm{Rg}\,A<m$ может быть представлена как

$$A=NDRP, \tag{1.24}$$

где N — матрица с ортонормированными столбцами y_j, принимающая вид

$$N=(y_1{}^0y_2{}^0\dots y_r{}^0\underbrace{O\dots O}_{m-r}),$$

D — диагональная матрица размера $[m\times m]$, у которой элементы $d_{ii}\neq 0$ при $i=1\div r$ и $d_{ii}=0$ при $i=(r+1)\div m$; R — матрица размера $[m\times m]$, имеющая следующее блочное представление:

$$R=\begin{pmatrix}R_{11}&R_{12}\\O_{21}&O_{22}\end{pmatrix}. \tag{1.25}$$

Здесь R_{11} — верхняя треугольная матрица размера $[r \times r]$ с единичной диагональю; R_{12} — матрица размера $[r \times (m-r)]$; O_{21} и O_{22} — нулевые матрицы соответственно размеров $[(m-r) \times r]$ и $[(m-r) \times (m-r)]$; P — матрица перестановок.

Доказательство. Представим исходную матрицу A в виде $(a_1 a_2 \dots a_j \dots a_m)$, где a_j — вектор-столбец. С помощью матрицы перестановки P осуществим перестановку столбцов в матрице A таким образом, чтобы векторы-столбцы a_{p_j} при $j = 1 \div r$ были линейно независимые, а a_{p_j} при $j = (r+1) \div m$ — линейно зависимые, т. е.

$$\hat{A} = AP^{\mathrm{T}} \Rightarrow \hat{A} = (a_{p_1} a_{p_2} \dots a_{p_r} a_{p_{r+1}} \dots a_{p_m}). \qquad (1.26)$$

Выполним теперь ортонормирование системы векторов $\{a_{p_j}\}_{j=1 \div r}$, например, методом Грама — Шмидта, в результате чего получим ортонормированную систему векторов $\{a^0{}_{p_j}\}_{j=1 \div r}$, образующую базис в пространстве $R_r \subset R_n$. Векторы $a_{p_{r+1}}$, $a_{p_{r+2}}$, ..., a_{p_m}, входящие в (1.26) и принадлежащие R_r, представим в виде линейной комбинации базисных векторов $\{a^0{}_{p_j}\}_{j=1 \div r}$:

$$a_{p_j} = \alpha_{1j} a_{p_1}{}^0 + \alpha_{2j} a_{p_2}{}^0 + \dots + \alpha_{rj} a_{p_r}{}^0 \quad \text{для} \quad j = (r+1) \div m.$$

При этом коэффициенты α_{ij} определяются как

$$\alpha_{ij} = (a_{p_j}, a_{p_i}{}^0) \text{ для } i = 1 \div r, \; j = (r+1) \div m. \qquad (1.27)$$

Выполненные преобразования можно записать следующим образом:

$$(a_{p_1} a_{p_2} \dots a_{p_r} a_{p_{r+1}} \dots a_{p_m}) = (a_{p_1}{}^0 a_{p_2}{}^0 \dots a_{p_r}{}^0 \underbrace{\bar{o} \dots \bar{o}}_{m-r}) S,$$

где $\bar{o}$ — вектор-столбец высотой n, S — матрица преобразования, имеющая вид

$$S = D \begin{pmatrix} R_{11} & R_{12} \\ O_{21} & O_{22} \end{pmatrix}.$$

Здесь R_{11} — верхняя треугольная матрица с единичной диагональю и элементами α_{ij}, вычисляемыми по формуле (1.13) для $i, j = 1 \div r$; R_{12} — матрица размера $[r \times (m-r)]$ с элементами α_{ij}, вычисляемыми по формуле (1.26) для $i = 1 \div r$ и $j = (r+1) \div m$.

С учетом формулы (1.25) получаем доказываемую формулу (1.24), в которой

$$N = (a_{p_1}{}^0 a_{p_2}{}^0 \dots a_{p_r}{}^0 \underbrace{o \dots o}_{m-r}).$$

Из формулы (1.24) вытекает следующая теорема об ортогональном разложении матрицы, которая широко используется в различных вычислительных алгоритмах.

Теорема 1.5. Пусть A — матрица размера $[n\times m]$, ранга r, причем $n>m$. Существуют ортогональная матрица Q размера $[n\times n]$ и матрица перестановок P размера $[m\times m]$, такие, что

$$QAP^{\mathrm{T}}=\begin{pmatrix} R_{11} & R_{12} \\ O_{(n-r)\times r} & O_{(n-r)\times(m-r)} \end{pmatrix}, \tag{1.28}$$

где R_{11} и R_{12} — матрицы, входящие в формулу (1.25).

Для доказательства теоремы выполним в матричном равенстве (1.24) умножение слева на ортонормированную по строкам матрицу F размера $[n\times n]$. При этом пусть матрица F имеет такую структуру, при которой первые r ее строк совпадают с первичными столбцами r матрицы N. Тогда после перемножения матриц F и N равенство (1.24) примет вид

$$FA=\hat{E}DRP, \tag{1.29}$$

где матрица $\hat{E}=FN$ имеет следующую блочную структуру:

$$\hat{E}=\begin{pmatrix} E_{r\times r} & O_{r\times(m-r)} \\ O_{(m-r)\times r} & O_{(m-r)\times(m-r)} \\ O_{(n-m)\times r} & O_{(n-m)\times(m-r)} \end{pmatrix}.$$

Здесь нижний индекс при матрицах обозначает их размерности.

Умножим равенство (1.29) слева на диагональную матрицу $\hat{D}$ размера $[n\times n]$, у которой элементы $\hat{d}_{ii}=1/d_{ii}$ при $i=1\div r$ и d_{ii} в общем случае не равны нулю для всех остальных $n-r$ значений i. В результате получим выражение

$$QA=\hat{E}RP,$$

которое после выполнения в нем матричных операций преобразуется к виду

$$QAP^{\mathrm{T}}=\hat{R}, \tag{1.30}$$

где $\hat{R}=\begin{pmatrix} R_{11} & R_{12} \\ O_{(n-r)\times r} & O_{(n-r)\times(m-r)} \end{pmatrix}$.

Изображая матрицу в виде прямоугольника соответствующего размера, где его заштрихованная часть отражает ту область матрицы, в которой ее элементы в общем случае не равны нулю, и разделяя подматрицы (блоки) утолщенными линиями, формулу (1.28) можно проиллюстрировать таким образом, как показано на рис. 1, *а*. Для случая, когда $m>n>r$,

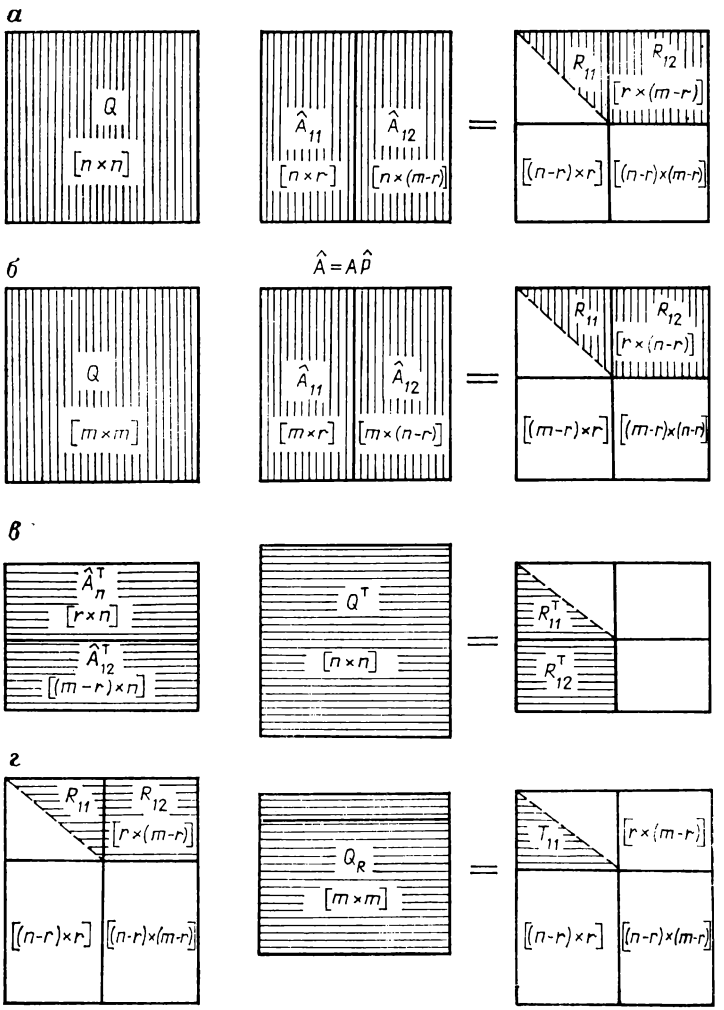


Рис. 1. Диаграммы ортогональных разложений матриц:
а — ортогональное разложение (1.28) матрицы A, для которой $n>m>r$; *б* — ортогональное разложение (1.28) матрицы A, для которой $m>n>r$; *в* — ортогональное разложение (1.31) матрицы A, для которой $n>m>r$; *г* — ортогональное разложение (1.32) матрицы A, для которой $n>m>r$

схематичное изображение формулы (1.28) приведено на рис. 1, *б*. Выполнив в равенстве (1.24) операцию транспонирования матриц, приходим к следующей формуле:

$$PA^{\mathrm{T}}Q^{\mathrm{T}}=\hat{R}^{\mathrm{T}}. \qquad (1.31)$$

Эту формулу можно изобразить в виде, показанном на рис. 1, *в*.

Дальнейшее упрощение структуры матрицы $\hat{R}$ в формуле (1.28) можно достигнуть за счет применения следующей теоремы.

Теорема 1.6. Пусть A — матрица размера $[n\times m]$ и ранга r, $n>m>r$. Найдутся такие ортогональные матрицы V размера $[n\times n]$ и W размера $[m\times m]$, что

$$VAW^{\mathrm{T}}=\begin{pmatrix} T_{11} & O \\ O & O \end{pmatrix}, \tag{1.32}$$

где T_{11} — нижняя треугольная матрица размера $[r\times r]$.

Доказательство. Рассмотрим равенство (1.31), в котором в качестве матрицы A^{T} принимают матрицу $\hat{R}$ из выражения (1.30). Тогда на основании теоремы 1.5 можно утверждать, что найдется ортогональная (по строкам) матрица $Q_R{}^{\mathrm{T}}$, при которой

$$RQ^{\mathrm{T}}{}_R=T.$$

Исходя из матричного равенства, представленного на рис. 1, *в*, приведенное равенство принимает вид, изображенный на рис. 1, *г*. Причем матрица T имеет такую же структуру, как и матрица, стоящая справа в равенстве (1.32).

После умножения слева матричного равенства (1.30) на матрицу $Q_R{}^{\mathrm{T}}$ получим доказываемое равенство

$$VAW^{\mathrm{T}}=T,$$

где матрица $V=Q$, а W^{T} — ортогональная матрица, равная $W=P^{\mathrm{T}}Q_R{}^{\mathrm{T}}$.

Заметим, что выбором V и W в уравнении (1.32) можно достигнуть, чтобы матрица T_{11} была и верхней треугольной.

Используя формулу (1.32), можно получить следующее разложение:

$$A=V^{\mathrm{T}}TW, \tag{1.33}$$

которое будем называть *ортогональным разложением матрицы A*.

Рассмотрим систему векторов $\{x_i\}_{i=1\div m}$, которая в общем случае может быть и линейно зависимой. Возникает задача установления зависимости или независимости заданной системы векторов, с которой непосредственно связаны такие важные для вычислительной практики задачи, как нахождение ранга матрицы и ортогональных преобразований матриц (см. теорему 1.4). При решении этих задач можно использовать функцию

$$G(x_1, x_2, \ldots, x_k)=$$

$$=\det\begin{pmatrix} (x_1, x_1) & (x_1, x_2) & \ldots & (x_1, x_j) & \ldots & (x_1, x_k) \\ (x_2, x_1) & (x_2, x_2) & \ldots & (x_2, x_j) & \ldots & (x_2, x_k) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ (x_i, x_1) & (x_i, x_2) & \ldots & (x_i, x_j) & \ldots & (x_i, x_k) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ (x_k, x_1) & (x_k, x_2) & \ldots & (x_k, x_j) & \ldots & (x_k, x_k) \end{pmatrix}, \tag{1.34}$$

получившую название *определителя Грама.* Матрица, составленная из скалярных произведений векторов системы $\{x_i\}_{i=1\div k}$, называется *матрицей Грама.*

Рассмотрим несколько свойств определителя Грама.

Свойство 1. Определитель Грама не изменяется при перемене любых двух векторов в системе $x_1, x_2, \ldots, x_k$. Так, если в системе $\{x_i\}_{i=1\div k}$ поменять местами какие-либо два вектора x_i и x_j, то в определителе Грама поменяются местами i-й и j-й столбцы, а также i-я и j-я строки. При этом определитель Грама два раза сменит знак и в результате останется без изменения.

Свойство 2. Определитель Грама не изменяется от прибавления к любому вектору системы $\{x_i\}_{i=1\div k}$ какой-либо линейной комбинации остальных векторов.

Для доказательства этого свойства достаточно разобрать случай, когда изменяется вектор x_1, так как остальные случаи, с учетом свойства 1, сводятся к данному. Рассмотрим в евклидовом пространстве систему векторов $y_1, x_2, \ldots, x_k$, где $y_1 = x_1 + \sum_{i=2}^{k} \alpha_i x_i$.

Определитель Грама для этой системы примет вид

$$G(y_1, x_2, \ldots, x_k) =$$

$$= \det \begin{pmatrix} (y_1, y_1) & (y_1, x_2) & & (y_1, x_k) \\ (x_2, y_1) & (x_2, x_2) & & (x_2, x_k) \\ (x_k, y_1) & (x_k, x_2) & & (x_k, x_k) \end{pmatrix}.$$

Обозначим i-ю строку определителя $G(y_1, x_2, \ldots, x_k)$ как b_i и преобразуем первую строку b_1:

$$b_1' = b_1 - \sum_{i=2}^{k} \alpha_i b_i.$$

Так как данное преобразование линейное, то по свойству определителей имеем

$$G(y_1, x_2, \ldots, x_k) = \det \begin{pmatrix} b_1' \\ b_2 \\ b_k \end{pmatrix},$$

где

$$b_1' = \left(\left(y_1 - \sum_{i=2}^{k} \alpha_i x_1,\ y_1\right), \left(y_1 - \sum_{i=2}^{k} \alpha_i x_i,\ x_2\right), \ldots, \left(y_1 - \sum_{i=2}^{k} \alpha_i x_i,\ y_k\right) \right).$$

Использовав свойства скалярных произведений и исходное преобразование вектора $y_1 = x_1 + \sum_{i=2}^{k} \alpha_i x_i$, получим

$$\det\begin{pmatrix} b_1' \\ b_2 \\ \dots \\ b_k \end{pmatrix} = \det\begin{pmatrix} (x_1, y_1) & (x_1, x_2) & \dots & (x_1, x_k) \\ (x_2, y_1) & (x_2, x_2) & \dots & (x_2, x_k) \\ & & & \\ (x_k, y_1) & (x_k, x_2) & & (x_k, x_k) \end{pmatrix}.$$

Обозначим в этом определителе j-й столбец как C_j и преобразуем первый столбец C_1 таким образом:

$$C_1' = C_1 - \sum_{j=2}^{k} \alpha_j C_j.$$

Тогда по свойству определителей имеем

$$G(y_1, x_2, \dots, x_k) = \det(C_1' C_2 \dots C_k).$$

Если для определения компонент вектора C_1' выполнить преобразования, аналогичные тем, что и для определения компонент вектора b_1', то окончательно получим

$$G(y_1, x_2, \dots, x_k) = \det(x_1, x_2, \dots, x_k).$$

В тех случаях, когда векторы системы $\{x_i\}_{i=1\div k}$ принадлежат унитарному пространству, столбцы определителя Грама умножаются на комплексные числа $\bar{\alpha}_2, \dots, \bar{\alpha}_m$. В остальном доказательство этого свойства остается без изменения.

Теорема 1.7. Для определителя Грама:

1. Выполняется неравенство

$$0 \leqslant G(x_1 x_2 \dots x_k) \leqslant \prod_{j=1}^{k} (x_j, x_j). \tag{1.35}$$

2. Равенство нулю достигается тогда и только тогда, когда система векторов $\{x_i\}_{i=1\div k}$ линейно зависимая.

Доказательство. Нетрудно заметить, что матрица Грама для системы $\{x_i\}_{i=1\div k}$ представляется как

$$G = A^{\tau} A, \tag{1.36}$$

где матрица $A = (x_1, x_2, \dots, x_k)$, x_i — вектор-столбец, принадлежащий евклидову (или унитарному) пространству x_n. Предположим, что система $\{x_i\}_{i=1\div k}$ состоит из линейно независимых векторов. Тогда $\mathrm{Rg}\, A = k$. На основании теоремы 1.4 матрицу A можно представить в виде

$$A = NDR,$$

где R — верхняя треугольная матрица размера $[k \times k]$. С учетом данной формулы и формулы (1.36) получаем для матрицы G выражение вида

$$G = R^{\tau} D^2 R.$$

Переходя теперь к вычислению определителя Грама, будем иметь

$$\det G = \det (R^T D^2 R).$$

Используя свойства определителя и тот факт, что R — треугольная матрица с единичной диагональю, получаем

$$\det G = \det^2 R \det D^2 = \det D^2. \tag{1.37}$$

Таким образом, для системы $\{x_i\}_{i=1 \div k}$, состоящей из линейно независимых векторов, определитель Грама больше нуля, так как $\operatorname{Rg} D = k$.

Предположим, что система $\{x_i\}_{i=1 \div k}$ линейно зависимая. Тогда $\operatorname{Rg} A = r < k$ и, исходя из теоремы 1.4, имеем $\det D = 0$, так как $d_{ii} = 0$ при $i = (r+1) \div k$. А следовательно, равен нулю и определитель Грама.

Если задать теперь $G(x_1, x_2, \ldots, x_k) = 0$, то из формулы (1.37) вытекает, что $\det D = 0$. Отсюда имеем: ранг диагональной матрицы D равен $r < k$, а это равносильно тому, что система векторов $\{x_i\}_{i=1 \div k}$ линейно зависимая. Тем самым доказано второе утверждение теоремы.

Исходя из формулы (1.37), определитель Грама представляется в виде

$$G(x_1, x_2, \ldots, x_k) = \| y_1 \|_E^2 \| y_2 \|_E^2 \ldots \| y_k \|_E^2, \tag{1.38}$$

где $\{y_i\}_{i=1 \div k}$ — система ортогональных векторов, которые связаны с векторами из системы $\{x_i\}_{i=1 \div k}$ по формулам (1.12), а именно

$$y_i = x_j - \sum_{i=1}^{j-1} \alpha_{ij} y_i,$$

где

$$\alpha_{ij} = (x_j y_i) / (y_i, y_i), \; j = 1 \div k.$$

Переходя к определению (y_j, y_j) и используя свойство $(y_i, y_j) = 0$ при $i \neq j$, получим

$$(y_j, y_j) = (x_j, y_j)$$

и после дальнейших преобразований

$$(y_j, y_j) = (x_j, x_j) - \sum_{i=1}^{j-1} (x_j, y_i)^2 / (y_i, y_i). \tag{1.39}$$

Таким образом, получили $\|y_j\| \leqslant \|x_j\|$. Отсюда с учетом (1.38) имеем

$$0 \leqslant G(x_1, x_2, \ldots, x_k) \leqslant \prod_{i=1}^{k} (x_i, x_i).$$

Следствие. Если каждый из векторов системы $\{x_i\}_{i=1 \div k}$ ортогонален ко всем предшествующим векторам, то для опре-

делителя Грама справедливо равенство

$$G(x_1, x_2, \ldots, x_k) = \prod_{i=1}^{k} (x_i, x_i).$$

Справедливость этого утверждения вытекает из формулы (1.39), в которой при заданных в следствии условиях $(x_j, y_i) = 0$, а значит и $(y_j, y_j) = (x_j, x_j)$.

§ 1.5. ЛИНЕЙНЫЕ ОТОБРАЖЕНИЯ И ПРЕОБРАЗОВАНИЯ

Определение 1.12. Пусть L и L' — два линейных пространства над одним и тем же полем P. Под *отображением* $\mathcal{A}$ пространства L в пространство L' понимается правило, по которому каждому вектору x из L сопоставлен единственный вектор x' из L'. При этом вектор x' будем называть образом, а вектор x — прообразом вектора x'.

Для записи отображения используют ряд обозначений, например $\mathcal{A}: L \to L'$ или $x' = \mathcal{A}(x)$, или $x' = \mathcal{A}x$. Последняя запись наиболее часто применяется в линейной алгебре для обозначения линейных отображений.

Часто при отображении многомерного пространства в многомерное используется термин *оператор*, а при отображении многомерного пространства на числовое множество — термин *функционал*.

Определение 1.13. Отображение $\mathcal{A}$ называется *линейным*, если для любых векторов x и y из L и любого числа $\alpha \in P$ выполняются условия:

$$1)\ \mathcal{A}(x+y) = \mathcal{A}x + \mathcal{A}y, \qquad (1.40)$$

$$2)\ \mathcal{A}(\alpha x) = \alpha \mathcal{A}x.$$

Отметим, что операции сложения векторов и умножения на скаляр, стоящие в левых и правых частях условий (1.40), — различные операции, одни из которых определены в пространстве L, а другие — в L'.

Зададим в пространстве L_n базис $\{e_i\}_{i=1 \div n}$, а в пространстве L'_m — базис $\{d_i\}_{i=1 \div m}$. Тогда если $\mathcal{A}$ — линейное отображение $L_n \to L'_m$, то образ произвольного вектора $x = x_1 e_1 + x_2 e_2 + \ldots + x_n e_n$, принадлежащего L_n, примет вид

$$\mathcal{A}(x) = x_1 \mathcal{A}(e_1) + x_2 \mathcal{A}(e_2) + \ldots + x_n \mathcal{A}(e_n). \qquad (1.41)$$

Исходя из определения линейного отображения, $\mathcal{A}e_j$ является вектором, принадлежащим L'_m, а следовательно, в заданном базисе $\{d_i\}_{i=1 \div m}$ он будет представлен через свои координаты a_{ij} следующим образом:

$$\mathcal{A}e_j = a_{1j} d_1 + a_{2j} d_2 + \ldots + a_{ij} d_j + \ldots + a_{mj} d_m. \qquad (1.42)$$

Подставив (1.42) в формулу (1.41), в силу единственности разложения вектора по базису, получаем выражение для i-й координаты вектора $\mathscr{A}(x)$ в базисе $\{d_i\}_{i=1\div m}$:

$$y_i=\sum_{j=1}^{n} a_{ij}x_j \text{ для всех } i=1\div m. \tag{1.43}$$

Если из чисел a_{ij} составить матрицу A, то равенства (1.43) могут быть записаны в матричной форме

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ y_i \\ \cdot \\ y_m \end{pmatrix}=\begin{pmatrix} a_{11}a_{12}\ldots a_{1j}\ldots a_{1n} \\ a_{21}a_{22}\ldots a_{2j}\ldots a_{2n} \\ a_{i1}a_{i2}\ldots a_{ij}\ldots a_{in} \\ a_{m1}a_{n2}\ldots a_{nj}\ldots a_{mn} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ x_j \\ \cdot \\ x_n \end{pmatrix}, \tag{1.44}$$

в которой координатный столбец образа $\mathscr{A}(x)$ в базисе $\{d_i\}_{i=1\div m}$ пространства L'_m выражен через произведение матрицы A на координатный столбец прообраза x в базисе $\{e_i\}_{i=1\div n}$ пространства L_n.

Матрица A, входящая в выражение (1.44), называется *матрицей линейного отображения* $\mathscr{A}$. Столбцами этой матрицы являются координатные столбцы соответственно векторов $\mathscr{A}e_1$, $\mathscr{A}e_2$, ..., $\mathscr{A}e_n$ в базисе $\{q_i\}_{i=1\div m}$.

Таким образом, установлено, что каждому линейному отображению $\mathscr{A}: L_n \to L'_m$ в заданных базисах для этих пространств соответствует определенная матрица размера $[m\times n]$. Имеет место и обратное утверждение: каждая матрица A порядка $[m\times n]$ может рассматриваться как матрица линейного отображения $\mathscr{A}$ линейного пространства L_n в L'_m в некоторых базисах этих пространств.

Определение 1.14. Пусть $\mathscr{A}$ — линейное отображение: $L_n \to L_m$. Тогда множество всех векторов x' из L_m вида $x'=\mathscr{A}(x)$, где $x\in L_n$, называется *областью значений отображения* $\mathscr{A}$, а множество N всех x из L_n, таких, что $\mathscr{A}x=0$, где 0 — нулевой элемент пространства L_m, — *ядром* L_n.

Область значений отображения $\mathscr{A}$ будем обозначать как $\mathscr{A}L_n$. Нетрудно видеть, что $\mathscr{A}L_n \subseteq L_m$, а $N\subseteq L_n$, т. е. $\mathscr{A}L_n$ и N являются подпространствами пространств L_m и L_n. Ядро L_n будем обозначать символом $\ker \mathscr{A}$.

Если ввести между линейными отображениями операции сложения, умножения на скаляр и умножения как соответствующие операции между матрицами, а именно

$$\mathscr{C}x=\mathscr{A}x+\mathscr{B}x,\quad \mathscr{D}x=(\alpha\mathscr{A})x,\quad \mathscr{M}x=\mathscr{B}(\mathscr{A}x), \tag{1.45}$$

то выполняются следующие свойства относительно этих операций.

Свойство 1. Отображения $\mathscr{C}$, $\mathscr{D}$, $\mathscr{M}$, являющиеся результатами введенных выше операций над линейными отображениями, есть снова линейные отображения.

Свойство 2. Множества линейных отображений и матриц относительно операций сложений и умножения на число образуют линейные изоморфные пространства.

Вследствие изоморфизма между пространствами линейных отображений и матрицами этих отображений с формальной точки зрения безразлично, осуществляются ли линейные операции и преобразования над отображениями или над матрицами этих отображений. В связи с этим символ $\mathscr{A}x$ можно понимать как отображение вектора или как произведение матрицы на вектор-столбец.

Таким образом, в дальнейшем не будем делать различия между операторными и матричными соотношениями и все новые понятия и факты, имеющие место в отношении матриц (отображений), будем распространять и на отображения (матрицы).

Важным подклассом линейных отображений являются *линейные преобразования.* Линейное отображение будем называть линейным преобразованием, если пространства L и L' совпадают.

На основании формулы (1.44) для линейного преобразования $\mathscr{A}$ при фиксированном базисе $\{e_i\}_{i=1\div n}$ пространства R_n имеем

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ y_i \\ \cdot \\ y_n \end{pmatrix} = \begin{pmatrix} a_{11}a_{12}\ldots a_{1n} \\ a_{21}a_{22}\ldots a_{2n} \\ a_{i1}a_{i2}\ldots a_{in} \\ a_{n1}a_{n2}\ldots a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ x_j \\ \cdot \\ x_n \end{pmatrix}, \tag{1.46}$$

где x_j и y_i — координаты векторов x и y, принадлежащих пространству R_n. Квадратную матрицу $A=\{a_{ij}\}$, входящую в данное выражение, будем называть матрицей линейного преобразования $\mathscr{A}$.

Рассмотрим, как изменяется матрица линейного отображения при замене базисов в пространствах L_n и L'_m.

Пусть в базисах $\{e_i\}_{i=1\div n}$ и $\{d_i\}_{i=1\div m}$ соответственно пространств L_n и L'_m отображение $\mathscr{A}: L_n \to L'_m$ определяется матрицей A, а в другой паре базисов $\{e_i'\}_{i=1\div n}$ и $\{d_i'\}_{i=1\div m}$ это же отображение определяется матрицей $\tilde{A}$. Наша задача — найти взаимосвязь между матрицами A и $\tilde{A}$.

Произвольный вектор $u \in L_n$ и его образ $V=\mathscr{A}(u)$ могут быть представлены в заданных базисах через свои координатные столбцы. Обозначим координатные столбцы вектора и в базисах $\{e_i\}_{i=1\div n}$ и $\{e_i'\}_{i=1\div n}$ соответственно через x и x', а координатные столбцы вектора V в базисах $\{d_i\}_{i=1\div m}$ и $\{d_i'\}_{i=1\div m}$ — через y и y'. Тогда если переход от старого базиса к новому в пространстве L_n задан матрицей линейного преобразования S, а в пространстве L_m—T, то имеем

$$x=Sx',\ y=Ty'.$$

Подставляя эти выражения в формулу $y=Ax$, которая определяет в матричном виде линейное отображение $\mathscr{A}$, после несложных матричных преобразований получим

$$y'=T^{-1}ASx'. \tag{1.47}$$

Заметим, что для матриц T и S всегда имеются обратные матрицы в силу независимости координатных столбцов (строк) базисных векторов.

Поскольку $y'=\tilde{A}x'$ по определению $\tilde{A}$, то окончательно получаем

$$\tilde{A}=T^{-1}AS. \tag{1.48}$$

В случае линейного преобразования матрицы T и S совпадают. В связи с этим формула (1.48) для линейных преобразований приобретает вид

$$\tilde{A}=T^{-1}AT. \tag{1.49}$$

Матрицы $\tilde{A}$ и A, связанные этим преобразованием, называются *подобными*, само преобразование — *преобразованием подобия*, а матрица T — *матрицей подобного преобразования.*

Отметим ряд свойств преобразования подобия.

Свойство 1. Подобие матриц является отношением эквивалентности.

Действительно, если $\tilde{A}=T^{-1}AT$, то $A=T\tilde{A}T^{-1}\Leftrightarrow A=S^{-1}AS$, где $S^{-1}=T$. Следовательно, преобразования подобия *симметричны.*

Далее, если подобны A и B, B и C, то подобны A и C. Это свойство *транзитивности* преобразования подобия. Предлагаем его проверить самостоятельно.

И наконец, легко проверяется и свойство *рефлексивности* преобразования подобия, т. е. то, что матрица подобна самой себе.

Свойство 2. Определители подобных матриц равны между собой. Перейдем от равенства матриц (1.49) к равенству определителей

$$\det\tilde{A}=\det(T^{-1}AT).$$

Отсюда после преобразования левой части получаем

$$\det\tilde{A}=\det A\det(T^{-1})\det\mathrm{T}\Leftrightarrow$$

$$\Leftrightarrow\det A\det(T^{-1}T)\Leftrightarrow\det A.$$

Введем следующее понятие. Для квадратной матрицы A рассмотрим сумму ее диагональных элементов, получившую название *следа матрицы* и обозначенную $\mathrm{tr}\,A$.

Свойство 3. Подобные матрицы имеют одинаковый след. Доказательство данного свойства вытекает из следующего соотношения:

$$\operatorname{tr}(AB) = \operatorname{tr}(BA). \tag{1.50}$$

В результате имеем цепочку эквивалентных преобразований

$$\operatorname{tr}\tilde{A} = \operatorname{tr}(T^{-1}AT) \Leftrightarrow \operatorname{tr}\tilde{A} =$$
$$= \operatorname{tr}(AT^{-1})T \Leftrightarrow \operatorname{tr}\tilde{A} = \operatorname{tr}AE,$$

что и доказывает свойство

$$\operatorname{tr}\tilde{A} = \operatorname{tr}A.$$

Пусть задана квадратная матрица M. Определим *многочлен от матрицы* $P(M)$ как:

$$P(M) = a_0M^k + a_1M^{k-1} + \ldots + a_{k-1}M + a_kE, \tag{1.51}$$

где a_i — элементы числового поля; E — единичная матрица того же размера, что матрица M.

Свойство 4. Для многочленов от подобных матриц выполняется равенство

$$P(T^{-1}MT) = T^{-1}P(M)T. \tag{1.52}$$

Доказательство этого равенства вытекает из следующих двух свойств:

1) $T^{-1}A_1T + T^{-1}A_2T + \ldots + T^{-1}A_pT = T^{-1}(A_1 + A_2 + \ldots + A_p)T$,

2) $(T^{-1}AT)^p = (T^{-1}AT)(T^{-1}AT)\underbrace{[(T^{-1}AT)(T^{-1}AT)\ldots(T^{-1}AT)]}_{p-2} \Leftrightarrow$

$\Leftrightarrow (T^{-1}AT)^p = (T^{-1}A^2T)(T^{-1}AT)\underbrace{[(T^{-1}AT)(T^{-1}AT)\ldots(T^{-1}AT)]}_{p-3} \Leftrightarrow$

$\Leftrightarrow (T^{-1}AT)^p = (T^{-1}A^pT)$.

Окончательный этап доказательства формулы (1.52) предлагаем выполнить читателю в порядке упражнения.

Выше были введены понятия линейного отображения и линейного преобразования как некоторое обобщение понятия функции. Если теперь предположить, что в линейных пространствах X и Y задана норма вектора, то можно рассматривать и такие свойства функции, как ее непрерывность и ограниченность.

Определение 1.15. Отображение $\mathcal{A}$, действующее из пространства X в Y, называется *непрерывным* в $x_0 \in X$, если из условия $x_k \to x_0$ следует что $\mathcal{A}x_k \to \mathcal{A}x_0$ для любой последовательности $\{x_k\}$ из X.

Это определение непрерывности (по Гейне) эквивалентно следующему определению (по Коши). Отображение $\mathcal{A}$ непре-

рывно в x_0, если для $\forall\varepsilon>0\exists\delta>0$ такое, что из $\|x-x_0\|<\delta \Rightarrow$ $\Rightarrow \|\mathscr{A}(x)-\mathscr{A}(x_0)\|<\varepsilon$. Если отображение $\mathscr{A}$ непрерывно для всех $x\in\tilde{X}$, где $\tilde{X}\subseteq X$, то тогда говорят, что $\mathscr{A}$ непрерывно на $\tilde{X}$. Для символической записи непрерывности отображения $\mathscr{A}$ в x_0 будем использовать обозначение $\mathscr{A}x\to\mathscr{A}x_0$ при $x\to x_0$.

О непрерывности линейного отображения на $\tilde{X}\subseteq X$, в котором оно определено, можно судить по непрерывности его в нуле пространства X. Для доказательства этого утверждения воспользуемся равенством $\mathscr{A}x-\mathscr{A}x_0=\mathscr{A}(x-x_0)$. Отсюда при $x\to x_0\Leftrightarrow z=(x-x_0)\to 0$, а следовательно, из условия непрерывности $\mathscr{A}$ в нуле $\mathscr{A}z\to 0$, но тогда $\mathscr{A}x-\mathscr{A}x_0\to 0$, что и требовалось доказать.

Теорема 1.8. Линейное отображение, действующее в конечномерных пространствах, является непрерывным.

Доказательство. Разложим вектор $x_0\in X$ и вектор $x_k\in X$ из последовательности $\{x_k\}$ по базису $\{e_i\}_{i=1\div n}$ пространства X. Тогда

$$x_0=\sum_{i=1}^{n}\xi_i^0 e_i \text{ и } x_k=\sum_{i=1}^{n}\xi_{k_i}e_i.$$

Найдем линейное отображение этих векторов

$$\mathscr{A}x_0=\sum_{i=1}^{n}\xi_i^0\mathscr{A}e_i \text{ и } \mathscr{A}x_k=\sum_{i=1}^{n}\xi_{k_i}\mathscr{A}e_i.$$

Предполагая, что $x_k\underset{k\to\infty}{\to}x_0$, и используя определение координатной сходимости последовательности векторов, имеем $\xi_{k_i}\to\xi_0$ для всех $i=1\div n$. Отсюда следует сходимость $\mathscr{A}x_k\underset{k\to\infty}{\to}\mathscr{A}x_0$ как координатная, так и по норме в пространстве Y.

Из доказанной теоремы следует непрерывность нормы вектора от координат вектора. Так, рассматривая норму вектора как отображение n-мерного пространства в одномерное R_1, нетрудно доказать, что это отображение есть линейный функционал. Следовательно, из предыдущей теоремы имеем

$$x_k\underset{k\to\infty}{\longrightarrow}x_0\Rightarrow\|x_k\|\underset{k\to\infty}{\longrightarrow}\|x_0\|$$

или

$$\text{для } x_k=(\xi_{k_1}\xi_{k_2}\ldots\xi_{k_i}\ldots\xi_{k_n}),$$
$$x_0(\xi_1^0\xi_2^0\ldots\xi_n^0) \text{ и всех } i=1\div n:$$
$$\xi_{k_i}\underset{k\to\infty}{\longrightarrow}\xi_i^0\Rightarrow\|x_k\|\underset{k\to\infty}{\longrightarrow}\|x_0\|. \qquad (1.53)$$

Определение 1.16. Отображение $\mathscr{A}$ из X в Y называется *ограниченным*, если существует такая числовая константа $k>0$, что для $\forall x\in X$:

$$\| \mathcal{A}x \| \leqslant k \| x \|. \tag{1.54}$$

Теорема 1.9. Для того чтобы линейное отображение было ограниченным, необходимо и достаточно, чтобы оно было непрерывным.

Доказательство достаточности. Пусть $\mathcal{A}$ определено и непрерывно в $\tilde{X} \subseteq X$. Тогда из определения непрерывности $\mathcal{A}$ в нуле X (по Коши) имеем: для фиксированного числа $\varepsilon_0 > 0$ найдется такое числовое значение $\delta_0 > 0$, что для всех $z \neq 0$, $z \in \tilde{X}$ и удовлетворяющих условию $\|z\| < \delta_0$ будет выполняться неравенство $\|\mathcal{A}z\| < \varepsilon_0$. Возьмем в качестве таких векторов

$$z = x\delta_0 / n \| x \|,$$

где $x \in \tilde{X}$ и $n > 1$.
Тогда для $\forall x \in \tilde{X}$ будет выполняться условие $\|z\| < \delta_0$, а следовательно

$$\| \mathcal{A}x \| \leqslant \frac{n}{\delta_0} \| x \|.$$

Доказательство необходимости. Пусть $\mathcal{A}$ ограничено. Тогда выполняется неравенство (1.54). Отсюда если $x \to 0$, то и $\mathcal{A}x \to 0$, т. е. $\mathcal{A}$ непрерывно в нуле, а следовательно, и в $\tilde{X}$, на котором $\mathcal{A}$ определено.

Из доказанной теоремы, в частности, следует, что любое линейное отображение, действующее в конечномерных пространствах, ограничено. Это утверждение сразу становится очевидным после применения теоремы (1.8), а затем теоремы (1.9).

Теорема 1.10. Любые две нормы в конечномерных линейных пространствах эквивалентны.

Доказательство. Пусть в пространстве X определены нормы $\varphi_1(x)$ и $\varphi_2(x)$ для $x \in X$. Тогда, в силу непрерывности отображения φ_1, действующего из X в R_1, найдется такое число $m_1 > 0$, что

$$\varphi_1(x) \leqslant m_1 \varphi_2(x),$$

аналогично для φ_2 найдется такое число $m = 1/m_2 > 0$, что

$$\varphi_2(x) \leqslant \frac{1}{m_2} \varphi_1(x).$$

Исходя из обоих неравенств, получаем

$$m_2 \varphi_1(x) \leqslant \varphi_1 \leqslant m_1 \varphi_2(x),$$

что и требовалось доказать.

Таким образом, все три ранее введенные векторные нормы $|*|_l$, $|*|_E$, $|*|_c$ являются попарно эквивалентными в n-мерном линейном пространстве.

§ 1.6. СОБСТВЕННЫЕ ЗНАЧЕНИЯ И СОБСТВЕННЫЕ ВЕКТОРЫ

В различных приложениях математики, а также в численных методах широко используются понятия собственных значений и собственных векторов. Напомним основные сведения о них.

Пусть в пространстве L действует линейное преобразование $\mathscr{A}$. Тогда может оказаться, что в результате данного преобразования образ некоторого ненулевого вектора $x \in L$ является коллинеарным прообразу x, что можно записать как

$$\mathscr{A}x_\lambda = \lambda x_\lambda \text{ или } Ax_\lambda = \lambda x_\lambda, \tag{1.55}$$

где λ называется *собственным значением*, а ненулевой вектор x_λ — *собственным вектором* линейного преобразования $\mathscr{A}$ (или матрицы A), соответствующим собственному значению λ.

Заметим, что если x_λ — собственный вектор, то любой коллинеарный вектор αx_λ при $\alpha \neq 0$ будет также собственным вектором, соответствующим собственному значению λ. Множество x_λ всех собственных векторов, соответствующих собственному значению λ, с присоединенным к нему нулевым вектором того же пространства L образует подпространство, которое получило название *собственного подпространства преобразования* $\mathscr{A}$. Понятия собственного подпространства и собственного подпространства линейного преобразования не тождественны. Так, если рассмотреть тождественное преобразование E в пространстве L, то это преобразование имеет только одно собственное значение, равное 1, а собственное подпространство преобразования E совпадает с пространством L. Тогда как собственное подпространство пространства L, как следует из определения, не должно совпадать с пространством L. Уравнение (1.55) может быть переписано в виде

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

При решении этого уравнения относительно x_λ получаем, что ненулевое решение его возможно только в случае, когда

$$\det(\lambda E - A) = 0. \tag{1.56}$$

Если раскрыть определитель $f(\lambda) = \det(\lambda E - A)$, то получим многочлен относительно λ, степень которого соответствует размерности пространства L:

$$f(\lambda) = a_0 + a_1\lambda + \ldots + a_i\lambda^i + \ldots + a_{n-1}\lambda^{n-1} + a_n\lambda^n. \tag{1.57}$$

Многочлен $f(\lambda)$ получил название *характеристического многочлена* линейного преобразования (матрицы). Коэффициенты a_i этого многочлена вычисляют по элементам матрицы A, а потому они принадлежат полю P, над которым задано линейное пространство L. Таким образом, для того чтобы число λ из поля P было собственным значением преобразования $\mathscr{A}$, необхо-

димо и достаточно, чтобы оно было корнем характеристического многочлена $f(\lambda)$ т. е. $f(\lambda)=0$. Уравнение $f(\lambda)=0$ называется *характеристическим уравнением.*

Как следует из свойств многочленов, его корни могут не принадлежать тому полю, в котором заданы коэффициенты многочлена. Отсюда, в частности, следует, что не всякое линейное преобразование имеет хотя бы один собственный вектор. Так, если линейное пространство L задано над полем вещественных чисел, а все значения λ принадлежат полю комплексных чисел, то в этом случае образ и прообраз такого линейного преобразования $\mathscr{A}$ не будут коллинеарными, а следовательно, преобразование $\mathscr{A}$ не будет иметь ни одного собственного вектора. В качестве геометрического примера рассмотрим вращение векторов в плоскости вокруг начала координат на фиксированный угол $0<\alpha<180°$. Ясно, что при выполнении указанной операции никакой вектор не станет коллинеарным самому себе.

Для исключения случаев, приведенных в предыдущем примере, будем считать, что коэффициенты a_i для любого характеристического многочлена $f(\lambda)$ задаются в том же поле, что и корни этого многочлена. Многочлены, удовлетворяющие этому условию, называются *алгебраически замкнутыми.* В частности отметим, что поле комплексных чисел замкнуто, а поля действительных и рациональных чисел не замкнуты.

В результате такого подхода отметим, что линейное преобразование, действующее в пространстве, заданном над алгебраически замкнутым полем, имеет хотя бы один собственный вектор. В дальнейшем будем считать, что линейное преобразование задано в комплексном пространстве C_n.

Приведем ряд свойств для собственных значений.

Свойство 1. Всякая матрица линейного преобразования, заданного в пространстве C_n, имеет ровно n собственных значений с учетом их кратности.

Доказательство этого свойства следует из того, что число корней алгебраически замкнутого многочлена равно его степени.

Свойство 2. Если у матрицы A собственное значение λ является комплексным числом кратности k, то и комплексно сопряженное число $\overline{\lambda}$ будет собственным значением матрицы A с той же кратностью.

Обоснование этого свойства предлагаем выполнить самостоятельно.

Свойство 3. Подобные матрицы имеют одинаковые собственные значения.

Рассмотрим два характеристических многочлена $\det(\lambda E-A)$ и $\det(\lambda E-B)$, где $B=T^{-1}AT$. Так как матрица подобного преобразования T невырожденная, то

$$Ax=\lambda x \Leftrightarrow T^{-1}AT(T^{-1}x)=\lambda(T^{-1}x) \Leftrightarrow By=\lambda y.$$

Таким образом, матрицы A и B имеют одинаковые собственные значения λ. Отсюда также следует, что

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

Свойство 4. Если λ_i — собственное значение матрицы A размера $[n\times n]$ и $P(x)$ — скалярный многочлен, то $P(\lambda_i)$ будет собственным значением многочлена от матрицы $P(A)$ для всех $i=1\div n$.

Доказательство. Если λ_i — собственное значение A, то имеем

$$Ax_i=\lambda_i x_i.$$

Умножим обе части этого уравнения на матрицу A, тогда

$$A^2x_i=\lambda_i Ax_i \Leftrightarrow A^2x_i=\lambda_i{}^2x_i.$$

Выполнив аналогичные преобразования k раз, получим

$$A^kx_i=\lambda^k{}_ix_i \text{ для } \forall\, i=1\div n \text{ и } \forall\, k\in Z^+, \tag{1.58}$$

где Z^+ — множество целых положительных чисел.

Пусть многочлен от матрицы $P(A)$ определен следующим образом:

$$P(A)=a_0A^n+a_1A^{n-1}+\ldots+a_{n-1}A+a_nE,$$

тогда для любого вектора-столбца $x\in R_n$, а следовательно, и для x_i можно записать

$$P(A)x_i=a_0A^nx_i+a_1A^{n-1}x_i+\ldots+a_{n-1}Ax_i+a_nx_i.$$

Подставляя в данное выражение вместо A^kx_i значения из формулы (1.58), получим для $\forall i=1\div n$

$$P(A)x_i=(a_0\lambda_i{}^n+a_1\lambda_i{}^{n-1}+\ldots+a_{n-1}\lambda_i+a_n)x_i$$

или

$$P(A)x_i=P(\lambda_i)x_i, \tag{1.59}$$

что и служит доказательством свойства 4.

Из свойства 4 достаточно просто получить известную теорему Гамильтона — Кэли, которую можно сформулировать следующим образом.

Теорема 1.11. Если $f(\lambda)$ — характеристический многочлен матрицы A, то многочлен от матрицы $f(A)=0$.

Действительно, из выражения (1.59) имеем

$$f(A)x_i=f(\lambda_i)x_i \text{ для } i=1\div n.$$

Так как λ_i — корень характеристического многочлена $f(\lambda)$, то $f(\lambda_i)=0$, а следовательно, и $f(A)=0$ с учетом того, что $x_i\neq 0$ как собственный вектор A.

В качестве примера найдем характеристический многочлен матрицы A с использованием теоремы Гамильтона — Кэли. Пусть

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

Тогда из теоремы 1.11 имеем $f(A)=0$, т. е.

$$A^3+a_1A^2+a_2A^1+a_3=0.$$

Умножив это равенство на произвольный вектор-столбец $\bar{b}_0\neq 0$, получаем систему линейных уравнений относительно коэффициентов характеристического многочлена матрицы A

$$(\bar{b}_2\bar{b}_1\bar{b}_0)\bar{a}=-\bar{b}_3,$$

где

$$\bar{b}_k=A^k\bar{b}_0 \text{ для } k=1\div 3,\ \bar{a}=(a_1a_2a_3)^{\mathrm{T}}.$$

Для сокращения арифметических операций компоненты вектора $\bar{b}_k$ будем вычислять по формуле

$$\bar{b}_k=A\bar{b}_{k-1} \text{ при } \bar{b}_k\neq\alpha\bar{b}_{k-1},$$

а в качестве b_0 возьмем вектор $b_0=(100)^{\mathrm{T}}$. В результате получаем систему

$$\begin{pmatrix} 3 & -1 & 1 \\ 0 & 2 & 0 \\ 2 & 0 & 0 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}=-\begin{pmatrix} 1 \\ 6 \\ 2 \end{pmatrix},$$

решением которой являются значения $a_1=-1$, $a_2=-3$, $a_3=-1$.

Данный способ получения характеристического многочлена легко распространяется и для матриц размера $[n\times n]$. При этом число операций умножения, необходимое для реализации этого способа с применением метода Гаусса для решения СЛАУ можно оценить по формуле $N_{\text{умн}}\cong 4/3\,n^3$ для $n>10$.

Теорема 1.12. Пусть S_r — сумма всех главных миноров порядка r матрицы A размера $[n\times n]$, где $1\leqslant r\leqslant n$. Тогда

$$f(\lambda)=\lambda^n-S_1\lambda^{n-1}+S_2\lambda^{n-2}+\ldots+(-1)^nS_n. \qquad (1.60)$$

Прежде чем переходить к доказательству этой теоремы, напомним определение главного минора. *Минором* матрицы A называют определитель матрицы, полученной из элементов $a_{i_k j_k}$ матрицы A, стоящих на пересечении ее $i_1, i_2, \ldots, i_p$ строк и $j_1, j_2, \ldots, j_p$ столбцов, где $1\leqslant p\leqslant n$. Когда $i_k=j_k$ для всех $k=1\div p$, миноры называются *главными* для матрицы A.

Теперь рассмотрим частный случай, когда A имеет размер $[2\times 2]$. Тогда

$$f_2(\lambda)=\det(A-\lambda E)=\begin{vmatrix} a_{11}-\lambda & a_{12} \\ a_{21} & a_{22}-\lambda \end{vmatrix}.$$

Воспользуемся свойством определителя

$$\det(\bar a_1 \ldots (\bar a_k+\bar b_k)\ldots \bar a_n)=$$
$$=\det(\bar a_1 \ldots \bar a_k \ldots \bar a_n)+\det(\bar a_1 \ldots \bar b_k \ldots \bar a_n)$$

для раскрытия выражения $f_2(\lambda)$:

$$f_2(\lambda)=\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}-\begin{vmatrix} \lambda & a_{12} \\ 0 & a_{22} \end{vmatrix}+\begin{vmatrix} \lambda & 0 \\ 0 & \lambda \end{vmatrix}-\begin{vmatrix} a_{11} & 0 \\ a_{21} & \lambda \end{vmatrix}$$

или

$$f_2(\lambda)=\det A-(a_{11}+a_{22})\lambda+\lambda^2.$$

Аналогичные преобразования, исходя из свойства 4 для собственных значений характеристического многочлена, можно выполнить и в общем случае, когда A имеет размер $[n\times n]$. В результате характеристический многочлен $f(\lambda)$ можно представить в виде, включающем сумму 2^n определителей n порядка:

$$(-1)^n f(\lambda)=b_0+b_1\lambda+b_2\lambda^2+\ldots+b_k\lambda^k+\ldots+b_{n-1}\lambda^{n-1}+b_n\lambda^n,$$

где коэффициенты b_k определяются по формулам

$$b_0=\det A, \tag{1.61}$$

$$b_k=(-1)^k\sum_{j_1=1}^{n-k+1}\ldots\sum_{i_s=i_{s-1}+1}^{n-k+s}\ldots\sum_{i_k=i_{k-1}+1}^{n}\det(\bar a_1\ldots\bar e_{j_1}\ldots\bar a_{j_s-1}\bar e_{j_k}\bar a_{j_s+1}\ldots\bar e_{j_k}\ldots\bar a_n).$$

Вектор $\bar e_{j_s}$ является единичным вектором, у которого j_s-компонента (координата) равна 1. Исследование приведенного выражения показывает, что число определителей, входящих в формулу для вычисления b_k, равно $\binom{n}{k}$. При этом каждый из этих определителей равен определителю, получающемуся из исходной матрицы путем вычеркивания всех имеющихся в ней единичных столбцов и строк с номерами, соответствующими номерам единичных столбцов. Полученный таким образом определитель, согласно данному выше определению, является главным минором. Отсюда имеем

$$b_k=(-1)^k S_{n-k}$$

и

$$(-1)^n f(\lambda)=\det A-S_{n-1}\lambda+S_{n-2}\lambda_2{}^2-\ldots+$$

$$+(-1)^k S_{n-k}\lambda^k+\ldots+(-1)^n S_0\lambda^n.$$

После умножения обеих частей этого равенства на $(-1)^n$, получаем окончательное выражение (1.61), так как $S_0=E$ и $S_n=\det A$.

В дополнение к приведенным выше четырем свойствам для собственных значений из доказанной теоремы можно получить еще ряд свойств для λ.

Свойство 5. Сумма всех собственных значений матрицы A с учетом их кратности равна следу этой матрицы.

Для доказательства данного свойства воспользуемся формулой Вьета, устанавливающей связь между коэффициентом a_{n-1} и корнями характеристического многочлена λ_i, т. е.

$$a_{n-1}=-(\lambda_1+\lambda_2+\ldots+\lambda_n).$$

С другой стороны, из формулы (1.60)

$$a_{n-1}=-S_{n-1}=-(a_{11}+a_{22}+\ldots+a_{nn}).$$

Таким образом, имеем

$$\sum_{i=1}^{n}\lambda_i=\operatorname{tr}A. \tag{1.62}$$

Свойство 6. Произведение всех собственных значений матрицы с учетом их кратности равно определителю этой матрицы.

Из формулы Вьета для коэффициента характеристического многочлена a_0 можно написать

$$a_0=(-1)^n\lambda_1\lambda_2\ldots\lambda_n.$$

Этот же коэффициент из выражения (1.61) представим в виде

$$a_0=(-1)^n S_n=(-1)^n\det A.$$

В результате получаем

$$\prod_{i=1}^{n}\lambda_i=\det A. \tag{1.63}$$

Свойство 7. Матрица A размера $[n\times n]$ и ранга $\operatorname{Rg}A=n-r$ имеет собственное значение $\lambda=0$ кратности $k\geqslant r$.

Доказательство. Если матрица A ранга $\operatorname{Rg}A=n-r$, то, как следует из определения ранга матрицы, наивысший порядок минора, отличного от нуля, равен $n-r$. Отметим, что таким минором не обязательно будет главный, а следовательно, все миноры матрицы A порядка выше $n-r$ равны нулю. Отсюда для характеристического многочлена $f(\lambda)$, представленного по формуле (1.60), имеем

$$S_{n-r+1}=\ldots=S_{n-1}=S_n=0$$

и характеристическое уравнение запишется в виде

$$\lambda^n - S_1\lambda^{n-1} + \ldots + (-1)^{n-r}S_{n-r}\lambda^r = 0.$$

Таким образом, если в написанном выше уравнении $S_{n-r} \neq 0$, то нуль будет корнем этого уравнения кратности $k=r$. Если же $S_{n-r}=S_{n-r-1}=\ldots=S_{n-r-t}=0$, где $0 \leqslant t \leqslant n-r-1$, то нуль будет корнем характеристического уравнения кратности $k=r+t+1>r$.

Приведем пример, когда ранг матрицы меньше кратности ее собственного значения, равного нулю. Пусть задана матрица

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

Для нее $\operatorname{Rg} A=2$, а $\det(\lambda E-A)=\lambda^3$, так что $r=1$ и матрица A имеет нуль собственным значением кратности $k=3$.

Свойство 7 лежит в основе доказательства следующей теоремы.

Теорема 1.13. Если матрица $(\lambda_i E-A)$ имеет ранг $\operatorname{Rg} A = n-r$, то корень характеристического многочлена $\det(\lambda_i E-A)$ будет иметь кратность $k \geqslant r$.

Доказательство. Рассмотрим характеристический многочлен матрицы $B=\lambda_i E-A$. Если μ — собственное значение матрицы B, то по свойству 7 и условию теоремы имеем $\mu^k=0$, где $k \geqslant r$. Запишем характеристический многочлен матрицы B в виде

$$\det(\mu E - B) = \det[(\lambda_i - \mu)E - A].$$

Отсюда получаем $\lambda=\lambda_i-\mu$. Следовательно, λ примет значение λ_i столько раз, сколько раз μ будет равно нулю, т. е. кратность λ_i будет равна $k \geqslant r$.

Теорема 1.14. Собственные векторы линейного преобразования (матрицы), соответствующие различным его (ее) собственным значениям, линейно независимы.

Доказательство. Пусть $\lambda_1, \lambda_2, \ldots, \lambda_j$ — различные собственные значения линейного преобразования $\mathcal{A}$, а $x_1, x_2, \ldots, \ldots, x_j$ — соответствующие им собственные векторы. Так как по определению собственные векторы ненулевые, то при $i=1$ теорема верна. Докажем, что теорема верна и при $i=2$. Для этого нужно проверить, при каких значениях a_1 и a_2 удовлетворяется равенство $a_1x_1+a_2x_2=0$. Нетрудно заметить, что оно удовлетворяется только при значениях $a_1=a_2=0$, так как в противном случае векторы x_1 и x_2 коллинеарны, а следовательно, не могут соответствовать двум различным собственным значениям.

Предположим, что теорема верна и для случая $j=m-1$, т. е. векторы $x_1, x_2, \ldots, x_{m-1}$ линейно независимы. Тогда, сле-

дуя принципу индукции, остается показать, что теорема верна и при $j=m$. Допустим, что это не так, т. е. найдутся такие числа $a_1, a_2, \ldots, a_m$, одновременно не равные нулю, для которых выполняется равенство

$$a_1x_1+a_2x_2+\ldots+a_mx_m=0. \qquad (1.64)$$

Так как нумерация собственных значений может быть выбрана произвольно, то можно считать, что $a_1\neq 0$. Применяя преобразование $\mathscr{A}$ к (1.64), имеем

$$a_1\lambda_1x_1+a_2\lambda_2x_2+\ldots+a_m\lambda_mx_m=0.$$

После умножения (1.64) на λ_m и вычитания полученного результата из предыдущего равенства находим

$$a_1(\lambda_1-\lambda_m)x_1+a_2(\lambda_2-\lambda_m)x_2+\ldots+a_{m-1}(\lambda_{m-1}-\lambda_m)x_{m-1}=0.$$

Из этого равенства и индуктивного предположения о независимости системы векторов $\{x_i\}_{i=1\div(m-1)}$, в частности, вытекает, что $a_1(\lambda_1-\lambda_m)=0$, что противоречит условию $\lambda_1\neq\lambda_m$ и $a_1\neq 0$. Следовательно, система векторов $x_1, x_2, \ldots, x_m$ линейно независима.

Определение 1.17. Линейное преобразование (матрица), действующее в n-мерном линейном пространстве, называется преобразованием (матрицей) *простой структуры*, если оно имеет n линейно независимых собственных векторов.

Преобразования простой структуры обладают весьма важным для приложений свойством. Только эти преобразования в некотором базисе имеют диагональные матрицы. Возьмем в качестве базиса пространства R_n собственные векторы $x_1, x_2, \ldots, \ldots, x_n$ и построим матрицу A для преобразования $\mathscr{A}$ в этом базисе. Как следует из (1.44), элементами столбцов матрицы A являются координаты образов векторов базиса, т. е. $\{Ax_j\}_{j=1\div n}$. Но, с другой стороны, $Ax_j=\lambda_jx_j$. Поэтому матрица A_λ преобразования $\mathscr{A}$ в базисе $\{x_j\}_{j=1\div n}$ примет следующий вид:

$$A_\lambda=\begin{pmatrix}\lambda_1 & 0 & \ldots & 0\\ 0 & \lambda_2 & \ldots & 0\\ \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & \ldots & \lambda_n\end{pmatrix}.$$

Предположим теперь, что диагональная матрица A_λ является матрицей линейного преобразования $\mathscr{A}$ в каком-то базисе $\{x_i\}_{i=1\div n}$, тогда векторы этого базиса являются собственными векторами $\mathscr{A}$, соответствующими собственным значениям $\{\lambda_i\}_{i=1\div n}$, среди которых могут быть и кратные. Следовательно, только линейным преобразованиям простой структуры могут соответствовать диагональные матрицы. При этом базисом ли-

нейного пространства, в котором действует линейное преобразование, являются лишь его собственные векторы.

Из приведенных рассуждений следует, что для матрицы A линейного преобразования простой структуры всегда найдется такая матрица S подобного преобразования, с помощью которой можно получить диагональную матрицу A_λ, подобную A, т. е.

$$S^{-1}AS = A_\lambda.$$

Диагональными элементами матрицы A_λ будут собственные значения матрицы A. Таким образом, необходимым и достаточным условием представления матрицы в диагональном виде является наличие у нее полной системы собственных векторов. Из теоремы следует, что данное условие всегда выполнимо, если матрица простой структуры. В том случае, когда среди собственных значений матрицы A размера $[n \times n]$ имеются кратные, в линейном n-мерном пространстве не всегда возможно построить базис из собственных векторов матрицы A. Так, если сумма размерностей всех собственных подпространств линейного преобразования $\mathscr{A}$ будет меньше размерности пространства L_n, в котором действует данное преобразование, то тогда из всей совокупности собственных векторов преобразования $\mathscr{A}$ нельзя построить базис пространства L_n.

В качестве примера рассмотрим блочно-диагональную матрицу I размера $[n \times n]$:

$$I = \begin{pmatrix} I_{m_1}(\lambda_1) & O & \dots & O & O \\ O & I_{m_2}(\lambda_2) & \dots & O & O \\ \dots & \dots & \dots & \dots & \dots \\ O & O & \dots & I_{m_{k-1}}(\lambda_{k-1}) & O \\ O & O & \dots & O & I_{m_k}(\lambda_k) \end{pmatrix}, \qquad (1.65)$$

где матрицы $I_{m_i}(\lambda_i)$, являющиеся блоками матрицы I размера $[m_i \times m_i]$, имеют следующий вид:

$$I_{m_i}(\lambda_i) = \begin{pmatrix} \lambda_i & 1 & o & \dots & o & o \\ o & \lambda_i & 1 & \dots & o & o \\ \dots & \dots & \dots & \dots & \dots & \dots \\ o & o & o & \dots & \lambda_i & 1 \\ o & o & o & \dots & o & \lambda_i \end{pmatrix}. \qquad (1.66)$$

Матрица $I_{m_i}(\lambda_i)$ получила название *жордановой клетки* порядка m_i, а матрица I — жордановой формы. Нетрудно убедиться, что для $\forall_i = 1 \div m_k$ каждая жорданова клетка $I_{m_i}(\lambda_i)$ соответствует одному собственному значению λ_i кратности m_i. При этом собственное подпространство X_{λ_i} матрицы I будет

одномерным с множеством векторов $x_i=(0\ldots0\,\alpha\,0\ldots)^{\mathrm{T}}$, у которых i-компонента равна произвольному числа α, отличному от нуля. Так как $k\leqslant m_1+m_2+\ldots+m_k=\sum\limits_{i=1}^{k} m_i=n$, то в случае, когда хотя бы одна из клеток $I_{m_i}(\lambda_i)$, входящая в (1.65), имеет размер больше двух, т. е. $\exists i$, для которого $m_i\geqslant 2$, имеем строгое неравенство $k<n$. Следовательно, невозможно в пространстве L_n построить базис лишь из собственных векторов матрицы I. В тех случаях, когда для $\forall i$ имеем $m_i=1$, матрица I приобретает форму обычной диагональной матрицы.

В ряде приложений и численных методах часто используется такое утверждение: *для всякого линейного преобразования, действующего в комплексном пространстве, найдется базис, при котором матрица этого преобразования принимает жорданову форму* (1.65). Доказательство его можно найти, например, в [1]. Из данного утверждения следует, что любая квадратная матрица с элементами из поля комплексных чисел может быть приведена подобным преобразованием к жордановой форме I, т. е. всегда найдется матрица S размера $[n\times n]$, что будет выполняться равенство

$$S^{-1}AS=I \text{ или } A=SIS^{-1}. \tag{1.67}$$

Заметим, что равенства (1.67) и (1.65) будут сохраняться для матрицы с вещественными элементами только в тех случаях, если все ее ненулевые собственные значения являются вещественными и их общее число с учетом их кратности равно рангу матрицы.

§ 1.7. СОПРЯЖЕННЫЕ ОТОБРАЖЕНИЯ И ПРЕОБРАЗОВАНИЯ

Определение 1.18. Пусть заданы унитарные пространства X и Y и линейное отображение (оператор) $\mathscr{A}: X\to Y$. Рассмотрим отображение $\mathscr{A}^*: Y\to X$, при котором для любых $x\in X$ и $y\in Y$ выполняется равенство

$$(\mathscr{A}x,\ y)=(x,\ \mathscr{A}^*y). \tag{1.68}$$

Тогда $\mathscr{A}^*$ называется *сопряженным* отображением (оператором) по отношению к отображению (оператору) $\mathscr{A}$.

Так как между линейными отображениями и прямоугольными матрицами устанавливается взаимно однозначное соответствие при любых фиксированных базисах, то равенство (1.68) можно представить в матричной форме

$$(Ax,\ y)=(x,\ A^*y), \tag{1.69}$$

где A и A^* — матрицы соответственно отображений $\mathscr{A}$ и $\mathscr{A}^*$ в каких-то фиксированных базисах $\{e_i\}_{i=1\div m}$ пространства X и $\{q_i\}_{i=1\div n}$ пространства Y.

Пусть в евклидовых пространствах X и Y заданы соответствующие им ортонормированные базисы $\{\hat{e}_i\}_{i=1\div m}$ и $\{\hat{q}_i\}_{i=1\div n}$. Если $\hat{A}$ — матрица отображения $\mathscr{A}$ в этих базисах, то в результате проверки можно установить равенство

$$(\hat{A}\hat{x},\ \hat{y})=(\hat{x},\ \hat{A}^{\tau}\hat{y}), \tag{1.70}$$

где $\hat{A}^{\tau}$ — матрица, транспонированная к матрице $\hat{A}$. Предлагаем эту проверку выполнить самостоятельно. Так как для каждой матрицы имеется только одна, транспонированная к ней, то из равенства (1.70) следует, что в ортонормированных базисах для любого линейного отображения $\mathscr{A}$ существует сопряженное отображение $\mathscr{A}^*$, и притом только одно. Данный вывод можно получить и для произвольных базисов, заданных в пространстве X и Y.

Теорема 1.15. При произвольно заданных базисах в евклидовых пространствах X и Y между матрицей отображения $\mathscr{A}$ и матрицей сопряженного отображения $\mathscr{A}^*$ устанавливается взаимно однозначная зависимость

$$A^*=G_e^{-1}A^{\tau}G_q, \tag{1.71}$$

в которой A^{τ} — матрица, транспонированная к A, а G_e и G_q — матрицы Грама: первая составлена для системы базисных векторов пространства X, а вторая — для системы базисных векторов пространства Y.

Доказательство. Рассмотрим скалярное произведение $(u,\ y)$, определенное в евклидовом пространстве Y, где вектор $u\in Y$ является образом отображения $\mathscr{A}x$. Если в евклидовом пространстве Y задан базис $\{q_i\}_{i=1\div n}$, то векторы $u,\ y$ и скалярное произведение записывают в виде

$$u=\sum_{i=1}^{n}u_iq_i,\ y=\sum_{j=1}^{n}y_jq_j,$$

$$(u,\ y)=\sum_{k=1}^{n}[\,(q_1,\ q_k)u_1+(q_2,\ q_k)u_2+\ \ldots\ +(q_n,\ q_k)u_n\,]y_k \tag{1.72}$$

или

$$(u,\ y)=\sum_{k=1}^{n}[\,(q_k,\ q_1)y_1+(q_k,\ q_2)y_2+\ \ldots\ +(q_k,\ q_n)y_n\,]u_k.$$

С учетом того, что для евклидовых пространств $(q_i,\ q_j)=(q_jq_i)$, формулы (1.72) можно представить как

$$(u,\ y)=(G_qu,\ y)=(u,\ G_qy). \tag{1.73}$$

В этой формуле G_q — матрица Грама для системы базисных векторов $\{q_i\}_{i=1\div n}$ пространства Y. Из аналогичных рассуждений

для скалярного произведения (x, v), определенного в евклидовом пространстве X, в котором задан базис $\{e_i\}_{i=1\div m}$, где вектор v является образом отображения $\mathscr{A}^*y$, получим

$$(x, v)=(G_e x_e, v_e)=(x_e, G_e v_e).$$

Выполним теперь в пространстве X переход от базиса $\{e_i\}_{i=1\div m}$ к ортонормированному базису $\{\hat{e}_i\}_{i=1\div m}$, а в пространстве Y — от базиса $\{q_i\}_{i=1\div n}$ к ортонормированному базису $\{\hat{q}_i\}_{i=1\div n}$. Тогда если за P_1 обозначить матрицу подобного преобразования при переходе в X от базиса $\{\hat{e}_i\}_{i=1\div m}$ к $\{e_i\}_{i=1\div m}$, а за P_2 — матрицу подобного преобразования при переходе в Y от базиса $\{\hat{q}_i\}_{i=1\div m}$ к $\{q_i\}_{i=1\div m}$, то, исходя из формулы (1.48), получим взаимосвязь между матрицей A отображения $\mathscr{A}$ в заданных базисах $\{e_i\}_{i=1\div m}$ и $\{q_i\}_{i=1\div n}$ и матрицей того же отображения, но в ортонормированных базисах $\{\hat{e}_i\}_{i=1\div m}$ и $\{\hat{q}_i\}_{i=1\div n}$:

$$\hat{A}=P_2^{-1}AP_1,\ \hat{u}=P_2^{-1}AP_1\hat{x}. \tag{1.74}$$

В результате при вычислении скалярного произведения векторов в пространстве Y с заданным ортонормированным базисом $\{\hat{e}_i\}_{i=1\div m}$ из формулы (1.73) получаем

$$(u, y)=(P_2\hat{u}, G_qP_2\hat{y}).$$

Используя преобразования (1.74), эту формулу запишем в виде

$$(u, y)=(AP_1\hat{x}, G_qP_2\hat{y}).$$

Применив формулу (1.71) к приведенному выше выражению, получаем

$$(u, y)=(\hat{x}, P_1^{\mathrm{T}}A^{\mathrm{T}}G_qP_2\hat{y}). \tag{1.75}$$

Из тех же соображений для вычисления скалярного произведения векторов x и v в базисе $\{\hat{e}_i\}_{i=1\div m}$ пространства X получаем следующее выражение:

$$(x, v)=(\hat{x}, P_1^{\mathrm{T}}G_eA^*P_2\hat{y}).$$

Так как в силу определения сопряженного отображения $(u, y)=(x, v)$, то

$$(\hat{x}, P_1^{\mathrm{T}}A^{\mathrm{T}}G_qP_2\hat{y})=(\hat{x}, P_1^{\mathrm{T}}G_eA^*P_2\hat{y}).$$

В связи с тем что это равенство должно выполняться для любых x и y, имеем

$$P_1^{\mathrm{T}}A^{\mathrm{T}}G_qP_2=P_1^{\mathrm{T}}G_eA^*P_2.$$

Отсюда окончательно получаем равенство (1.71), т. е.

$$G_e^{-1}A^{\mathrm{T}}G_q=A^*.$$

Доказанную теорему можно распространить и на случай унитарных пространств X и Y. В этом случае для произвольных базисов в X и Y формула (1.71) принимает вид

$$A^* = \overline{G}_e^{-1} (\overline{A})^{\tau} \overline{G}_q, \tag{1.76}$$

где черта над матрицей указывает на комплексное сопряжение между соответствующими элементами матрицы с чертой и без черты. Если базисы ортонормированные, то $A^* = (\overline{A})^{\tau} = (\overline{A^{\tau}})$, т. е. $a^*_{ij} = \overline{a}_{ji}$ для всех i и j.

Исходя из формулы (1.76) и учитывая, что между линейным оператором (отображением) и его матрицей устанавливается взаимно однозначное соответствие в фиксированном базисе, нетрудно получить следующие соотношения для сопряженных операторов:

$$(\mathscr{A}+B)^* = \mathscr{A}^* + B^*, \ (\alpha\mathscr{A})^* = \overline{\alpha}\mathscr{A}^*, \ (\mathscr{A}B)^* = B^*\mathscr{A}^*,$$

$$(\mathscr{A}^*)^* = \mathscr{A}, \ (\mathscr{A}^*)^{-1} = (\mathscr{A}^{-1})^*. \tag{1.77}$$

В порядке упражнения предлагается доказательство приведенных соотношений выполнить самостоятельно.

Аналогично сопряженному отображению можно ввести понятие сопряженного преобразования в линейном унитарном пространстве X.

Определение 1.19. Если в унитарном пространстве X задано линейное преобразование $\mathscr{A}x = y$ и для любых x, y из этого пространства выполняется равенство

$$(\mathscr{A}x, \ y) = (x, \ \mathscr{A}^* y), \tag{1.78}$$

то *линейное преобразование* $\mathscr{A}^*$ *называется сопряженным* по отношению к преобразованию $\mathscr{A}$.

Для матрицы сопряженного преобразования $\mathscr{A}^*$ в произвольно заданном базисе пространства X формула (1.76) принимает вид

$$A^* = \overline{G}_e^{-1} (\overline{A})^{\tau} \overline{G}_e, \tag{1.79}$$

а при задании ортонормированного базиса в X получаем

$$A^* = (\overline{A})^{\tau}.$$

Теорема 1.16. Если в базисе $\{q_i\}_{i=1 \div n}$ евклидова пространства X линейное преобразование имеет матрицу B_q, то сопряженное ему преобразование в другом базисе $\{e_i\}_{i=1 \div n}$ имеет матрицу

$$B_e^* = P^{-1} G_q^{-1} B_q^{\tau} G_q P, \tag{1.80}$$

где P — матрица подобного преобразования.

Доказательство. На основании формулы (1.79) матрица сопряженного преобразования в базисе $\{q_i\}_{i=1\div n}$ примет вид

$$B_q{}^* = G_q{}^{-1} B_q{}^{\tau} G_q.$$

При переходе от базиса $\{q_i\}_{i=1\div n}$ к базису $\{e_i\}_{i=1\div n}$ воспользуемся преобразованием подобия. Тогда приведенная выше формула преобразуется следующим образом:

$$B_e{}^* = (P^{-1} G_q{}^{-1} P)(P^{-1} B_q{}^{\tau} P)(P^{-1} G_q P).$$

Здесь P — матрица подобного преобразования $u_q = P u_e$, где u_q, $u_e \in X$. И после несложных преобразований получаем формулу (1.80).

Определение 1.20. Базис $\{e_i\}_{i=1\div n}$ унитарного пространства называется *двойственным* по отношению к базису $\{q_i\}_{i=1\div n}$ того же пространства, если

$$(q_i, e_j{}^q) = \begin{cases} 0 \text{ при } i \neq j, \\ 1 \text{ при } i = j, \end{cases}$$

где $e_j{}^q$ — вектор e_j, разложенный по базису $\{q_i\}_{i=1\div n}$.

Взаимно двойственные базисы называются *биортонормированными.*

Найдем матрицу перехода P от базиса $\{q_i\}_{i=1\div n}$ к двойственному базису $\{e_i\}_{i=1\div n}$. Обозначив элементы этой матрицы как p_{ij} для $i, j = 1\div n$, получаем выражение для вектора e_j через его координаты в базисе $\{q_i\}_{i=1\div n}$:

$$e_j{}^q = (p_{1j} p_{2j} \ldots p_{nj}).$$

При этом скалярное произведение $(q_i, e_j{}^q)$ будет равно

$$(q_i, e_j{}^q) = (q_i q_1) p_{1j} + (q_i q_2) p_{2j} + \ldots$$
$$\ldots + (q_i q_k) p_{kj} + \ldots + (q_i q_n) p_{nj}.$$

Рассматривая множество всевозможных скалярных произведений данного вида для $i, j = 1\div n$, замечаем, что его можно описать как

$$\begin{pmatrix} c_{11} & c_{12} & \ldots & c_{1n} \\ c_{21} & c_{22} & \ldots & c_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ c_{n1} & c_{n2} & & c_{nn} \end{pmatrix} = \begin{pmatrix} (q_1, q_1) & (q_1, q_2) & \ldots & (q_1, q_n) \\ (q_2, q_1) & (q_2, q_2) & \ldots & (q_2, q_n) \\ \ldots & \ldots & \ldots & \ldots \\ (q_n, q_1) & (q_n, q_2) & \ldots & (q_n, q_n) \end{pmatrix} \times$$
$$\times \begin{pmatrix} p_{11} & p_{12} & \ldots & p_{1n} \\ p_{21} & p_{22} & \ldots & p_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ p_{n1} & p_{n2} & \ldots & p_{nn} \end{pmatrix},$$

где $c_{ij}=(q_i, e_i^q)$ — элементы матрицы C. Исходя из определения двойственного базиса имеем $C=E$, а следовательно

$$G_q P=E. \tag{1.81}$$

Таким образом, матрица подобного преобразования при переходе от заданного базиса к двойственному ему базису равна матрице, обратной к матрице Грама, составленной из векторов исходного (заданного) базиса. Из формулы (1.81) также вытекает, что любой базис имеет двойственный по отношению к данному и притом только один.

Из теоремы 1.16 вытекает ряд следствий.

Следствие 1. Если в некотором базисе евклидова пространства линейное преобразование $\mathscr{A}$ имеет матрицу B, то в базисе, двойственном к данному, сопряженное преобразование $\mathscr{A}^*$ имеет матрицу B^{T}.

Доказательство этого следствия непосредственно следует после подстановки в формулу (1.80) выражения, определяемого по формуле (1.81).

Следствие 2. Спектры сопряженных преобразований в евклидовом пространстве совпадают, и равные собственные значения имеют одинаковые кратности.

Действительно, если $\mathscr{A}x=\lambda x$, а $\mathscr{A}^*y=\mu y$, то спектр оператора $\mathscr{A}$ определяется собственными значениями матрицы A, а спектр оператора $\mathscr{A}^*$ — собственными значениями сопряженной матрицы $A^*=T^{-1}A^{\mathrm{T}}T$, где $T=GP$ — невырожденная матрица. Так как матрица A^* подобна матрице A^{T}, а спектры матриц A^{T} и A совпадают и $\mathrm{Rg}(A-\lambda E)=\mathrm{Rg}(A^{\mathrm{T}}-\lambda E)$, следовательно, совпадают спектры матриц A^* и A и равны кратности их одинаковых собственных значений.

Если пространство, в котором задано линейное преобразование, — унитарное, то для любого собственного значения λ_i преобразования $\mathscr{A}$ найдется такое собственное значение μ_k преобразования $\mathscr{A}^*$, для которого $\mu_k=\overline{\lambda}_i$.

Отсюда, в частности, следует, что если преобразование $\mathscr{A}$ — простой структуры, то и сопряженное ему преобразование $\mathscr{A}^*$ — простой структуры.

Следствие 3. Базисные системы собственных векторов преобразований $\mathscr{A}$ и $\mathscr{A}^*$ можно выбрать таким образом, чтобы они были биортонормированными.

Для доказательства рассмотрим скалярные произведения $(\mathscr{A}x_i, y_i)=\lambda_i(x_i, y_i)$ и $(x_i, \mathscr{A}^*y_j)=\mu_j(x_i, y_j)$. Предположим, что $\lambda_i\neq\mu_j$. Тогда в силу того, что $(\mathscr{A}x_i, y_j)=(x_i, \mathscr{A}^*y_j)$, имеем $(\lambda_i-\mu_j)(x_i, y_j)=0$. Отсюда следует, что любой собственный вектор преобразования $\mathscr{A}$, соответствующий собственному значению λ, ортогонален и любому собственному вектору преобразования $\mathscr{A}^*$, соответствующему собственному значению $\mu\neq\lambda$. Если теперь нормировать собственные векторы преобразований

$\mathscr{A}$ и $\mathscr{A}^*$ и образовать из них две последовательности векторов $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m)$ и $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m)$, то собственные векторы $\hat{x}_i$ и $\hat{y}_i$ преобразований $\mathscr{A}$ и $\mathscr{A}^*$, соответствующие одним и тем же собственным значениям λ_i, образуют две последовательности биортонормированных векторов.

§ 1.8. НОРМАЛЬНЫЕ ПРЕОБРАЗОВАНИЯ И СИНГУЛЯРНОЕ РАЗЛОЖЕНИЕ МАТРИЦ

В различных приложениях математики, в том числе и в вычислительной практике, нашли широкое применение операторы, которые в унитарном пространстве имеют ортонормированные базисные системы, состоящие из собственных векторов. Изучение данного класса операторов связано с понятием нормального преобразования (оператора).

Определение 1.21. Линейное преобразование $\mathscr{A}$ называется нормальным, если оно перестановочно со своим сопряженным, т. е.

$$\mathscr{A}^*\mathscr{A} = \mathscr{A}\mathscr{A}^*. \tag{1.82}$$

Отметим одно из свойств нормального преобразования, которое будет использоваться при доказательстве теоремы 1.17.

Если матрица нормального преобразования в каком-то фиксированном ортонормированном базисе унитарного пространства является треугольной, то она диагональна. Доказательство этого свойства вытекает из условия равенства нулю диагональных элементов матрицы $(AA^{\mathrm{T}} - A^{\mathrm{T}}A) = 0$.

Теорема 1.17. Для того чтобы преобразование в евклидовом пространстве было нормальным, необходимо и достаточно, чтобы оно имело базисную систему ортонормированных собственных векторов.

Доказательство. Пусть $\mathscr{A}$ — нормальное преобразование. Согласно теореме 1.6, в евклидовом пространстве, в котором задано это преобразование, найдется такой ортонормированный базис, в котором преобразованию $\mathscr{A}$ будет соответствовать треугольная матрица R вида

$$R = \begin{pmatrix} R_{11} & O \\ O & O \end{pmatrix}.$$

В том же базисе сопряженному преобразованию $\mathscr{A}^*$ будет соответствовать матрица

$$R^{\mathrm{T}} = \begin{pmatrix} R^{\mathrm{T}}{}_{11} & O \\ O & O \end{pmatrix}.$$

Так как преобразование $\mathscr{A}$ нормальное, то из условия (1.82) следует

$$RR^{\mathrm{T}} = R^{\mathrm{T}}R.$$

Отсюда на основании отмеченного выше свойства для треугольных матриц имеем, что матрица R диагональная. Таким образом, ортонормированный базис, в котором преобразованию $\mathscr{A}$ соответствует матрица R, целиком составлен из собственных векторов этого преобразования.

Предположим теперь, что преобразование $\mathscr{A}$ имеет базисную систему ортонормированных собственных векторов. Тогда в этом базисе матрица преобразования $\mathscr{A}$ будет диагональной. Аналогично будет диагональной в том же базисе и матрица сопряженного преобразования $\mathscr{A}^*$. Так как произведение диагональных матриц обладает свойством коммутативности, то выполняется и условие (1.82).

Доказанная теорема справедлива и для унитарных пространств, так как для них выполняется и теорема 1.6.

Одним из примеров нормальных преобразований является *унитарное преобразование*. Напомним, что унитарным преобразованием называется преобразование, у которого сопряженное преобразование совпадает с обратным, т. е.

$$\mathscr{A}\mathscr{A}^* = \mathscr{A}^*\mathscr{A} = E.$$

В случае, когда преобразование, удовлетворяющее этому условию, задано в евклидовом пространстве, оно называется *ортогональным*.

Важнейшим свойством унитарных преобразований является сохранение скалярного произведения. Действительно

$$(\mathscr{A}x,\ \mathscr{A}y) = (x,\ \mathscr{A}^*\mathscr{A}y) = (\mathscr{A}^*\mathscr{A}x,\ y) = (x,\ y). \qquad (1.83)$$

Из данного свойства вытекает такое утверждение: собственные значения унитарных преобразований по модулю равны 1. Так из определения собственного вектора и значения преобразования имеем $\mathscr{A}x = \lambda x$. Переходя к скалярному произведению собственных векторов, получаем

$$(\mathscr{A}x,\ \mathscr{A}x) = \lambda^2 (x,\ x).$$

Отсюда, с учетом равенства (1.83), имеем

$$\lambda = |\,1\,|.$$

Другим примером нормального преобразования является *самосопряженное* (эрмитово) *преобразование*, являющееся линейным преобразованием, для которого выполняется условие $\mathscr{A}^* = \mathscr{A}$.

Из данного определения следует, что в евклидовом пространстве матрица самосопряженного преобразования симметрическая. Доказательство этого утверждения вытекает из матричного равенства

$$A = A^* = A^{\mathrm{T}},$$

выполняемого для ортонормированного базиса евклидова пространства. В унитарном пространстве с ортонормированным базисом матрицы самосопряженных преобразований связаны соотношением

$$\bar{A}^{\mathrm{T}}=A. \tag{1.84}$$

т. е. $\bar{a}_{ij}=a_{ji}$, где верхняя черта свидетельствует о комплексном сопряжении чисел. Матрица A, удовлетворяющая условию (1.84), называется *эрмитовой*.

Особое место в практике обработки результатов измерений и экспериментов занимают преобразования вида $\mathscr{A}^*\mathscr{A}$ и $\mathscr{A}\mathscr{A}^*$. Остановимся на их рассмотрении несколько подробнее. Во-первых, отметим, что если $\mathscr{A}$ — оператор, действующий из евклидова (унитарного) пространства X размерности m в пространство Y размерности n, то $\mathscr{A}^*\mathscr{A}$ определяет преобразование в пространстве Y, а $\mathscr{A}\mathscr{A}^*$ — в X. Нетрудно проверить, что оба этих преобразования самосопряженные. Кроме того, они неотрицательные. Действительно, если $x\in X$ и $y\in Y$, то соответствующие в X и Y квадратичные формы примут вид

$$(\mathscr{A}^*\mathscr{A}x,\ x)=(\mathscr{A}x,\ \mathscr{A}x)\geqslant 0,$$

$$(\mathscr{A}\mathscr{A}^*y,\ y)=(\mathscr{A}^*y,\ \mathscr{A}^*y)\geqslant 0,$$

причем равенство нулю в этих выражениях достигается тогда и только тогда, когда $\mathscr{A}x=0$ или $\mathscr{A}^*y=0$, т. е.

$$\mathscr{A}^*\mathscr{A}x=0\Leftrightarrow\mathscr{A}x=0 \text{ и } \mathscr{A}\mathscr{A}^*y=0\Leftrightarrow\mathscr{A}^*y=0. \tag{1.85}$$

$$\operatorname{Ker}\mathscr{A}=\operatorname{Ker}\mathscr{A}^*\mathscr{A} \text{ и } \operatorname{Ker}\mathscr{A}^*=\operatorname{Ker}\mathscr{A}\mathscr{A}^*.$$

В силу самосопряженности преобразований $\mathscr{A}^*\mathscr{A}$ и $\mathscr{A}\mathscr{A}^*$ имеем в ортонормированных базисах пространств X и Y равенство между матрицами самосопряженных преобразований, т. е. $S^{\mathrm{T}}=S$, где S — матрица преобразования $\mathscr{A}^*\mathscr{A}$. Так как ранги матриц S^{T} и S равны, следовательно, равны и ранги преобразований $\mathscr{A}^*\mathscr{A}$ и $\mathscr{A}\mathscr{A}^*$, т. е.

$$\operatorname{Rg}\mathscr{A}^*\mathscr{A}=\operatorname{Rg}\mathscr{A}\mathscr{A}^*=\operatorname{Rg}\mathscr{A}=\operatorname{Rg}\mathscr{A}^*. \tag{1.86}$$

Теорема 1.18. Ненулевые собственные значения преобразований $\mathscr{A}^*\mathscr{A}$ и $\mathscr{A}\mathscr{A}^*$ положительны и совпадают с учетом их кратности, а собственные векторы каждого из этих преобразований образуют систему ортонормированных собственных векторов.

Доказательство. Обозначим $S_1=\mathscr{A}^*\mathscr{A}$. Тогда в пространстве X скалярное произведение $(S_1x_i,\ x_i)=\lambda_i(x_i,\ x_i)$, где x_i — собственный вектор преобразования S_1, соответствующий собственному значению λ_i. Отсюда

$$\lambda_i=\frac{(S_1x_i,\ x_i)}{(x_i,\ x_i)}\Leftrightarrow\lambda_i=\frac{(\mathscr{A}x_i,\ \mathscr{A}x_i)}{(x_i,\ x_i)}\geqslant 0. \tag{1.87}$$

Причем $\lambda_i=0$ тогда и только тогда, когда $\mathcal{A}x_i=0$. Если же $\mathcal{A}x_i\neq 0$, то имеем

$$\mathcal{A}\mathcal{A}^*(\mathcal{A}x_i)=\mathcal{A}(\mathcal{A}^*\mathcal{A}x_i)\Leftrightarrow$$
$$\Leftrightarrow \mathcal{A}(\lambda_i x_i)=\lambda_i(\mathcal{A}x_i).$$

Таким образом, $\mathcal{A}x_i$ является собственным вектором преобразования $\mathcal{A}\mathcal{A}^*$, соответствующим собственному значению λ_i. Поэтому ненулевые собственные значения преобразований $\mathcal{A}^*\mathcal{A}$ и $\mathcal{A}\mathcal{A}^*$ совпадают с учетом их кратности. Кратности нулевых собственных значений рассматриваемых преобразований при $m\neq n$ отличаются. Так, если преобразование $\mathcal{A}^*\mathcal{A}$ имеет r ненулевых собственных значений с учетом их кратности, то кратность нулевых собственных значений у этого преобразования будет равна $m-r$, а у преобразования $\mathcal{A}\mathcal{A}^*-(n-r)$.

Для доказательства второго утверждения рассмотрим скалярное произведение (S_1x_i, x_j), где x_i и x_j — собственные векторы преобразования $\mathcal{A}^*\mathcal{A}$, соответствующие различным собственным значениям $\lambda_i\neq\lambda_j$. Тогда

$$(S_1x_i, x_j)=\lambda_i(x_i, x_j)\Leftrightarrow (x_i, S_1x_j)=\lambda_j(x_i, x_j).$$

Так как $(S_1x_i, x_j)=(x_i, S_1x_j)$, то $(x_i, x_j)=0$. Если преобразование S_1 нормальное, то на основании теоремы 1.18 в пространстве X можно построить ортонормированный базис из собственных векторов $x_1, x_2, \ldots, x_r, \ldots, x_m$ преобразования S. При этом r первых значений будут соответствовать собственным значениям $\lambda_1\geqslant\lambda_2\geqslant\ldots\lambda_r>0$ и принадлежать множеству значений T^* преобразования $\mathcal{A}^*$, а остальные собственные векторы, соответствующие $\lambda_i=0$, где $i=(r+1)\div m$, принадлежат $\ker\mathcal{A}$. Построенный таким образом базис получил название *первого сингулярного базиса отображения* $\mathcal{A}$.

Теперь в пространстве Y построим ортонормированный базис $\{y_i\}_{i=1\div n}$. При этом для $i=1\div r$ за вектор y_i можно выбрать

$$y_i=\frac{\mathcal{A}x_i}{(\mathcal{A}x_i, \mathcal{A}x_i)^{1/2}}=\frac{\mathcal{A}x_i}{\sqrt{\lambda_i}}, \qquad (1.88)$$

так как из доказанного выше $\mathcal{A}x_i$ — собственные векторы эрмитова преобразования $\mathcal{A}\mathcal{A}^*$, следовательно, они образуют ортогональный базис в области значений преобразования $\mathcal{A}$. За остальные y_k для $k=(r+1)\div n$ возьмем любые векторы, принадлежащие $\ker\mathcal{A}^*$ и образующие в нем ортонормированный базис. Эти векторы тоже являются собственными для преобразования $\mathcal{A}\mathcal{A}^*$ и соответствуют нулевым собственным значениям этого преобразования. Построенный таким образом базис $\{y_i\}_{i=1\div n}$ получил название *второго сингулярного базиса отображения* $\mathcal{A}$.

Сингулярные базисы связаны с отображениями $\mathcal{A}$ и $\mathcal{A}^*$ следующими соотношениями:

$$\mathscr{A}x_k = \begin{cases} \rho_k y_k & \text{для } k=1 \div r, \\ 0 & \text{для } k=(r+1) \div n, \end{cases}$$

$$\mathscr{A}^* y_k = \begin{cases} \rho_k x_k & \text{для } k=1 \div r, \\ 0 & \text{для } k=(r+1) \div m. \end{cases} \tag{1.89}$$

Здесь числа $\rho_k = \sqrt{\lambda_k}$ называются *сингулярными числами* отображения $\mathscr{A}$.

Из зависимостей (1.89) непосредственно следует теорема.

Теорема 1.19. В паре сингулярных базисов матрица отображения $\mathscr{A}$, имеющая ранг r, принимает диагональный вид

$$D = \begin{pmatrix} D_r & O \\ O & O \end{pmatrix}. \tag{1.90}$$

Здесь D_r — диагональная матрица размера $[r \times r]$, по диагонали которой расположены сингулярные числа отображения $\mathscr{A}$ в порядке их возрастания (убывания).

Из приведенной теоремы вытекает важное для вычислительных методов следствие о сингулярном разложении матрицы.

Следствие 1. Пусть A — вещественная матрица размера $[m \times n]$ и ранга r. Тогда существуют ортогональные матрицы U размера $[m \times m]$, V размера $[n \times n]$ и матрица D размера $[m \times n]$, такие, что

$$\text{а) } U^{\mathrm{T}} A V = D, \qquad \text{б) } A = U D V^{\mathrm{T}} \tag{1.91}$$

Причем матрицу D можно выбрать таким образом, чтобы она имела вид (1.90). Равенство (1.91, *б*) получило название *сингулярного разложения матрицы.*

§ 1.9. ОПЕРАТОРНЫЕ УРАВНЕНИЯ. РЕШЕНИЕ УРАВНЕНИЙ МЕТОДОМ НАИМЕНЬШИХ КВАДРАТОВ

Напомним, что понятие линейного оператора равносильно понятию линейного отображения $\mathscr{A}$ для многомерных линейных пространств.

Пусть $\mathscr{A}$ — оператор, действующий из m-мерного линейного пространства X в n-мерное линейное пространство Y, заданных над одним и тем же полем P. Определим все векторы $x \in X$, которые при заданных операторе $\mathscr{A}$ и векторе $y \in Y$ удовлетворяют уравнению

$$\mathscr{A}x = y. \tag{1.92}$$

Это уравнение называется *операторным уравнением*, вектор x — *его решением*, а вектор y — *правой частью.*

Если теперь в пространствах X и Y зафиксировать базисы, то уравнение (1.92) будет соответствовать известному матричному уравнению $Ax=y$, которое также называется системой линейных алгебраических уравнений (СЛАУ). Естественно, что все свойства систем уравнений распространяются и на операторные уравнения и наоборот.

Пусть пространства X и Y унитарные. Тогда определен оператор $\mathscr{A}^*$, для которого можно составить уравнение вида

$$\mathscr{A}^*u=v, \tag{1.93}$$

которое принято называть *сопряженным уравнением*. Если правые части в уравнениях (1.92) и (1.93) равны нулю, то тогда эти уравнения называются *однородными*.

При решении СЛАУ могут возникать такие случаи: 1) имеется единственное решение; 2) имеется множество решений; 3) не имеется ни одного решения. В первых двух случаях говорят, что операторное уравнение разрешимо (или система уравнений — совместная), а в третьем случае — операторное уравнение неразрешимо (или система линейных уравнений — несовместная).

Теорема 1.20 (Фредгольма). Операторное уравнение разрешимо тогда и только тогда, когда его правая часть ортогональна всем решениям сопряженного однородного уравнения.

Доказательство. Из условия разрешимости уравнения (1.92) следует, что его правая часть принадлежит области значения T оператора $\mathscr{A}$. Рассмотрим скалярное произведение в пространстве Y между векторами $u\in\ker\mathscr{A}^*$ и y. Тогда

$$(y, u)=(\mathscr{A}x, u) \Leftrightarrow (y, u)=(x, \mathscr{A}^*u).$$

И так как $\mathscr{A}^*u=0$, то имеем $(y, u)=0$.

Теперь пусть задано $(y, u)=0$ для всех $u\in\ker\mathscr{A}^*$. Тогда $y\perp\ker\mathscr{A}^*$, а следовательно, $y\in T$. Отсюда вытекает, что $\exists x\in X$, для которого $\mathscr{A}x=y$, т. е. неоднородное уравнение разрешимо.

К решению уравнения (1.92) можно подойти и несколько с других позиций. При произвольно заданных операторе $\mathscr{A}$ и векторе y может оказаться, что равенство (1.92) не выполняется ни при каких значениях x, что равносильно несовместимости системы уравнений $\mathscr{A}x=y$. Для того чтобы от уравнения (1.92) перейти к строгому равенству, введем в рассмотрение вектор $v=\mathscr{A}x-y$, принадлежащий пространству Y. Вектор v будем называть *вектором невязки*. Случай, когда $v=0$, соответствует разрешимости исходного уравнения $Ax=y$. С другой стороны, на основании определения нормы в линейном пространстве вектору $v=0$ соответствует значение $\|v\|=0$, которое является наименьшим из всех значений норм. Таким образом, разрешимость операторного уравнения (1.92) равносильна условию $\|v\|=0$. На основании указанного в общем случае за решение уравне-

ния (1.92) можно принять такие векторы $\hat{x}\in X$, для которых функция

$$F(x)=\| v \|=\| \mathscr{A}x-y \| \tag{1.94}$$

достигает наименьшего значения.

Каждый из элементов $\hat{x}\in X$, для которого выполняется при заданных $\mathscr{A}$, y и $\|*\|$ равенство

$$\| y-\mathscr{A}\hat{x} \|=\inf \| \mathscr{A}x-y \|, \tag{1.95}$$

где символ inf обозначает наименьшее значение и называется *наилучшим решением операторного уравнения,* а $u_0=\mathscr{A}\hat{x}$ при этом получило название *элемента наилучшего приближения* для $y\in Y$. Таким образом, нахождение решения уравнения (1.92) свелось к задаче поиска в $T\subseteq Y$ элемента наилучшего приближения по норме к элементу $y\in Y$.

Для случая, когда в качестве нормы вектора берется c-норма, равенство (1.95) преобразуется к виду

$$\| y-\mathscr{A}\hat{x} \|_c=\inf \sup | \mathscr{A}x-y |. \tag{1.96}$$

Задача нахождения $\hat{x}$, удовлетворяющего приведенному условию, получила название *чебышевского приближения* (или равномерного приближения).

Если в равенстве (1.95) в качестве нормы вектора берется евклидова норма, то в этом случае задача определения x называется *приближением по методу наименьших квадратов* или просто методом наименьших квадратов. В практике обработки результатов геодезических и фотограмметрических измерений метод наименьших квадратов нашел самое широкое применение.

Для нахождения элемента наилучшего приближения по этому методу важное значение приобретает следующая теорема.

Теорема 1.21. Если в линейном подпространстве T унитарного пространства Y существует элемент u_0, дающий наилучшее приближение к элементу $y\in Y$ по методу наименьших квадратов, то элемент $(y-u_0)$ ортогонален ко всем элементам подпространства T.

Доказательство. Допустим противное, т. е. существование элемента $u_1\in T$, для которого $(y-u_0, u_1)=\alpha\|u_1\|_2^2\neq 0$. Рассмотрим элемент $u_2=u_0+\alpha u_1$ и оценим норму $\|y-u_2\|_E$:

$$\| y-u_2 \|_E^2=(y-u_2, y-u_2)=(y-u_0, y-u_0)-$$
$$-\alpha(u_1, y-u_0)-\bar{\alpha}(y-u_0, u_1)+$$
$$+\alpha\bar{\alpha}\| u_1 \|_E^2=\| y-u_0 \|_E^2-\alpha\bar{\alpha}\| u_1 \|_E^2-$$
$$-\alpha\bar{\alpha}\| u_1 \|_E^2+\alpha\bar{\alpha}\| u_1 \|_E^2=\| y-u_0 \|_E^2-|\alpha|^2\| u_1 \|_E^2.$$

Отсюда получаем

$$\| y-u_2 \|_E^2<\| y-u_0 \|_E^2,$$

но это невозможно, так как по условию u_0 — элемент наилучшего приближения.

Следствие. В одном линейном подпространстве может существовать только один элемент наилучшего приближения к фиксированному элементу унитарного пространства.

Действительно, если бы существовало для $y \in Y$ два элемента наилучшего приближения в T, например u_0 и u'_0, то на основании предыдущей теоремы для $\forall u \in T$

$$(y - u_0, u) = 0 \text{ и } (y - u_0', u) = 0.$$

А следовательно, эти выражения выполнялись бы и для $u = u_0 - u'_0$, т. е.

$$(y - u_0, u_0 - u_0') = 0 \text{ и } (y - u_0', u_0 - u_0') = 0.$$

Тогда, вычитая эти выражения и используя свойства скалярного произведения, получим

$$(u_0 - u_0', u_0 - u_0') = 0,$$

а это означает, что $u_0 = u'_0$.

В формулировках теоремы и следствия к ней указывается только на возможность существования элемента наилучшего приближения в линейном подпространстве унитарного пространства. Возможность существования такого элемента вытекает из ограниченности снизу $(F(x) \geqslant 0)$ числовой функции $F(x)$, определяемой из (1.94). На вопрос же о существовании элемента $u_0 \in T$, для которого имеет место равенство (1.93), дает ответ теорема о существовании элемента наилучшего приближения.

Теорема 1.22. Для любого элемента нормированного пространства в конечномерном линейном подпространстве данного пространства существует элемент наилучшего приближения.

С доказательством этой теоремы предлагаем познакомиться в [22]. Здесь только обратим внимание на то, что по условию теоремы ограничение на размерность нормированного пространства не накладывается, т. е. оно может быть и бесконечномерным.

Таким образом, если в (1.92) оператор $\mathscr{A}$ действует из конечномерного пространства X_m в конечномерное Y_n, то всегда найдется такой единственный вектор $u_0 \in \mathscr{A}x$, для которого выполняются условие (1.95) и следующая теорема.

Теорема 1.23. Для любого операторного уравнения (1.92) имеется решение по методу наименьших квадратов $\hat{x} \in X$, удовлетворяющее равенству

$$\mathscr{A}^* \mathscr{A} \hat{x} = \mathscr{A}^* y. \tag{1.97}$$

Доказательство. Пусть T — область значения оператора $\mathscr{A}$, тогда $T \subseteq Y$, причем $T = Y \Leftrightarrow m = n = \operatorname{Rg} \mathscr{A}$. На основании

теоремы 1.22 в T найдется элемент $u_0 \in T$, для которого выполняется равенство

$$\| y - u_0 \|_E = \inf \| \mathcal{A}x - y \|_E.$$

При этом, согласно теореме 1.21, для элемента u_0 и любого элемента $u \in T$ должно выполняться условие

$$(y - u_0, u) = 0,$$

которое в нашем случае примет вид

$$(y - \mathcal{A}\hat{x}, \mathcal{A}x) = 0.$$

Отсюда

$$(\mathcal{A}^* y - \mathcal{A}^* \mathcal{A}\hat{x}, x) = 0.$$

Так как это равенство выполняется для любых $x \in X$, то получаем доказываемое равенство (1.97).

Теперь остается доказать, что $\hat{x}$ всегда найдется. Это равносильно тому, что операторное уравнение (1.97) разрешимо. Рассмотрим скалярное произведение между правой частью уравнения (1.97) и любым вектором $v \in \ker \mathcal{A}^* \mathcal{A}$. С учетом того, что $\ker \mathcal{A} = \ker \mathcal{A}^* \mathcal{A}$, имеем

$$(\mathcal{A}^* y, v) = (y, \mathcal{A}v) \Leftrightarrow (\mathcal{A}^* y, v) = (y, \mathcal{A}^* \mathcal{A} v).$$

Но так как $\mathcal{A}^* \mathcal{A} v = 0$, то и $(\mathcal{A}^* y, v) = 0$ и на основании теоремы Фредгольма получаем, что уравнение (1.97) разрешимо.

При произвольно заданных базисах в евклидовых пространствах X и Y уравнение (1.97) представляется в виде системы линейных алгебраических уравнений, которая с учетом (1.71) имеет следующую матричную форму записи:

$$G_e^{-1} A^T G_q A \hat{x} = G_e^{-1} A^T G_q y,$$

после несложных преобразований получаем

$$A^T G_q A \hat{x} = A^T G_q y, \tag{1.98}$$

где G_q — матрица Грама для системы базисных векторов в пространстве Y. Для случая, когда в Y задана ортонормированная система базисных векторов, $G_q = E$ и $A^T A \hat{x} = A^T y$.

В геодезии выражение (1.98) получило название *нормальной системы уравнений,* а выражение

$$Ax - y = v \tag{1.99}$$

называется *системой уравнений поправок.* Вектор решения системы нормальных уравнений $\hat{x}$ называется *псевдорешением* уравнений поправок.

Так как нормальная система уравнений совместная, то она имеет или единственное решение, или множество решений.

Когда $\operatorname{Rg} A=m$, где m — размерность пространства X, псевдорешение единственно, когда же $\operatorname{Rg} A<m$, то нормальная система уравнений имеет множество решений. Если решение нормальной системы уравнения не единственно, то при отсутствии какой-либо дополнительной информации о решении системы уравнений поправок, которая, как правило, может быть получена из физической сущности задачи или по дополнительным измерениям, за ее решение берется псеводорешение наименьшей длины, называемое *нормальным псевдорешением.*

Теорема 1.24. Каждое линейное операторное уравнение имеет одно и только одно нормальное псевдорешение.

Доказательство. Пусть $\hat{x}$ — псевдорешение. Так как $\hat{x}\in X$, его можно разложить по первому сингулярному базису, т. е. если x_i — собственные векторы $\mathscr{A}^*\mathscr{A}$, то

$$\hat{x}=\sum_{i=1}^{r}\gamma_i x_i+\sum_{k=r+1}^{m}\gamma_k x_k.$$

Отсюда

$$\|\hat{x}\|_E^2=\sum_{i=1}^{r}\gamma_i^2+\sum_{k=r+1}^{m}\gamma_k^2. \tag{1.100}$$

Разложив теперь по второму сингулярному базису вектор $y\in Y$ и учитывая, что

$$\mathscr{A}^*\mathscr{A}\hat{x}=\mathscr{A}^*\mathscr{A}\sum_{i=1}^{r}\gamma_i x_i+\mathscr{A}^*\mathscr{A}\sum_{k=r+1}^{m}\gamma_k x_k \Leftrightarrow$$

$$\Leftrightarrow \mathscr{A}^*\mathscr{A}\hat{x}=\sum_{i=1}^{r}\gamma_i\rho_i^2 x_i+\sum_{k=r+1}^{m}\gamma_k\rho_k^2 x_k,$$

имеем

$$\sum_{i=1}^{r}\gamma_i\rho_i^2 x_i+\sum_{k=r+1}^{m}\gamma_k\rho_k^2 x_k=A^*\sum_{i=1}^{n}\beta_i y_i \Leftrightarrow$$

$$\Leftrightarrow \sum_{i=1}^{r}\gamma_i\rho_i^2 x_i+\sum_{k=r+1}^{m}\gamma_k\rho_k^2 x_k=$$

$$=\sum_{i=1}^{r}\beta_i\rho_i x_i+\sum_{k=r+1}^{n}\beta_i\rho_i x_i.$$

Из этого равенства при условии, что для $i\geqslant r+1$ сингулярные числа $\rho_i=0$, получаем $\gamma_i=\beta_i/\rho_i$ для $i=1\div r$. Подставим полученные значения γ_i в (1.100), тогда

$$\|\hat{x}\|^2_E=\sum_{i=1}^{r}\left(\frac{\beta_i}{\rho_i}\right)^2+\sum_{k=r+1}^{m}\gamma^2_k. \tag{1.101}$$

Так как $\rho_i=0$ для $i=(r+1)\div m$ и значения γ_k для $k=(r+1)\div m$

можно выбирать произвольные, то нормальное псевдорешение x^0 будет соответствовать

$$\min \|\hat{x}\|^2_E = \sum_{i=1}^{r} \left(\frac{\beta_i}{\rho_i}\right)^2,$$

и, как следствие этого, в сингулярном базисе для x^0 получим разложение

$$x^0 = \begin{cases} \sum\limits_{i=1}^{r} (\beta_i / \rho_i) x_i, \\ 0 \qquad \text{для } i \geqslant r+1, \end{cases} \tag{1.102}$$

которое для заданного уравнения единственно.

Следствие 1. Нормальное псевдорешение уравнения (1.92) ортогонально ядру оператора $\mathscr{A}$.

Для доказательства разложим вектор $z \in \ker \mathscr{A}$ по первому сингулярному базису:

$$z = z_1 x_1 + z_2 x_2 + \ldots + z_m x_m.$$

Тогда

$$\mathscr{A}z = z_1 \mathscr{A}x_1 + z_2 \mathscr{A}x_2 + \ldots + z_m \mathscr{A}x_m \Leftrightarrow$$
$$\Leftrightarrow z_1\rho_1 x_1 + z_2\rho_2 x_2 + \ldots + z_r\rho_r x_r +$$
$$+ z_{r+1}\rho_{r+1}x_{r+1} + \ldots + z_m\rho_m x_m = 0.$$

Так как система собственных векторов $\{x_i\}_{i=1\div m}$ независимая, то значения $z_i\rho_i = 0$ для всех $i = 1 \div m$. Из условия $\rho_i \neq 0$ для $i = 1 \div r$ следует, что $z_i = 0$ для $i = 1 \div r$. Вычисляя скалярное произведение (x^0, z) в сингулярном базисе пространства X, получаем

$$(x^0, z) = \sum_{i=1}^{r} (\beta_i / \rho_i) z_i + \sum_{k=r+1}^{m} 0 \cdot z_k \Rightarrow (x^0, z) = 0. \tag{1.103}$$

Следствие 2. Сопряженное уравнение (1.93), правая часть которого — нормальное псевдорешение уравнения (1.92), является совместным, т. е. выполняется равенство

$$\mathscr{A}^* u = x^0. \tag{1.104}$$

На основании теоремы Фредгольма, если уравнение $\mathscr{A}^* u = x^0$ совместное, то $(x^0, z) = 0$, где z — решение однородного сопряженного уравнения $\mathscr{A}z = 0$, т. е. $z \in \ker \mathscr{A}$. Но так как условие $(x^0, z) = 0$ вытекает из предыдущего следствия, то тем самым доказано равенство (1.104).

Используя следствие 1 можно получить общее решение нормального уравнения в виде

$$\hat{x} = x^0 + qz. \tag{1.105}$$

Таблица 1

Ранг	Тип системы	Отношение между		
		1 ($m=n$)		
		x	$\hat{x}$	x^0
а	Совместная	x	$\hat{x}=x$	$x^0=\hat{x}$
($r=\inf(n, m)$)	Несовместная	—	—	—
б	Совместная	$\{x\}$	$\{\hat{x}\}=\{x\}$	x_c^0
($r<\inf(n, m)$)	Несовместная	$\varnothing$	$\{\hat{x}\}$	$x^0_н$

Здесь x^0 — нормальное псевдорешение, z — решение однородной системы $\mathscr{A}z=0$. Так как векторы x^0 и z ортогональны, то значение q можно вычислить по формуле

$$q=(\hat{x}z)\,/\,\|z\|_E^2.$$

Из равенства (1.105) следует, что $\hat{x}=x^0$ при $z=0$, и в силу единственности x^0 имеем единственным и решение нормального уравнения (1.97). Таким образом, невырожденность оператора $\mathscr{A}^*\mathscr{A}$ обеспечивает единственное решение нормального уравнения $\mathscr{A}^*\mathscr{A}x=\mathscr{A}^*y$ (или единственное решение нормальной системы уравнений (1.97) обеспечивается невырожденностью матрицы оператора $\mathscr{A}^*\mathscr{A}$, имеющей вид $\mathscr{A}^{\mathrm{T}}G_q\mathscr{A}$).

Обобщим теперь результаты, полученные для решения операторного уравнения (1.92) по методу наименьших квадратов. В зависимости от соотношений между размерностями пространств X и Y и рангом оператора A все задачи этого метода разбиваются на случаи 1*а*, 1*б*, 2*а*, 2*б*, 3*а*, 3*б* (табл. 1). В этой же таблице в колонках x, $\hat{x}$ и x^0 приведены решения уравнений (1.92) и (1.97) для всех шести случаев. При этом $\varnothing$ обозначает пустое множество решений, $\{x\}$ и $\{\hat{x}\}$ — множество решений уравнения (1.92) и соответствующего нормального уравнения (1.97), содержащие больше одного элемента. Если от операторного уравнения (1.92) перейти к рассмотрению системы линейных уравнений

$$Ax=y, \tag{1.106}$$

размерностями пространств X и Y

2 ($m>n$)			3 ($m<n$)		
x	$\hat{x}$	x^0	x	$\hat{x}$	x^0
$\{x\}$	$\{\hat{x}\}=\{x\}$	x_c^0	x	$\hat{x}=x$	$x_c^0=\hat{x}$
$\varnothing$	$\{\hat{x}\}$	$x_н^0$	$\varnothing$	$\{\hat{x}\}$	$x_н^0=\hat{x}$
$\{x\}$	$\{\hat{x}\}=\{x\}$	x_c^0	$\{x\}$	$\{\hat{x}\}=\{x\}$	x_c^0
$\varnothing$	$\{\hat{x}\}$	$x_н^0$	$\varnothing$	$\{x\}$	$x_н^0$

где A — $[n\times m]$-матрица оператора $\mathscr{A}$ и x, y — векторы-столбцы соответственно размерности m и n, то все данные таблицы относятся и к выражению (1.106), для которого нормальной системой уравнений будет система (1.98).

§ 1.10. ПСЕВДООБРАТНЫЙ ОПЕРАТОР. АЛГОРИТМЫ ПСЕВДООБРАЩЕНИЯ МАТРИЦЫ

Пусть $\mathscr{A}$ — оператор, действующий из пространства X в Y. Тогда на основании теоремы 1.24 для $\forall y\in Y$ найдется вектор $x^0\in X$, являющийся нормальным псевдорешением уравнения (1.92). Отмеченное отображение Y в X определяет некоторый оператор, получивший название *псевдообратного оператора*, обозначаемого в дальнейшем $\mathscr{A}^+$. Таким образом, из введенного определения для $\forall y\in Y$ имеем

$$x^0=\mathscr{A}^+y. \tag{1.107}$$

Из теоремы 1.24 также следует, что каждый оператор имеет один и только один псевдообратный.

Предположим, что заданы два операторных уравнения $\mathscr{A}u=b$ и $\mathscr{A}v=d$, а u^0 и v^0 — их нормальные псевдорешения. Рассмотрим теперь уравнение вида $\mathscr{A}w=\gamma b+\mu d$ и найдем его нормальное псевдорешение w^0. Применив формулу (1.102), получаем

$$u^0=\begin{cases}\sum\limits_{i=1}^{r}\frac{b_i}{\rho_i}x_i,\\ 0 \text{ для } i\geqslant r+1;\end{cases}\qquad v^0=\begin{cases}\sum\limits_{i=1}^{r}\frac{d_i}{\rho_i}x_i,\\ 0 \text{ для } i\geqslant r+1;\end{cases}$$

$$w^0=\begin{cases}\gamma\sum\limits_{1}^{r}\frac{b_i}{\rho_i}x_i+\gamma\sum\limits_{1}^{r}\frac{d_i}{\rho_i}x_i,\\ 0 \text{ для } i\geqslant r+1.\end{cases}$$

Отсюда имеем

$$w^0=\gamma u^0+\gamma v^0.$$

Таким образом, доказана линейность псевдообратного оператора.

Как и для любого линейного отображения, для оператора $\mathscr{A}^+$ существует сопряженный оператор $(\mathscr{A}^+)^*$. При этом оператор $(\mathscr{A}^+)^*$ обладает следующим свойством:

$$(\mathscr{A}^+)^*=(\mathscr{A}^*)^+. \tag{1.108}$$

Для доказательства этого свойства рассмотрим скалярное произведение между вектором x^0, являющимся нормальным псевдорешением уравнения (1.92), и вектором $z\in\ker\mathscr{A}$. Тогда на основании формул (1.103) и (1.68) получим

$$(x^0, z)=0 \Leftrightarrow (\mathscr{A}^+y, z)=0 \Leftrightarrow (y, (\mathscr{A}^+)^*z)=0.$$

С другой стороны, из теоремы Фредгольма

$$(\mathscr{A}^*y, z)=0 \Leftrightarrow (y, (\mathscr{A}^*)^+z)=0.$$

Сопоставляя полученные равенства, имеем для $\forall z$

$$(y, ((\mathscr{A}^+)^*-(\mathscr{A}^*)^+)z)=0,$$

и так как $y\neq 0$, то в результате получим доказываемое равенство (1.108).

Теорема 1.25. Сингулярными числами оператора $\mathscr{A}^*\mathscr{A}$ (или $\mathscr{A}\mathscr{A}^+$) являются $\rho_i=1$ для $i=1\div r$ и $\rho_i=0$ для $i\geqslant r+1$.

Доказательство. В равенстве (1.107) вектор x^0 представим по формуле (1.102), а вектор y разложим по второму сингулярному базису

$$\sum_{i=1}^{r}\frac{\beta_i}{\rho_i}x_i+\sum_{k=1}^{m}0\cdot x_k=\mathscr{A}^+\left(\sum_{i=1}^{r}\beta_i y_i+\sum_{j=1}^{n}\beta_j y_j\right). \tag{1.109}$$

Преобразуя правую часть этого равенства с учетом формулы (1.89), получим

$$\sum_{i=1}^{r}\frac{\beta_i}{\rho_i}x_i+\sum_{k=1}^{m}0\cdot x_k=\mathscr{A}^+\mathscr{A}\left(\sum_{i=1}^{r}\frac{\beta_i}{\rho_i}x_i+\sum_{j=1}^{n}\beta_j\cdot 0\cdot x_j\right).$$

Отсюда

$$\mathscr{A}^{+}\mathscr{A}x_i=\begin{cases}1\cdot x_i & \text{для } i=1\div r,\\ 0\cdot x_i & \text{для } i\geqslant r+1.\end{cases} \tag{1.110}$$

Преобразуя аналогичным образом левую часть равенства (1.109), получим

$$\mathscr{A}\mathscr{A}^{+}y_i=\begin{cases}1\cdot y_i & \text{для } i=1\div r,\\ 0\cdot y_i & \text{для } i\geqslant r+1.\end{cases} \tag{1.111}$$

Из доказанной теоремы вытекает ряд свойств для преобразования вида $\mathscr{A}^{+}\mathscr{A}$ и оператора $\mathscr{A}^{+}$. Остановимся на некоторых из них.

Свойство 1. Преобразования $\mathscr{A}^{+}\mathscr{A}$ и $\mathscr{A}\mathscr{A}^{+}$ являются эрмитовыми.

Действительно, так как собственные векторы этих преобразований образуют ортонормированную базисную систему векторов для пространств X и Y, а их собственные значения — неотрицательные числа, то это и служит доказательством данного свойства.

Свойство 2. Для преобразований $\mathscr{A}^{+}\mathscr{A}$ и $\mathscr{A}\mathscr{A}^{+}$ удовлетворяются равенства

$$(\mathscr{A}^{+}\mathscr{A})^{*}=\mathscr{A}^{+}\mathscr{A} \text{ и } (\mathscr{A}\mathscr{A}^{+})^{*}=\mathscr{A}\mathscr{A}^{+}. \tag{1.112}$$

Предоставляем читателю в порядке упражнения выполнить доказательство этого свойства самостоятельно.

Свойство 3. Имеют место следующие равенства:

$$\mathscr{A}\mathscr{A}^{+}\mathscr{A}=\mathscr{A} \text{ и } \mathscr{A}^{*}\mathscr{A}\mathscr{A}^{+}=\mathscr{A}^{*}. \tag{1.113}$$

Чтобы проверить эти равенства, достаточно левые части выражения (1.110) умножить на оператор $\mathscr{A}$ и выражение (1.111) — на оператор A*.

Свойство 4. Псевдообратный оператор осуществляет переход от второго сингулярного базиса к первому сингулярному базису по формуле

$$\mathscr{A}^{+}y_i=\begin{cases}\rho_i^{-1}x_i & \text{для } i=1\div r,\\ 0 & \text{для } i\geqslant r+1.\end{cases} \tag{1.114}$$

Доказательство этого свойства непосредственно следует из формулы (1.89). Кроме того, из данного свойства следует, что если ρ_i для $i=1\div r$ являются ненулевыми сингулярными числами $\mathscr{A}$, то ρ_i^{-1} — ненулевые сингулярные числа $\mathscr{A}^{+}$.

Свойство 5. Псевдообратный оператор к оператору $\mathscr{A}^{+}$ равен оператору $\mathscr{A}$, т. е.

$$(\mathscr{A}^{+})^{+}=\mathscr{A}. \tag{1.115}$$

Для доказательства данного свойства в формуле (1.89) зафиксируем k, для которого $\rho_k\neq 0$, в результате чего получим равен-

ство $\mathscr{A}^{+}y_k=\rho_k^{-1}x_k$. Из этого равенства получаем $y_k=$ $=(\mathscr{A}^{+})^{+}\rho_k^{-1}x_k$. Так как $k\in N_r$, где N_r — множество натуральных чисел $i=1\div r$, то на основании формулы (1.89) имеем $\rho_k^{-1}\mathscr{A}x_k=$ $=(\mathscr{A}^{+})^{+}\rho_k^{-1}x_k$. Отсюда получаем (1.115).

Пусть оператор $\mathscr{A}$ — невырожденное линейное преобразование в пространстве X, тогда уравнение

$$\mathscr{A}x=b, \tag{1.116}$$

в котором x и b принадлежат X, имеет единственное решение, т. е. $x=\hat{x}=x^0$ (табл. 1, случай 1*а*). Отсюда

$$x=\mathscr{A}^{+}b.$$

В рассматриваемом случае оператор $\mathscr{A}^{+}$ называется *обратным преобразованием* к $\mathscr{A}$ и обозначается $\mathscr{A}^{-1}$. При этом решение уравнения (1.116) запишется в виде

$$x=\mathscr{A}^{-1}b. \tag{1.117}$$

Из равенств (1.117) и (1.116) и условий, что $x\neq 0$ и $b\neq 0$, следует

$$\mathscr{A}\mathscr{A}^{-1}=\mathscr{A}^{-1}\mathscr{A}=E. \tag{1.118}$$

Свойство 6. Для унитарного преобразования Q выполняется равенство

$$Q^{+}=Q^{-1}=Q^{*}. \tag{1.119}$$

Действительно, из определения унитарного преобразования и равенства (1.118) имеем $QQ^{*}=QQ^{-1}=E$, а отсюда следует (1.119).

Свойство 7. Если Q — унитарное преобразование в пространстве X, P — унитарное преобразование в Y и $\mathscr{A}$ — оператор, действующий из X в Y, то выполняется равенство

$$(P\mathscr{A}Q)^{+}=Q^{*}\mathscr{A}^{+}P^{*}. \tag{1.120}$$

Для доказательства этого свойства рассмотрим уравнение $P\mathscr{A}Qx=y$, нормальным псевдорешением которого является $x^0=$ $=(P\mathscr{A}Q)^{+}y$. Так как P — унитарное преобразование, то имеем $\mathscr{A}Qx=P^{*}y$, нормальным псевдорешением которого является $z^0=\mathscr{A}^{+}P^{*}y$, где $z^0=Qx^0$. Отсюда $x^0=Q^{*}\mathscr{A}^{+}P^{*}y$, что и доказывает равенство (1.120).

Как и для любого линейного отображения, для псевдообратного оператора в заданных базисах пространств X и Y существует матрица, которую называют *псевдообратной матрицей* оператора $\mathscr{A}^{+}$. Будем обозначать ее A^{+}.

Исходя из формулы (1.107), нормальное псевдорешение системы линейных уравнений (1.106), определяется как

$$x^0=A^{+}y. \tag{1.121}$$

Таблица 2

Матрица	Векторы			
$A=\begin{pmatrix}1&1&0\\1&0&-1\\0&1&1\\2&2&0\end{pmatrix}$	$b_1=\begin{pmatrix}1\\1\\0\end{pmatrix}$	$b_2=\begin{pmatrix}1\\0\\-1\end{pmatrix}$	$b_3=\begin{pmatrix}0\\1\\1\end{pmatrix}$	$b_4=\begin{pmatrix}2\\2\\0\end{pmatrix}$
$A^{\mathrm{T}}A=\begin{pmatrix}6&5&-1\\5&6&1\\-1&1&2\end{pmatrix}$	$a_1^{+}=\begin{pmatrix}1/11\\1/11\\0\end{pmatrix}$	$a_2^{+}=\begin{pmatrix}7/33\\-4/33\\-1/3\end{pmatrix}$	$a_3^{+}=\begin{pmatrix}-4/33\\7/33\\1/3\end{pmatrix}$	$a_4^{+}=\begin{pmatrix}2/11\\2/11\\0\end{pmatrix}$

Это же равенство может служить и для нахождения элементов псевдообратной матрицы A^{+}. Пусть a^{+}_{ij} — элементы матрицы A^{+}, причем $i=1\div m$ и $j=1\div n$. Тогда, обозначая j-й столбец матрицы A^{+} как $a^{+}_{j}=(a^{+}_{1j}a^{+}_{2j}\dots a^{+}_{mj})^{\mathrm{T}}$ и ставя в соответствие каждому j-му столбцу матрицы A^{+} в качестве вектора b единичный вектор $e_j=(0\dots010\dots0)^{\mathrm{T}}$, j-я компонента которого есть 1, получим

$$a_j^{+}=x_j^{0} \text{ для } j=1\div n, \tag{1.122}$$

где x_j^{0} — нормальное псевдорешение уравнения $Ax=e_j$. Отсюда для нахождения элементов псевдообратной матрицы по формуле (1.122) необходимо найти нормальные псевдорешения n систем уравнений с одной и той же матрицей системы и с n различными единичными столбцами.

Для определения элементов обратной матрицы в случае, когда матрица A размера $[n\times n]$ невырожденная, воспользуемся равенством (1.118). Тогда если a^{-}_{ij} обозначают элементы обратной матрицы, то их можно определить из решения n систем уравнений вида

$$Aa_j^{-}=e_j \text{ для } j=1\div n, \tag{1.123}$$

где $$a_j^{-}=(a_{1j}^{-}a_{2j}^{-}\dots a_{nj}^{-})^{\mathrm{T}}.$$

Таким образом, элементы псевдообратной или обратной матрицы можно найти из решения n систем уравнений с неизвестным вектором-столбцом $\tilde{a}_j$

$$A\tilde{a}_j=e_j \text{ для } j=1\div n. \tag{1.124}$$

При этом только для случая 1а табл. 1 в результате решений этих систем будут получены элементы обратной матрицы A^{-1},

во всех же остальных пяти случаях получаются элементы псевдообратной матрицы A^+.

В качестве примера найдем псевдообратную матрицу A^+ к матрице A, представленной в табл. 2.

Решение. Предположим, что матрицы A и A^τ определяют операторы $\mathscr{A}$ и $\mathscr{A}^*$ в ортонормированных базисах пространств X и Y. Тогда система нормальных уравнений (1.98) примет вид $A^\tau Ax = A^\tau y$. Сначала вычислим элементы матрицы $A^\tau A$, а затем — столбцы вида $b_j = A^\tau e_j$ для $j = 1 \div 4$. Результаты этих вычислений сведены в табл. 2. Найдем теперь псевдорешение для каждой из четырех полученных нормальных систем уравнений. Если решение этих систем проводить по схеме Гаусса, то, например, для первой из них получим

$$\begin{pmatrix} 6 & 5 & -1 \\ 5 & 6 & 1 \\ -1 & 1 & 2 \end{pmatrix} \begin{pmatrix} \hat{x}_{11} \\ \hat{x}_{12} \\ \hat{x}_{13} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \Leftrightarrow$$

$$\Leftrightarrow \begin{pmatrix} 1 & 5/6 & -1/6 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{x}_{11} \\ \hat{x}_{12} \\ \hat{x}_{13} \end{pmatrix} = \begin{pmatrix} 1/6 \\ 1/11 \\ 0 \end{pmatrix}.$$

Отсюда находим общее псевдорешение $\hat{x}_1$:

$$\hat{x}_{11} = 1/11 + \alpha, \; \hat{x}_{12} = 1/11 - \alpha, \; \hat{x}_{13} = \alpha.$$

Для определения нормального псевдорешения системы $A^\tau Ax = A^\tau b_1$ требуется вычислить значение α^0, при котором достигается

$$\inf \sum_{i=1}^{3} x_{1i}^2 = \inf [\,(1/11+\alpha)^2 + (1/11-\alpha)^2 + \alpha^2\,].$$

Отсюда получаем $\alpha^0 = 0$ и $x^0{}_{11} = 1/11$, $x^0{}_{12} = 1/11$, $x^0{}_{13} = 0$. Аналогичным образом определяем $\hat{x}_2$, $\hat{x}_3$, $\hat{x}_4$:

$$\hat{x}_2 = \begin{pmatrix} 6/11+\alpha \\ -5/11-\alpha \\ \alpha \end{pmatrix}, \quad \hat{x}_3 = \begin{pmatrix} -5/11+\alpha \\ 6/11-\alpha \\ \alpha \end{pmatrix}, \quad \hat{x}_4 = \begin{pmatrix} 2/11+\alpha \\ 2/11-\alpha \\ \alpha \end{pmatrix}$$

и нормальные псевдорешения $x_2{}^0$, $x_3{}^0$, $x_4{}^0$. Столбцы $a_1{}^+$, $a_2{}^+$, $a_3{}^+$, $a_4{}^+$ матрицы A^+, равные соответствующим векторам $x_1{}^0$, $x_2{}^0$, $x_3{}^0$, $x_4{}^+$, расположены в табл. 2 под столбцами b_1, b_2, b_3, b_4. Описанный способ нахождения элементов псевдообратной матрицы хотя и носит общий характер, но из-за своей сложности применение его ограничивается только матрицами низкой размерности.

Следующая теорема позволяет по более рациональному вычислительному алгоритму найти псевдообратную матрицу в двух широко применяемых на практике случаях, а именно когда для матрицы A размера $[n\times m]$

1) $n>m$ и $\operatorname{Rg} A=m$;

2) $n<m$ и $\operatorname{Rg} A=n$.

Теорема 1.26. Если $\mathscr{A}$ — оператор, отображающий m-мерное пространство X в n-мерное Y, то:

1) для $n>m$ и $\operatorname{Rg}\mathscr{A}=m$

$$\mathscr{A}^{+}=(\mathscr{A}^{*}\mathscr{A})^{-1}\mathscr{A}^{*}; \tag{1.125}$$

2) для $n<m$ и $\operatorname{Rg}\mathscr{A}=n$

$$\mathscr{A}^{+}=\mathscr{A}^{*}(\mathscr{A}\mathscr{A}^{*})^{-1}. \tag{1.126}$$

Доказательство. Рассмотрим сначала случай 1, для которого нормальное уравнение принимает вид

$$\mathscr{A}^{*}\mathscr{A}\hat{x}=\mathscr{A}^{*}y \Leftrightarrow T\hat{x}=\mathscr{A}^{*}y,$$

где $T=\mathscr{A}^{*}\mathscr{A}$ — линейное преобразование в пространстве X, причем ранг этого преобразования равен размерности пространства X, т. е. $\operatorname{Rg}\mathscr{A}^{*}\mathscr{A}=m$. Тогда из формулы (1.117) имеем

$$\hat{x}=T^{-1}\mathscr{A}^{*}y \Leftrightarrow \hat{x}=(\mathscr{A}^{*}\mathscr{A})^{-1}\mathscr{A}^{*}y.$$

Так как $\hat{x}=x^{0}$, то для первого случая

$$\mathscr{A}^{+}=(\mathscr{A}^{*}\mathscr{A})^{-1}\mathscr{A}^{*}.$$

Пусть в операторном уравнении $\mathscr{A}^{*}u=v$ для оператора $\mathscr{A}^{*}$ выполняются условия 2. При этом $u\in Y$, а $v\in X$. Нормальное уравнение для данного уравнения примет вид

$$\mathscr{A}\mathscr{A}^{*}\hat{u}=\mathscr{A}v \Leftrightarrow R\hat{u}=\mathscr{A}v,$$

где $R=\mathscr{A}\mathscr{A}^{*}$ — линейное преобразование в пространстве Y, ранг которого равен размерности пространства Y. Отсюда, с учетом того, что $\hat{u}=u^{0}$, получаем

$$u=R^{-1}\mathscr{A}v \Leftrightarrow u^{0}=(\mathscr{A}\mathscr{A}^{*})^{-1}\mathscr{A}v.$$

Так как $u^{0}=(\mathscr{A}^{*})^{+}v$, то полученное равенство приводится к виду (1.126) в результате следующих преобразований:

$$(\mathscr{A}^{*})^{+}=(\mathscr{A}\mathscr{A}^{*})^{-1}\mathscr{A} \Leftrightarrow (\mathscr{A}^{+})^{*}=(\mathscr{A}\mathscr{A}^{*})^{-1}\mathscr{A} \Leftrightarrow \mathscr{A}^{+}=$$
$$=\mathscr{A}^{*}(\mathscr{A}\mathscr{A}^{*})^{-1}.$$

Из доказанной теоремы следует, что если в евклидовых пространствах X и Y зафиксированы базисы и A — матрица оператора $\mathscr{A}$, имеющая размер $[n\times m]$ и $\operatorname{Rg} A=m$, то псевдообратная матрица A^{+} в тех же базисах, исходя из формулы (1.71), будет

$$A^{+}=(A^{\mathrm{T}}G_{q}A)^{-1}A^{\mathrm{T}}G_{q}. \tag{1.127}$$

Для случая, когда $\operatorname{Rg} A=n$

$$A^{+}=G_e^{-1}A^{\tau}(AG_e^{-1}A^{\tau})^{-1}. \qquad (1.128)$$

В полученных формулах G_e и G_q — матрицы Грама, первая из которых составлена для системы базисных векторов в X, а вторая — в Y.

Если же базисы в пространствах X и Y ортонормированные, то формулы (1.127) и (1.128) упрощаются и принимают вид

$$A^{+}=(A^{\tau}A)^{-1}A^{\tau},\ A^{+}=A^{\tau}(AA^{\tau})^{-1}. \qquad (1.129)$$

При использовании формул (1.127) и (1.128) можно выполнить псевдообращение любой ненулевой матрицы размера $[n\times m]$. Прежде чем переходить к описанию такого алгоритма, введем следующее понятие.

Определение 1.22. Разложение матрицы A размера $[n\times m]$ и ранга $r>0$ на множители в виде $A=BC$, где B и C имеют размеры соответственно $[n\times r]$ и $[r\times m]$, называется *скелетным разложением матрицы* A.

Приведем один из способов скелетного разложения матрицы A размера $[n\times m]$ и ранга $0<r<\min(m, n)$. Представим матрицу A в виде $A=(\bar{a}_1\bar{a}_2\ldots\bar{a}_j\ldots\bar{a}_m)$, где $\bar{a}_j$ — вектор-столбец с компонентами a_{ij} при $i=1\div n$. Так как ранг матрицы A равен r, то среди $\bar{a}_j$ имеется система из r векторов, являющаяся линейно независимой. Используя матрицу перестановок столбцов P, перейдем от матрицы A к матрице $A_p=AP$, в которой $A_p=(\bar{a}^p{}_1\bar{a}^p{}_2\ldots\bar{a}^p{}_r\bar{a}^p{}_{r+1}\ldots\bar{a}^p{}_m)$, причем система векторов $\{\bar{a}^p{}_j\}_{j=1\div r}$ является линейно независимой. Тогда в силу линейной зависимости векторов $a^p{}_j$ при $j\geqslant r+1$ их можно разложить по базису $\{\bar{a}^p{}_j\}_{j=1\div r}$, т. е.

$$\bar{a}_j{}^p=\alpha_{1j}\bar{a}_1{}^p+\alpha_{2j}\bar{a}_2{}^p+\ldots+\alpha_{rj}\bar{a}_r{}^p \text{ для } i=(r+1)\div m. \qquad (1.130)$$

Составим матрицы $B=(\bar{a}^p{}_1\bar{a}^p{}_2\ldots\bar{a}^p{}_r)$ и $C=(\bar{c}_1\bar{c}_2\ldots\bar{c}_r\bar{c}_{r+1}\ldots\bar{c}_m)$, где c_j — вектор-столбец с компонентами c_{ij} при $i=1\div r$. Если в матрице C положить $\bar{c}_1=\bar{e}_1$, $\bar{c}_2=\bar{e}_2,\ldots,\bar{c}_r=\bar{e}_r$, где $\bar{e}_j$ — единичный вектор-столбец с j-й компонентой, равной единице, а $\bar{c}_{r+1}=\bar{\alpha}_1$, $\bar{c}_{r+2}=\bar{\alpha}_2$, …, $\bar{c}_m=\bar{\alpha}_{m-r}$, где $\bar{\alpha}_j$ — вектор-столбец с компонентами α_{ij} при $i=1\div r$, входящими в разложение (1.130), то получим

$$A_p=BC \text{ и } A=BC_p, \qquad (1.131)$$

где B — матрица размера $[n\times r]$, $C_p=CP$ — матрица размера $[r\times m]$.

Теорема 1.27. Если в пространствах X и Y соответственно размерностей m и n задан оператор $\mathscr{A}$ ранга $r>0$, отображающий X в Y, и заданы операторы $\mathscr{B}$ и $\mathscr{C}$, первый из которых отоб-

ражает подпространство $X_r \subseteq X$ размерности r в Y, а второй — X в подпространство X_r, причем $\mathscr{A}=\mathscr{B}\mathscr{C}$, то для псевдообратного оператора $\mathscr{A}^+$ имеет место равенство

$$\mathscr{A}^+=\mathscr{C}^+\mathscr{B}^+=\mathscr{C}^*(\mathscr{C}\mathscr{C}^*)^{-1}(\mathscr{B}^*\mathscr{B})^{-1}\mathscr{B}^*. \qquad (1.132)$$

Доказательство. Из формул (1.113) следует

$$\mathscr{A}^*\mathscr{A}\mathscr{A}^+=\mathscr{A}^* \Leftrightarrow \mathscr{C}^*\mathscr{B}^*\mathscr{B}\mathscr{C}\mathscr{A}^+=$$
$$=\mathscr{C}^*\mathscr{B}^* \Leftrightarrow (\mathscr{C}\mathscr{C}^*)(\mathscr{B}^*\mathscr{B})\mathscr{C}\mathscr{A}^+=(\mathscr{C}\mathscr{C}^*)\mathscr{B}^*.$$

Так как $(\mathscr{C}\mathscr{C}^*)$ и $(\mathscr{B}^*\mathscr{B})$ — невырожденные преобразования, то имеем

$$\mathscr{C}\mathscr{A}^+=(\mathscr{B}^*\mathscr{B})^{-1}\mathscr{B}^*.$$

Применим теперь свойство 3 и формулу (1.108) для сопряженного оператора $\mathscr{A}^*$, тогда

$$\mathscr{A}\mathscr{A}^*(\mathscr{A}^*)^+=\mathscr{A} \Leftrightarrow \mathscr{B}\mathscr{C}\mathscr{C}^*\mathscr{B}^*(\mathscr{A}^+)^*=$$
$$=\mathscr{B}\mathscr{C} \Leftrightarrow \mathscr{B}^*(\mathscr{A}^+)^*=(\mathscr{C}\mathscr{C}^*)^{-1}\mathscr{C},$$
$$\mathscr{A}^+\mathscr{B}=\mathscr{C}^*(\mathscr{C}\mathscr{C}^*)^{-1}.$$

Перемножая операторы $\mathscr{A}^+\mathscr{B}$ и $\mathscr{C}\mathscr{A}^+$ и применяя свойство 5 и свойство 3 к оператору $\mathscr{A}^+$, получаем

$$\mathscr{A}^+\mathscr{B}\mathscr{C}\mathscr{A}^+=\mathscr{A}^+\mathscr{A}\mathscr{A}^+ \Leftrightarrow (\mathscr{A}^+)(\mathscr{A}^+)^+(\mathscr{A}^+)=$$
$$=\mathscr{A}^+ \Leftrightarrow \mathscr{A}^+\mathscr{A}\mathscr{A}^+=\mathscr{A}^+,$$
$$\mathscr{A}^+=\mathscr{C}^*(\mathscr{C}\mathscr{C}^*)^{-1}(\mathscr{B}^*\mathscr{B})^{-1}\mathscr{B}^*.$$

Отсюда следует теорема 1.27, в которой

$$\mathscr{A}^+=\mathscr{C}^+\mathscr{B}^+.$$

Следствие 1. Если $A=BC$ — скелетное разложение матрицы A и A, B, C — матрицы операторов $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$ в произвольно заданных базисах пространств X, Y, X_r, то псевдообратная матрица A^+ имеет вид

$$A^+=G_e^{-1}C^{\mathrm{T}}(CG_e^{-1}C^{\mathrm{T}})^{-1}(B^{\mathrm{T}}G_qB)^{-1}B^{\mathrm{T}}G_q. \qquad (1.133)$$

Для доказательства этого равенства достаточно в формулу (1.132) вместо C^* и B^* подставить их выражения, определяемые из формулы (1.71).

Если матрицы A, B, C соответствуют операторам $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$ в ортонормированных базисах пространств X, Y, X_r, то равенство (1.133) преобразуется к виду

$$A^+=C^{\mathrm{T}}(CC^{\mathrm{T}})^{-1}(B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}}. \qquad (1.134)$$

В качестве примера найдем псевдообратную матрицу A^+ для матрицы A, заданной в табл. 2, используя метод скелетного разложения исходной матрицы. За матрицу B можно взять первые

два столбца матрицы A, для которой $\mathrm{Rg}\,A=2$. Представляя матрицу C в виде

$$C=\begin{pmatrix}1 & 0 & \alpha_{11}\\ 0 & 1 & \alpha_{21}\end{pmatrix},$$

определим значения α_{11} и α_{21} из решения системы уравнений вида

$$\begin{pmatrix}0\\-1\\1\\0\end{pmatrix}=\alpha_{11}\begin{pmatrix}1\\1\\0\\2\end{pmatrix}+\alpha_{21}\begin{pmatrix}1\\0\\1\\2\end{pmatrix}\Leftrightarrow\begin{pmatrix}0\\-1\end{pmatrix}=\alpha_{11}\begin{pmatrix}1\\1\end{pmatrix}+\alpha_{21}\begin{pmatrix}1\\0\end{pmatrix}.$$

В результате решения этой системы имеем $\alpha_{11}=-1$, $\alpha_{21}=1$. Скелетным разложением матрицы A будет

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

Для получения матрицы A^+ можно теперь воспользоваться формулой (1.134), тогда

$$A^+=\begin{pmatrix}1 & 0\\ 0 & 1\\ -1 & 1\end{pmatrix}\begin{pmatrix}2 & -1\\ -1 & 2\end{pmatrix}^{-1}\begin{pmatrix}6 & 5\\ 5 & 6\end{pmatrix}^{-1}\begin{pmatrix}1 & 1 & 0 & 2\\ 1 & 0 & 1 & 2\end{pmatrix},$$

и после соответствующих вычислений получим матрицу A^+, аналогичную той, которая представлена в табл. 2.

Весьма эффективный алгоритм псевдообращения матрицы можно получить в результате сочетания ортогонального и скелетного разложений матрицы. Так, на основании теоремы 1.6 любую матрицу A размера $[n\times m]$ и ранга $r>0$ можно разложить таким образом:

$$A=V^{\mathrm{T}}RW,$$

где V и W — ортогональные матрицы соответственно размеров $[n\times n]$ и $[m\times m]$; R — матрица размера $[n\times m]$:

$$R=\begin{pmatrix}R_{11} & 0\\ 0 & 0\end{pmatrix}.$$

Здесь R_{11} — треугольная матрица размера $[r\times r]$. Псевдообратная матрица A^+ с учетом указанного разложения и формулы (1.120) примет вид

$$A^+=W^{\mathrm{T}}R^+V. \qquad (1.135)$$

Для нахождения элементов матрицы R^+ применим метод скелетного разложения матрицы R

$$R=\begin{pmatrix} R_{11} \\ 0 \end{pmatrix}(E_r 0),$$

где E_r — единичная матрица размера $[r\times r]$. В результате по формуле (1.134) находим R^+

$$R^+=\begin{pmatrix} E_r \\ 0 \end{pmatrix}(R_{11}{}^{\mathrm{T}}R_{11})^{-1} \Leftrightarrow R^+=\begin{pmatrix} R_{11}{}^{-1}(R_{11}{}^{-1})^{\mathrm{T}} \\ 0 \end{pmatrix}. \qquad (1.136)$$

Таким образом, алгоритм нахождения матрицы A^+ будет состоять в следующем:

1) ортогональное разложение матрицы A по формуле (1.33);
2) вычисление элементов матрицы $R^{-1}{}_{11}$;
3) вычисление элементов псевдообратной матрицы R^+ по формуле (1.136);
4) вычисление псевдообратной матрицы A^+ по формуле (1.135).

В заключение этого параграфа отметим, что в тех случаях, когда элементы матрицы A^+ определены, решение уравнения (1.106) по методу наименьших квадратов сводится к вычислению вектора x^0 по формуле (1.121). При этом будут охвачены все случаи решения, сведенные в табл. 1.

§ 1.11. НОРМЫ МАТРИЦ И ПРЕДЕЛ ПОСЛЕДОВАТЕЛЬНОСТИ МАТРИЦ

Рассмотрим множество матриц M_{nm} размера $[n\times m]$. Пусть в этом множестве введены обычные операции сложения матриц и умножения матрицы на число. Причем это число и элементы матрицы, принадлежащей множеству M_{nm}, берутся из одного поля P. Тогда данное множество M_{nm} с введенными в нем операциями представляет собой конечномерное линейное пространство над полем P. Дадим этому пространству такое же обозначение, что и множеству матриц — M_{nm}.

Зададим в пространстве M_{nm} числовую функцию $\varphi(A)$, которая каждой матрице из M_{nm} сопоставляет некоторое вещественное число, удовлетворяющее всем условиям определения 1.3. Эту функцию будем называть *нормой* в пространстве матриц, а ее значение на матрице A — *нормой матрицы*.

Часто вместо обозначения $\varphi(A)$ используют обозначение $\|A\|$, как и для нормы вектора. Таким образом, для нормы матрицы должны выполняться следующие условия:

$$\begin{aligned} &1.\ \|A\|>0 \text{ для любой ненулевой матрицы из } M_{nm}.\\ &2.\ \|\alpha A\|=|\alpha|\,\|A\|, \text{ где } \alpha \text{ — число из поля } P. \qquad (1.137)\\ &3.\ \|A+B\|\leqslant\|A\|+\|B\|, \text{ где } A, B\in M_{nm}. \end{aligned}$$

Рассматривая теперь M_{nm} как нормированное конечномерное пространство размерности $[n\times m]$, элементами которого являются a_{ij}, а норма удовлетворяет условиям (1.137), получаем следующие свойства для нормы матрицы.

Свойство 1. Норма матрицы непрерывно зависит от элементов матрицы. Это свойство является прямым следствием непрерывности нормы вектора.

Свойство 2. Любые две нормы матрицы, принадлежащие пространству M_{nm}, являются попарно эквивалентными, т. е. для них выполняется условие

$$m_2\varphi_2(A) \leqslant \varphi_1(A) \leqslant m_1\varphi_2(A), \tag{1.138}$$

где m_1 и m_2 — числовые константы больше нуля и зависят от выбора норм $\varphi_1(A)$ и $\varphi_2(A)$. Данное свойство является следствием теоремы 1.10.

Определение 1.23. Если в пространстве M_{n^2} норма матрицы удовлетворяет для любых матриц A и B условию

$$\|AB\| \leqslant \|A\|\,\|B\|, \tag{1.139}$$

то ее будем называть *мультипликативной нормой* матрицы.

Свойство 3. Если матрица $A\in M_{n^2}$ и k — натуральное число, для мультипликативных норм матриц имеет место неравенство

$$\|A^k\| \leqslant \|A\|^k. \tag{1.140}$$

Действительно, если $\|A\|$ — мультипликативная норма матрицы A, то на основании условия (1.139) имеем

$$\|A^k\| = \|AA^{k-1}\| \leqslant \|A\|\,\|A^{k-1}\| =$$
$$= \|A\|\,\|AA^{k-2}\| \leqslant \|A\|^2\,\|A^{k-2}\|.$$

Продолжая аналогичные преобразования, в итоге получим неравенство (1.140).

Свойство 4. Любая мультипликативная норма единичной матрицы не меньше единицы, т. е.

$$\|E\| \geqslant 1. \tag{1.141}$$

Это свойство непосредственно вытекает из неравенства

$$\|AE\| \leqslant \|A\|\,\|E\|.$$

Свойство 5. Если матрица A невырожденная, то для любой ее мультипликативной нормы выполняется неравенство

$$\|A^{-1}\| \geqslant \|A\|^{-1}. \tag{1.142}$$

Доказательство этого свойства можно получить, используя определение обратной матрицы и свойства 5. В качестве упражнения проведите его самостоятельно.

Приведем наиболее часто используемые в практике вычислений нормы матриц, принадлежащие пространству M_{nm}.

$$1.\ \|A\|_l = \max_j \sum_{i=1}^{n} |a_{ij}|, \tag{1.143}$$

т. е. максимальная из сумм модулей элементов матрицы A по столбцам. Число таких сумм равно числу столбцов m.

$$2.\ \|A\|_c = \max_i \sum_{i=1}^{m} |a_{ij}|, \tag{1.144}$$

т. е. максимальная из сумм модулей элементов матрицы A по строкам. Число таких сумм равно числу строк n.

$$3.\ \|A\|_{E,U} = \left(\sum_i^m \sum_j^n |a_{ij}|^2\right)^{1/2}, \tag{1.145}$$

получившая название *евклидовой нормы матрицы*, если a_{ij} — вещественные числа, и название *унитарной нормы матрицы*, когда a_{ij} — комплексные числа.

Евклидову (или унитарную) норму матрицы можно определить и таким образом:

$$\begin{aligned} \|A\|_E &= \sqrt{\operatorname{tr}(A^{\mathrm{T}}A)} = \sqrt{\operatorname{tr}(AA^{\mathrm{T}})}, \\ \|A\|_U &= \sqrt{\operatorname{tr}(\bar{A}^{\mathrm{T}}A)} = \sqrt{\operatorname{tr}(A\bar{A}^{\mathrm{T}})}. \end{aligned} \tag{1.146}$$

Эквивалентность определений евклидовых норм матрицы по формулам (1.145) и (1.146) непосредственно следует из выражения для q_{jj} элемента матрицы $A^{\mathrm{T}}A$:

$$q_{jj} = \sum_{i=1}^{n} a_{ij}^2 \text{ при } j = 1 \div m.$$

$$4.\ M(A) = (mn)^{1/2} \max_{i,j} |a_{ij}|, \tag{1.147}$$

которая получила название M-нормы. Отметим, что и выражение

$$\|A\|_M = \max_{i,j} |a_{ij}|$$

будет тоже нормой матрицы A.

$$5.\ \|A\|_2 = \rho_1, \text{ где } \rho_1 = \sqrt{\max |\lambda_{A^{\mathrm{T}}A}|}, \tag{1.148}$$

получившая название *спектральной нормы матрицы*. Числа $\rho_i = \sqrt{\lambda_i}$, где λ_i — собственные значения матрицы $A^{\mathrm{T}}A$, являются *сингулярными числами* матрицы A.

Если вместо пространства M_{nm} рассматривать пространство квадратных матриц M_{n^2}, то нормы матриц, представленные фор-

мулами (1.143) — (1.148), будут и мультипликативными нормами матриц, в то время как норма матрицы

$$\|A\|_M = \max_{i,j} |a_{ij}|$$

не является мультипликативной. Для доказательства приведем пример. Пусть нормы матриц A и B, принадлежащие M_{n^2}, имеют вид

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix},$$

тогда $\|AB\|_M = 2$, $\|A\|_M = 1$, $\|B\|_M = 1$. Отсюда нетрудно проверить, что нарушается условие (1.139).

Для доказательства того, что выражения, определяемые по формулам (1.143) — (1.148), являются нормами матриц, требуется выполнить проверку всех трех условий (1.137), входящих в определение нормы матрицы.

Практически любой вычислительный процесс в линейных нормированных пространствах X и Y связан с отображениями вида $\mathscr{A}x = y$, которые при фиксированных базисах в пространстве X и Y определяются как $Ax = y$, где A — матрица размера $[n \times m]$, x — вектор-столбец размера m и y — вектор-столбец размера n, т. е. x и y принадлежат арифметическим пространствам соответственно R_m и R_n. В пространствах R_m и R_n можно ввести нормы $\varphi_m(x)$ и $\varphi_n(y)$, которые должны удовлетворять всем трем условиям определения 1.3. Таким образом, в R_n норма для вектора y будет определяться как

$$\varphi_n(y) = \| Ax \|. \tag{1.149}$$

Возникает теперь вопрос, можно ли найти такую числовую функцию $\varphi_c(A) \geqslant 0$, для которой при $\forall x \in X$, $\forall y \in Y$ и любой $A \in M_{nm}$ будет выполняться условие

$$\varphi_n(y) \leqslant \varphi_c(A)\varphi_m(x). \tag{1.150}$$

Если на этот вопрос можно ответить положительно, то функцию $\varphi_c(A)$, удовлетворяющую условиям (1.137), будем называть *согласованной нормой* матрицы A с нормами в арифметических пространствах R_m и R_n.

Представляя матрицу A как

$$(\bar{a}_1 \bar{a}_2 \ldots \bar{a}_i \ldots \bar{a}_n)^{\mathrm{T}},$$

где a_i — вектор-строка с компонентами a_{ij} при $j = 1 \div m$. Тогда

$$Ax = (\gamma_1 \gamma_2 \ldots \gamma_i \ldots \gamma_n)^{\mathrm{T}},$$

где $\gamma_i = (\bar{a}_i, x)$.

И для евклидовой нормы в пространстве R_n получим

$$\| Ax \|_E = (\sum_{i=1}^{n} \gamma_i^2)^{1/2} = (\sum_{i=1}^{n} (\bar{a}_i, x)^2)^{1/2}.$$

Задав евклидову норму в пространстве R_m на основании неравенства Коши - Буняковского, получим

$$(a_i, x)^2 \leqslant \| a_i \|_E^2 \| x \|_E^2.$$

Отсюда

$$\| Ax \|_E \leqslant (\sum_{i=1}^{n} \| \bar{a}_i \|_E) \| x \|_E \Leftrightarrow \| Ax \|_E \leqslant$$
$$\leqslant \| A \|_E \| x \|_E. \qquad (1.151)$$

Здесь введенная функция $\varphi_c(A) = \|A\|_E$ отвечает всем требованиям, предъявляемым к норме матрицы. В силу теоремы 1.10 об эквивалентности норм в арифметическом пространстве неравенство (1.151) распространяется и на другие нормы, вводимые в арифметических пространствах. Итак, установлена возможность всегда найти такую норму $\varphi_c(A)$, для которой в арифметических пространствах R_m и R_n выполняется условие (1.150).

Свойство 6. Норма матрицы $\|A\|_l$ согласована с l-нормами в арифметических пространствах R_m и R_n, $\|A\|_c$ согласована с c-нормами в R_m и R_n, $\|A\|_E$ и $\|A\|_2$ — с евклидовыми нормами в R_m и R_n, M — норма матрицы согласована с l-, c-нормой и евклидовой нормой арифметических пространств R_m и R_n.

Доказательство этого свойства для евклидовой нормы матрицы следует из неравенства (1.151).

Выполним доказательство этого свойства для l-нормы матрицы. Обозначим i-ю компоненту вектора-столбца $y = Ax$ через $\alpha_i = \sum_{i=1}^{m} a_{ij} x_j$. Тогда

$$\| Ax \|_l = \sum_{i=1}^{n} | \alpha_i | = \sum_{i=1}^{n} | \sum_{j=1}^{m} a_{ij} x_j | \Leftrightarrow$$
$$\Leftrightarrow \| Ax \|_l \leqslant \sum_{j=1}^{m} (\sum_{i=1}^{n} | a_{ij} |) | x_j |.$$

Если в полученном неравенстве заменить $\sum_{i=1}^{n} |a_{ij}|$ на $\max_j \sum_{i=1}^{n} |a_{ij}|$, то

$$\| Ax \|_l \leqslant (\max_j \sum_{i=1}^{n} | a_{ij} |) \sum_{j=1}^{m} | x_j | \Leftrightarrow$$
$$\Leftrightarrow \| Ax \|_l \leqslant \| A \|_l \| x \|_l.$$

Доказательство свойства для других норм предлагаем выполнить самостоятельно.

Свойство 7. Модули собственных значений любого линейного преобразования не превосходят любой согласованной нормы матрицы этого преобразования, т. е.

$$|\lambda| \leqslant \|A\|. \tag{1.152}$$

Рассмотрим преобразование вида $Ax=\lambda x$, для которого $\|Ax\|=|\lambda|\,\|x\|$. Если $\|A\|$ — согласованная норма, то $|\lambda|\|x\| \leqslant \leqslant \|A\|\|x\|$. И так как $\|x\| \neq 0$, то $|\lambda| < \|A\|$.

Свойство 8. Любая мультипликативная норма матрицы является согласованной.

Пусть в M_{n^2} заданы матрицы A, B, S, причем $S=AB$. Тогда, как следует из формулы (1.150), всегда найдется такая норма $\varphi_c(S)$, для которой $\varphi_n(Sx) \leqslant \varphi_c(S)\varphi_n(x)$, где $x \in R_n$. Если $\varphi_c(S)$ — мультипликативная норма, то

$$\varphi_n(Sx) \leqslant \|A\|\,\|B\|\,\varphi_n(x).$$

С другой стороны, обозначая $Bx=z$, из (1.150) получаем неравенство

$$\varphi_n(Az) \leqslant \varphi_c(A)\varphi_n(z) \Leftrightarrow \varphi_n(Sx) \leqslant \varphi_c(A)\varphi_c(B)\varphi_n(x).$$

Сопоставляя данное неравенство с предыдущим, убеждаемся в справедливости сформулированного свойства.

Следует, однако, обратить внимание на тот факт, что не всякая согласованная норма является мультипликативной. Например, рассмотрим норму матрицы $\|A\|_M$. Эта норма для матриц, принадлежащих множеству M_{nm}, является согласованной с векторными l-нормой, заданной в пространстве R_m, и c-нормой, заданной в пространстве R_n. Данное утверждение легко проверяется, если рассмотреть неравенство

$$\left|\sum_{j=1}^{m} a_{ij}x_j\right| \leqslant \max_{ij} |a_{ij}| \sum_{j=1}^{m} |x_j|,$$

выполняющееся для любого $i=1 \div n$. В то же время $\|A\|_M$ не является мультипликативной.

Рассмотрим норму вектора $\|Ax\|$ как функцию $f_A(x)$ на множестве $G_x=\{x, \|x\|=1\}$. Так как норма вектора непрерывна от координат вектора, то $f_A(x)$ — непрерывная функция на G_x. Кроме того, G_x — единичная сфера в R_m, являющаяся замкнутым и ограниченным множеством, а следовательно, $f(x)$ — ограниченная функция, причем на большее значение функции $f_A(x)$ достигается при каком-то значении $x_0 \in G_x$. Обозначая наибольшее значение $f_A(x)$ как

$$\sup_{\|x\|=1} f_A(x) = \sup_{\|x\|=1} \|Ax\|,$$

полученный выше результат можно представить в виде

$$f_A(x_0)=\sup_{\|x\|=1}\|Ax\|. \tag{1.153}$$

Теорема 1.28. Наибольшее значение функции $f_A(x)=\|Ax\|$, заданной на единичной сфере $\|x\|=1$, где $x\in R_m$, $Ax\in R_n$ и $A\in M_{nm}$, является согласованной нормой матрицы и не превосходит любой другой согласованной нормы при заданных векторных нормах в пространствах R_n и R_m, т. е. выполняется неравенство

$$\sup_{\|x\|=1}\|Ax\|\leqslant\varphi_c(A). \tag{1.154}$$

Доказательство первой части теоремы начнем с проверки выполнения условий (1.137).

1. В связи с тем что $f_A(x)=\|Ax\|$ — неотрицательная функция и $|A\cdot 0\|\Leftrightarrow 0$, выполняется первое условие в определении нормы матрицы.

2. Так как $(\lambda A)x=\lambda(Ax)$, то $\|(\lambda A)x\|=|\lambda|\,\|Ax\|$. Отсюда на основании (1.153) имеем

$$|\lambda|\,f_A(x_0)=\sup_{\|x\|=1}\|(\lambda A)x\|\Leftrightarrow|\lambda|\sup_{\|x\|=1}\|Ax\|=\sup_{\|x\|=1}\|(\lambda A)x\|,$$

т. е. выполняется и второе условие в определении нормы матрицы.

3. Заметим, что для любых A, B и $x\in Q_x$ и при $C=A+B$ по определению нормы вектора

$$\|Cx\|=\|(Ax)+(Bx)\|\leqslant\|Ax\|+\|Bx\|.$$

Если $f_c(x)=\|Cx\|$ принимает в точке $x_c\in Q_x$ наибольшее значение, то приведенное выше неравенство с учетом (1.153) преобразуется к виду

$$f_C(x_c)\leqslant f_A(x_c)+f_B(x_c).$$

Принимая во внимание, что $f_A(x_c)\leqslant f_A(x_a)$ и $f_B(x_c)\leqslant f_B(x_b)$, где x_a и x_b — точки из Q_x, в которых функции f_A и f_B принимают наибольшие значения, получаем

$$f_C(x_c)\leqslant f_A(x_a)+f_B(x_b)\Leftrightarrow\sup_{\|x\|=1}\|(A+B)x\|\leqslant$$

$$\leqslant\sup_{\|x\|=1}\|Ax\|+\sup_{\|x\|=1}\|Bx\|.$$

Тем самым устанавливается выполнение третьего условия в определении нормы матрицы.

Таким образом, доказано, что $\sup\limits_{\|x\|=1}\|Ax\|$ является нормой матрицы A и не зависит от x. Поэтому правомочно следующее обозначение:

$$\varphi_и(A)=\sup_{\|x\|=1}\|Ax\|. \tag{1.155}$$

Доказательство того, что $\varphi_{\text{и}}(A)$ — согласованная норма матрицы, связано с проверкой выполнения неравенства $\|Ax\| \leqslant \leqslant \varphi_{\text{и}}(A)\|x\|$. Так как $\|x\|=1$, то $\|Ax\| \leqslant \varphi_{\text{и}}(A)$, а это неравенство выполняется для $\forall x \in Q_x$ в силу того, что $\varphi_{\text{и}}(A)$ — наибольшее значение $f_A(x)$.

Для доказательства второй части теоремы рассмотрим неравенство $\|Ax\| \leqslant \varphi_{\text{с}}(A)\|x\|$, которое выполняется для всех $x \in R_m$ и любой согласованной нормы матрицы с векторными нормами $\|*\|$, заданными в R_n и R_m. Так как это неравенство выполняется для всех $x \in R_m$, то для $x_0 \in Q_x$, в котором $f_A(x)=\|Ax\|$ принимает наибольшее значение, получим

$$\|Ax_0\| \leqslant \varphi_{\text{с}}(A)\|x_0\| \Rightarrow \varphi_{\text{и}}(A) \leqslant \varphi_{\text{с}}(A). \tag{1.156}$$

Определение 1.24. Норма матрицы в пространстве M_{nm}, определяемая формулой (1.155), называется *индуцированной нормой в пространствах* R_n и R_m или просто *индуцированной нормой матрицы*.

Это определение эквивалентно определению для индуцированной нормы матрицы, в котором для всех $z \neq 0$ из R_m выполняется

$$\varphi_{\text{и}}(A)=\sup_{z \neq 0}(\|Az\|/\|z\|). \tag{1.157}$$

Так, если в формуле (1.155) положить $x=z/\|z\|$, то получаем формулу (1.157), а если теперь в (1.157) положить $z=x$ и $\|x\|=1$, то получим (1.155).

Свойство 9. Индуцированная норма матрицы не превосходит согласованной нормы той же матрицы. Это свойство непосредственно вытекает из теоремы 1.28.

Свойство 10. Любая индуцированная норма в M_{n^2} является и мультипликативной нормой.

Пусть A и B принадлежат M_{n^2} и $C=AB$. Тогда по определению (1.139) и с учетом согласованности индуцированных норм получим

$$\varphi_{\text{и}}(C)=\sup_{\|x\|=1}\|A(Bx)\| \Rightarrow \varphi_{\text{и}}(C) \leqslant \sup_{\|x\|=1}(\varphi_{\text{и}}(A)\|Bx\|$$

и далее, так как

$$\sup(\varphi_{\text{и}}(A))\|(Bx)\|=\varphi_{\text{и}}(A)\sup_{\|x\|=1}\|Bx\| \Rightarrow \varphi_{\text{и}}(C) \leqslant \varphi_{\text{и}}(A)\varphi_{\text{и}}(B).$$

Свойство 11. Любая индуцированная норма единичной матрицы равна единице.

Доказательство данного свойства в порядке упражнения предлагаем выполнить самостоятельно. Заметим, что для евклидовой нормы матрицы $\|E\|_E=\sqrt{n}$, а для нормы, определяемой по формуле (1.147), имеем $M(E)=n$.

Свойство 12. Норма матрицы $\|A\|_l$ в пространстве M_{nm} является индуцированной l-нормами, $\|A\|_c$ — индуцированной C-нормами и $\|A\|_2$ — индуцированной евклидовыми нормами в пространствах R_n и R_m.

Выполним доказательство этого свойства для $\|A\|_2$ — спектральной нормы матрицы. Введем для евклидовых пространств R_m и R_n, в которых $\mathscr{A}$: $R_m \to R_n$, отношение

$$\rho(x) = \| Ax \|_E / \| x \|_E, \tag{1.158}$$

получившее название *отношения Релея*. В качестве базиса в R_m возьмем сингулярный базис $\{x_k\}_{k=1\div m}$. При этом предположим, что соответствующие сингулярные числа расположены в невозрастающем порядке $\rho_1 \geqslant \rho_2 \geqslant \ldots \geqslant \rho_n$.

Разложим вектор x по базису

$$x = \sum_{k=1}^{m} \alpha_k x_k.$$

Так как система $\{x_k\}_{k=1\div m}$ ортонормированная, то

$$\rho(x) = \frac{(x,\ A^{\mathrm{T}}Ax)^{1/2}}{(x,\ x)^{1/2}} \Leftrightarrow \rho(x) = \left(\frac{\sum\limits_{k=1}^{m} \alpha_k^2 \rho_k^2}{\sum\limits_{k=1}^{m} \alpha_k^2} \right)^{1/2}, \tag{1.159}$$

здесь ρ_k — сингулярные числа отображения $\mathscr{A}$. Положив $\rho_k = \rho_1$ для $k = 1 \div m$, где ρ_1 — наибольшее сингулярное число отображения $\mathscr{A}$, получаем

$$\| Ax \|_E / \| x \|_E \leqslant \rho_1. \tag{1.160}$$

Причем для $x = x_1$, где $\|x_1\|_E = 1$, $\rho(x_1) = \rho_1$. Таким образом, имеем

$$\sup_{\|x\|_E = 1} \| Ax \|_E = \rho_1,$$

а следовательно

$$\| A \|_2 = \varphi_{\text{и}}(A).$$

В отличие от норм матриц, входящих в формулировку свойства 12, евклидова норма $\|A\|_E$ и норма $M(A)$ не являются индуцированными нормами. Это утверждение следует из замечания к свойству 11.

Свойство 13. Спектральная норма матрицы не меняется при ортогональном разложении матрицы, т. е.

$$\| A \|_2 = \| QAR \|_2. \tag{1.161}$$

Пусть Q и R — ортогональные матрицы соответственно размеров $[n \times n]$ и $[m \times m]$. По определению нормы матрицы, ин-

дуцированной евклидовыми нормами векторов в пространствах R_n и R_m, имеем

$$\varphi_и(A) = \sup_{\|x\|_E=1} \|Ax\|_E \text{ и } \varphi_и(QAR) = \sup_{\|x\|_E=1} \|QARx\|_E.$$

Преобразуем вторую из приведенных формул:

$$\varphi_и(QAR) = \sup_{\|x\|_E=1} (QARx, QARx)^{1/2} \Leftrightarrow$$

$$\Leftrightarrow \sup_{\|x\|_E=1} (x, R^тA^тARx) \Leftrightarrow \sup_{\|R^Tz\|_E=1} (z, A^тAz), \text{ где } z=Rx.$$

В связи с тем что евклидова норма вектора не меняется при ортогональных преобразованиях, $\|R^тz\|_E=\|z\|_E$. Отсюда для $\forall z\in Q_x$ и $\forall x\in Q_x$ выполняется равенство

$$\sup_{\|x\|_E=1} \|Ax\|_E = \sup_{\|z\|_E=1} (z, A^тAz) \Leftrightarrow \varphi_и(A)=\varphi_и(QAR),$$

где $\varphi_и$ — любая из норм, индуцированная евклидовыми нормами в R_n и R_m. И так как спектральная норма матрицы — индуцированная евклидовыми нормами в R_n и R_m, то получаем доказываемое равенство (1.161).

Свойство 14. Евклидова норма матрицы A сохраняется при ее ортогональном разложении и равна корню квадратному из суммы квадратов сингулярных чисел отображения $\mathcal{A}$, действующего из m-мерного пространства X в n-мерное Y.

На основании формулы (1.146) имеем

$$\|Q^тAR\|_E = \operatorname{tr}(Q^тAR)^т(QAR) = \operatorname{tr}(R^тA^тAR),$$
$$\|Q^тAR\|_E = \operatorname{tr}(Q^тAR)(Q^тAR)^т = \operatorname{tr}(Q^тA^тAQ).$$

Так как в приведенных формулах Q и R — произвольные ортогональные матрицы, то, положив $Q=E$, получим

$$\|AR\|_E = \operatorname{tr}(R^тA^тAR) = \operatorname{tr}(A^тA).$$

Сравнивая эти равенства с приведенными выше, окончательно получаем

$$\|Q^тAR\|_E = \|AR\|_E = \|Q^тA\|_E = \|A\|_E. \qquad (1.162)$$

Доказательство второй части данного свойства вытекает из формулы (1.91, *б*), примененной для матрицы $(A^тA)$. Тогда если r — ранг матрицы A, то

$$\operatorname{tr}(A^тA) = \sum_{i=1}^{r} \rho_i^2 \Leftrightarrow \|A\|_E = \sqrt{\sum_{i=1}^{r} \rho_i^2}. \qquad (1.163)$$

Из полученной выше формулы также следует, что

$$\rho_1 \leqslant \|A\|_E \leqslant \rho_1\sqrt{r},$$

где ρ_1 — наибольшее сингулярное число матрицы A. Так как

$$\rho_1 = \| A \|_2 \text{ и } r \leqslant \min(m, n),$$

в результате получим

$$\| A \|_2 \leqslant \| A \|_E \leqslant \sqrt{t}\, \| A \|_2 \text{ при } t = \min(m, n). \qquad (1.164)$$

При проведении теоретических исследований удобнее использовать понятие нормы оператора, чем нормы матрицы. Введение нормы в пространстве линейных операторов осуществляется так же, как и в любом линейном пространстве, а следовательно, и как в пространстве матриц. Поэтому норму оператора можно определять аксиоматически через его матрицу. Для этого, фиксируя базисы в пространствах, в которых действует оператор $\mathscr{A}$, и тем самым задавая матрицу оператора, поставим этой матрице в соответствие число $\|A\|$, определяемое условиями (1.137) и называемое нормой матрицы. Если теперь каждому оператору в заданном пространстве операторов поставить в соответствие норму его матрицы, то таким образом вводится и норма в пространстве операторов. При этом ясно, что условия (1.137) будут выполняться и для операторов. Число $\|\mathscr{A}\| = \|A\|$, удовлетворяющее условиям (1.137) в пространстве операторов, где A — матрица оператора $\mathscr{A}$, будем называть *нормой оператора*. Верно и обратное утверждение, что любая норма оператора порождает при фиксированных базисах норму матрицы. Аксиоматически можно потребовать соблюдение согласованности и индуцированности нормы оператора, для чего должны выполняться условия (1.150) и (1.155) соответственно.

Рассмотрим вопрос о сходимости последовательности матриц.

Пусть элементами этой последовательности являются матрицы, принадлежащие линейному пространству M_{nm}. Если в этом пространстве введена норма, например по формулам (1.137), то пространство M_{nm} будет нормированным. Следовательно, в нем, как и в любом линейном нормированном пространстве, можно определить сходимость последовательности векторов (в данном случае — матриц), принадлежащих пространству M_{nm}. Для этого достаточно воспользоваться определением 1.5.

Определение 1.25. Последовательность матриц $\{A_k\}$, принадлежащая нормированному пространству M_{nm}, сходится к матрице $A_0 \in M_{nm}$, если для каждого $\varepsilon > 0$ найдется такое $k_0(\varepsilon)$, что

$$\| A_k - A_0 \| < \varepsilon \text{ для всех } k \geqslant k_0.$$

Если теперь элементы матрицы $A \in M_{nm}$ представить в виде одного столбца (или строки) размера $[n \times m]$, то M_{nm} можно рассматривать как арифметическое пространство. Матрица, принадлежащая данному пространству, будет его вектором, а эле-

менты матрицы — координатами этих векторов. В результате в образованном арифметическом пространстве матриц можно по аналогии с координатной сходимостью последовательности векторов определить поэлементную сходимость последовательности матриц. При этом, исходя из теоремы 1.2, в нормированном пространстве M_{nm} устанавливается эквивалентность двух введенных определений сходимости матричных последовательностей, т. е. имеет место следующая теорема.

Теорема 1.29. Последовательность матриц $\{A_k\}$ сходится к матрице A_0 поэлементно тогда и только тогда, когда она сходится к A_0 по какой-либо матричной норме.

Из приведенной теоремы вытекает ряд следствий.

Следствие 1. Для того чтобы степенная последовательность матриц A^m, принадлежащая пространству M_{n^2}, сходилась к нулевой матрице того же пространства, достаточно, чтобы мультипликативная норма матрицы A была строго меньше единицы.

Для доказательства используем формулу (1.140)

$$\| A^m \| \leqslant \| A \|^m.$$

Из нее следует, что если $\|A\| \leqslant q$, где $q<1$, то при $m \to \infty$ правая часть этого неравенства стремится к 0. Отсюда имеем $\lim\limits_{m\to\infty} \|A^m\| = O$ и из теоремы 1.29 следует, что $\lim\limits_{m\to\infty} A^m = O$, где O — нулевая матрица размера $[n \times n]$.

Следствие 2. Для сходимости по согласованной норме матриц степенной последовательности матриц A^m, принадлежащей пространству M_{n^2}, к нулевой матрице O необходимо и достаточно, чтобы все собственные значения λ_i матрицы A по модулю были строго меньше единицы.

Необходимость. Пусть $A_k \to 0$ при $k \to \infty$. Тогда из определения 1.23 следует, что $\|A^k\| \to 0$ и на основании предыдущего следствия $\|A\| < 1$. С другой стороны, для согласованной нормы матрицы выполняется формула (1.152), т. е.

$$|\lambda| \leqslant \| A \| < 1.$$

Достаточность. Из выражения (1.55) следует

$$Ax = \lambda x \Leftrightarrow A^2 x = \lambda A x \Leftrightarrow A^2 x = \lambda^2 x.$$

Выполнив подобные преобразования m раз, получим

$$A^m x = \lambda^m x. \tag{1.165}$$

Переходя к согласованным нормам, будем иметь

$$\| A^m x \| = |\lambda|^m \| x \|.$$

И далее из условия, что $|\lambda| < 1$, следует

$$\lim_{m\to\infty} \| A^m x \| = 0,$$

так как $x \neq 0$, то окончательно получаем $A^m \longrightarrow 0$ при $m \longrightarrow \infty$.

Если задана последовательность матриц $\{A_k\}$, принадлежащая пространству M_{n^2}, то матрицу $S_N = \sum_{k=1}^{N} A_k$ будем называть N-й *частичной суммой* матричного ряда $\sum_{k=1}^{\infty} A_k$, а числовую матрицу $S = \lim_{N\to\infty} S_N$ — *суммой* этого ряда.

Теорема 1.30. Для сходимости степенного матричного ряда

$$\sum_{m=0}^{\infty} A^m = E + A + A^2 + \ldots + A^m + \ldots,$$

где матрица $A \in M_{n^2}$, необходимо и достаточно, чтобы $A^m \longrightarrow 0$ при $m \longrightarrow \infty$. При этом будет выполняться равенство

$$\sum_{m=0}^{\infty} A^m = (E - A)^{-1}. \tag{1.166}$$

Доказательство теоремы приведено в [5].

Глава 2

ОШИБКИ ОКРУГЛЕНИЯ ПРИ ВЫЧИСЛЕНИЯХ НА ЭВМ И ИХ ОЦЕНКИ

§ 2.1. АБСОЛЮТНЫЕ И ОТНОСИТЕЛЬНЫЕ ОШИБКИ. ОСНОВНЫЕ ИСТОЧНИКИ ОШИБОК ВЫЧИСЛЕНИЙ

При выполнении различных видов измерений в результате необходимо получить числовую информацию либо о самих измеряемых величинах, либо о величинах, являющихся функциями измеряемых величин.

Для любой величины существует объективная количественная характеристика в виде числа A, которую будем называть *точным значением* рассматриваемой величины.

Так как никакие измерения и их обработка на практике не могут быть выполнены абсолютно точно, то точное значение величины остается неизвестным и его приходится заменять другим числовым значением, отличающимся от точного. Число, принимаемое за значение определяемой величины и отличающееся от ее точного значения A, будем называть *приближенным значением* этой величины и обозначать его a.

Отличие точного значения величины от ее приближенного значения называется *ошибкой приближенного числа,* которая вычисляется по формуле

$$\Delta a = a - A. \tag{2.1}$$

Из этой формулы следует, что при $A > a$ значение $\Delta a < 0$, а при $A \leqslant a$ значение $\Delta a \geqslant 0$. Однако в целом ряде случаев отношение порядка между значениями A и a установить невозможно, поэтому вводится понятие *абсолютной ошибки приближенного числа a.* Вычисление абсолютной ошибки можно выполнить по формуле

$$\Delta_a = |\Delta a| = |a - A|. \tag{2.2}$$

Дать характеристику приближенному значению определяемой величины только по абсолютной ошибке приближенного числа обычно нельзя. Так, если абсолютная ошибка некоторой измеряемой величины получилась равной 1 мм, то без указания объекта измерения и размера измеряемой величины еще не известно, хорошо или плохо проведено измерение. Если эта ошибка получена при измерении расстояний в 100 и 10 м между двумя пунктами на местности, то ясно, что в первом случае измерения выполнены с более высокой точностью, чем во втором.

Поэтому для приближенного значения величины, точное значение которой не равно нулю, вводится оценка, которая получила название *относительной ошибки приближенного числа.* Обозначив относительную ошибку приближенного числа a как ε_a, получим для $A \neq 0$

$$\varepsilon_a = \frac{\Delta_a}{|A|} \Leftrightarrow \varepsilon_a = \left| \frac{a}{A} - 1 \right|. \tag{2.3}$$

Таким образом, если известно точное значение величины, то для приближенного значения величины по формулам (2.2) и (2.3) можно дать абсолютную и относительную оценки точности. Однако, как правило, точное значение определяемой величины не известно. В этом случае вместо неопределенных значений абсолютной и относительной ошибок приближенного числа для характеристики приближенного значения определяемой величины вводят понятия *предельной абсолютной и предельной относительной ошибок приближенного числа.* За предельную абсолютную ошибку приближенного числа определяемой величины будем принимать такое число $\Delta_{\text{пр}}$, которое удовлетворяет неравенству

$$\Delta_{\text{пр}} \geqslant \Delta_a \Leftrightarrow a - \Delta_{\text{пр}} \leqslant A \leqslant a + \Delta_{\text{пр}}, \tag{2.4}$$

Здесь число $(a - \Delta_{\text{пр}})$ является приближением числа A по недостатку, а $(a + \Delta_{\text{пр}})$ — приближением A по избытку. Исходя из формулы (2.4), для числа A принята следующая запись:

$$A = a \pm \Delta_{\text{пр}}, \tag{2.5}$$

отражающая тот факт, что $A \in [(a - \Delta_{пр}), (a + \Delta_{пр})]$. Как правило, значение $\Delta_{пр}$ находится из теоретических предпосылок или из характеристик физических устройств, участвующих в измерениях и обработке определяемой величины. Отметим также, что значение $\Delta_{пр}$ не ограничено сверху, но ограничено снизу значением Δ_a. Отсюда если Δ_1 — предельная ошибка числа a и $\Delta_2 \geqslant \Delta_1$, то и Δ_2 является предельной ошибкой приближенного числа a.

За предельную относительную ошибку приближенного числа будем принимать число $\varepsilon_{пр}$, удовлетворяющее неравенству

$$\varepsilon_{пр} \geqslant \varepsilon_a. \tag{2.6}$$

Из этого неравенства следует, что если ε_1 — предельная относительная ошибка числа a и $\varepsilon_2 \geqslant \varepsilon_1$, то и ε_2 — предельная относительная ошибка приближенного числа a. Отсюда при условии, что $A \cong a$ и $a \neq 0$, получим формулы

$$a(1 - \varepsilon_{пр}) \leqslant A \leqslant a(1 + \varepsilon_{пр}) \text{ и } \Delta_{пр} = \varepsilon_{пр} |a| \tag{2.7}$$

и на основании записи (2.7) число A можно представить в виде

$$A = a(1 \pm \varepsilon_{пр}). \tag{2.8}$$

В процессе решения задач с применением вычислительных средств приходится иметь дело с двумя основными видами ошибок: *ошибками*, содержащимися *в исходной информации*, и *ошибками вычисления*. Каждую из этих ошибок можно представить в абсолютной и относительной формах.

Ошибки в исходной информации возникают в результате неточностей измерений, неадекватности математической модели, описывающей реальный физический процесс, или из-за невозможности представить какую-либо физическую или математическую величину конечной дробью. Ошибки данного вида существуют независимо от того, каким методом и на каком вычислительном устройстве проводятся вычисления.

Ошибки вычислений можно подразделить на ошибки, вызванные выбором численного метода, и ошибки, связанные с представлением чисел в ЭВМ. Второй вид ошибок получил название *ошибок округления*.

Существуют такие численные методы, которые в предположении, что ошибки в исходной информации и округления отсутствуют и вычислительный процесс конечен (т. е. реализуется за конечное число элементарных арифметических операций), не приводят к ошибкам вычислений. Такие численные методы будем называть *прямыми точными методами*. Численные методы, которые даже при отсутствии ошибок округления и в исходной информации приводят к ошибкам вычисления независимо от того, конечен или бесконечен вычислительный процесс, будем называть *приближенными методами*. Отсюда, в частности, сле-

дует, что любой итерационный численный метод является приближенным.

В общем случае ошибки, вызванные выбором численного метода (алгоритма), существенно зависят от ошибок округления. Так, в некоторых алгоритмах небольшая ошибка, допущенная на каком-либо его шаге, может сильно возрастать и в результате получится большая вычислительная ошибка. Такие алгоритмы называются *неустойчивыми*. Естественно, что при выборе численного метода решения какой-либо задачи необходимо избегать неустойчивых алгоритмов.

В качестве примера неустойчивого алгоритма рассмотрим вычисление по рекуррентному соотношению

$$y_n = a y_{n-1} + b,$$

$$\text{где } a > 1,\ y_0 = \bar{y}_0 + \delta_0.$$

Тогда если δ_0 — ошибка округления $\bar{y}_0$, а $\delta_1, \delta_2, \ldots, \delta_n$ — ошибки вычисления, то

$$y_1 = \bar{y}_1 + \delta_1,\ y_2 = \bar{y}_2 + \delta_2,\ \ldots,\ y_n = \bar{y}_n + \delta_n,$$

где ошибка вычисления значения $\bar{y}_n$ будет определяться как $\delta_n = a^n \delta_0$. Таким образом, даже малая ошибка округления δ_0 при пренебрежении другими источниками ошибок с увеличением n приводит к быстрому нарастанию ошибки δ_n. Так, при $a = 1{,}6$ и $n = 5$ ошибка в определении y_n по приведенной формуле возрастет более чем на порядок по сравнению с δ_0. Приведенный пример также показывает, насколько осторожно надо относиться к рекуррентным вычислительным алгоритмам.

§ 2.2. ПРЕДСТАВЛЕНИЕ ЧИСЕЛ И ИХ ОКРУГЛЕНИЕ В ЭВМ

Если предположить, что исходная информация не содержит никаких ошибок и численные методы, используемые при обработке результатов измерений, точные, то все равно и в этом случае присутствуют ошибки вычислений, вызванные ошибками округления. Ошибки округления в основном определяются тем, какая система записи чисел принята в вычислительном устройстве, на котором будет производиться обработка исходной информации.

В настоящее время во всех вычислительных устройствах принята запись чисел, основанная, как и наша десятичная система счисления, на позиционной системе счисления.

Определение 2.1. Пусть заданы целое число $P > 1$ и множество целых чисел $\{\alpha_i\}_{i=0 \div (p-1)} = \{\alpha_0, \alpha_1, \ldots, \alpha_{p-1}\}$. Если любое число A может быть представлено в виде ряда

$$A = \pm (b_n p^n + b_{n-1} p^{n-1} + \ldots + b_0 + b_{-1} p^{-1} + b_{-2} p^{-2} \ldots), \tag{2.9}$$

где каждый из коэффициентов b_i может принимать одно из значений $\{\alpha_i\}_{i=0\div(p-1)}$, то запись вида

$$A = \pm b_n b_{n-1} \ldots b_0 b_{-1} b_{-2} \ldots \qquad (2.10)$$

называется *позиционной системой счисления.*

Число P в (2.9) называется *основанием системы счисления*, числа $\alpha_0, \alpha_1, \ldots, \alpha_{p-1}$ — *базисными числами*, b_i — *коэффициентом* i разряда числа A.

Как правило, рассматриваются такие позиционные системы, у которых базисные числа образуют совокупность $\{i\}_{i=0\div(P-1)} = \{0, 1, 2, \ldots, (P-1)\}$. Любое вещественное число может быть представлено в виде (2.9), если базисные числа образуют совокупность $\{i\}_{i=0\div(P-1)}$. Если $P=10$ и базисными числами являются 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, то число, представленное в виде (2.9) или (2.10), будет называться десятичной дробью. Если же $P=2$ и базисные числа — $\{0, 1\}$, то число, представленное в виде (2.9) или (2.10), будет называться двоичной дробью, а система счисления, в которой представлено это число, — двоичной системой счисления. В тех случаях, когда возникает неопределенность при изображении числа, будем рядом с числом указывать основание системы счисления, в котором оно представлено. Например, $A_1 = (1011,\ 11)_2$ — число, представленное в двоичной системе счисления, $A_2 = (1011,\ 11)_{10}$ — число, представленное в десятичной системе счисления. Хотя оба из приведенных чисел состоят из одних и тех же базисных чисел, но они совершенно различные: так, первое из них в десятичной системе счисления $A_1 = 9{,}75$.

В современных ЭВМ числа, как правило, представляются в двух формах — с фиксированной и плавающей запятой. Пусть под любое число отводится $(t+1)$ разрядов, из которых r разрядов отводится под целую часть числа. Такое представление чисел называется *с фиксированной запятой.* Если число представлено с фиксированной запятой, то оно записывается в виде

$$a = \pm b_{r-1} b_{r-2} \ldots b_1 b_0,\ b_{-1} b_{-2} \ldots b_{r-t}, \qquad (2.11)$$

т. е. под дробную часть числа отводится $(t-r)$ разрядов.

Найдем диапазон представления чисел, заданных с фиксированной запятой. Пусть число изображается в виде (2.9). Тогда максимальное по модулю число будет то, для которого $b_{r-i} = (P-1)$ при всех $i = 1 \div t$. В результате из формулы (2.9) получаем

$$\max |a| = (P-1)(P^{r-1} + P^{2-2} + \ldots + P^0 + P^{-1} + \ldots + P^{r-t}).$$

С учетом того, что второй сомножитель этого выражения является суммой членов геометрической прогрессии, имеем

$$\max |a| = P^r(1 - P^{-t}). \qquad (2.12)$$

Отсюда

$$-P^r \leqslant a < P^r.$$

В качестве примера рассчитаем диапазон представления чисел с фиксированной запятой в вычислительных машинах типа ЕС. Как известно, в ЕС ЭВМ в форме с фиксированной запятой представляются только целые числа, для которых предусмотрены две длины: стандартная — 4 байта и нестандартная — 2 байта. Заметим, что 1 байт равен 8 бит, а 1 бит соответствует одному двоичному разряду информации. Отсюда для представления чисел с фиксированной запятой в ЕС ЭВМ имеем $P=2$, $t=r$, один разряд отводится под знак числа и получаем, что

$$-(P^r) \leqslant a \leqslant (P^r - 1). \tag{2.13}$$

Увеличение диапазона для изображения в ЭВМ отрицательных целых чисел на единицу по сравнению с расчетами, выполненными по формуле (2.12), связано со спецификой представления отрицательных чисел в ЭВМ (отрицательные числа представлены в ЭВМ в дополнительном коде). Таким образом, в ЕС ЭВМ диапазон измерения целых чисел стандартной длины определяется как

$$-2^{31} \leqslant a \leqslant (2^{31}-1) \Rightarrow |a| \leqslant 2147483647,$$

а для нестандартной длины — как

$$-2^{15} \leqslant a \leqslant (2^{15}-1) \Rightarrow -32768 \leqslant a \leqslant 32767.$$

Рассмотрим теперь числа с плавающей запятой. Представление ненулевого числа будем называть *с плавающей запятой*, если оно описывается в виде

$$a = mP^n, \tag{2.14}$$

где m — мантисса числа, n — целое число, предназначенное для задания порядка чисел. Если на мантиссу числа, представленного в виде (2.14), накладывается ограничение

$$1/P \leqslant |m| < 1, \tag{2.15}$$

то такое представление числа называется *нормализованным с плавающей запятой.* Число 0 в указанном представлении — это число с нулевой мантиссой при любом порядке. Если модуль числа a меньше единицы, то порядок числа $n \leqslant 0$, и если $|a| \geqslant 1$, то $n \geqslant 1$.

Вещественные числа в ЭВМ в основном представляются в нормализованном виде с плавающей запятой. Найдем диапазон изменения этих чисел. Предположим, что под порядок числа отводится $r+1$ разрядов в системе счисления q, а число в ЭВМ представляется в P системе счисления. Тогда максималь-

ные и минимальные по модулю числа, которые могут быть представлены в нормализованном виде, будут иметь вид

$$\max |a| = 1P^{q^{(r-1)}}, \min |a| = (1/P)P^{-(q^r)}.$$

Здесь $(q^r - 1)$ — максимальный положительный порядок и $(-q^r)$ — минимальный отрицательный порядок, определяемые на основании формулы (2.13). Таким образом, получаем

$$P^{-(q^r+1)} \leqslant |a| \leqslant P^{(q^r-1)}. \qquad (2.16)$$

Применительно к ЕС ЭВМ имеем $P=16$, т. е. мантисса числа в ЭВМ задается в 16-й системе счисления, $q=2$, $r=6$. Отсюда в ЕС ЭВМ диапазон представления вещественных чисел в нормализованном виде определяется из выражения

$$10^{-79} \leqslant |a| \leqslant 10^{75}.$$

Отметим, что диапазон представления вещественных чисел в ЭВМ не зависит от длины задания этих чисел в ЕС ЭВМ. Напомним, что в ЕС ЭВМ вещественные числа могут иметь стандартную длину — 4 байта и нестандартную — 8 байтов.

Как отмечалось выше, в реальных вычислительных устройствах числа задаются конечной дробью вида (2.11) вместо (2.10). Операцию, в результате которой осуществляется замена исходного числа таким числом, все разряды которого в заданной системе счисления, начиная с $(s-1)$-го, являются нулевыми, будем называть *округлением* числа до s разрядов в принятой системе счисления. Разность между округленным и округляемым числами называется *ошибкой округления.*

Операции округления могут быть реализованы различными способами. Однако наибольшее применение в вычислительной практике нашли два способа округления:

1) с усечением числа до s разрядов;
2) симметричное.

Усечением числа A до s разрядов называется его округление до s разрядов, при котором сохраняются все разряды числа слева до s-го включительно. В технической литературе этот способ округления называют *отбрасыванием младших разрядов.* Сначала оценим точность округления способом усечения нормализованных чисел. Для этого нормализованное число запишем как

$$a = m_1 P^n + m_2 P^{n-t}, \qquad (2.17)$$

где m_1 — мантисса нормализованного числа, на которую отведено t разрядов в P системе счисления, т. е. $1/P \leqslant m_1 < 1$, для m_2 отводятся остальные разряды, т. е. $0 \leqslant m_2 < 1$. Например, если $a = 0{,}23140092 \cdot 10^3$ и $t=5$, то $0{,}23140092 \cdot 10^3 = 0{,}23140 \cdot 10^3 + 0{,}09200 \cdot 10^{-2}$. Положив $s=t$, получаем, что максимальная абсолютная ошибка округления возникает при $m_2 = 1$ и

$$\overline{\Delta}_{\text{окр}} = |a - A| = 1P^{n-t}. \tag{2.18}$$

Отсюда наибольшая относительная ошибка округления при $m_1 = 1/P$ будет

$$\overline{\varepsilon}_{\text{окр}} = \frac{1P^{n-t}}{(1/P)\,P^n} = P^{-t+1}. \tag{2.19}$$

Из приведенной формулы следует, что предельная относительная ошибка округления нормализованных чисел по способу усечения не зависит от числа, а зависит от количества разрядов, отводимых под мантиссу числа.

Определим ошибку округления чисел, представленных с фиксированной запятой, при условии, что под число отводится t разрядов в P системе счисления, а под целую часть числа — r разрядов. Из формул (2.9) и (2.11) следует для $A>0$:

$$A - a = b_{(r-t)-1}P^{(r-t)-1} + b_{(r-t)-2}P^{(r-t)-2} + \ldots$$

$$\ldots \leqslant (P-1)(P^{s-1}\left(1 + \frac{1}{P} + \frac{1}{P^2} + \ldots\right),$$

где $S = r - t => S \leqslant 0$. После алгебраических преобразований получаем

$$a - A \leqslant -P^s. \tag{2.20}$$

Если $A<0$, то эта формула преобразуется к виду

$$a - A \geqslant P^s. \tag{2.21}$$

В общем случае можно написать

$$|a - A| \leqslant P^s. \tag{2.22}$$

Анализ выражения (2.22) показывает, что предельная абсолютная ошибка округления по способу усечения для чисел, представленных с фиксированной запятой, не зависит от значения самого числа, а зависит только от разности количества разрядов, отводимых под число и его целую часть. Кроме того, из выражений (2.20) и (2.21) следует, что знак ошибки округления приближенного числа (независимо от того, представлено оно с фиксированной или с плавающей запятой) противоположен знаку исходного (точного) числа, что может привести к такому нежелательному явлению, как быстрое накопление ошибок вычислений.

Рассмотрим теперь симметричный способ округления, который можно описать следующим образом:

$$|\hat{a}| = \begin{cases} |a_s|, & \text{если } |a_s - A| < (1/2)P^s, \\ |a_s| + P^s, & \text{если } |a_s - A| \geqslant (1/2)P^s, \end{cases} \tag{2.23}$$

где a_s — приближенное число при округлении A по способу усечения. Предполагая, что распределение ошибок округления

подчиняется равномерному закону распределения, получаем для предельной абсолютной ошибки симметричного округления выражение

$$\hat{\Delta}_{\text{окр}} = |\hat{a} - A| \leqslant \frac{1}{2} P^s. \qquad (2.24)$$

Напомним, что $s = r - t => s \leqslant 0$. Формула (2.24) служит для оценки точности симметричного округления чисел, представленных с фиксированной запятой.

Оценивать точность симметричного округления чисел, представленных в нормализованном виде, будем предельной относительной ошибкой округления, имеющей на основании формул (2.19) и (2.24) вид

$$\hat{\varepsilon}_{\text{окр}} \leqslant (1/2) P^{-t+1}. \qquad (2.25)$$

Из сравнения способов симметричного округления и с отбрасыванием младших разрядов (усечения) следует, что способ симметричного округления приводит во всех случаях к меньшим ошибкам вычисления, чем способ с отбрасыванием младших разрядов. Однако большинство трансляторов алгоритмических языков (например, ФОРТРАН, БЭЙСИК, ПЛ/1), существующих на современных ЭВМ, организовано таким образом, что рабочие программы, составленные с их помощью, производят округление чисел по способу отбрасывания младших разрядов. Это вызвано тем, что техническая реализация способа усечения предельно простая, не требующая никаких дополнительных программных ресурсов, а следовательно, и затрат машинного времени. В случае необходимости реализации способа симметричного округления требуется составление отдельной программы по алгоритму, в основе которого должны лежать зависимости (2.23).

В качестве примера рассмотрим округление чисел по способу усечения в ЕС ЭВМ. Так как в ЕС ЭВМ числами с фиксированной запятой являются только целые числа, то независимо от их длины

$$\Delta_{\text{окр}} \leqslant (2)^0 \Rightarrow \Delta_{\text{окр}} \leqslant 1.$$

Для тех случаев, когда округляются вещественные числа стандартной длины $t = 3$ байта (или $t = 6$ при $P = 16$),

$$\varepsilon_{\text{окр}} \leqslant 16^{-6+1} \Rightarrow \varepsilon_{\text{окр}} \leqslant 16^{-5} \Rightarrow \bar{\varepsilon}_{\text{окр}} \cong 10^{-6},$$

а при округлении вещественных чисел нестандартной длины $t = 7$ байт (или $t = 14$ при $p = 16$)

$$\varepsilon_{\text{окр}} \leqslant 16^{-13} \Rightarrow \bar{\varepsilon}_{\text{окр}} \cong 10^{-16}. \qquad (2.26)$$

При организации ввода и вывода числовых данных для ЭВМ в ряде случаев требуется задание форматов данных и, в частно-

сти, должно быть указано количество позиций (десятичных разрядов), отводимых под запись каждого вводимого (или выводимого) числа. Как правило, необходимое количество таких позиций намного меньше максимально возможного количества позиций, предусматриваемых транслятором алгоритмического языка под запись чисел. Задание формата числа с учетом только необходимого количества позиций не приводит к таким дополнительным вычислительным затратам, как набивка ненужных символов в исходной информации, печать лишних символов при выдаче на печатающее устройство ЭВМ результатов счета и др.

На основании указанного возникает задача определения необходимого количества десятичных цифр в записи числа (2.11) при его вводе (или выводе) в ЭВМ, если подготовка и печать числовой информации производятся в обычной десятичной системе счисления.

В записи (2.11) все базисные числа, начиная с первого слева ненулевого, будем называть *значащими цифрами.* Количество значащих цифр можно увеличить или уменьшить путем приписывания или отбрасывания нулей в младших разрядах числа. Например, число 0,00300800 имеет 6 значащих цифр, а число 0,003008—4 значащие цифры и, следовательно, эти два числа не равнозначны.

Для значащих цифр приближенного числа, ошибка приближения которого известна, вводится понятие верных значащих цифр. Значащую цифру b_s в числе (2.11) будем называть *верной значащей цифрой,* если для абсолютной ошибки приближенного числа a выполняется неравенство

$$\Delta_a \leqslant \Omega P^s, \tag{2.27}$$

где параметр Ω устанавливается равным 1 или $^1/_2$ в зависимости от способа округления чисел. Отметим, что в данном определении неважен источник искажения точного числа A. Так как округление чисел на ЭВМ производится по способу усечения, то в формуле (2.27) будем считать $\Omega = 1$. Если базисное число b_s при записи приближенного числа в виде (2.11) — верная значащая цифра, то очевидно, что и все предшествующие значащие цифры верные.

Обозначим длину последовательности значащих цифр в разложении (2.9) за t и пусть $P = 10$. Тогда при выполнении условия

$$\Delta_a \leqslant 10^{n-t+1} \tag{2.28}$$

десятичные цифры b_n, b_{n-1}, ..., b_{n-t+1} являются верными. Например, приближенное значение $a = 3{,}14$ числа π имеет все три верные значащие цифры, так как для него выполняется условие $|\pi - a| \leqslant 10^{(0-3+1)}$. При записи же приближенного значения чис-

ла π в виде $a=3{,}140$ последняя значащая цифра не является верной, так как $|\pi-a| \geqslant 10^{(0-4+1)}$.

Теорема 2.1. Если ε — относительная ошибка приближенного числа a и t — целочисленное решение неравенства

$$\varepsilon \leqslant \frac{1}{b_n+1}\ 10^{-t+1}, \tag{2.29}$$

то приближенное число a, начиная с первой значащей цифры b_n, имеет по крайней мере t верных значащих цифр.

С доказательством этой теоремы можно познакомиться в работе [4]. Приведенная теорема позволяет рассчитать необходимое для удержания количество значащих цифр в приближенном числе и тем самым дает возможность обеспечить рациональную запись исходной числовой информации для ввода в ЭВМ и печать числовых результатов, получаемых в процессе счета на ЭВМ.

Как следует из формулы (2.26), в ЕС ЭВМ возможно записать без потери точности вещественные числа, представленные в обычной десятичной системе счисления с 16 верными значащими числами. Естественно, что такую точность при вводе (или выводе) чисел в ЭВМ обеспечивать не надо, если само исходное число является приближенным, полученным в результате, например, каких-то измерений с меньшей точностью. Так, если при вычислениях (или измерениях) расстояния между двумя точками получено число с семью значащими числами $L=981{,}1576$ и известно, что относительная ошибка вычисления (измерения) не превосходит $\bar{\varepsilon}=10^{-5}$, то на основании формулы (2.29) находим необходимое количество значащих цифр, которое требуется удержать в числе L:

$$10^{-5}=\frac{1}{10}\ 10^{-t+1} \Rightarrow t=5.$$

Таким образом, полученный результат может быть представлен как $L=981{,}15$.

§ 2.3. ОШИБКИ ОКРУГЛЕНИЯ ПРИ ВЫПОЛНЕНИИ АРИФМЕТИЧЕСКИХ ОПЕРАЦИЙ И ИХ РАСПРОСТРАНЕНИЕ

Одним из важнейших вопросов в численном анализе является вопрос о том, как ошибки, возникающие в определенном месте в ходе вычислений, распространяются дальше, т. е. становится ли их влияние больше или меньше по мере того, как производятся последующие операции. В качестве первого шага при рассмотрении этого вопроса необходимо оценить ошибки, возникающие при выполнении элементарных арифметических операций: «+», «—», «*», «/».

Предположим, что числа A, B, C точные, причем C — результат выполнения элементарной арифметической операции над числами A и B. Тогда при выполнении над A и B тех же операций в ЭВМ получим:

1) если $|C|<m$, где m — минимальное по модулю число, которое может быть записано в ЭВМ, то $A+B=0$, т. е. результатом является *машинный нуль*;

2) если $|C|>M$, где M — максимальное по модулю число, которое может быть записано в ЭВМ, то результат операции не определен. В этом случае будем говорить, что имеет место *переполнение;*

3) если $m\leqslant|C|\leqslant M$, то результатом в ЭВМ будет число $\hat{c}$, являющееся приближением для C, за счет округления результата. В этом случае ошибка выполняемой арифметической операции на ЭВМ будет равна ошибке округления числа $\hat{c}$. Следовательно, если результат операции представлен в ЭВМ с фиксированной запятой, то ошибка этой операции из-за округления чисел будет определяться по формуле (2.18), а для результата, представляемого в ЭВМ с плавающей запятой,— по формуле (2.19). Таким образом, при получении результата операции в виде числа с фиксированной запятой будем иметь

$$\hat{c}=C\pm\Delta_{\text{окр}}, \tag{2.30}$$

где $\Delta_{\text{окр}}\leqslant P^s$.

В случае, когда A, B, C — целые числа, операции «+», «—», «*» выполняются точно, а при выполнении операции деления имеем $\Delta_{\text{окр}}\leqslant 1$. Именно этот случай имеет место при выполнении арифметических операций над числами с фиксированной запятой в ЕС ЭВМ.

При выполнении арифметических операций над числами с плавающей запятой имеем

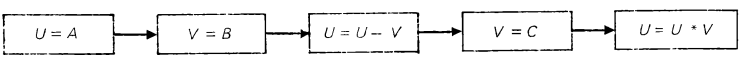
$$\hat{c}=C(1\pm\bar{\varepsilon}_{\text{окр}}), \tag{2.31}$$

где $\bar{\varepsilon}_{\text{окр}}=P^{-t+1}$.

Как было указано выше, значение t равно количеству разрядов в P системе счисления, отводимых под мантиссу числа. Трансляторы современных ЭВМ допускают работу с числовой информацией при различных значениях t. Так, на ЕС ЭВМ при задании вещественных чисел можно использовать стандартную длину ($t=3$ байта) и нестандартную длину ($t=7$ байт). Естественно, что ошибка округления приближенного числа, представленного в нестандартной длине, будет существенно меньше, чем ошибка округления для чисел, представленных в стандартной длине. Однако при этом для записи вещественных чисел нестандартной длины требуется в 2 раза больше ячеек памяти,

чем для вещественных чисел, представленных в стандартной длине. Следовательно, все используемые в программе вещественные числа представлять в нестандартной длине нерационально. Поэтому для сохранения t верных значащих цифр при выполнении вычислений над исходными числами с t верными значащими цифрами при программировании поступают следующим образом. Вводятся дополнительные переменные, под мантиссу которых отводится число разрядов, необходимое для получения результата с t верными значащими цифрами. В ЕС ЭВМ дополнительные переменные должны быть описаны как вещественные переменные нестандартной длины (8 байт). Режим вычислений, при котором все промежуточные вычисления осуществляются с $(2t+1)$ разрядами, отводимыми под мантиссу числа, и лишь конечные результаты вычислений округляются до t верных значащих цифр, получил название *режима накопления.*

Рассмотрим следующий пример. Пусть требуется вычислить значение $K=(A—B)C$ с четырьмя верными значащими цифрами при условии, что A, B, C заданы также с четырьмя верными значащими цифрами $A=0{,}1011\cdot 10^2$, $B=0{,}9999\cdot 10^1$, $C=0{,}2134\times$ $\times 10^2$. При вычислениях, производимых без накопления, т. е. для нашего примера с четырьмя значащими цифрами, получим результат $K_1=0{,}2560\cdot 10^{-3}$. Если же ввести дополнительные переменные U и V удвоенной точности, представляемые с восьмью верными значащими цифрами, то алгоритм вычисления K можно записать в виде



Знак «=» является символом присвоения, служащим для присвоения результата арифметической операции, стоящей справа от него, переменной, стоящей слева от этого символа. Результатом вычисления K по приведенному алгоритму будет значение $K_2=0{,}2368\cdot 10^{-3}$. Все значащие цифры в этом случае являются верными, тогда как в первом случае результат K_1 получен только с одной верной значащей цифрой.

Нужно отметить, что приближенные арифметические операции обладают совсем иными свойствами, чем точные операции. Например, они не ассоциативны, для них не выполняется закон дистрибутивности. Так, в предыдущем примере при вычислениях с четырьмя значащими цифрами $(A—B)C=0{,}2560\cdot 10^{-3}$, а $AC—BC=0{,}2400$, т. е. получили $(A—B)C\neq AC—BC$. Из-за нарушения законов ассоциативности и дистрибутивности для операций сложения и умножения, выполняемых над вещественными числами в ЭВМ, расстановка скобок в арифметическом выражении может существенно влиять на результат его вычис-

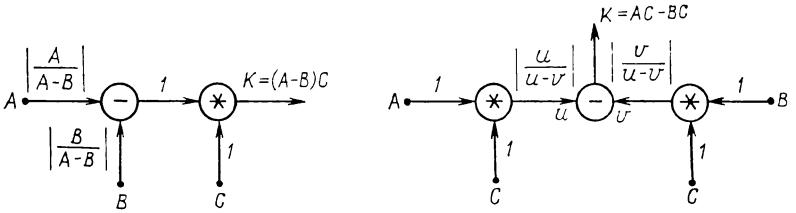


Рис. 2. Схема вычислений выражений $K=(A—B)C$ и $K=AC—BC$

ления. С другой стороны, именно расстановка скобок определяет алгоритм вычисления арифметического выражения. Для удобства анализа алгоритмов вычисления арифметических выражений будем использовать графический способ их описания. При этом любую бинарную операцию будем изображать в виде кружка, внутри которого заключен символ операции (например, для арифметических операций такими символами будут «+», «—», «*», «/»), двух стрелок, входящих в этот кружок, у которых началом, считая слева направо, являются соответствующие операнды данной операции, и одной стрелки, выходящей из кружка, для результата операции. Кроме бинарных операций в состав арифметических выражений могут входить и различные функции. Для их изображения будем использовать кружочек с вписанным внутри символом этой операции, стрелки, входящие в указанный кружок, началом которых являются аргументы этой функции, и стрелку, выходящую из кружка и соответствующую результату вычисления заданной функции. На рис. 2 изображены алгоритмы вычисления арифметических выражений $K=(A—B)C$ и $K=AC—BC$, где $u=AC$, $v=BC$.

Графический способ описания алгоритма вычисления арифметического выражения весьма удобен для подсчета общей ошибки окончательного результата. С этой целью около каждой стрелки ставится коэффициент, равный коэффициенту, стоящему при предельной относительной ошибке операнда, являющегося началом соответствующей стрелки, при определении предельной относительной ошибки рассматриваемой операции (или в общем случае функции). Таким образом, предельная относительная ошибка результата любой операции или функции (кружка) входит в результат следующей операции или функции, умножаясь на коэффициент, стоящий у стрелки, соединяющей эти два кружка.

Следовательно, любое арифметическое выражение, состоящее более чем из двух арифметических операций, связано с вычислениями над приближенными числами. Поэтому возникает задача определения ошибки функции, если известны ошибки ее аргументов (исходных данных). Причем из соображений, изложенных в начале этого параграфа, будем находить предельные

абсолютные и относительные ошибки функций. Такой анализ ошибок относится к *прямому анализу ошибок*, обобщенное определение которого дается в § 2.5.

Пусть задана функция $Y=F(X_1, X_2, \ldots, X_n)$, точные значения аргументов которой не известны, а известны лишь их приближенные значения $x_1, x_2, \ldots, x_n$ и предельные абсолютные и относительные ошибки этих значений. Требуется найти приближенное значение функции y и дать оценку точности полученного приближенного значения. Наложим следующие ограничения на функцию F. Предположим, что, во-первых, она непрерывно дифференцируема в рассматриваемой области изменения приближенных значений x_i, во-вторых, абсолютные ошибки значений x_i существенно меньше самих приближенных значений, так что всегда можно ограничиться членами, линейными относительно ошибок аргументов, пренебрегая членами более высокого порядка малости. Исходя из сделанных предположений, после разложения функции F в ряд Тейлора в окрестности точки $(x_1, x_2, \ldots, x_n)$ имеем

$$F(X_1, X_2, \ldots, X_n)=F(x_1, x_2, \ldots, x_n)+ \\ +\sum_{i=1}^{n}\frac{\partial}{\partial X_i}\ F(\overline{\theta}_i)\,\Delta x_i. \qquad (2.32)$$

Здесь $\overline{\theta}$ — точка, принадлежащая отрезку $[\overline{X}, \overline{x}]$ и $\Delta x_i=X_i-x_i$. Предполагая, что, в-третьих, замена $\frac{\partial}{\partial X_i}F(\theta)\Delta x_i$ на $\frac{\partial}{\partial X_i}\ F(\bar{x})\Delta x_i$ приводит к ошибкам только высокого порядка малости (т. е. выше первого), полученную формулу (2.32) можно преобразовать к виду

$$\Delta_y=|\,y-Y\,|=|\sum_{i=1}^{n}F_{x_i}'(\bar{x})\Delta_{x_i}|.$$

Отсюда, переходя к предельным абсолютным ошибкам, получим

$$\Delta_y\leqslant\sum_{i=1}^{n}|\,F_{X_i}'(\bar{x})\,|\,\overline{\Delta}_{x_i}, \qquad (2.33)$$

где $\overline{\Delta}_{x_i}$ обозначают предельные абсолютные ошибки x_i. Если отбросить третье предположение, накладываемое на функцию F, то для предельной абсолютной ошибки приближенного значения y получим формулу

$$\overline{\Delta}_y=\sum_{i=1}^{n}M_i\overline{\Delta}_{x_i}, \qquad (2.34)$$

где M_i — наибольшее абсолютное значение производной F'_{x_i} на отрезке $[\overline{X}, \overline{x}]$, которое в силу первого предположения, при-

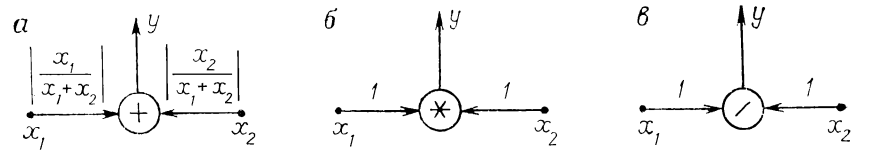


Рис. 3. Схема выполнения арифметических операций:
а — сложение; *б*— умножение; *в* — деление

менительно к функции F, всегда достигается на $[\overline{X}, \overline{x}]$ для любого значения $i=1 \div n$.

При получении предельной относительной ошибки приближенного значения y через предельную абсолютную ошибку воспользуемся формулами (2.3) и (2.34), тогда

$$\varepsilon_y \leqslant \overline{\Delta}_y / \| F(\overline{x}) \|. \tag{2.35}$$

Применим теперь формулы, полученные для оценки точности вычисленного приближенного значения функции, для оценки точности выполнения элементарных арифметических операций, операндами которых являются приближенные значения величин. Пусть x_1 и x_2 — приближенные значения к величинам X_1 и X_2, имеющие предельные абсолютные ошибки $\overline{\Delta}_{x_1}$ и $\overline{\Delta}_{x_2}$, тогда вместо точного сложения $Y=X_1+X_2$ получим приближенное значение y, равное сумме чисел x_1 и x_2. При этом предельная абсолютная ошибка y будет определяться как

$$\overline{\Delta}_y = \overline{\Delta}_{x1} + \overline{\Delta}_{x2} + \overline{\Delta}_{\text{окр}}. \tag{2.36}$$

После перехода к предельным относительным ошибкам получим

$$\overline{\varepsilon}_y = \left|\frac{x_1}{x_1+x_2}\right| \overline{\varepsilon}_{x_1}' + \left|\frac{x_2}{x_1+x_2}\right| \overline{\varepsilon}_{x_2} + \overline{\varepsilon}_{\text{окр}}. \tag{2.37}$$

В этой формуле $\overline{\varepsilon}_{\text{окр}}$ для сложения нормализованных действительных чисел, представленных с плавающей запятой, будет находиться по формуле (2.19). На основании формулы (2.37) графическое изображение операции сложения примет вид, показанный на рис. 3, *а*. Для вычитания чисел x_1 и x_2 получим те же оценки, что и для сложения двух чисел, только x_2 будем брать с противоположным знаком. Из анализа формулы (2.37) следует, что если x_1 и x_2 имеют противоположные знаки и значения $|x_1|$, $|x_2|$ близки, то относительная ошибка ε_y может принять большое значение, а это может привести к значительной ошибке последующих вычислений, выполняемых со значениями y.

При выполнении операции умножения над приближенными числами x_1 и x_2 для оценки точности произведения y получим

$$\Delta_y \leqslant |x_2| \overline{\Delta}_{x1} + |x_1| \overline{\Delta}_{x2} + \overline{\Delta}_{\text{окр}}.$$

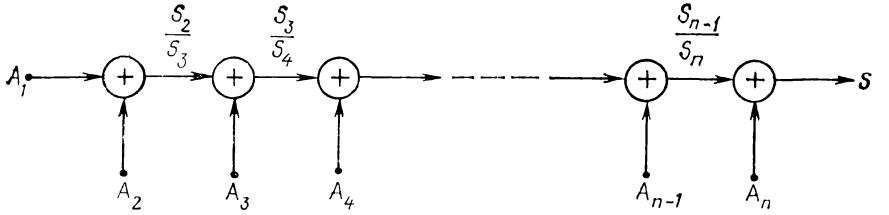


Рис. 4. Схема выполнения последовательного суммирования

Отсюда для предельной относительной ошибки произведения имеем

$$\bar{\varepsilon}_y = \bar{\varepsilon}_{x_1} + \bar{\varepsilon}_{x_2} + \bar{\varepsilon}_{\text{окр}}. \tag{2.38}$$

Графическое изображение операции умножения дано на рис. 3, *б*. Для оценки точности результата деления числа x_1 на x_2 имеем

$$\Delta_y \leqslant \frac{1}{|x_2|}\bar{\Delta}_{x_1} + \left|\frac{x_1}{x_2^2}\right|\bar{\Delta}_{x_2} + \bar{\Delta}_{\text{окр}} \tag{2.39}$$

и далее для предельной относительной ошибки y получаем

$$\bar{\varepsilon}_y = \bar{\varepsilon}_{x_1} + \bar{\varepsilon}_{x_2} + \bar{\varepsilon}_{\text{окр}}. \tag{2.40}$$

Графическое изображение операции деления приведено на рис. 3, *в*.

§ 2.4. АЛГОРИТМЫ СУММИРОВАНИЯ И ПЕРЕМНОЖЕНИЯ ПОСЛЕДОВАТЕЛЬНОСТИ ЧИСЕЛ

Оценим влияние ошибок округления на точность результата вычисления на примере суммирования конечной последовательности точных величин. Пусть задана последовательность точных величин $\{A_1, A_2, \ldots, A_n\}$, в которой $A_i \geqslant 0$ для всех $i = 1 \div n$. Требуется вычислить

$$S = \sum_{i=1}^{n} A_i.$$

В качестве алгоритма для вычисления S выбираем алгоритм последовательного суммирования

$$s_k = s_{k-1} + A_k,$$

где $k = 2 \div n$, $s_1 = A_1$ и $S = s_n$.

Графическое изображение этого алгоритма приведено на рис. 4.

Из приведенного алгоритма следует

$$\bar{\varepsilon}^0_{s_k} = \frac{s_{k-1}}{s_k}\bar{\varepsilon}_{s_{k-1}} + \bar{\varepsilon}_{\text{окр}}, \tag{2.41}$$

где $k=2 \div n$, $\varepsilon_{s_1}=\varepsilon_{A_1}=0$,

причем $\varepsilon_{A_i}=0$, так как A_i — точные значения. Распишем формулу (2.41) для $k=2, 3, 4$, выражая значение ε_{s_4} через ε_{s_3}, а ε_{s_3} — через ε_{s_2}, т. е.

$$\bar{\varepsilon}^0_{s_2}=\bar{\varepsilon}_{\text{окр}}, \quad \bar{\varepsilon}^0_{s_3}=\frac{s_2}{s_3}\bar{\varepsilon}_{s_2}+\varepsilon_{\text{окр}} \Rightarrow \bar{\varepsilon}^0_{s_3}=\left(\frac{s_2}{s_3}+1\right)\bar{\varepsilon}_{\text{окр}},$$

$$\bar{\varepsilon}^0_{s_4}=\frac{s_3}{s_4}\left(\frac{s_2}{s_3}+1\right)\bar{\varepsilon}_{\text{окр}}+\bar{\varepsilon}_{\text{окр}} \Rightarrow \bar{\varepsilon}^0_{s_4}=\left(\frac{s_2}{s_4}+\frac{s_3}{s_4}+1\right)\bar{\varepsilon}_{\text{окр}}.$$

Используя метод индукции, получаем формулу предельной относительной ошибки округления при суммировании n точных значений величин по алгоритму последовательного суммирования

$$\bar{\varepsilon}^0_{s_n}=\left(\frac{s_2}{s_n}+\frac{s_3}{s_n}+\ldots+\frac{s_{n-1}}{s_n}+1\right)\bar{\varepsilon}_{\text{окр}} \Rightarrow$$

$$\Rightarrow \bar{\varepsilon}^0_{s_n}=\frac{\sum\limits_{k=1}^{n-1} s_{k+1}}{s_n}\bar{\varepsilon}_{\text{окр}}. \qquad (2.42)$$

Проанализируем эту формулу. Нетрудно заметить, что, во-первых, с увеличением n относительная ошибка ε_{s_n} увеличивается, во-вторых, величина ошибки ε_{s_n} существенным образом зависит от расположения чисел в последовательности $\{A_i\}_{i=1 \div n}$. Причем если числа в $\{A_i\}_{i=1 \div n}$ располагаются в порядке возрастания, т. е. $A_1 \leqslant A_2 \leqslant \ldots \leqslant A_{n-1} \leqslant A_n$, то относительная ошибка округления при суммировании этих чисел по алгоритму последовательного суммирования будет наименьшей. Для предельной относительной ошибки $\bar{\varepsilon}^0_{s_n}$ можно получить и более простое выражение, если при вычислении всех частичных сумм s_k положить $A_i=\bar{A}$, где $\bar{A}$ — наибольшее значение в последовательности $\{A_i\}_{i=1 \div n}$. Тогда из формулы (2.42) получаем

$$\bar{\varepsilon}^0_{s_n} \leqslant \frac{(2+3+\ldots+(n-1)+n)}{n}\bar{\varepsilon}_{\text{окр}} \Rightarrow \bar{\varepsilon}^0_{s_n} \leqslant \frac{(n-1)(n+2)}{2n}\bar{\varepsilon}_{\text{окр}}$$

и после несложных преобразований имеем

$$\bar{\varepsilon}^0_{s_n}=\frac{n+2}{2}\bar{\varepsilon}_{\text{окр}}. \qquad (2.43)$$

Если значения A_i, входящие в $\{A_i\}_{i=1 \div n}$, изменяются незначительно, то для оценки точности их суммы S можно также использовать формулу (2.42).

В случае, когда последовательность $\{A_k\}^*_{k=1\div n}$ — убывающая арифметическая прогрессия, и при большом значении n получим

$$\bar{\varepsilon}^0_{s_n} = \frac{n+2}{3}\,\bar{\varepsilon}_{\text{окр}}. \tag{2.44}$$

Таким образом, если члены последовательности $\{A_i\}_{i=1\div n}$ изменяются в незначительных пределах или по степенному закону с $P \leqslant 1$, то упорядоченность членов A_i в порядке их возрастания не может привести к заметному повышению точности вычисления суммы $s_n = \sum_{i=1}^{n} A_i$.

Рассмотрим случай, когда члены в последовательности $\{A_k\}^*_{k=1\div n}$ изменяются по геометрической прогрессии со знаменателем $q>1$. Тогда на основании формулы (2.42) для $n>10$ имеем

$$\bar{\varepsilon}^0_{s_n} = \left(\frac{n}{q^n} + \frac{q}{q-1}\right)\bar{\varepsilon}_{\text{окр}} \Rightarrow \bar{\varepsilon}^0_{s_n} = \frac{q}{q-1}\,\bar{\varepsilon}_{\text{окр}}. \tag{2.45}$$

Из анализа полученной формулы следует, что с возрастанием q точность суммирования членов геометрической прогрессии возрастает и при $q>10$ она практически будет определяться только точностью выполнения одной операции сложения. Аналогичные выводы следуют и для суммирования членов упорядоченной последовательности $\{A_k\}^*_{k=1\div n}$, изменение которых происходит по степенному закону с $P \geqslant 2$ и $A_1>1$. Таким образом, в тех случаях, когда изменение членов последовательности $\{A_i\}_{i=1\div n}$ после их упорядочения нелинейно возрастает, переход от $\{A_i\}_{i=1\div n}$ к последовательности $\{A_k\}^*_{k=1\div n}$ оправдан, так как он может привести к существенному уменьшению ошибки суммирования $\bar{\varepsilon}^0_{s_n}$.

Рассмотрим алгоритм суммирования членов последовательности $\{A_i\}_{i=1\div n}$ при $n=2^m$, графическое изображение которого для $n=8$ приведено на рис. 5.

Обозначая предельную относительную ошибку округления от суммирования двух чисел как $\bar{\varepsilon}_{\text{окр}}$, получаем для предельной относительной ошибки суммирования четырех чисел

$$\bar{\varepsilon}^0_{s_{14}} = \bar{\varepsilon}^0_{s_{58}} = \frac{s_{12}}{s_{14}}\,\bar{\varepsilon}_{\text{окр}} + \frac{s_{34}}{s_{14}}\,\bar{\varepsilon}_{\text{окр}} + \bar{\varepsilon}_{\text{окр}} \Rightarrow \varepsilon^0_{s_{14}} = 2\bar{\varepsilon}_{\text{окр}}.$$

Из аналогичных рассуждений при суммировании восьми чисел имеем

$$\varepsilon^0_{s_{18}} = \frac{s_{14}}{s_{18}}\,(2\bar{\varepsilon}_{\text{окр}}) + \frac{s_{58}}{s_{18}}\,(2\bar{\varepsilon}_{\text{окр}}) + \bar{\varepsilon}_{\text{окр}} = 3\bar{\varepsilon}_{\text{окр}}.$$

По методу индукции находим ε_{s_n} при суммировании $n=2^m$ чисел

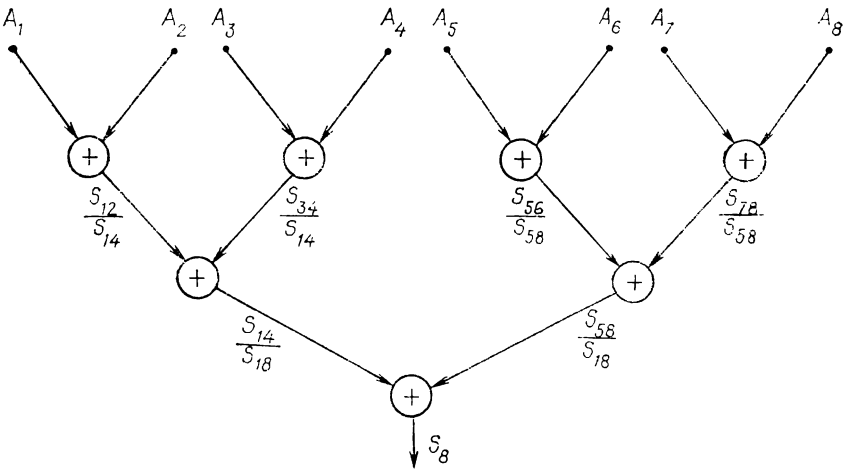


Рис. 5. Схема выполнения суммирования по парам чисел

$$\varepsilon^0 s_n = \bar{\varepsilon}^0 s_m = m\bar{\varepsilon}_{\text{окр}} \Rightarrow \bar{\varepsilon}^0 s_n = \bar{\varepsilon}_{\text{окр}} \log_2 n. \qquad (2.46)$$

Сравнивая полученную оценку с оценкой, определяемой по формуле (2.43), убеждаемся, что при суммировании близких по значению величин с использованием алгоритма суммирования *по парам чисел* повышается точность определения суммы последовательности чисел по сравнению с алгоритмом последовательного суммирования примерно в $n/(2\log_2 n)$ раз. Так, для $n=512$ это уменьшение может достигнуть 30 раз, что приведет к сохранению двух значащих цифр.

Рассмотрим задачу суммирования последовательности приближенных чисел. Ранее указывалось, что источниками приближенных чисел могут быть измерения величин и результаты предыдущих вычислений. Например, при вычислении суммы тригонометрического ряда $S=\sum_{k=1}^{n} a_k$, где $a_k=\sin kx$, значения a_k являются приближенными числами, даже если x — точное число.

Пусть a_k — приближенное число, предельная относительная ошибка которого равна $\bar{\varepsilon}_{a_k}$. Найдем сумму $s_n=\sum_{k=1}^{n} a_k$ и оценим ошибку вычисления этой суммы. Если предположить, что ошибки округления при выполнении операции суммирования отсутствуют, то предельная относительная ошибка вычисления s_n будет определяться по формуле

$$\bar{\varepsilon}^n_{s_n} = \sum_{k=1}^{n} a_k \bar{\varepsilon}_{a_k} / s_n. \qquad (2.47)$$

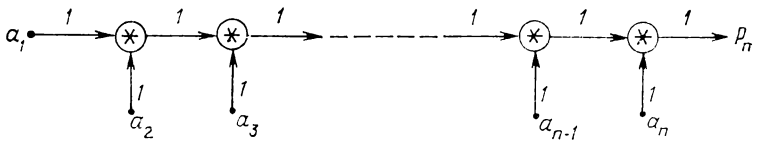


Рис. 6. Схема вычислений произведения

Причем эта формула не зависит от алгоритма суммирования. Ее вывод в порядке упражнения предлагаем сделать самостоятельно. При условии, что $\bar{\varepsilon}_{a_k} = \bar{\varepsilon}_a = \text{const}$ для всех k, имеем $\bar{\varepsilon}^n{}_{s_n} = \bar{\varepsilon}_a$. Таким образом, ошибки в исходных данных не приводят к накоплению ошибок при суммировании этих данных.

Общую ошибку суммирования n приближенных чисел с учетом ошибок округления можно оценить по предельной относительной ошибке

$$\bar{\varepsilon}_{s_n} = \bar{\varepsilon}^n{}_{s_n} + \bar{\varepsilon}^0{}_{s_n}. \tag{2.48}$$

При этом от накопления ошибок зависит только значение $\bar{\varepsilon}^0{}_{s_n}$, которое, как было показано выше, существенно зависит от выбора алгоритма суммирования.

Другой часто встречающейся математической процедурой в различных вычислительных процессах является нахождение произведения последовательности чисел, т. е.

$$p = \prod_{i=1}^{n} a_i = a_1 a_2 \ldots a_{n-1} a_n. \tag{2.49}$$

Вычисление значения p будем выполнять по алгоритму

$$p_i = p_{i-1} a_i,$$

где $i = 2 \div n$, $p_1 = a_1$, $p_n = p$, графическое изображение которого представлено на рис. 6.

Оценим точность вычисленного значения p, для этого сначала найдем предельную относительную ошибку значения p_i при $i = 2, 3$:

$$\bar{\varepsilon}_{p_2} = \bar{\varepsilon}_{a_1} + \bar{\varepsilon}_{a_2} + \bar{\varepsilon}_{\text{окр}}, \qquad \bar{\varepsilon}_{p_3} = \sum_{i=1}^{3} \bar{\varepsilon}_{a_i} + 2\bar{\varepsilon}_{\text{окр}}.$$

И далее по индукции имеем

$$\bar{\varepsilon}_{p_n} = \sum_{i=1}^{n} \bar{\varepsilon}_{a_i} + (n-1)\bar{\varepsilon}_{\text{окр}}. \tag{2.50}$$

Если $\bar{\varepsilon}_{a_i} = \varepsilon_a = \text{const}$, то

$$\bar{\varepsilon}_{p_n} = n\bar{\varepsilon}_a + (n-1)\bar{\varepsilon}_{\text{окр}}.$$

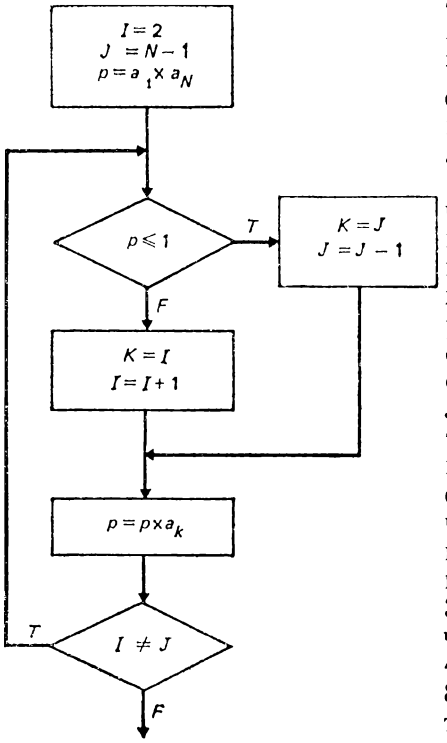


Рис. 7. Блок-схема алгоритма вычисления произведения

Таким образом, процесс перемножения n приближенных чисел приводит к накоплению ошибок и из-за ошибок исходных данных и ошибок округления.

Другими источниками вычислительных ошибок, как отмечалось в начале § 2.3, являются машинный нуль и переполнение. В качестве примера рассмотрим умножение вещественных чисел в десятичной системе счисления, представленных с фиксированной запятой (отметим, что приводимые ниже рассуждения будут относиться и для нормализованных чисел, представленных с плавающей запятой). Пусть под целую часть числа отводятся 3 разряда а под дробную часть числа — 4 разряда и {0,0050, 4,0000; 0,0040; 0,0500; 15,0000; 80,0000; 20,0000} — последовательность, числа которой требуется перемножить. Если эту последовательность чисел перемножать по алгоритму, изображенному на рис. 6, то после перемножения первых трех чисел получим машинный нуль, а следовательно, и результат произведения всех чисел данной последовательности будет равен нулю. Если же перемножение осуществлять с конца последовательности , то получим переполнение разрядной сетки уже после умножения двух последних чисел этой последовательности. Точный же результат перемножения чисел, входящих в исходную последовательность, равен значению 0,096, которое умещается в заданную разрядную сетку. Поэтому необходимо использовать такой алгоритм перемножения последовательности чисел, при котором конечный результат перемножения не равен машинному нулю, не возникает переполнения разрядной сетки представления чисел и имеется возможность получить конечный результат перемножения с точностью, определяемой формулой (2.50).

Блок-схема такого алгоритма приведена на рис. 7. В ней символ T над стрелкой указывает на то, что в случае выполнения условия, стоящего в ромбике, из которого выходит поме-

ченная стрелка, вычислительный процесс продолжается вдоль этой стрелки, в противном случае (когда условие не выполняется) — вдоль стрелки, помеченной символом F.

Возвращаясь к описанному выше примеру, при использовании алгоритма, приведенного на рис. 7, последовательность вычислений будет следующая. Сначала получаем $\{a_k\}^*{}_{k=1\div 7} = \{0{,}0040;\ 0{,}0050;\ 0{,}0500;\ 4{,}0000;\ 15{,}0000;\ 20{,}0000;\ 80{,}0000\}$. Затем вычисляем $p_1 = a_1 a_7 a_6 = 6{,}4000$, $Q_1 = p_1 a_2 = 0{,}0320$, $p_2 = Q_1 a_5 a_4 = 1{,}9200$ и $Q_3 = p_2 a_3 = 0{,}0960$.

Одной из часто используемых вычислительных процедур в численных методах, и особенно в линейной алгебре, является нахождение скалярного произведения по формуле

$$\mathrm{scl} = (x,\ y) = \sum_{i=1}^{n} x_i y_i.$$

Поэтому точность вычисления значения скалярного произведения в ряде алгоритмов фактически определяет и точность всего вычислительного процесса, выполняемого на вычислителе. В основу алгоритма вычисления значения scl положим формулу суммирования $s_n = \sum_{i=1}^{n} a_i$, где $a_i = x_i y_i$. В результате для предельной относительной ошибки значения scl получим

$$\bar{\varepsilon}_{\mathrm{scl}} = \sum_{i=1}^{n} a_i \bar{\varepsilon}_{a_i} / S_n + \bar{\varepsilon}^0{}_{s_n},\quad \bar{\varepsilon}_{a_i} = \bar{\varepsilon}_{x_i} + \bar{\varepsilon}_{y_i} + \bar{\varepsilon}_{\mathrm{окр}}, \qquad (2.51)$$

где $\bar{\varepsilon}^0{}_{s_n}$ — ошибка, вызванная округлением при вычислении суммы из n чисел, зависящая от выбора алгоритма суммирования. Предполагая, что $\varepsilon_{x_i} = \varepsilon_{y_i} = \varepsilon = \mathrm{const}$, имеем

$$\bar{\varepsilon}_{\mathrm{scl}} = 2\varepsilon + \bar{\varepsilon}_{\mathrm{окр}} + \bar{\varepsilon}^0{}_{s_n}.$$

Следовательно, чтобы ошибки округления не оказывали решающего влияния на точность вычисления $\bar{\varepsilon}_{\mathrm{scl}}$, необходимо промежуточные вычисления производить с удвоенной точностью, т. е. в режиме накопления. В тех случаях, когда вычислительные средства не позволяют реализовать режим накопления, необходимо использовать высокоточные алгоритмы суммирования, например суммирование по парам чисел.

§ 2.5. ОБРАТНЫЙ АНАЛИЗ ОШИБОК. ВОЗМУЩЕНИЯ ОПЕРАТОРА

В предыдущем параграфе было показано, как от выбора алгоритма может существенно измениться конечный результат даже при вычислении самых простых арифметических выражений.

Для более сложных математических преобразований эти изменения могут быть еще значительней. Если через A обозначить входные данные задачи, а через B — результат их обработки по исходному алгоритму φ, причем алгоритм φ конечен, т. е. не приводит к бесконечным результатам, то

$$B=\varphi(A). \tag{2.52}$$

При реализации исходного алгоритма на ЭВМ в силу специфики выполнения машинных операций и других вычислительных ограничений он будет заменен на другой машинный алгоритм $\hat{\varphi}$, вследствие чего вместо B будет получен результат $\hat{B}$, для которого имеем

$$\hat{B}=\varphi(\hat{A}).$$

Будем считать, что процессы обработки данных A по алгоритмам φ и $\hat{\varphi}$ являются конечными и не приводят к бесконечным числовым результатам. Предполагая, что B и $\hat{B}$ принадлежат одному и тому же множеству решений, можно получить количественную оценку вычисленного решения после подходящим образом введенной на этом множестве метрики $\rho(B, \hat{B})$. Такой подход к оценке суммарного влияния ошибок вычислений получил название *прямого анализа ошибок.* С подобным понятием мы уже встречались при оценке ошибки вычисления функции через известные ошибки ее аргументов или параметров. Другими примерами прямого анализа ошибок могут служить оценки ошибок округления, возникающих при сложении и умножении последовательности чисел. Однако при вычислении более сложных математических выражений прямой анализ ошибок может стать весьма трудоемким. Кроме того, в ряде случаев, например при цифровом моделировании физических процессов (измерений) планирования экспериментов или в задачах автоматического проектирования, необходимо знать отдельное влияние суммарных ошибок вычисления и ошибок измерения на решение задачи. Для этого целесообразно использовать такой подход к оценке влияния ошибок, при котором вычисленное решение $\hat{B}$ рассматривается как результат обработки возмущенных входных данных $\hat{A}$ по исходному алгоритму φ, т. е.

$$\hat{B}=\varphi(\hat{A})=\varphi(A+\delta A). \tag{2.53}$$

В этом случае ошибка вычисленного решения будет зависеть от элемента $\delta A=\hat{A}-A$, который принято называть *эквивалентным возмущением.* Предполагая, что A и $\hat{A}$ принадлежат одному множеству входных данных, и введя на этом множестве подходящую метрику, можно получить количественную оценку эквивалентного возмущения через $\rho(A, \hat{A})$, а затем и оценить

ошибку результата вычисления $\hat{B}$. Описанный подход получил название *обратного анализа ошибок.*

Реализация задачи обратного анализа ошибок связана со следующими тремя этапами. Первый относится к переходу от приближенного алгоритма $\hat{\varphi}$ к точному алгоритму φ за счет внесения эквивалентных возмущений в исходную информацию (входные данные). На этом же этапе требуется и обоснование такого перехода. На втором этапе основным является выбор нормы в пространстве входных данных для оценки эквивалентного возмущения $\|\delta A\|$. На третьем этапе осуществляется оценка результата вычислений. Причем она может быть выполнена в некоторых случаях только качественно, без вычисления ее количественного значения.

На практике входные данные всегда содержат определенные ошибки, связанные с измерениями или предварительными вычислениями. Обратный анализ ошибок показывает, что во многих случаях влияние ошибок последующих вычислений (округлений) равносильно дополнительному внесению ошибок во входные данные. Сравнение первичных ошибок входных данных и эквивалентных возмущений позволяет правильно соизмерить точность входных данных с точностью самих вычислений. Так, если в результате такого сравнения окажется, что влияние ошибок округления равносильно первичным ошибкам, то можно считать, что сами вычисления проведены достаточно точно, и не вычислять ошибку округления конечного результата.

Обратный анализ ошибок можно рассмотреть и на примере выполнения в ЭВМ одной арифметической операции над вещественными числами. Тогда, как следует из формулы (2.31), результат любой такой операции может быть представлен в виде

$$\hat{C}=C(1+\varepsilon), \tag{2.54}$$

где ε можно характеризовать как суммарное относительное возмущение операндов A и B. Причем если A и B представлены в нормализованном виде с плавающей запятой, то для ε имеется оценка $|\varepsilon|\leqslant p^{-t+1}$. В связи с тем что в ЭВМ для выполнения алгебраических действий над числами имеются лишь элементарные арифметические операции «+», «—», «*», «/», то любое по виду и сложности арифметическое выражение, состоящее из совокупности арифметических операций и различных функций, при реализации на ЭВМ аппроксимируется другим по виду выражением, состоящим из совокупности только предусмотренных в ЭВМ арифметических операций. На основании указанного, при разработке исходного алгоритма целесообразно использовать такие численные методы, в которых для вычисления арифметических выражений используются только элементарные арифметические операции с определенной

последовательностью их выполнения. Если теперь предположить, что арифметические выражения исходного алгоритма φ состоят только из элементарных операций, то реализацию φ этого алгоритма на ЭВМ, исходя из формулы (2.31), можно рассматривать как исходный алгоритм с возмущенными входными данными арифметических выражений. Таким образом, для оценки точности численного результата, получаемого в итоге математических вычислений над вещественными числами, во многих случаях можно использовать подход обратного анализа ошибок. Он особенно эффективен для оценки устойчивости вычислительного алгоритма, критерием которой может служить число, характеризующее степень роста ошибки результата от величины эквивалентных возмущений, предварительно вносимых во входные данные. Для этого прежде всего требуется установить связь между возмущениями входных данных и возмущениями (ошибками) результата вычисления.

Рассмотрим решение сформулированной задачи на примере оценки возмущения псевдообратного оператора. Пусть $\mathscr{A}$ — исходный оператор. Определим возмущенный оператор $\hat{\mathscr{A}}=\mathscr{A}+Q$ и возмущение приближенного псевдообратного оператора $\delta_+ = \hat{\mathscr{A}}^+ - \mathscr{A}^+$. Наша цель — определить зависимость δ_+ от Q и получить оценки $\|\delta_+\|$ через $\|\mathscr{A}\|$ и $\|Q\|$.

Теорема 2.2. Оператор δ_+ может быть представлен в виде

$$\delta_+ = G_1 + G_2 + G_3, \tag{2.55}$$

где $G_1 = -\hat{\mathscr{A}}^+ Q \mathscr{A}^+$, $G_2 = \hat{\mathscr{A}}^+ (E - \mathscr{A}\mathscr{A}^+)$,

$$G_3 = -(E - \hat{\mathscr{A}}^+ \hat{\mathscr{A}}) \mathscr{A}^+.$$

При этом для G_1, G_2, G_3 справедливы оценки

$$\| G_1 \|_2 \leqslant \| Q \|_2 \| \mathscr{A}^+ \|_2 \| \hat{\mathscr{A}}^+ \|_2, \quad \| G_2 \|_2 \leqslant \| Q \|_2 \| \hat{\mathscr{A}}^+ \|_2^2,$$
$$\| G_3 \|_2 \leqslant \| Q \|_2 \| \mathscr{A}^+ \|_2^2, \tag{2.56}$$

где берутся спектральные нормы матриц.

Доказательство. Первая часть теоремы доказывается проверкой формулы (2.55)

$$\delta_+ = -\hat{\mathscr{A}}^+ Q \mathscr{A}^+ + \hat{\mathscr{A}}^+ (E - \mathscr{A}\mathscr{A}^+) - (E - \hat{\mathscr{A}}^+ \hat{\mathscr{A}}) \mathscr{A}^+ \Leftrightarrow$$
$$\Leftrightarrow \delta_+ = -\hat{\mathscr{A}}^+ (\hat{\mathscr{A}} - \mathscr{A}) \mathscr{A}^+ + \hat{\mathscr{A}}^+ - \hat{\mathscr{A}}^+ \mathscr{A}\mathscr{A}^+ -$$
$$- \mathscr{A}^+ + \hat{\mathscr{A}}^+ \hat{\mathscr{A}} \mathscr{A}^+ \Leftrightarrow \delta_+ = \hat{\mathscr{A}}^+ - \mathscr{A}^+.$$

Перейдем к доказательству формул (2.56). Так

$$\| G_1 \|_2 = \| \hat{\mathscr{A}}^+ Q \mathscr{A}^+ \|_2 \Rightarrow \| G_1 \|_2 \leqslant \| \hat{\mathscr{A}}^+ \|_2 \| Q \|_2 \| \mathscr{A}^+ \|_2.$$

Для получения оценок $\|G_2\|_2$ и $\|G_3\|_2$ сначала докажем равенство

$$\| E - TT^+ \|_2 = 1, \tag{2.57}$$

выполняемое при условии $(E-TT^+)\neq 0$. Оператор $P=(E-TT^+)$ имеет собственные значения, равные единице и нулю, так как сингулярные числа оператора TT^+ равны единице и нулю. Кроме того, из свойств $(\mathscr{A}\mathscr{A}^+)^*=\mathscr{A}\mathscr{A}^+$ вытекает равенство $P=P^*$. Операторы с описанными выше свойствами называются *проекционными операторами*, а соответствующие им матрицы — *проекционными матрицами*. Если P — проекционный оператор и $P\neq 0$, то наибольшее сингулярное число этого оператора равно единице, так как по определению P — эрмитовый оператор, у которого $\max\lambda_p=1$. Отсюда $\|P\|_2=1$ при $P\neq 0$. Следовательно, $\|E-TT^+\|_2=\|E_n-\mathscr{A}\mathscr{A}^+\|_2=1$, если $\mathscr{A}\mathscr{A}^+\neq E_n$. Аналогично для $\mathscr{A}^+\mathscr{A}\neq E_m$ имеем $\|E_m-\mathscr{A}^+\mathscr{A}\|_2=1$. В том случае, когда $\mathscr{A}^+=\mathscr{A}^{-1}$, операторы $G_2=G_3=0$ и $\|G_2\|=\|G_3\|=0$. Рассмотрим случай когда $\mathscr{A}^+\neq\mathscr{A}^{-1}$. Выразим G_2 и G_3 через возмущение Q. Используя (1.108) и (1.120) имеем

$$G_2=\hat{\mathscr{A}}^+(\hat{\mathscr{A}}^*)^+\hat{\mathscr{A}}^*(E-(\mathscr{A}^*)^+\mathscr{A}^*)\Leftrightarrow G_2=$$
$$=\hat{\mathscr{A}}^+(\hat{\mathscr{A}}^*)^+\hat{\mathscr{A}}^*-\hat{\mathscr{A}}^+(\hat{\mathscr{A}}^*)^+\hat{\mathscr{A}}^*(\mathscr{A}^*)^+\mathscr{A}^*\Leftrightarrow$$
$$\Leftrightarrow G_2=\hat{\mathscr{A}}^+(\hat{\mathscr{A}}^*)^+(\hat{\mathscr{A}}^*-(\mathscr{A}^*+Q)(\mathscr{A}^*)^+\mathscr{A}^*)\Leftrightarrow$$
$$\Leftrightarrow G_2=\hat{\mathscr{A}}^+(\hat{\mathscr{A}}^*)^+Q^*(E-\mathscr{A}\mathscr{A}^+). \qquad (2.58)$$

Выполнив подобные по характеру преобразования, для G_3 имеем

$$G_3=(E-\hat{\mathscr{A}}^+\hat{\mathscr{A}})Q^*(\mathscr{A}^*)^+\mathscr{A}^+. \qquad (2.59)$$

Переходя к вычислению норм в равенствах (2.58) и (2.59), получаем оценки $\|G_2\|_2$ и $\|G_3\|_2$, определяемые по формулам (2.56).

В оценках, вычисляемых по формулам (2.56), необходимо определение $\|\hat{\mathscr{A}}^+\|$, что в ряде случаев приводит к дополнительным вычислительным затратам. В теореме, сформулированной ниже, получены оценки, не включающие $\|\hat{\mathscr{A}}^+\|$.

Теорема 2.3. Пусть G_1, G_2, G_3 — операторы, определенные в теореме 2.2. Предположим, что $q_2=\|Q\|_2\|\mathscr{A}^+\|_2<1$ и $\operatorname{Rg}\hat{\mathscr{A}}\leqslant\leqslant\operatorname{Rg}\mathscr{A}$. Тогда $\operatorname{Rg}\hat{\mathscr{A}}=\operatorname{Rg}\mathscr{A}$ и

$$а)\quad \|G_1\|_2\leqslant\frac{q_2\|\mathscr{A}^+\|_2}{1-q_2},\qquad \|G_2\|_2\leqslant\frac{q_2\|\mathscr{A}^+\|_2}{1-q_2},$$
$$\|G_3\|_2\leqslant q_2\|\mathscr{A}^+\|_2, \qquad (2.60)$$
$$б)\quad \|\delta_+\|_2\leqslant\frac{cq_2\|\mathscr{A}^+\|_2}{1-q_2}.$$

При $\operatorname{Rg}\mathscr{A}<\min(m,n)$ значение $c\cong 1{,}62$, а при $\operatorname{Rg}\mathscr{A}=\min(m,n)$ значение $c\cong 1{,}41$. Доказательство теоремы приведено в [11].

Из теоремы вытекают важные для вычислительной практики следствия.

Следствие 1. Пусть в пространстве X_n действует преобразование (оператор) $(E+Q)$, где $\|Q\|_2<1$. Тогда существует преобразование $(E+Q)^{-1}$, для которого выполняется оценка

$$\|(E+Q)^{-1}\|_2 \leqslant \frac{1}{1-\|Q\|_2}. \tag{2.61}$$

Доказательство того, что оператор $(E+Q)^{-1}$ при $\|Q\|_2<1$ является невырожденным, следует из первой части теоремы 2.3.

Рассмотрим теперь невырожденное преобразование $\mathscr{A}$, действующее в линейном пространстве X, и возмущенное преобразование $(\mathscr{A}+Q)=\hat{\mathscr{A}}$.

Следствие 2. Возмущенное преобразование $\hat{\mathscr{A}}$ будет невырожденным при всех возмущениях Q невырожденного преобразования $\mathscr{A}$, если выполняется условие

$$\|Q\|_2 < \|\mathscr{A}^{-1}\|_2^{-1}. \tag{2.62}$$

Действительно, так как оператор $\mathscr{A}^{-1}$ невырожденный, то преобразование

$$\mathscr{A}+Q=\mathscr{A}(E+\mathscr{A}^{-1}Q)$$

будет невырожденным тогда и только тогда, когда будет невырожденным преобразование $(E+\mathscr{A}^{-1}Q)$, а это, как вытекает из предыдущего следствия, равносильно тому, что выполняется неравенство

$$\|\mathscr{A}^{-1}Q\|_2 < 1.$$

Тем более условие невырожденности преобразования $(E+\mathscr{A}^{-1}Q)$ будет выполняться, если $\|\mathscr{A}^{-1}\|\,\|Q\|_2<1$, что и доказывает утверждение следствия.

Обозначим через

$$\varepsilon_A=\frac{\|Q\|_2}{\|\mathscr{A}\|_2}, \qquad \varepsilon_2^{+}=\frac{\|\delta_+\|_2}{\|\mathscr{A}\|_2} \tag{2,63}$$

величины относительных возмущений операторов $\mathscr{A}$ и $\mathscr{A}^{+}$. Тогда при принятых обозначениях и выполнении условия теоремы 2.3, что $q_2<1$, формула (2.63) преобразуется к виду

$$\varepsilon_2^{+} \leqslant \frac{c\varepsilon_A K_2^{+}}{1-\varepsilon_A K_2^{+}}, \tag{2.64}$$

где значение

$$K_2^{+}=\|\mathscr{A}^{+}\|_2\,\|\mathscr{A}\|_2 \tag{2.65}$$

называется *спектральным числом обусловленности оператора* $\mathscr{A}$.

Если $\mathscr{A}$ является невырожденным преобразованием (оператором) в пространстве X_n, то формула (2.64) при обозначениях

$$\varepsilon_2^- = \frac{\|\hat{\mathscr{A}}^{-1} - \mathscr{A}^{-1}\|_2}{\|\mathscr{A}^{-1}\|_2}, \qquad K_2^- = \|\mathscr{A}^{-1}\|_2 \|\mathscr{A}\|_2 \tag{2.66}$$

и условии $q_2 = \varepsilon_A K_2^- < 1$ примет вид

$$\varepsilon_2^- \leqslant \frac{\varepsilon_A K_2^-}{1 - \varepsilon_A K_2^-}. \tag{2.67}$$

Отметим, что при оценках δ_+ и $\mathscr{A}^+$ везде использовалась спектральная норма оператора. Это вызвано тем, что для получения нормального псевдорешения (т. е. псевдообращения) необходимо минимизировать евклидовы нормы сначала вектора невязки, а затем вектора псевдорешения. Но только спектральная норма оператора индуцируется евклидовыми нормами в пространствах R_n и R_m. Если для оценки δ_+ и $\mathscr{A}^+$ использовать согласованные нормы операторов, то тогда, кроме спектральных норм, эти оценки можно осуществлять и по евклидовой норме оператора, являющейся согласованной с евклидовыми нормами векторов в пространствах R_n и R_m. С учетом формулы (1.164), т. е.

$$\|\mathscr{A}\|_2 \leqslant \|\mathscr{A}\|_E \leqslant \sqrt{t}\,\|\mathscr{A}\|_2 \text{ при } t = \min(m, n),$$

получим оценки $\|\delta_+\|_E$, $\|\mathscr{A}^+\|_E$, $\|\mathscr{A}\|_E$ и других операторов, используя оценки этих же операторов через их спектральные нормы. Так, оценка для оператора G_1, полученная в теореме 2.3, преобразуется для евклидовой нормы оператора следующим образом:

$$\|G_1\|_E \leqslant \sqrt{t}\,\|G_1\|_2 \leqslant \sqrt{t}\,\|Q\|_2 \|\mathscr{A}^+\|_2 \|\hat{\mathscr{A}}^+\|_2 \leqslant$$
$$\leqslant \sqrt{t}\,\|Q\|_E \|\mathscr{A}^+\|_E \|\hat{\mathscr{A}}^+\|_E.$$

Аналогичные преобразования для $\|G_2\|_2$ и $\|G_3\|_2$ приведут к оценкам вида

$$\|G_1\|_E \leqslant \sqrt{t}\,q_E \|\hat{\mathscr{A}}^+\|_E, \quad \|G_2\|_E \leqslant \sqrt{t}\,\|\hat{\mathscr{A}}^+\|^2_E,$$
$$\|G_3\|_E \leqslant \sqrt{t}\,q_E \|\mathscr{A}^+\|_E, \tag{2.68}$$

где
$$q_E = \|Q\|_E \|\mathscr{A}^+\|_E. \tag{2.69}$$

Условие $q_2 = \|Q\|_2 \|\mathscr{A}^+\|_2 < 1$, входящее в формулировку теоремы 2.3, при переходе к евклидовой норме оператора может быть заменено другим условием

$$q_E = \|Q\|_E \|\mathscr{A}^+\|_E < 1, \tag{2.70}$$

которое усиливает условие $\|Q\|_2 \|\mathscr{A}^+\|_2 < 1$. В результате первое утверждение теоремы 2.3 остается в силе, а формула (2.60, *б*) преобразуется к виду

$$\| \delta_+ \|_E \leqslant \frac{c\sqrt{l}\, q_E \| \mathscr{A}^+ \|_E}{1 - q_E}. \tag{2.71}$$

Оценивая величины относительных возмущений операторов $\mathscr{A}$ и $\mathscr{A}^+$ через евклидовы нормы оператора, на основании формулы (2.65) получим неравенство

$$\varepsilon_E^+ \leqslant \frac{c\sqrt{l}\, \varepsilon_A K_E^+}{1 - \varepsilon_A K_E^+}, \tag{2.72}$$

в котором

$$\varepsilon_E^+ = \| \delta_+ \|_E / \| \mathscr{A}^+ \|_E,$$
$$\varepsilon_A = \| Q \|_E / \| \mathscr{A} \|_E,$$
$$K_E^+ = \| \mathscr{A}^+ \|_E \| \mathscr{A} \|_E.$$

Если $\mathscr{A}$ является невырожденным преобразованием в X_n, то для оценки $\mathscr{A}^{-1}$ можно ввести любую согласованную норму. При этом формула (2.60, *б*) при условии $q_c = \|Q\|_c \|\mathscr{A}^{-1}\|_c < 1$ примет вид

$$\| \delta_- \|_c \leqslant \frac{\| Q \|_c \| \mathscr{A}^{-1} \|_c^2 \| E \|_c}{1 - \| Q \|_c \| \mathscr{A}^{-1} \|_c}. \tag{2.73}$$

Отсюда при тех же условиях для оценки относительного возмущения ε^- получим

$$\varepsilon_c^- \leqslant \frac{\varepsilon_A K_c^- \| E \|_c}{1 - \varepsilon_A K_c}, \tag{2.74}$$

где число K_c^-, определяемое по формуле

$$K_c^- = \| \mathscr{A}^{-1} \|_c \| \mathscr{A} \|_c, \tag{2.75}$$

называется *числом обусловленности невырожденного преобразования $\mathscr{A}$ по согласованной норме.*

Для любой согласованной нормы оператора при условии, что $\|Q\|_c < 1$, формула (2.61) представляется в виде

$$\| (E+Q)^{-1} \|_c \leqslant \| E \|_c / (1 - \| Q \|_c). \tag{2.76}$$

Формулировка следствия 2 для согласованных норм оператора не изменяется, т. е., если $\|Q\|_c < \|\mathscr{A}^{-1}\|_c^{-1}$, то из невырожденности $\mathscr{A}$ следует невырожденность преобразования $\hat{\mathscr{A}} = \mathscr{A} + Q$.

Оценки, полученные для любого оператора $\mathscr{A}$, совпадают с соответствующими оценками для матриц этого оператора, так как $\|\mathscr{A}\| = \|A\|$ по определению нормы оператора.

§ 2.6. УСТОЙЧИВОСТЬ РЕШЕНИЯ ОПЕРАТОРНЫХ УРАВНЕНИЙ. АЛГОРИТМЫ РЕШЕНИЯ ПЛОХО ОБУСЛОВЛЕННЫХ СИСТЕМ

Предположим, что решается операторное уравнение (1.92) общего вида, т. е. $m \neq n$ и $\operatorname{Rg} \mathscr{A} \leqslant \min(m, n)$:

$$\mathscr{A} x = b.$$

Как правило, из-за ошибок в исходной информации, связанных как с ошибками измерений, так и с ошибками предыдущих вычислений, на практике решается возмущенная система

$$(\mathcal{A}+Q^{(1)})x^{(1)}=b+db^{(1)}. \tag{2.77}$$

Возмущения Q и db можно оценить, учитывая точность измерений или проводя анализ ошибок предыдущих вычислений.

В результате решения возмущенной системы (2.77) будет получено решение $x^{(1)}=x+dx^{(1)}$, где $dx^{(1)}$ — ошибка решения, вызванная ошибками в исходной информации.

Независимо от существования возмущений $Q^{(1)}$ и $db^{(1)}$ при записи исходной информации в ЭВМ (или в другом вычислительном устройстве), а также при решении системы (2.77) в ЭВМ появляются новые ошибки, связанные с невозможностью точного представления вещественных чисел конечноразрядными машинными числами. Эти ошибки были названы ошибками округления, которые приводят к тому, что вместо точной системы решается возмущенная система

$$(\mathcal{A}+Q^{(2)})x^{(2)}=b+db^{(2)}.$$

Как было показано на примере суммирования последовательности чисел, ошибки округления существенно зависят от выбора алгоритма. Таким образом, вместо идеальной системы $\mathcal{A}x=b$ реально решается возмущенная система

$$(\mathcal{A}+Q)\tilde{x}=b+db, \tag{2.78}$$

где $\tilde{x}$ можно рассматривать как точное решение этой системы. Здесь Q и db будут эквивалентными возмущениями метода (алгоритма) решения системы (1.92). Пусть для относительных возмущений $\varepsilon_{\mathcal{A}}$ и ε_b получаются оценки вида

$$\varepsilon_{\mathcal{A}}\leqslant f_{\mathcal{A}}(n)\bar{\varepsilon}_{\text{окр}},\qquad \varepsilon_b\leqslant f_b(n)\bar{\varepsilon}_{\text{окр}}, \tag{2.79}$$

где $\bar{\varepsilon}_{\text{окр}}$ характеризует относительную точность машинной операции, т. е. $\bar{\varepsilon}_{\text{окр}}\leqslant 2^{-t+1}$, $f_{\mathcal{A}}(n)$, $f_b(n)$ — функции, достаточно медленно растущие с возрастанием n, например типа полинома третьей степени от n с коэффициентом при кубическом члене $a_0<2$. Если алгоритм решения уравнения (1.92) приводит к возмущениям Q и db, оцениваемым по формулам (2.79), то данный алгоритм будем считать *устойчивым по Уилкинсону.*

К вопросу об оценке устойчивости алгоритма решения операторного уравнения можно подойти и с других позиций. Рассмотрим его на примере невырожденного операторного уравнения $\mathcal{A}x=b$, точное решение которого обозначим $x^{\text{т}}=\mathcal{A}^{-1}b$. Пусть это уравнение решается на ЭВМ по заданному алгоритму, тогда исходное уравнение можно аппроксимировать уравнением вида (2.78). При этом предположим, что выполняется условие $q_c=\|\mathcal{A}^{-1}\|_c\|Q\|_c<1$. Тогда на основании следствия 2

теоремы 2.3 получаем, что возмущенное операторное уравнение (1.92) тоже будет невырожденным. Отсюда при $\tilde{x}=x^{\tau}+\Delta x$ независимо от выбранного алгоритма решения имеют место такие преобразования:

$$(\mathscr{A}+Q)\tilde{x}-\mathscr{A}x^{\tau}=db \Leftrightarrow (\mathscr{A}+Q)\Delta x+Qx^{\tau}=db \Leftrightarrow$$
$$\Leftrightarrow \Delta x=(\mathscr{A}+Q)^{-1}db-(\mathscr{A}+Q)^{-1}Qx^{\tau}.$$

С учетом того, что

$$(\mathscr{A}+Q)^{-1}=(\mathscr{A}+\mathscr{A}\mathscr{A}^{-1}Q)^{-1}=(E+\mathscr{A}^{-1}Q)^{-1}\mathscr{A}^{-1},$$

выражение для Δx принимает вид

$$\Delta x=(E+\mathscr{A}^{-1}Q)^{-1}\mathscr{A}^{-1}db-(E+\mathscr{A}^{-1}Q)^{-1}\mathscr{A}^{-1}Qx^{\tau}.$$

Оценивая теперь вектор Δx по норме $\|\Delta x\|$ и используя для оценки операторов, входящих в это выражение, нормы, согласованные с векторными нормами, с учетом формулы (2.76) получим

$$\|\Delta x\| \leqslant \frac{\|E\|_c\|\mathscr{A}^{-1}\|_c}{1-q_c}\|db\|+\frac{\|E\|_c q_c}{1-q_c}\|x^{\tau}\|.$$

Отсюда, обозначая $\|\Delta x\|/\|x^{\tau}\|=\varepsilon_x$, $\|db\|/\|b\|=\varepsilon_b$, это неравенство преобразуем к виду

$$\varepsilon_x \leqslant \frac{\|E\|_c}{1-q_c}\left(\frac{\|\mathscr{A}^{-1}\|_c\|db\|}{\|x\|\ \|\mathscr{A}\|_c}\|\mathscr{A}\|_c+q_c\right).$$

Так как $K_c^-=\|\mathscr{A}^{-1}\|_c\|\mathscr{A}\|_c$ и $q_c=\varepsilon_A K_c^-$, имеем

$$\varepsilon_x \leqslant \frac{K_c^-\|E\|_c}{1-\varepsilon_A K_c^-}(\varepsilon_b+\varepsilon_A). \tag{2.80}$$

Таким образом, получили результат, который можно сформулировать в виде следующего утверждения. Пусть оператор $\mathscr{A}$ и вектор b, составляющие операторное уравнение $\mathscr{A}x=b$, получили возмущения Q и db, причем в некоторой согласованной норме выполняется условие $q=\|\mathscr{A}^{-1}\|_c\|Q\|_c<1$. Тогда возмущенная система вида (2.78) имеет единственное решение и относительное возмущение ε_x решения исходного (точного) уравнения оценивается через относительные возмущения ε_A и ε_b по формуле (2.80), в которой K^- — число обусловленности оператора $\mathscr{A}$ (матрицы A) в заданной норме.

Если $\varepsilon_A K_c^- \leqslant 1$ и $\|E\|_c=1$, то из формул (2.80) и (2.79) следует неравенство

$$\varepsilon_x \leqslant K_c^-(\varepsilon_b+\varepsilon_A) \leqslant K_c^- f(n)\bar{\varepsilon}_{\text{окр}}. \tag{2.81}$$

Отсюда устойчивость решения различных операторных уравнений по заданному алгоритму целиком определяется числом обусловленности исходных операторов. Причем даже в тех слу-

чаях, когда применяемый алгоритм решения дает «плохую» устойчивость по Уилкинсону, т. е. не выполняются неравенства (2.79), он может обладать все же вполне «хорошей» устойчивостью, если выполняется условие

$$K_c^- < \min\left(\frac{1}{\varepsilon_A}, \frac{\bar{\varepsilon}_x}{\varepsilon_A}\right), \tag{2.82}$$

где $\bar{\varepsilon}_x$ — предельная относительная ошибка решения операторного уравнения. Алгоритм решения операторных уравнений, приводящий к выполнению условия (2.82), будем называть *устойчивым по критерию обусловленности.*

Анализ формулы (2.80) показывает, что с увеличением числа обусловленности K^- возрастает и относительное возмущение решения при условии сохранения возмущений исходных данных. Однако так как число обусловленности определяется не только оператором, но и выбором нормы, то по формуле (2.80) будем получать разные оценки относительного возмущения ε_x, более грубые или более точные, в зависимости от выбора норм оператора, что равносильно нормам матриц соответствующих операторов.

Приведем некоторые свойства числа обусловленности невырожденных операторов.

Свойство 1. Независимо от выбора нормы матрицы (оператора) имеет место равенство

$$K^-(A) = K^-(A^{-1}). \tag{2.83}$$

Оно вытекает непосредственно из определения $K^-(A) = \|A^{-1}\| \times \times \|A\|$.

Свойство 2. Для мультипликативной нормы матрицы выполняется неравенство

$$\max\left\{\frac{K^-(A)}{K^-(B)}, \frac{K^-(B)}{K^-(A)}\right\} \leqslant K^-(C) \leqslant K^-(A)K^-(B), \tag{2.84}$$

где матрица (оператор) $C = AB$.

Если $C = AB$, то для мультипликативной нормы имеем

$$\|C\| \leqslant \|A\|\,\|B\| \text{ и } \|C^{-1}\| \leqslant \|A^{-1}\|\,\|B^{-1}\|.$$

Отсюда доказывается правая часть неравенства (2.84).

Для доказательства левой части этого неравенства воспользуемся правой частью неравенства (2.84) для $B = A^{-1}C$. И с учетом свойства 1 получим

$$K^-(B) \leqslant K^-(A^{-1})K^-(C) \Leftrightarrow K^-(C) \geqslant \frac{K^-(B)}{K^-(A)}.$$

Таким же образом из выражения $A = CB^{-1}$ получаем

$$K^-(C) \geqslant \frac{K^-(A)}{K^-(B)}.$$

Из сопоставления обоих полученных выше неравенств для $K^-(C)$ следует доказательство и левой части неравенства (2.84).

Свойство 3. Для любой согласованной нормы матрицы выполняется неравенство

$$K_c^-(A) \geqslant 1. \tag{2.85}$$

Действительно, из определения согласованной нормы матрицы имеем

$$\| x \| = \| AA^{-1}x \| \Rightarrow \| x \| \leqslant$$
$$\leqslant \| A \|_c \| A^{-1}x \| \Rightarrow \| x \| \leqslant \| A \|_c \| A^{-1} \|_c \| x \|.$$

Отсюда получаем неравенство (2.85).

Из этого свойства следует, что если норма матрицы является индуцированной соответствующими векторными нормами, то

$$K^-(A) \geqslant 1.$$

Оператор, для которого спектральное число обусловленности равно единице, называется *идеально обусловленным*. К идеально обусловленным операторам относятся ортогональные операторы. Действительно, так как Q — ортогональный оператор, то

$$K_2^-(Q) = \| Q \|_2 \| Q^{-1} \|_2 \Leftrightarrow K_2^-(Q) = \| Q \|_2 \| Q^{\mathrm{T}} \|_2,$$

где индекс 2 указывает на то, что берется спектральная норма оператора, которая для ортогональной матрицы равна 1. Следовательно, $K_2^-(Q) = 1$ для спектральной нормы оператора.

Свойство 4. Ортогональное преобразование (разложение) оператора оставляет без изменения евклидово и спектральное его число обусловленности, т. е. если G и P — ортогональные операторы, то

$$K^-_{2,E}(G\mathscr{A}P) = K^-_{2,E}(\mathscr{A}). \tag{2.86}$$

Доказательство этого свойства следует из формул (1.161) и (1.162). Так, для спектральной нормы оператора (матрицы)

$$K_2^-(G\mathscr{A}P) = \| G\mathscr{A}P \|_2 \| P'\mathscr{A}^{-1}G' \|_2 \Leftrightarrow$$
$$\Leftrightarrow K_2^-(G\mathscr{A}P) = \| \mathscr{A} \|_2 \| \mathscr{A}^{-1} \|_2 = K_2^-(\mathscr{A}).$$

Аналогично доказательство и для евклидовой нормы оператора (матрицы).

Отметим, что свойства 1—4, доказанные для невырожденных операторов, выполняются и для числа обусловленности K^+, вычисляемого для любого оператора. Напомним, что при вычислении $K^+(\mathscr{A})$ используются только спектральные или евклидовы нормы матриц.

Вычисление числа обусловленности K^- можно выполнить исходя из определения, т. е. по формуле (2.75). При этом наиболее сложной машинной процедурой является нахождение об-

ратной матрицы, на которую, как правило, отводится больше затрат машинного времени (числа операций), чем на решение операторного уравнения.

Часто возникает необходимость в вычислении спектрального числа обусловленности, например при уравнивании геодезических и фотограмметрических сетей по методу наименьших квадратов. Пусть ρ_1 и ρ_n — наибольшее и наименьшее сингулярные числа матрицы (оператора). Тогда $\|A\|_2=\rho_1$, а $\|A^{-1}\|_2=\rho_n^{-1}$. Отсюда следует, что спектральное число обусловленности находится по формуле

$$K_2^-(A)=\rho_1/\rho_n. \tag{2.87}$$

Из этой формулы, в частности, следует, что плохая обусловленность невырожденной матрицы не связана с величиной определителя матрицы A, как могло показаться при «беглом» анализе формулы (2.75). В подтверждение этого сравним диагональные матрицы A_1 с элементами $a_{ii}^{(1)}=10$ для $i=1\div(n-1)$, $a_{nn}^{(1)}=10^{-n+1}$ и A_2 с элементами $a_{ii}^{(2)}=1/10$ для $i=1\div n$. Тогда для матрицы A_1 получим $\det A_1=1$ и $K_2^-(A_1)=10^n$, а для матрицы A_2 соответственно — $\det A_2=10^{-n}$ и $K_2^-(A_2)=1$.

Из принятого выше для устойчивого решения операторного уравнения должно выполняться условие (2.82). Естественно, что при нарушении условия невырожденности возмущенного оператора $\hat{\mathcal{A}}=\mathcal{A}+Q$ решение операторного уравнения становится неустойчивым. Это связано с тем, что в окрестности вырожденного оператора или в более общем случае оператора неполного ранга, т. е. когда $\operatorname{Rg}A<\min(m,n)$, нарушается непрерывность решения системы, что видно и из анализа формулы (2.80). Таким образом, если для возмущения Q невырожденного преобразования $\mathcal{A}$ условие $\|Q\|_2\|\mathcal{A}^{-1}\|_2<1$ не выполняется, то решение уравнения $\mathcal{A}x=b$ может стать неустойчивым. Пусть ρ_1 и ρ_n — наибольшее и наименьшее сингулярные числа оператора $\mathcal{A}$. Тогда ρ_1^2 и ρ_n^2 — наибольшее и наименьшее сингулярные числа $\mathcal{A}^{\mathrm{T}}\mathcal{A}$. Отсюда для спектрального числа обусловленности оператора $\mathcal{A}^{\mathrm{T}}\mathcal{A}$ имеем

$$K_2(\mathcal{A}^{\mathrm{T}}\mathcal{A})=(K_2(\mathcal{A}))^2.$$

В результате может оказаться, что возмущение G оператора $\mathcal{A}^{\mathrm{T}}\mathcal{A}$ не удовлетворяет условию (2.62), т. е. $\|G\|_2\geqslant 1/\rho_n^2$, и в то же время для возмущения Q оператора $\mathcal{A}$ будем иметь $\|Q\|_2\leqslant 1/\rho_n$. Отсюда следует, что решение уравнения $\mathcal{A}^{\mathrm{T}}\mathcal{A}x=\mathcal{A}^{\mathrm{T}}b$ может стать неустойчивым, тогда как нормальное псевдорешение исходного уравнения $x^0=\mathcal{A}^+b$ будет устойчивым в связи с выполнением условия $\|Q\|_2\|\mathcal{A}^+\|_2<1$.

При этом оценку точности решения нормального псевдорешения можно выполнить по формулам

$$\|\Delta_x\|_E \leqslant \|A^+\|_2\left(\frac{\|Q\|_2\|x^0\|_E}{1-q_2}+\frac{\|\Delta b\|_E}{1-q_2}+\frac{q_2\|V\|_E}{1-q_2}+\|Q\|_2\|x^0\|_E\right), \tag{2.88}$$

$$\varepsilon_x=\frac{\|\Delta x\|_2}{\|x^0\|_E}\leqslant \hat{K}^+[(2+K_2{}^+\rho)\varepsilon_A+\gamma\varepsilon_b],$$

в которых приняты следующие обозначения:

$$\hat{K}^+=\frac{K_2{}^+}{1-K_2{}^+\varepsilon_A},\qquad \rho=\frac{\|V\|_E}{\|\mathscr{A}x^0\|_E},\qquad \gamma=\frac{\|b\|_E}{\|\mathscr{A}x^0\|_E},\qquad V=b-\mathscr{A}x,$$

где $K_2{}^+$ — спектральное число обусловленности для произвольного оператора $\mathscr{A}$. Причем для вычисления $K_2{}^+$ можно воспользоваться формулой, аналогичной (2.87), т. е.

$$K_2{}^+(\mathscr{A})=\rho_1/\rho_n. \tag{2.89}$$

С выводом приведенных формул (2.88) можно познакомиться в [11]. Здесь только отметим, что они получены при выполнении двух условий: $\|Q\|_2\|\mathscr{A}^+\|_2<1$ и $\operatorname{Rg}\tilde{A}\leqslant\operatorname{Rg}\mathscr{A}$, где $\tilde{A}=\mathscr{A}+Q$, причем второе условие всегда выполнимо, если оператор $\mathscr{A}$ полного ранга. Кроме того, при $m=n=\operatorname{Rg}\mathscr{A}$ формулы (2.88) принимают вид (2.80). Таким образом, из проведенного анализа следует, что для случаев, когда оператор $\mathscr{A}^T\mathscr{A}$ является почти вырожденным, а оператор $\mathscr{A}$ — полного ранга, с вычислительной точки зрения рациональнее вычислять нормальное псевдорешение без предварительного перехода к нормальной системе уравнений. Причем если оператор $\mathscr{A}$ отображает $R_m\to R_n$, где $n>m$ и $\operatorname{Rg}\mathscr{A}=m$, то нормальное псевдорешение уравнения $\mathscr{A}x=b$ совпадает с псевдорешением, которое будет в этом случае единственным.

Если же операторное уравнение имеет более одного псевдорешения, т. е. оператор $\mathscr{A}$ — неполного ранга, то возмущения исходного уравнения (независимо от их причины) могут приводить к большим возмущениям в нормальном псевдорешении. Например, это будет иметь место в тех случаях, когда сингулярные числа оператора $\mathscr{A}$ становятся соизмеримыми по величине с ошибками округления, что в конечном счете, как следует из (2.89), может привести к условию $K_2{}^+\varepsilon_A\geqslant 1$. Возникают такие вопросы: существует ли в принципе устойчивое нормальное псевдорешение в указанных случаях и если оно имеется, то как его найти? К рассмотрению этих вопросов мы и приступим.

Пусть оператор $\mathscr{A}$ действует из $R_m\to R_n$ и $r=\operatorname{Rg}(\mathscr{A})<\min(m,n)$. Для операторного уравнения $\mathscr{A}x=b$ наряду с его нормальным псевдорешением $x^0=\mathscr{A}^+b$ будем находить решение уравнения

$$(\mathscr{A}^*\mathscr{A}+\lambda^2E)x_\lambda=\mathscr{A}^*b, \tag{2.90}$$

где $\mathcal{A}^*$ — оператор, сопряженный оператору $\mathcal{A}$.

Теорема 2.4. Для любого оператора $\mathcal{A}$ и значения $\lambda \neq 0$, во-первых, существует обратный оператор $(\mathcal{A}^*\mathcal{A}+\lambda^2 E)^{-1}$, а во-вторых, при $\lambda \to 0$ решение уравнения (2.90) стремится к нормальному псевдорешению уравнения $\mathcal{A}x=b$.

Доказательство. Покажем, что оператор $(\mathcal{A}^*\mathcal{A}+\lambda^2 E)$ невырожденный. Для $\forall z \neq 0$ из R_m можно найти скалярное произведение

$$((\mathcal{A}^*\mathcal{A}+\lambda^2 E)z, z)=(\mathcal{A}^*\mathcal{A}z, z)+\lambda^2(z, z).$$

При $\lambda \neq 0$ значение этого скалярного произведения будет строго положительным, а следовательно, $(\mathcal{A}^*\mathcal{A}+\lambda^2 E)$ — эрмитовый и положительно определенный оператор. Отсюда

$$x_\lambda=(\mathcal{A}^*\mathcal{A}+\lambda^2 E)^{-1}\mathcal{A}^*b. \tag{2.91}$$

Для доказательства второй части теоремы, как видно из формулы (2.91), достаточно показать, что

$$\lim_{\lambda \to 0}(\mathcal{A}^*\mathcal{A}+\lambda^2 E)^{-1}\mathcal{A}^*=\mathcal{A}^+. \tag{2.92}$$

Зафиксируем в пространствах R_m и R_n сингулярные базисы. В этих базисах оператору $\mathcal{A}$ будет соответствовать матрица D размера $[n \times m]$ с элементами $d_{ij}=\rho_i$ для $i=j$ и $d_{ij}=0$ для всех $i \neq j$, где ρ_i — сингулярные числа матрицы (оператора) A при $i=1 \div r$, не равные нулю, а при $i>r$ равные нулю. Оператору $(\mathcal{A}^*\mathcal{A}+\lambda^2 E)^{-1}$ в указанных базисах будет соответствовать диагональная матрица B размера $[m \times m]$ с элементами $b_{ii}=(\rho i^2+\lambda^2)^{-1}$ при $i=1 \div r$, $b_{kk}=\lambda^{-2}$ при $k>r$. Произведением матриц BD^{T} будет матрица размера $[m \times n]$, ненулевыми элементами которой являются только элементы $c_{ii}=\rho_i(\rho_i^2+\lambda^2)^{-1}$ при $i=1 \div r$. Элементы c_{ii} при $\lambda \to 0$ будут стремиться к значениям ρ_i^{-1} для всех $i=1 \div r$. Таким образом, получили поэлементную сходимость матрицы C при $\lambda \to 0$ к матрице D^+. В связи с эквивалентностью понятий для матриц поэлементной сходимости со сходимостью по норме вытекает соотношение (2.92) и тем самым доказано второе утверждение теоремы.

В уравнении (2.90) x_λ называется *регуляризованным решением* системы $\mathcal{A}x=b$, а определение нормального псевдорешения с помощью регуляризованных решений называется *методом регуляризации.* В качестве приближения к нормальному псевдорешению в методе регуляризации берется решение операторного уравнения (2.90) при подходящем значении параметра λ. Из второго утверждения теоремы следует, что при точном задании уравнения (2.90), отсутствии ошибок округления и с уменьшением значения λ уменьшается евклидова норма $\|x_\lambda - x^0\|_E$. Однако при малых λ и в том случае, если оператор A неполного ранга, уравнение (2.90) становится плохо обус-

ловленным. В связи с этим вычисленное значение $\tilde{x}_\lambda$ и точное значение решения x_λ могут сильно отличаться, а это, в свою очередь, приведет к большой величине оценки $\|\tilde{x}_\lambda - x^0\|_E$. Поэтому возникает задача нахождения такого значения $\hat{\lambda}$, при котором обеспечивается минимальная или по крайней мере близкая к минимальной оценка $\|\tilde{x}_\lambda - x^0\|_E$.

Пусть имеется возмущенная система $\tilde{A}x = \tilde{b}$, у которой матрица $\tilde{A}$ и столбец свободных членов $\tilde{b}$ удовлетворяют неравенствам

$$\|\tilde{A} - A\|_E < \delta \text{ и } \|\tilde{b} - b\|_E < \delta.$$

Тогда имеет место утверждение: *нормальное псевдорешение исходной системы* $Ax = b$ *может быть определено с ошибкой порядка* $(\delta)^\tau$. *Причем если исходная система совместна, то* $\tau \geqslant 2/3$, а в *противном случае* — $\tau \geqslant 1/2$. Параметр λ, обеспечивающий с такой точностью приближение $\tilde{X}_\lambda$ к X^0, зависит не только от возмущений δ, но и от таких величин, как $\|A\|$ и ρ_i. В связи с этим для его нахождения требуется задавать дополнительную информацию к решаемой точной задаче. С доказательством приведенного утверждения можно познакомиться, например, в [5].

В вычислительной практике для устойчивого определения нормального псевдорешения нашли широкое применение методы, состоящие в замене исходного уравнения на уравнение $\hat{\mathscr{A}}\hat{x} = b$, где $\hat{\mathscr{A}}$ — оператор заданного (или определяемого в процессе вычислений) ранга $k \leqslant \operatorname{Rg} \mathscr{A}$, получившего название *псевдоранга.* При этом обязательным требованием для уравнения $\hat{\mathscr{A}}\hat{x} = b$ является выполнение условия $\|\hat{\mathscr{A}}^+\|_2 \|Q\|_2 < 1$.

Значительная часть методов вычисления решения СЛАУ общего вида (т. е. систем неполного ранга) основана на ортогональном разложении матрицы системы.

Пусть задана система $Ax = b$, в которой матрица A имеет размер $[n \times m]$ и $\operatorname{Rg} A = r$, где $r \leqslant \min(m, n)$. Если v — вектор невязки, равный $v = Ax - b$, то $\hat{x}$ — псевдорешение исходной системы, обеспечивающее $\min(v, v)$. Используя ортогональное разложение матрицы (1.33), исходную систему можно записать как

$$H^\tau R W x - b = v \Leftrightarrow RWx - Hb = Hv, \tag{2.93}$$

где H, W — ортогональные матрицы соответственно размеров $[m \times m]$ и $[n \times n]$ и R — матрица размера $[m \times n]$, имеющая вид

$$R = \begin{pmatrix} R_{11} & 0 \\ 0 & 0 \end{pmatrix},$$

где R_{11} — верхняя треугольная матрица размера $[r + r]$ и ранга

$\operatorname{Rg} R_{11}=r$. Введя обозначения $Wx=y$, $Hb=\bar{b}$, $Hv=\bar{v}$, перепишем систему (2.94) в виде

$$Ry-\bar{b}=\bar{v}.$$

Псевдорешение этой системы $\hat{y}$ будет находиться из решения нормальной системы уравнений

$$R^{\tau}R\hat{y}=R^{\tau}\bar{b}. \tag{2.94}$$

При этом обеспечивается $\min(\bar{v}, \bar{v})$. Так как

$$(\bar{v}, \bar{v})=(Hv, Hv) \Leftrightarrow (\bar{v}, \bar{v})=(v, v),$$

следовательно, $\hat{y}=W\hat{x} \Leftrightarrow W^{\tau}\hat{y}=\hat{x}$. С учетом вида матрицы R систему (2.95) можно представить как

$$\begin{pmatrix} R_{11}^{\tau} & O \\ O & O \end{pmatrix}\begin{pmatrix} R_{11} & O \\ O & O \end{pmatrix}\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \end{pmatrix}=\begin{pmatrix} R_{11}^{\tau}\bar{b}_1 \\ O \end{pmatrix}.$$

Отсюда получаем

$$R^{\tau}_{11}R_{11}\hat{y}_1=R^{\tau}_{11}\bar{b}_1. \tag{2.95}$$

Таким образом, общее решение нормальной системы уравнений (2.95) будет определяться вектором $\hat{y}=(\hat{y}_1, \hat{y}_2)^{\tau}$, для которого подвектор $\hat{y}_1$ находится из уравнения (2.95), а $\hat{y}_2$ — произвольный подвектор размерности $n-r$. В связи с тем что матрица (R_{11}^{τ}, R_{11}) невырожденная, система (2.95) имеет единственное решение $\hat{y}_1$. Следовательно, нормальное псевдорешение системы (2.94), соответствующее $\min\|\hat{y}\|$, получим, когда $\|\hat{y}_2\|=0$, т. е.

$$y^0=(\hat{y}_1\bar{o}_{n-r})^{\tau}. \tag{2.96}$$

Установим взаимосвязь между нормальным псевдорешением исходной системы и системы (2.94). Так как $\hat{x}=W^{\tau}\hat{y}$, то

$$\hat{x}=W^{\tau}y_1+W^{\tau}\hat{y}_2 \text{ и } \min(\hat{x}, \hat{x})=$$
$$=\min(W^{\tau}\hat{y}, W^{\tau}\hat{y}) \Leftrightarrow \min(\hat{x}, \hat{x})=\min(\hat{y}, \hat{y}).$$

Отсюда имеем

$$x^0=W^{\tau}y^0. \tag{2.97}$$

На основании полученной формулы строится алгоритм вычисления псевдорешения СЛАУ общего вида при условии, что ранг матрицы системы известен и выполняется неравенство $\|A^{+}\|_2\|Q\|_2<1$. В этом алгоритме должны последовательно реализовываться следующие этапы вычислений:

1) ортогональное разложение матрицы,
2) решение системы (2.95),

3) вычисление нормального псевдорешения x^0 по формуле (2.97).

Естественно, что по приведенному алгоритму, если $r < \min(m, n)$, вычисляется только нормальное псевдорешение. Для нахождения же общего псевдорешения требуется дополнительная информация при постановке точной задачи.

Если кроме исходной системы $Ax=b$ будем также находить нормальное псевдорешение для систем вида

$$Au_j = e_j,$$

где $e_j = (0 \ldots 010 \ldots 0)^\tau$, $j = 1 \div n$, то по описанному выше алгоритму можно вычислить и псевдообратную матрицу A^+. При этом первый этап вычисления, связанный с ортогональным разложением матрицы A, будет для всех значений j один и тот же.

Рассмотрим теперь случай, когда $r \leqslant \min(m, n)$ и $\|A^+\|_2 \|Q\|_2 > 1$, т. е. исходная система в общем неполного ранга и плохо обусловлена. Тогда в отличие от приведенного выше алгоритма на этапе ортогонального разложения матрицы A для каждого образованного вектора-столбца матрицы R будем производить сравнение евклидовой нормы этого вектора с заданным пороговым значением ε, т. е. выполнять операцию

$$r^2_{1j} + r^2_{2j} + \ldots + r^2_{jj} < \varepsilon \text{ для } j = 1 \div m, \tag{2.98}$$

где r_{ij} — элементы матрицы R. Если условие (2.98) выполняется, то координаты r_{ij} вектора r_j приравниваются нулю. В результате будет получена система

$$H^\tau \tilde{R} W z - b = v \Leftrightarrow \tilde{R} W z - Hb = Hv, \tag{2.99}$$

где $\tilde{R}$ — матрица размера $[m \times n]$, имеющая вид

$$\tilde{R} = \begin{pmatrix} \tilde{R}_{11} & O \\ O & O \end{pmatrix}.$$

Здесь $\tilde{R}_{11}$ — верхняя треугольная матрица размера $[k \times k]$ и ранга k, причем k является псевдорангом матрицы A, т. е. $k \leqslant \operatorname{Rg} A$. Если r — истинный ранг исходной системы, то $r \geqslant k$ и элементы r_{ij} матрицы R_{11} будут связаны с элементами $\tilde{r}_{ij}$ матрицы $\tilde{R}_{11}$ следующими соотношениями:

$$\tilde{r}_{ij} = r_{ij} \text{ для } j = 1 \div k,$$
$$\tilde{r}_{ij} = 0 \text{ для } i > j \text{ или } j > k.$$

Решение системы (2.99) по методу наименьших квадратов приводит к нормальной системе уравнений

$$\tilde{R}^\tau \tilde{R} \hat{u} = \tilde{R}^\tau \bar{b}, \tag{2.100}$$

где $W\hat{z} = \hat{u}$, откуда $\hat{z} = W^\tau \hat{u}$. Представляя вектор $\hat{u}$ через под-

векторы $\hat{u}_1$ размерностью k и $\hat{u}_2$ размерностью $n-k$, т. е. как $u=(\hat{u}_1, \hat{u}_2)$, получаем

$$\hat{z}=W^{\tau}\hat{u}_1+W^{\tau}\hat{u}_2. \tag{2.101}$$

Причем подвектор $\hat{u}_1$ находится из уравнения

$$\bar{R}^{\tau}{}_{11}\bar{R}_{11}\hat{u}_1=\bar{R}^{\tau}{}_{11}\bar{b}_1.$$

Исходя из (2.100) и (2.101), имеем

$$\text{а) } u^0=(\hat{u}_1,\ o_{n-k}), \quad \text{б) } z^0=W^{\tau}u^0. \tag{2.102}$$

Сравнение величин u^0 и z^0 с величинами y^0 и x^0, определяемыми по формулам (2.96) и (2.97), дает

$$u_i^0=y_i^0 \text{ и } x_i^0=z_i^0 \text{ для } i=1\div k. \tag{2.103}$$

Таким образом, для плохо обусловленной системы $Ax=b$ общего вида можно устойчиво определить k координат вектора, соответствующего ее нормальному псевдорешению. Для нахождения же общего псевдорешения этой системы требуется дополнительная информация при постановке исходной задачи. Алгоритм для вычисления k координат нормального псевдорешения плохо обусловленной системы состоит, как и при определении нормального псевдорешения для системы неполного и заданного ранга, из трех этапов. На первом этапе осуществляется ортогональное разложение исходной матрицы системы с учетом проверки условия (2.98), на втором вычисляется u_0 и на третьем — z^0.

Наиболее трудоемким с вычислительной точки зрения является этап ортогонального разложения матрицы системы. Процесс сингулярного разложения матрицы будем также причислять к процессу ортогонального разложения. При этом на сингулярное разложение требуется большее число арифметических операций, чем на ортогональное разложение, приводящее к треугольной матрице, например QR — алгоритм. Однако сингулярное разложение позволяет найти сингулярные числа матрицы, что весьма важно для решения ряда вычислительных задач, например при оценке точности и псевдообращении. При решении задачи псевдообращения сингулярному разложению целесообразнее придать несколько видоизмененную форму, чем (1.91), которая описывается как $A=UDV^{\tau}$, где D — матрица размера $[n\times m]$, у которой $d_{ii}=\rho_i$ и $d_{ij}=0$ при $i\neq j$, причем $\rho_i\neq 0$ при $i\leqslant k$ и $\rho_i=0$ при $i>k$. Для этого матрицу D представим в виде

$$D=(E_{k\times k}O_{(n-k)\times k})^{\tau}D_k(E_{k\times k}O_{k(m-k)}), \tag{2.104}$$

где D_k — диагональная матрица размера $[k\times k]$ с элементами $d_{ii}=\rho_i\neq 0$. Подставляя в сингулярное разложение (1.91) выра-

жение, определяемое D по формуле (2.104), получим разложение вида

$$A=U_1(D_kV_1^{\mathrm{T}}), \tag{2.105}$$

где $U_1=U(E_{k\times k}O_{(n-k)\times k})^{\mathrm{T}}$ — матрица размера $[n\times k]$ с ортогональными столбцами, $V_1^{\mathrm{T}}=(E_{k\times k}O_{k\times(m-k)})V^{\mathrm{T}}$ — матрица размера $[k\times m]$ с ортогональными строками. Так как размер матрицы $D_kV_1^{\mathrm{T}}$, входящей в качестве второго сомножителя в разложение (2.105), равен $[k\times m]$, то разложение (2.105) является скелетным разложением матрицы A. Воспользовавшись формулами (1.120) и (1.132), для псевдообратной матрицы A^+ получим

$$A^+=(V_1^{\mathrm{T}})^+D_k^+U_1^+ \Leftrightarrow A^+=V_1D_k^{-1}U_1^{\mathrm{T}}. \tag{2.106}$$

Таким образом, если для матрицы A получено сингулярное разложение (1.91), то нахождение матрицы A^+ по формулам (2.106) не приводит к заметным вычислительным затратам.

Другой способ нахождения псевдообратной матрицы, основанный на ортогональном разложении матрицы, приводящем к виду (1.32), был рассмотрен в § 1.10.

Естественно, что после определения псевдообратной матрицы вычисление псевдорешения становится несложной задачей. Еще раз отметим, что если матрица A неполного ранга или плохо обусловлена, то вычисляемая для нее псевдообратная матрица будет иметь псевдоранг, равный k и определяемый в процессе ортогонального разложения согласно проверке условия (2.98) или проверке условия $\rho_i<\varepsilon$ при сингулярном разложении.

Оценку решения по методу наименьших квадратов задачи $\mathcal{A}x=b$ неполного ранга или плохо обусловленной, если вычисления выполняются на ЭВМ с предельной относительной ошибкой округления $\bar{\varepsilon}_{\text{окр}}$ по алгоритму ортогонального разложения матрицы системы, можно выполнить по формуле, полученной в [11]:

$$\bar{\varepsilon}\frac{\|x^0-z^0\|_E}{\|x^0\|_E}\leqslant(6m-6k+43)k\bar{\varepsilon}_{\text{окр}}. \tag{2.107}$$

где k — псевдоранг матрицы A.

Приведем несколько простых приемов, позволяющих в ряде случаев повысить устойчивость и вычислительную точность решения СЛАУ.

Пусть задана система $Ax\cong b$, тогда можно выполнить масштабирование строк в исходной системе за счет перехода к системе вида

$$GAx-Gb=Gr \tag{2.108}$$

и решать эту систему по методу наименьших квадратов при условии минимизации величин

$$\| r \|^2_E = \| Gr \|^2_E. \tag{2.109}$$

Если G — диагональная матрица, то условие (2.109) соответствует

$$\min \sum_{i=1}^{m} g_{ii}^2 r_i^2$$

и псевдорешение уравнения (2.108) находится как решение нормальной системы уравнений вида

$$A^T W A x = A^T W x. \tag{2.110}$$

При этом матрица $W = G^T G$ задается или из известной статистической информации относительно неопределенности в векторе b, или из анализа величин элементов исходной матрицы системы. Например, стремятся сделать так, чтобы евклидовы нормы каждого вектора-строки исходной матрицы имели одинаковый порядок по величине.

В исходной матрице системы $Ax \cong b$ можно предусмотреть при необходимости и масштабирование строк. В этом случае ее решение заменяется на решение системы вида

$$\tilde{A}\tilde{x} \cong \tilde{b}, \tag{2.111}$$

где $\tilde{A} = AG, \quad \tilde{b} = b - A\eta, \quad x = G\tilde{x} + \eta.$

Предполагая, что G — матрица размера $[m \times m]$ и $\mathrm{Rg}\, G = m$, получим

$$\| \tilde{A}\tilde{x} - \tilde{b} \|_E = \| Ax - b \|_E.$$

Отсюда следует, что вместо решения системы $Ax \cong b$ по методу наименьших квадратов можно решать систему $\tilde{A}\tilde{x} \cong b$, а затем получать решение исходной системы в результате преобразования

$$x = G\tilde{x} + \eta.$$

Так, если столбцы матрицы A значительно различаются по величине их евклидовой нормы, то, задавая матрицу G в виде диагональной матрицы, можно получить такую матрицу $\tilde{A} = AG$, у которой евклидовы нормы столбцов будут по величине уже одного порядка.

Пример. Методом Гаусса решим систему уравнений

$$42x_1 + 2x_2 = 6,$$

$$56x_1 + 3x_2 = -5,$$

сохраняя при вычислениях три знака после запятой и при округлении используя метод «отбрасывания». После вычислений получим $x_1 = 1{,}796$ и $x_2 =$

$=35{,}195$. Сравнение этих результатов с точными значениями $\hat{x}_1=2$ и $\hat{x}_2=$ $=-39$ дает относительные ошибки вычисленных значений $\varepsilon_{x_1}=10\%$ и $\varepsilon_{x_2}=$ $=10\%$. Выполним теперь масштабирование столбцов в приведенной системе по формулам (2.111). При этом положим

$$G=\begin{pmatrix}0{,}1 & 0\\ 0 & 1\end{pmatrix},\quad \eta=0.$$

В результате получим новую систему вида

$$4{,}2y_1+2y_2=6,$$
$$5{,}6y_1+3y_2=-5,$$

которую будем решать при условиях, заданных в начале. Тогда получим $y_1=19{,}893$, $y_2=38{,}793$. Отсюда по формуле $x=Gy$ получаем решение $x_1=$ $=1{,}989$ и $x_2=38{,}793$ с относительными ошибками $\varepsilon_{x_1}=0{,}5\%$ и $\varepsilon_{x_2}=0{,}5\%$.

Глава 3

ПРЯМЫЕ МЕТОДЫ ЛИНЕЙНОЙ АЛГЕБРЫ

Прямыми вычислительными методами будем называть такие, которые при предположении отсутствия ошибок округления приводят к точным решениям. Естественно, что сделанное предположение является чисто гипотетическим, так как при реальных вычислениях оно никогда не выполняется. Поэтому с точки зрения точности решения нельзя отдать предпочтение какому-либо прямому методу по отношению, например, к итерационному или другому прямому методу без предварительных оценок ошибок округления.

§ 3.1. МАТРИЧНЫЕ СИСТЕМЫ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ. ПРЯМЫЕ МЕТОДЫ ОБРАЩЕНИЯ МАТРИЦ

Пусть задана система линейных алгебраических уравнений, у которой $m=n$ и определитель матрицы A отличен от нуля. Решение такой системы будет единственным и иметь вид

$$x=A^{-1}b. \tag{3.1}$$

Как известно из курса линейной алгебры, при поставленных выше условиях решение СЛАУ может быть найдено по формулам Крамера

$$x_i=\det A_i/\det A, \tag{3.2}$$

где $\det A_i$ — определитель, полученный из определителя матрицы A путем замены его i-го столбца столбцом свободных членов b.

Формулы Крамера имеют довольно простой вид, однако при их использовании необходимо производить большой объем вычислений $n!(n^2+1)+n$, что делает их практически неприменимыми при решении систем с числом уравнений более 5.

Решение системы в матричной форме (3.1) связано с нахождением обратной матрицы A^{-1}. Вычисление элементов обратной матрицы называется *обращением матрицы.*

В общем курсе линейной алгебры рассматривается метод обращения матриц, основанный на вычислении элементов a'_{ij} матрицы A^{-1} по формуле

$$a'_{ij}=A_{ji}/\det A, \tag{3.3}$$

где $i, j=1\div n$, A_{ks} — алгебраические дополнения соответствующих элементов a_{ks} матрицы A.

Обращение матриц по формуле (3.3) становится неэффективным при $n>5$ из тех же соображений, что и для решения систем по методу Крамера.

Как правило, процесс обращения матриц является более трудоемким, т. е. требующим выполнения большего числа операций, чем вычислительные процессы решения СЛАУ. Иногда рационально процесс обращения матриц свести к решению систем алгебраических уравнений. Рассмотрим один из таких методов.

На основании формулы (1.118) можно написать следующее уравнение:

$$AX=E, \tag{3.4}$$

где X — матрица, обратная матрице A, т. е. $X=A^{-1}$; E — единичная матрица.

Уравнение (3.4) является матричной системой линейных алгебраических уравнений. Представляя каждый столбец матрицы X и E как вектор-столбец, получим

$$A(\bar{x}_1\bar{x}_2\ldots\bar{x}_j\ldots\bar{x}_n)=(\bar{e}_1\bar{e}_2\ldots\bar{e}_j\ldots\bar{e}_n),$$

$$\text{где } \bar{x}_j=(x_{1j}x_{2j}\ldots x_{ij}\ldots x_{nj})^{\mathrm{T}},$$

$$\bar{e}_j=(0\,0\ldots1\ldots0)^{\mathrm{T}},\ j=1\div n.$$

Умножая в этих уравнениях каждый вектор-столбец x_j на матрицу A и используя понятие равенства векторов, имеем

$$A\bar{x}_j=\bar{e}_j,\quad j=1\div n. \tag{3.5}$$

Таким образом, для вычисления элементов матрицы X требуется решить n систем линейных алгебраических уравнений вида (3.5).

§ 3.2. ОБРАЩЕНИЕ МАТРИЦ СВЕДЕНИЕМ МАТРИЦЫ К ПРОИЗВЕДЕНИЮ МАТРИЦ ТРЕУГОЛЬНОГО ВИДА

В начале рассмотрим вопрос об обращении матриц специального вида — треугольных матриц. В курсах линейной алгебры доказывается, что сумма и произведение двух верхних (нижних) треугольных матриц одного и того же порядка есть также верхняя (нижняя) треугольная матрица того же порядка и обратная матрица к невырожденной верхней (нижней) треугольной матрице есть также верхняя (нижняя) треугольная матрица. Пользуясь последним обстоятельством и определением обратной матрицы, имеем

$$\begin{pmatrix} a'_{11} & a'_{12} & \dots & a'_{1n} \\ 0 & a'_{22} & \dots & a'_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a'_{nn} \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}, \tag{3.6}$$

где a_{ij}' — элементы матрицы A^{-1}, обратной матрице $A=\{a_{ij}\}$. Пользуясь правилом перемножения двух матриц и определением равенства двух матриц, получаем из (3.6) систему линейных уравнений относительно a_{ij}' для верхней треугольной матрицы. Решением этой системы уравнений является

$$\begin{gathered} a'_{ij}=0 \quad \text{при } i>j, \quad a'_{ij}=a_{ii}^{-1} \quad \text{при } i=j, \\ a'_{ij}=-a_{jj}^{-1}\sum_{k=i}^{j-1} a_{kj}a'_{ik} \text{ при } i<j. \end{gathered} \tag{3.7}$$

Аналогичным образом можно найти формулы для обращения нижней треугольной матрицы $C=\{c_{ij}\}$. Если элементы матрицы C^{-1} обозначить c_{ij}', то формулы примут вид

$$\begin{gathered} c'_{ij}=0 \text{ при } i<j, \quad c'_{ij}=c_{ii}^{-1} \text{ при } i=j, \\ c'_{ij}=-c_{ii}^{-1}\sum_{k=j}^{i-1} c_{ik}c'_{kj} \text{ при } i>j. \end{gathered} \tag{3.8}$$

Как следует из формул (3.7) и (3.8), порядок определения элементов обратной матрицы A^{-1} (или C^{-1}) состоит в последовательном вычислении сначала элементов обратной матрицы, расположенных в первой строке, затем во второй строке и т. д. до n-й строки, в результате чего для вычисления элементов a_{ij}',

$j=i\div n$ потребуется выполнить $N_i=1+2+\ldots+(n+1-i)$ операций умножения и деления. Тогда общее число операций умножения и деления, необходимое для обращения треугольной матрицы, будет

$$N_{\text{обр.т}}=\sum_{i=1}^{n}[1+2+\ldots+(n+1-i)].$$

Учитывая, что выражение, стоящее в квадратных скобках, является арифметической прогрессией, получаем

$$N_{\text{обр. т}}=\frac{(n+1)n+n(n-1)+\ldots+(n+1-k)(n-k)}{2}+\ldots+1.$$

Преобразуя данное выражение с учетом

$$\sum_{k=1}^{n}k^2=\frac{n^3}{3}+\frac{n^2}{2}+\frac{n}{6}, \tag{3.9}$$

получаем формулу для подсчета $N_{\text{обр.т}}$

$$N_{\text{обр. т}}=\frac{1}{2}\sum_{k=1}^{n}(n+2-k)(n+1-k)=\frac{(n+2)(n+1)n}{6}. \tag{3.10}$$

Теорема 3.1. Любую квадратную матрицу, имеющую отличные от нуля главные диагональные миноры, т. е.

$$\Delta_1=a_{11}\neq 0,\ \Delta_2=\begin{vmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\end{vmatrix}\neq 0$$

$$\Delta_k=\begin{vmatrix}\Delta_{11}&a_{12}&\ldots&a_{1k}\\a_{21}&a_{22}&\ldots&a_{2k}\\\ldots&\ldots&\ldots&\ldots\\a_{k1}&a_{k_2}&\ldots&a_{kk}\end{vmatrix}\neq 0,\ \ldots,\ \det A\neq 0$$

можно представить в виде произведения верхней (нижней) треугольной матрицы на нижнюю (верхнюю) треугольную матрицу. Причем это разложение будет единственным, если зафиксировать отличными от нуля числами диагональные элементы одной из треугольных матриц.

На основании данной теоремы представим матрицу A, главные диагональные миноры которой отличны от нуля, в виде произведения двух треугольных матриц

$$\begin{pmatrix}a_{11}&a_{12}&\ldots&a_{1n}\\a_{21}&a_{22}&\ldots&a_{2n}\\\ldots&\ldots&\ldots&\ldots\\a_{n1}&a_{n2}&\ldots&a_{nn}\end{pmatrix}=$$

$$=\begin{pmatrix}c_{11}&0&\ldots&0\\c_{21}&c_{22}&\ldots&0\\\ldots&\ldots&\ldots&\ldots\\c_{n1}&c_{n2}&\ldots&c_{nn}\end{pmatrix}\begin{pmatrix}b_{11}&b_{12}&\ldots&b_{1n}\\0&b_{22}&\ldots&b_{2n}\\\ldots&\ldots&\ldots&\ldots\\0&0&\ldots&b_{nn}\end{pmatrix}. \tag{3.11}$$

После перемножения матриц $C=\{c_{ij}\}$ и $B=\{b_{ij}\}$ переходим к системе линейных уравнений относительно неизвестных c_{ij} и b_{ij}, состоящей из n^2 уравнений и $N=2(1+2+3\ldots n)=n^2+n$ неизвестных. Так как число уравнений меньше на n числа неизвестных, то зафиксируем n неизвестных, например приняв $b_{ii}=1$ для любого i. Тогда решением полученной системы будет

$$b_{ij}=1 \quad \text{при } i=j,$$

$$b_{1j}=a_{1j}a_{11}^{-1}, \qquad b_{ij}=(a_{ij}-\sum_{k=1}^{i-1}c_{ik}b_{kj})c_{ii}^{-1}$$

$$\text{при } i<j, \quad i=2\div n, \quad j=3\div n. \tag{3.12}$$

$$c_{i1}=a_{i1}, \qquad c_{ij}=a_{ij}-\sum_{k=1}^{j-1}c_{ik}b_{kj}$$

$$\text{при } i\geqslant j,\ i,\ j=2\div n.$$

Последовательность определения элементов c_{ij} и b_{ij} по формулам (3.12) указывается с помощью пронумерованных стрелок в схеме, изображенной на рис. 8.

Так, при нечетных **шагах** $2k$—1 последовательно вычисляются элементы c_{ik}, а при четных шагах $2k$ последовательно вычисляются элементы b_{kj}.

Число операций умножения и деления, которое необходимо выполнить для представления матрицы A в виде произведения двух треугольных матриц C и B, подсчитываем по формуле

$$N_{CB}=2[(n-1)+2(n-2)+3(n-3)+\ldots+(n-1)]=$$

$$=2\sum_{k=1}^{n}(n-k)k=\frac{n^3-n}{3}. \tag{3.13}$$

Пользуясь формулой (3.11), находим матрицу A^{-1}, обратную матрице A по формуле

$$A^{-1}=B^{-1}C^{-1}. \tag{3.14}$$

Таким образом, обращение квадратной матрицы A, главные диагональные миноры которой отличны от нуля, сводится к следующему:

представление матрицы A в виде произведений двух треугольных матриц ($A=CB$) по формулам (3.12);

обращение матрицы C по формулам (3.8);

обращение матрицы B по формулам (3.7);

перемножение треугольных матриц B^{-1} и C^{-1} по формулам (3.16).

Одним из этапов обращения матриц является перемножение двух треугольных матриц. В общем случае перемножение двух

прямоугольных матриц $A=BC$ выполняется по формуле

$$a_{ij}=(\bar{b}_i,\ \bar{c}_j),\quad i=1\div m,\quad j=1\div l, \tag{3.15}$$

где a_{ij} — элементы матрицы A, $\bar{b}_i$ — i-й вектор-строка матрицы B; $\bar{c}_j$ — j-й вектор-столбец матрицы C. Если размер векторов b_i и c_i равен n, то для вычисления элементов a_{ij} необходимо выполнить $N_у=mnl$ операций умножения, а при $m=n=l$ число этих операций будет равно n^3.

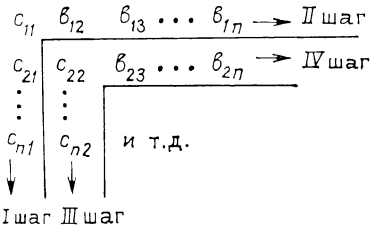


Рис. 8. Схема вычисления элементов c_{ij} и b_{ij}

При перемножении верхней и нижней треугольных матриц число выполняемых операций по сравнению с перемножением матриц общего вида можно сократить за счет исключения операций умножения на нулевые элементы, характеризующие эти матрицы. В результате при перемножении матриц B^{-1} и C^{-1} для элементов a_{ij}' матрицы A^{-1} получим

$$\begin{aligned} a'_{ij}&=\sum_{k=j}^{n} b'_{ik}c'_{kj} \quad \text{при } i\leqslant j,\\ a'_{ij}&=\sum_{k=i}^{n} b'_{ik}c'_{kj} \quad \text{при } i>j. \end{aligned} \tag{3.16}$$

Число операций умножения, необходимое выполнить для определения всех элементов i-й строки матрицы, получающейся от перемножения верхней и нижней треугольных матриц, имеют вид

$$N_i=[(n+1-i)+(n-i)+\ldots+1]+(i-1)(n+1-i)=$$
$$=\frac{n(n+1)-i^2+i}{2}.$$

В результате число операций умножения для перемножения верхней и нижней треугольных матриц будет определяться по формуле

$$N_{у.т}=\sum_{i=1}^{n} N_i=\frac{2n^3+3n^2+n}{6}=\frac{(2n+1)(n+1)}{6}. \tag{3.17}$$

Теперь можно найти общее число операций умножения и деления, которое потребуется для того, чтобы обратить матрицу A по приведенному выше алгоритму:

$$T_{обр}=N_{св}+2N_{обр.т}+N_{у.т}=$$
$$=\frac{2n^3+3n^2+n}{2}=\frac{(2n+1)(n+1)n}{2}. \tag{3.18}$$

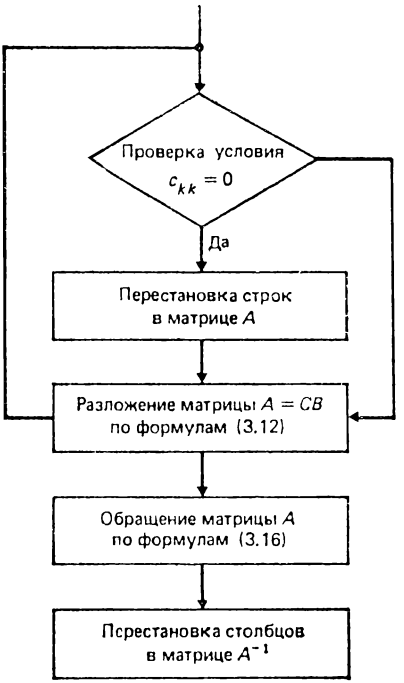


Рис. 9. Блок-схема алгоритма обращения матрицы разложением ее на произведение двух треугольных матриц

Однако обращение матриц по описанному алгоритму, как было отмечено выше, можно выполнить только при условии, что главные диагональные миноры матрицы A отличны от нуля. Если же это условие не выполняется, например для k-го главного диагонального минора, то при вычислении элемента c_{kk} матрицы C по формулам (3.12) он станет равен нулю, что, в свою очередь, приведет к невозможности вычисления элементов b_{kj} матрицы B. Поэтому необходимо исключить подобную ситуацию. Одним из таких путей является осуществление проверки условий неравенства нулю главных диагональных миноров до разложения $A=CB$. Однако такая проверка приведет к вычислениям, превышающим вычисления, непосредственно связанные с разложением $A=CB$. Поэтому с вычислительной точки зрения рациональнее проводить проверку условия $c_{kk}=0$ непосредственно в процессе вычисления элементов c_{ij} и b_{ij}. Тогда при выполнении этого условия требуется переставлять соответствующую k-ю строчку матрицы A с последующими $k+s$ строчками, где $s=1\div(n-k)$, до тех пор пока не будет выполнено условие $c_{k+s,k+s}\neq 0$. Отметим, что при невырожденности матрицы A такая перестановка всегда найдется. Этот процесс можно описать как

$$P_{k,k+s}A=\bar{A}, \tag{3.19}$$

где $P_{k,k+s}$ — матрица перестановок. В результате будет обращаться матрица $\bar{A}$. Переход от матрицы $\bar{A}^{-1}$ к матрице A^{-1} можно выполнить по формуле

$$A^{-1}=\bar{A}^{-1}P_{k,k+s}. \tag{3.20}$$

Таким образом, матрица A^{-1} будет получена после перестановки соответствующих столбцов в матрице $\bar{A}^{-1}$. Блок-схема алгоритма обращения любой невырожденной матрицы приведена на рис. 9.

Рассмотрим вопрос об обращении симметричных матриц.

Если матрица A симметричная, то, как известно из курса линейной алгебры, ее можно представить в виде произведения двух транспонированных между собой треугольных матриц, т. е.

$$A = T'T, \tag{3.21}$$

где

$$T = \begin{pmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ 0 & t_{22} & \dots & t_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & t_{nn} \end{pmatrix}, \quad T' = \begin{pmatrix} t_{11} & 0 & \dots & 0 \\ t_{12} & t_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ t_{1n} & t_{2n} & \dots & 0 \end{pmatrix}.$$

Перемножая матрицы T' и T, получим

$$t^2_{1i} + t^2_{2i} + \dots + t^2_{ii} = a_{ii},$$
$$t_{1i}t_{1j} + t_{2i}t_{2j} + \dots + t_{ii}t_{ij} = a_{ij}, \quad j > i,$$
$$t_{ij} = 0, \quad j < i,$$

где t_{ij} — элементы матрицы T.

После решения этой системы имеем

$$t_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} t^2_{ki}},$$

$$t_{ij} = (a_{ij} - \sum_{k=1}^{i-1} t_{ki}t_{kj})t_{ii}^{-1}, \quad j > i; \quad t_{ij} = 0, \quad j < i. \tag{3.22}$$

Если матрица A является симметричной и положительно определенной, то подкоренные выражения в формулах (3.22) положительные.

Чтобы вычислить элементы t_{ij}, потребуется выполнить n операций извлечения квадратного корня и $N_{к.к}$ операций умножения и деления:

$$N_{к.к} = \frac{n^3 + 3n^2 + 2n}{6} = \frac{(n+2)(n+1)n}{6}. \tag{3.23}$$

Для получения элементов матрицы A^{-1} воспользуемся формулой

$$A^{-1} = (T^{-1})(T^{-1})'. \tag{3.24}$$

В частности, из этой формулы следует, что если матрица A симметричная, то и матрица A^{-1} симметричная. Следовательно, для определения элементов матрицы A^{-1} достаточно вычислить a_{ij}' только при $j \geqslant i$. Поэтому при перемножении матриц T^{-1} и $(T^{-1})'$ можно воспользоваться формулами (3.16), но при $j \geqslant i$.

В этом случае число операций умножения при перемножении этих матриц будет определяться выражением

$$N_{\text{у. к. к}} = \frac{n^3 + 3n^2 + 2n}{6} = \frac{(n+2)(n+1)}{2}. \qquad (3.25)$$

Общее число операций умножения и деления, выполняемых при обращении симметричных матриц по изложенному алгоритму, подсчитывается по формуле

$$N_{\text{обр к.к}} = N_{\text{к.к}} + N_{\text{обр.т}} + N_{\text{у.к.к}} = \\ = \frac{n^3 + 3n^2 + 2n}{2} = \frac{(n+2)(n+1)n}{2}. \qquad (3.26)$$

Кроме того, при обращении симметричных матриц вычислительный процесс организуется таким образом, чтобы в ЭВМ держать на фиксированном поле памяти только информацию о ненулевых элементах треугольной матрицы, т. е. объем этого поля будет определяться $n(n+1)/2$ ячейками памяти ЭВМ.

§ 3.3. КЛЕТОЧНЫЙ МЕТОД ОБРАЩЕНИЯ МАТРИЦ

При решении систем линейных уравнений и обращении матриц высокого порядка, когда вся исходная информация не может быть размещена в оперативной памяти ЭВМ, целесообразно вычислительный процесс построить таким образом, чтобы обработка исходной информации осуществлялась по частям. Это можно осуществить, применяя клеточные методы решения систем и обращения матриц.

Пусть дана квадратная матрица A. Разобьем ее на матрицы низших порядков (клетки) с помощью горизонтальных и вертикальных перегородок, идущих вдоль всей матрицы. Причем будем рассматривать только такие разбиения, при которых диагональные клетки являются квадратными матрицами. Пусть

$$A = \left(\begin{array}{cc|c|cc} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ \hline a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ \hline a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{array}\right)$$

при приведенном способе разбиения имеет вид

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}.$$

Пусть в результате разбиения двух матриц одного и того же порядка $A[n\times n]$ и $B[n\times n]$ имеем

$$A=\begin{pmatrix} A_{11} & A_{12} & \dots & A_{1s} \\ A_{21} & A_{22} & \dots & A_{2s} \\ A_{s1} & A_{s2} & \dots & A_{ss} \end{pmatrix}, \quad B=\begin{pmatrix} B_{11} & B_{12} & \dots & B_{1s} \\ B_{21} & B_{22} & \dots & B_{2s} \\ B_{s1} & B_{s2} & \dots & B_{ss} \end{pmatrix},$$

причем диагональными клетками в данном разложении являются матрицы $A_{ii}[n_i\times n_i]$ и $B_{ii}[n_i\times n_i]$, где $\sum\limits_{i=1}^{s} n_i=n$. Тогда под суммой и произведением клеточных матриц A и B будем соответственно принимать матрицы C и D:

$$C=A+B, \quad D=AB,$$

где клетки C_{ij} и D_{ij} матриц C и D определяются по формулам

$$C_{ij}=A_{ij}+B_{ij}, \quad D_{ij}=\sum_{k=1}^{s} A_{ik}D_{kj}. \tag{3.27}$$

Пусть матрица A представлена в клеточном виде

$$A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

где диагональными клетками матрицы A являются матрицы $A_{11}[(n-r)\times(n-r)]$ и $A_{22}[r\times r]$. Будем искать матрицу A^{-1}, обратную матрице A, в виде клеточной матрицы

$$A^{-1}=\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \tag{3.28}$$

в которой диагональными клетками являются матрицы $C_{11}[(n-r)\times(n-r)]$ и $C_{22}[r\times r]$. На основании определения обратной матрицы имеем

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}=\begin{pmatrix} E_{n-r} & O_{n-r,\,r} \\ O_{r,\,n-r} & E_r \end{pmatrix}.$$

Выполнив перемножение и сложение матриц, стоящих в левой части этого равенства, по формулам (3.27) получаем систему матричных уравнений, неизвестными в которой являются матрицы C_{ij}:

$$\begin{aligned} &A_{11}C_{11}+A_{12}C_{21}=E_{n-r}, \\ &A_{11}C_{12}+A_{12}C_{22}=O_{n-r,r}, \\ &A_{21}C_{11}+A_{22}C_{21}=O_{r,n-r}, \\ &A_{21}C_{12}+A_{22}C_{22}=E_r. \end{aligned} \tag{3.29}$$

Из второго уравнения этой системы получаем

$$C_{12} = -A_{11}^{-1}A_{12}C_{22}.$$

Подставив данное выражение для C_{12} в последнее уравнение системы (3.29), решим его относительно C_{22}, тогда

$$C_{22} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}.$$

Аналогично из первого и третьего уравнений системы (3.29) находим

$$C_{11} = A_{11}^{-1} - A_{11}^{-1}A_{12}C_{21}, \quad C_{21} = -C_{22}A_{21}A_{11}^{-1}.$$

Таким образом, клетки матрицы A^{-1} могут быть определены последовательно по формулам

$$\begin{gathered} P_1 = A_{11}^{-1}A_{12}, \quad P_2 = A_{22} - A_{21}P_1, \quad P_3 = A_{21}A_{11}^{-1}, \\ C_{22} = P_2^{-1}, \quad C_{12} = -P_1C_{22}, \quad C_{21} = -C_{22}P_3, \\ C_{11} = A_{11}^{-1} - P_1C_{21}. \end{gathered} \tag{3.30}$$

Если порядок матриц A_{11} и P_2 в формулах (3.30) остается все еще большим, чтобы можно было их обратить с использованием только оперативной памяти ЭВМ, то для обращения этих матриц будем использовать метод разбиения матриц на клетки и затем по описанной выше процедуре осуществлять обращение матриц A^{11} и P_2. Этот процесс продолжается до тех пор, пока для обращения матриц можно будет использовать только оперативную память ЭВМ.

В клеточном методе обращения матриц большое значение имеет способ разбиения матриц на клетки. При выборе этого способа обычно руководствуются двумя соображениями: во-первых, обеспечением минимального числа обращений к внешнему запоминающему устройству ЭВМ, в котором хранятся элементы основной матрицы A, а также результаты промежуточных вычислений, и, во-вторых, выделением с помощью разбиения такой матрицы A_{11}, которую легко можно было бы обратить с помощью простых вычислительных операций, например если A_{11} — диагональная или треугольная матрица. Если в матрице A нет возможности за счет ее разбиения на клетки выделить особый тип матрицы A_{11}, который имеет простую вычислительную форму обращения, то в силу первого соображения разбиение должно быть таким, чтобы матрицы A_{11} и P_2, а следовательно, и A_{22} были приблизительно одной размерности.

Когда матрица A_{11} является вырожденной или когда наиболее простой для обращения клеткой является клетка $A_{kl} \neq A_{11}$, следует от матрицы A перейти к матрице $\tilde{A}$, в которой $\tilde{A}_{11} = A_{kl}$, за счет блочной перестановки

$$\tilde{A} = P_{1k}AQ_{1l}, \tag{3.31}$$

где P_{1k} и Q_{1l} — матрицы блочной перестановки, первая из которых переставляет клетки, расположенные на первой и k-й строках матрицы A, а вторая — клетки, расположенные после строчной перестановки в первом и l столбцах. Затем обращается матрица $\tilde{A}$ по формулам (3.30) и осуществляется переход к матрице A^{-1} по формуле

$$A^{-1}=Q_{1l}\tilde{A}^{-1}P_{1k}.$$

В качестве примера рассмотрим матрицу

$$A=\left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ -1 & 0 & -1 & 0 & 1 & 2 \\ \hline 1 & 0 & 0 & 1 & 2 & 1 \\ 0 & 1 & 0 & -1 & 1 & 2 \\ 0 & 0 & 1 & 2 & 0 & 1 \end{array}\right),$$

в которой после разбиения ее на клетки видно, что клетка A_{21} является единичной. Перейдем к матрице $\tilde{A}$, в которой клетка $\tilde{A}_{11}=A_{21}$:

$$\tilde{A}=\begin{pmatrix} O_{3\times 3} & E_{3\times 3} \\ E_{3\times 3} & O_{3\times 3} \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}=\begin{pmatrix} E & A_{22} \\ A_{11} & A_{12} \end{pmatrix}=$$

$$=\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 2 & 1 \\ 0 & 1 & 0 & -1 & 1 & 2 \\ 0 & 0 & 1 & 2 & 0 & 1 \\ \hline 1 & 2 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ -1 & 0 & -1 & 0 & 1 & 2 \end{array}\right).$$

После обращения матрицы $\tilde{A}$ по формулам (3.30), причем $\tilde{A}_{11}^{-1}=E$, находим A^{-1}

$$A^{-1}=\begin{pmatrix} \tilde{C}_{11} & \tilde{C}_{12} \\ \tilde{C}_{21} & \tilde{C}_{22} \end{pmatrix}\begin{pmatrix} O_{3\times 3} & E_{3\times 3} \\ E_{3\times 3} & O_{3\times 3} \end{pmatrix}.$$

Здесь $\tilde{C}_{ij}$ — клетки матрицы $\tilde{A}^{-1}$, вычисляемые по формулам

$$P_1=A_{22},\quad P_2=A_{12}-A_{11}P_1,\quad P_3=A_{11},\ \tilde{C}_{22}=P_2^{-1},$$
$$\tilde{C}_{12}=-P_1\tilde{C}_{22},\quad \tilde{C}_{21}=-\tilde{C}_{22}P_3,\quad \tilde{C}_{11}=E-P_1\tilde{C}_{21}.$$

Непосредственные вычисления по этим формулам предлагаем выполнить самостоятельно.

§ 3.4. АЛГОРИТМ ГАУССА

В предыдущих параграфах этой главы были рассмотрены некоторые прямые методы обращения матрицы. Если обратная матрица найдена, то решение СЛАУ сводится к умножению матрицы на вектор, т. е.

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

При этом количество операций умножения и деления, необходимое для решения СЛАУ, будет

$$N=N_{\text{обр.м}}+N_{\text{у.м.в}}, \tag{3.32}$$

где $N_{\text{обр.м}}$ — количество операций умножения и деления, необходимое для обращения матрицы. Оно при прямом методе обращения пропорционально n^3; $N_{\text{у.м.в}}$ — количество операций умножения, необходимое для умножения матрицы размера $[n\times n]$ на вектор, которое равно n^2.

Однако существуют прямые методы решения СЛАУ, в которых не осуществляется непосредственного обращения матрицы. Наиболее известен метод последовательного исключения неизвестных, получивший название алгоритма Гаусса, и его модификации.

Алгоритм Гаусса состоит из двух основных этапов. На первом из них осуществляется переход от исходной системы $Ax=b$ к эквивалентной

$$Tx=d, \tag{3.33}$$

где T — верхняя треугольная матрица с единичной диагональю. Расширенную матрицу системы (3.33) представим в виде

$$\left\langle\begin{array}{ccccccccc}
1 & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1j}^{(1)} & \dots & a_{1n}^{(1)} & a^{(1)}{}_{1,\,n+1} \\
0 & 1 & a_{23}^{(2)} & \dots & a_{2j}^{(2)} & \dots & a_{2n}^{(2)} & a^{(2)}{}_{2,\,n+1} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & 1 & \dots & a_{ij}^{(i)} & \dots & a_{in}^{(i)} & a^{(i)}{}_{i,\,n+1} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
 & & 0 & & & 1 & a^{(n-1)}{}_{n-1,\,n} & a^{(n-1)}{}_{n-1,\,n+1} \\
 & & & & & & 1 & a^{(n)}{}_{n,\,n+1}
\end{array}\right\rangle \tag{3.34}$$

Этап перехода от исходной системы к эквивалентной (3.33) принято называть *прямым ходом* в алгоритме Гаусса.

Прямой ход в алгоритме Гаусса состоит из n последовательных шагов исключения. На первом шаге переходим от исходной системы к эквивалентной $A^{(1)}x=d^{(1)}$, расширенная матрица которой имеет вид

$$\left(\begin{array}{c|ccc|c} 1 & a_{12}^{(1)} & \dots & a_{1n}^{(1)} & a^{(1)}{}_{1n+1} \\ 0 & & & & \\ \vdots & & a_{ij}^{(1)} & & \\ 0 & & (i=2\div n,\ j=2\div(n+1)) & & \end{array}\right)$$

Здесь элементы $a_{ij}^{(1)}$ находятся по формулам

$$a^{(1)}{}_{1j}=a_{1j}a_{11}^{-1} \text{ для } j=1\div(n+1),$$

$$a^{(1)}{}_{ij}=a_{ij}-a_{i1}a^{(1)}{}_{1j} \text{ для } i=2\div n,\ j=2\div(n+1).$$

на k-м шаге исключения осуществляем переход от системы $A^{(k-1)}x=d^{(k-1)}$ к эквивалентной $A^{(k)}x=d^{(k)}$, расширенную матрицу которой можно представить в виде

$$\left(\begin{array}{cccccc} 1 & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1n}^{(1)} & a^{(1)}{}_{1,\,n+1} \\ & 1 & a_{23}^{(2)} & \dots & a_{2n}^{(2)} & a^{(2)}{}_{2,\,n+1} \\ & & \dots & \dots & \dots & \dots \\ & & & 1 & a_{k_n}^{(k)} & a^{(k)}{}_{k,\,n+1} \\ 0 & & & & a_{ij}^{(k)} & \\ & & & & (i=(k+1)\div n, & \\ & & & & j=(k+1)\div(n+1) & \end{array}\right), \qquad (3.35)$$

где элементы $a_{ij}^{(k)}$ вычисляются по формулам

$$a_{kj}^{(k)}=a_{kj}^{(k-1)}/a_{kk}^{(k-1)} \qquad \text{для } j=(k+1)\div(n+1), \qquad (3.36)$$

$$a_{ij}^{(k)}=a_{ij}^{(k-1)}-a_{ik}^{(k-1)}a_{kj}^{(k)} \text{ для } i=(k+1)\div n,\ j=(k+1)\div(n+1).$$

И, наконец, после $k=n$ шагов исключения получим систему (3.33).

Таким образом, с вычислительной точки зрения прямой ход связан с получением значений $a_{ij}^{(k)}$ по формулам (3.36) для всех $k=1\div n$, причем $a_{ij}^{0}=a_{ij}$. Из формул (3.36) следует, что реализация описанного алгоритма возможна только в случае, когда элементы $a_{kk}^{(k-1)}$, называемые *ведущими элементами*, отличны от нуля при всех значениях k. Для обеспечения работоспособности алгоритма и в случаях $a_{ss}^{(s-1)}=0$ необходимо на каждом таком шаге $k=s$ предусмотреть перестановку строки s со строкой $(s+p)$, где $p=1\div(n-(p+s))$ в расширенной матрице системы. Отметим, что если матрица A невырожденная, то всегда найдется такая перестановка строк s и $s+p$, при которой $a_{s+p,\,s+p}^{(s-1)}\neq 0$.

Подсчитаем необходимое число $N_{\text{пр.х}}$ операций умножения и деления для реализации прямого хода в алгоритме Гаусса. Как следует из формулы (3.36), таких операций на 1-м шаге требуется $[n+n(n-1)]$, на 2-м шаге — $[(n-1)+(n-1)(n-2)]$ и т. д., на k-м шаге — $[(n-k+1)+(n-k+1)(n-k)]$ и, наконец, на n-м шаге требуется одна такая операция. Отсюда общее число будет

$$N_{\text{пр.х}}=\sum_{k=1}^{n}(n+1-k)^2=\sum_{k=1}^{n}k^2=\frac{n^3}{3}+\frac{n^2}{2}+\frac{n}{6}. \qquad (3.37)$$

Второй этап алгоритма Гаусса состоит в решении системы (3.33) в следующей последовательности. Сначала находится значение x_n из последнего уравнения системы, затем x_{n-1} из $n-1$ уравнения и т. д. до значения x_1, которое определяется из первого уравнения системы. Данная процедура, получившая название *обратного хода* алгоритма Гаусса, описывается зависимостями вида

$$x_n=a^{(n)}_{n,n+1},\ \ldots,\ x_k=a^{(k)}_{k,n+1}-$$
$$-\sum_{j=k+1}^{n}a^{(k)}_{kj}x_j \text{ для } k=n-1,\ n-2,\ldots,1. \qquad (3.38)$$

Число операций умножения, приходящихся на обратный ход:

$$N_{\text{обр.х}}=0+1+2+\ldots+(n-1)=\frac{n(n-1)}{2}. \qquad (3.39)$$

На основании формул (3.37) и (3.39) получаем общее число операций умножения и деления, необходимых для реализации алгоритма Гаусса:

$$N_{\text{г}}=N_{\text{пр.х}}+N_{\text{обр.х}}=\frac{n^3+3n^2-n}{3}. \qquad (3.40)$$

Проведем теперь исследование устойчивости решения СЛАУ по алгоритму Гаусса. В [5] показано, что вычисленное по алгоритму Гаусса приближенное решение $\tilde{x}$ будет точным для возмущенной системы

$$(A+Q)x=b,$$

причем для матрицы эквивалентного возмущения Q выполняется оценка

$$\|Q\|_c\leqslant f(n)g(A)\|A\|_c\bar{\varepsilon}_{\text{окр}}. \qquad (3.41)$$

Переходя к относительной оценке возмущения, получим

$$\varepsilon_A\leqslant f(n)g(A)\bar{\varepsilon}_{\text{окр}}. \qquad (3.42)$$

В этой оценке $f(n)$ — степенная функция, пропорциональная n^3.

В отличие от оценки вида (2.80) в ней присутствует член $g(A)$, смысл которого будет рассмотрен ниже. Введем обозначения

$$a_0=\max_{i,\,j}\,|\,a_{ij}\,|,\ a_k=\max_{ij}\,|\,a_{ij}^{(k)}\,|,$$

где $a_{ij}^{(k)}$ — элементы матрицы $A^{(k)}$, вычисляемые по формулам (3.36). Тогда для $g(A)$ имеем

$$g(A)=\max_{0\leqslant k\leqslant n-1}(a_k/a_0). \tag{3.43}$$

Таким образом, $g(A)$ показывает, насколько могут возрасти элементы матрицы $A^{(k)}$ в ходе гауссова исключения по сравнению с величиной коэффициентов, составляющих матрицу A. Поэтому величину $g(A)$ называют *коэффициентом роста* матрицы A при гауссовом исключении. В тех случаях, когда ведущий элемент $a_{kk}{}^{(k-1)}\neq 0$ алгоритма Гаусса находится в достаточно малой окрестности нуля, величина a_k может принять очень большое значение, вследствие чего алгоритм Гаусса станет неустойчивым.

Существует ряд модификаций алгоритма Гаусса, в которых предусмотрено ограничение роста $g(A)$ при исключениях. В наиболее употребительной из модификаций осуществляется выбор в качестве ведущего элемента на каждом шаге исключения, например k-м, такого элемента, который равен максимальному по модулю элементу среди $a_{ij}^{(k-1)}$, элементов, где $i\leqslant k$, расположенных в k-м столбце матрицы $A^{(k-1)}$. Пусть таким элементом будет $a_{ks}{}^{(k-1)}$, т. е.

$$|\,a_{ks}{}^{(k-1)}\,|=\max_{i\geqslant k}|a_{ik}{}^{(k-1)}\,|.$$

Теперь для того, чтобы этот элемент сделать ведущим, требуется осуществить переход от системы $A^{k-1}x=d^{k-1}$ к эквивалентной системе $\tilde{A}^{k-1}x=\tilde{a}^{k-1}$, в которой $\tilde{a}_{kk}{}^{k-1}=a_{ks}{}^{k-1}$. Этот переход связан с перестановкой k-й и s-й строк в расширенной матрице системы $A^{k-1}x=d^{k-1}$ и может быть описан таким образом:

$$P_{ks}A^{(k-1)}x=P_{ks}d^{(k-1)}.$$

Далее по формулам (3.36) производят k-й шаг исключения, ведущим элементом в котором будет $\tilde{a}_{kk}{}^{k-1}$. Описанный алгоритм называется *алгоритмом Гаусса с выбором главного элемента по столбцам*, при этом элементы $\tilde{a}_{kk}{}^{(k-1)}$ называются *главными элементами*. Если матрица A исходной системы является невырожденной, то данный алгоритм всегда реализуем, так как для любых значений k главные элементы не равны нулю.

Выбор главного элемента на каждом k-м шаге исключения можно осуществлять не только среди элементов k-столбца, но

и среди всех элементов a_{ij}^{k-1}, где $i \geqslant k$ и $j \geqslant k$. В этом случае главным элементом будет a_{lm}^{k-1}, для которого выполняется условие

$$|a_{lm}^{(k-1)}| = \max_{i \geqslant k,\ j \geqslant k} |a_{ij}^{(k-1)}|.$$

Далее осуществляется переход от системы $A^{(k-1)}x = d^{(k-1)}$ к эквивалентной системе $\tilde{A}^{(k-1)}x = \tilde{d}^{(k-1)}$ по следующим формулам

$$P_{kl}A^{(k-1)}P_{km}y = P_{kl}d^{(k-1)}, \quad y = P_{km}x \tag{3.44}$$

и осуществляется k-й шаг исключения по формулам (3.36). Описанный алгоритм получил название *алгоритма Гаусса с выбором главного элемента в результате полного перебора.* Этот алгоритм по сравнению с алгоритмом, в котором главные элементы выбираются по столбцам, приводит к большим вычислительным затратам, но обладает хорошей вычислительной устойчивостью за счет значительного ограничения коэффициента роста $g(A)$.

§ 3.5. ВЫЧИСЛИТЕЛЬНАЯ СХЕМА ХОЛЕЦКОГО. РЕШЕНИЕ СИСТЕМ НОРМАЛЬНЫХ УРАВНЕНИЙ МЕТОДОМ КВАДРАТНОГО КОРНЯ

Если при решении СЛАУ матрица системы A является невырожденной, то на основании теоремы 3.1 и формулы 3.19 эту матрицу можно разложить на множители

$$\tilde{A} = PA = \tilde{C}\tilde{B} \Leftrightarrow A = P\tilde{C}\tilde{B}, \tag{3.45}$$

где P — матрица перестановок строк, служащая для перехода от матрицы A к матрице $\tilde{A}$, у которой главные диагональные миноры не равны нулю; C и B — соответственно ниже и верхние треугольные матрицы, элементы которых $\tilde{c}_{ij}$ и $\tilde{b}_{ij}$ вычисляются через элементы $\tilde{a}_{ij}$ по формулам (3.12), причем будем иметь $\tilde{c}_{kk} \neq 0$. Подставляя выражение A из (3.45) в исходное уравнение, получим

$$\tilde{C}\tilde{B}x = \tilde{b}. \tag{3.46}$$

Это уравнение равносильно двум матричным уравнениям

$$\tilde{C}y = \tilde{b}, \quad \tilde{B}x = y,$$

решениями первого из которых являются

$$y_1 = \frac{\tilde{a}_{1,\,n+1}}{\tilde{c}_{11}}, \qquad y_k = \frac{1}{\tilde{c}_{kk}}\left(\tilde{a}_{k,\,n+1} - \sum_{j=2}^{k} \tilde{c}_{k,\,j-1}y_{j-1}\right) \text{ для } k = 2 \div n, \tag{3.47}$$

а решениями второго

$$x_n = y_n, \quad x_k = y_k - \sum_{i=k+1}^{n} \tilde{b}_{kj} x_j$$

$$\text{для } k = (n-1), (n-2), \ldots, 1. \tag{3.48}$$

Таким образом, решение СЛАУ по приведенному алгоритму, который называется *вычислительной схемой Холецкого*, состоит из прямого хода, связанного с разложением матрицы на две треугольные по формулам (3.12), вычислением значений y_j по формулам (3.47), и обратного хода, описываемого формулами (3.48).

Общее число операций умножения и деления, необходимое для решения СЛАУ по вычислительной схеме Холецкого:

$$N_{\text{хл}} = N_{\text{св}} + N_{\text{у}} + N_{\text{обр.х}},$$

где $N_{\text{у}}$ — число операций умножения и деления, приходящихся на вычисление значений y_j по формулам (3.47) и определяемых по формуле $N_{\text{у}} = n(n+1)/2$. С учетом формул (3.13) и (3.39) получаем

$$N_{\text{хл}} = (n^3 + 3n^2 - n)/3. \tag{3.49}$$

Основным преимуществом вычислительной схемы Холецкого перед алгоритмом Гаусса является более высокая его вычислительная устойчивость к ошибкам округления благодаря возможности реализации в нем счета с удвоенной точностью. Эта возможность связана с тем, что как при прямом, так и при обратном ходе производятся вычисления скалярных произведений, а они могут быть выполнены в режиме накопления.

Как отмечалось ранее, большой круг геодезических и инженерно-технических задач связан с решением систем нормальных уравнений. Матрица системы нормальных уравнений является симметричной и положительно определенной. Для решения таких систем может быть применен *метод квадратного корня*, основанный на том же принципе, что и вычислительная схема Холецкого.

Так как матрица A симметричная, то, как следует из формулы (3.21), ее можно представить в виде произведения двух треугольных матриц, трансформированных относительно друг друга, т. е.

$$A = T'T.$$

Подставляя это выражение для матрицы A в исходную систему нормальных уравнений, получим, как и в методе Холецкого, эквивалентную систему, состоящую из двух уравнений вида

$$T'y = b \text{ и } Tx = y. \tag{3.50}$$

Решениями этих систем будут

$$y_1 = a_{1,n+1} t_{11}^{-1}, \quad y_k = (a_{k,n+1} - \sum_{i=2}^{k} t_{j-1,k} y_{j-1}) t_{kk}^{-1} \text{ для } k = 2 \div n,$$

$$x_n = y_n t_{nn}^{-1}, \quad x_k = (y_m - \sum_{j=m+1}^{n} t_{mj} x_j) t_{kk}^{-1} \tag{3.51}$$

$$\text{для } k = (n-1), (n-2), \ldots, 1.$$

При решении нормальной системы уравнений по методу квадратного корня необходимо выполнить n раз извлечение квадратного корня и $N_{\text{р.к.к}}$ операций умножения и деления, число которых определится как

$$N_{\text{р.к.к}} = N_{\text{к.к}} + 2N_{\text{у}}.$$

Исходя из формул (3.23) и (3.51) для вычисления y_j, получим

$$N_{\text{р.к.к}} = n(n+1)(n+8)/6. \tag{3.52}$$

При решении СЛАУ высокого порядка, когда вся исходная информация не может быть размещена в оперативной памяти ЭВМ, целесообразно вычислительный процесс решения СЛАУ организовать таким же образом, как и разбиение матрицы системы A на клетки. При этом вектор-столбец x и вектор-столбец b должны быть разбиты соответствующим образом на подвекторы, а при разбиении матрицы A ее диагональные клетки должны быть квадратными матрицами. В этом случае исходная система уравнений будет представлена в клеточном виде

$$\begin{pmatrix} A_{11} & A_{12} & \ldots & A_{1s} \\ A_{21} & A_{22} & \ldots & A_{2s} \\ \ldots & \ldots & \ldots & \ldots \\ A_{s1} & A_{s2} & \ldots & A_{ss} \end{pmatrix} \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \ldots \\ \bar{x}_s \end{pmatrix} = \begin{pmatrix} \bar{b}_1 \\ \bar{b}_2 \\ \ldots \\ \bar{b}_s \end{pmatrix}, \tag{3.53}$$

где A_{ii} — матрицы размерностей $[n_i \times n_i]$; $\bar{x}_i$ и $\bar{b}_i$ — подвекторы размерности n_i.

Решение системы (2.53) можно осуществить с помощью ранее рассмотренных прямых методов, только вместо действий над числами надо проводить действия над матрицами.

Применим рассмотренный выше метод квадратного корня для решения нормальной системы уравнений, представленных в блочном (клеточном) виде (3.53). Тогда T и T' примут вид

$$T = \begin{pmatrix} T_{11} & T_{12} & \ldots & T_{1,s-1} & T_{1s} \\ O & T_{22} & \ldots & T_{2,s-1} & T_{2s} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ O & O & \ldots & O & T_{ss} \end{pmatrix},$$

$$T'=\begin{pmatrix} T'_{11} & O & \dots & O \\ T'_{12} & T'_{22} & \dots & O \\ \dots & \dots & \dots & \dots \\ T'_{1s} & T'_{2s} & \dots & T'_{ss} \end{pmatrix},$$

в которых клетки T_{ii} имеют размерность n_i.

Произведя умножение матриц T' и T в клеточной форме и приравнивая соответствующие клетки, получим матричные уравнения

$$T'_{ii}T_{ii}=A_{ii}-\sum_{k=1}^{i-1}T'_{ki}T_{ki} \text{ для } i=1\div s,$$

$$T'_{ii}T_{ij}=A_{ij}-\sum_{k=1}^{i-1}T'_{ki}T_{kj} \text{ для } j>i. \qquad (3.54)$$

Поиск клеток T_{ij} осуществляется последовательно от строки к строке, т. е. сначала находятся все клетки T_{ij}, расположенные на i-й строке, а затем все клетки $T_{i+1,\,j}$, расположенные на следующей по порядку $i+1$ строке и т. д. При этом элементы матрицы T_{ii} находятся обычным методом квадратного корня по формулам (3.22), а элементы матрицы T_{ij} находятся из уравнений

$$T'_{ii}\bar{t}_p^{(ij)}=\bar{c}_p^{(ij)}, \qquad (3.55)$$

где $t_p^{(ij)}$ — вектор-столбец, образованный из p-го столбца матрицы T_{ij}, с компонентами $t_{qp}^{(ij)}$ при $q=1\div n_i$; $c_p^{(ij)}$ — вектор-столбец, образованный из p-го столбца матрицы $A_{ij}-\sum_{k=1}^{i-1}T'_{ki}T_{kj}$ для $j>i$ с компонентами $c_{qp}^{(ij)}$ при $q=1\div n_i$. Число уравнений вида (3.55) равно n_j, так как $p=1\div n_j$.

Так как в этих уравнениях матрицы T'_{ii} — нижние треугольные, то вычисление значений $t_{qp}^{(ij)}$ можно осуществить по формулам, имеющим такой же вид, что и формулы (3.51), а именно

$$t_{1p}^{(ij)}=c_{1p}^{(ij)}/t_{11}^{(ii)},\quad t_{kp}^{(ij)}=(c_{kp}^{(ij)}-\sum_{m=2}^{k}t^{(ii)}_{m-1,\,k})/t^{(ij)}_{m-1,\,p})/t^{(ii)}_{kk} \qquad (3.56)$$

для всех $p=1\div n_j$ и $k=2\div n_i$.

После вычисления элементов матриц T_{ij} решение системы (3.53) сводится, как и в методе квадратного корня, к последовательному решению двух матричных систем вида

$$\begin{pmatrix} T'_{11} & 0 & & \\ T'_{12} & T'_{22} & & \\ \dots & \dots & \dots & \dots \\ T'_{1s} & T'_{2s} & \dots & T'_{ss} \end{pmatrix}\begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \dots \\ \bar{y}_s \end{pmatrix}=\begin{pmatrix} \bar{b}_1 \\ \bar{b}_2 \\ \dots \\ \bar{b}_s \end{pmatrix},$$

$$\begin{pmatrix} T_{11} & T_{12} & \dots & T_{1s} \\ & T_{22} & \dots & T_{2s} \\ \dots & \dots & \dots & \dots \\ 0 & & & T_{ss} \end{pmatrix} \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \dots \\ \bar{x}_s \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \dots \\ \bar{y}_s \end{pmatrix}.$$

Эти системы можно записать в виде

$$T'_{ll}\bar{y}_l = \bar{d}_l, \quad T_{ll}\bar{x}_l = \bar{z}_l \text{ для } l = 1 \div s, \tag{3.57}$$

где $\bar{d}_l$ и $\bar{z}_l$ — подвекторы, определяемые как

$$d_1 = b_1, \quad d_l = b_l - \sum_{k=2}^{l} T_{k-1,l}\bar{y}_{k-1} \text{ для } l = 2 \div s,$$

$$\bar{z}_s = \bar{y}_s, \quad \bar{z}_l = \bar{y}_l - \sum_{k=l+1}^{s} T_{lk}x_k \text{ для } l = (s-1)(s-2), \dots, 1. \tag{3.58}$$

Так как в системе (3.57) матрицы T_{ll} и T'_{ll} треугольные, то вычисление компонент подвекторов $\bar{y}_l$ и $\bar{x}_l$ можно выполнить по формулам, аналогичным по виду формулам (3.51), т. е.

$$y_{1l} = d_{1l}/t_{11}^{(ll)}, \quad y_{il} = (d_{il} -$$

$$- \sum_{k=2}^{i} t^{(ll)}_{k-1,\,i} y_{k-1,\,l})/t_{kk}^{(ll)} \text{ для } i = 2 \div n_l, \quad l = 1 \div s, \tag{3.59}$$

$$x_{sl} = t_{sl}^{(ll)}/t_{ss}^{(ll)}, \quad x_{il} = (z_{il} -$$

$$- \sum_{k=i+1}^{s} t_{ik}^{(ll)} x_{kl})/t_{kk}^{(ll)} \text{ для } i = (s-1), (s-2), \dots, 1, \quad l = 1 \div s.$$

Нижний индекс l в приведенных формулах указывает на то, что рассматривается l — компонента подвектора.

Решение нормальной системы уравнений блочным (клеточным) методом квадратного корня может быть выполнено таким образом:

1) приведение системы $Ax = b$ к блочному виду;

2) вычисление элементов матрицы T_{ii} методом квадратного корня, т. е. по формулам (3.22);

3) вычисление элементов матриц T_{ij} по формулам (3.56);

4) вычисление компонент вектора y по формулам (3.58) и (3.59);

5) вычисление компонент вектора решения x по формулам (3.58) и (3.59).

Одной из важнейших характеристик в блочных методах решения СЛАУ на ЭВМ является число операций обращений к внешним запоминающим устройствам. Эти операции являются наиболее долговременными машинными операциями, а следова-

тельно, от их числа в значительной степени зависит время решения СЛАУ. В блочном методе квадратного корня при условии, что обмен информацией между внешним запоминающим устройством и оперативным запоминающим устройством ЭВМ в процессе решения СЛАУ производится целыми блоками (клетками), число таких обращений будет

$$N_{\text{обр}} = \frac{s^3 + 12s^2 + 2s}{3}. \tag{3.60}$$

§ 3.6. ОРТОГОНАЛЬНЫЕ ПРЕОБРАЗОВАНИЯ И ИХ ПРИМЕНЕНИЕ В АЛГОРИТМАХ ОРТОГОНАЛЬНЫХ РАЗЛОЖЕНИЙ МАТРИЦ

Напомним, что ортогональное разложение матрицы описывается формулой (1.33), т. е.

$$A = V^{\text{т}} T W.$$

Следовательно, для получения этого разложения требуется найти такие ортогональные матрицы V и W, для которых выполняется равенство (1.32)

$$VAW^{\text{т}} = T.$$

Это равенство можно преобразовать как

$$V\hat{A} = R, \quad WR^{\text{т}} = T', \tag{3.61}$$

где T' — матрица, транспонированная к матрице T, матрица $\hat{A}$ получена из матрицы A после перестановки в матрице $\hat{A}$ строк таким образом, что первые k строк в матрице A линейно независимые. Причем если $n > m$, то матрица R принимает вид, представленный на рис. 1, *а*. Из формул (3.61) вытекает, что для ортогонального разложения (1.33) можно построить алгоритм с двумя однотипными процедурами, суть которых состоит в приведении исходной матрицы, например B, за счет ее ортогонального преобразования к матрице, имеющей структуру, аналогичную матрице R. Такая вычислительная процедура получила название *QR-алгоритма*. Этот алгоритм основан на реализации следующего равенства:

$$QB = R, \tag{3.62}$$

в котором матрица B размера $[n \times m]$ и ранга k имеет k линейно независимых первых строк.

Представим матрицу B в виде последовательности векторов-столбцов, т. е. как $B = (\bar{b}_1 \bar{b}_2 \ldots \bar{b}_j \ldots \bar{b}_m)$, где $b_j = (b_{1j} b_{2j} \ldots b_{nj})^{\text{т}}$ для $j = 1 \div m$, а матрицу R — в виде $R = (\bar{r}_1 \bar{r}_2 \ldots \bar{r}_j \ldots \bar{r}_k \bar{r}_{k+1} \ldots r_m)$, где $\bar{r}_j = (r_{1j} r_{2j} \ldots r_{jj} \underbrace{0 \ldots 0}_{n-j})^{\text{т}}$ для $j = 1 \div k$ и $\bar{r}_j = (r_{1j} r_{2j} \ldots r_{kj} \underbrace{0 \ldots 0}_{n-k})^{\text{т}}$

для $j=(k+1)\div m$. Введем в рассмотрение матрицу Q_p, обеспечивающую преобразование вида

$$Q_1\bar{b}_1=\bar{r}_1, \tag{3.63}$$

$$Q_p\bar{b}_p{}^{(p-1)}=\bar{r}_p \text{ при } \bar{b}_j{}^{(p-1)}=Q_{p-1}\bar{b}_j{}^{(p-2)}$$

$$\text{для } p=2\div m \text{ и } j=p\div m,$$

при этом предполагается, что $\bar{b}_j{}^{(0)}=\bar{b}_j$ и $\bar{b}_j{}^{(p-1)}=(b_{1j}{}^{(p-1)}b_{2j}{}^{(p-1)}\ldots$ $\ldots b_{pj}{}^{(p-1)}\ldots b_{nj}{}^{(p-1)})$. Кроме того, матрица Q_p должна удовлетворять следующим условиям: быть ортогональной и обеспечивать преобразование

$$Q_p\bar{r}_s=\bar{r}_s \text{ для } s=1\div(p-1), \quad p=2\div m. \tag{3.64}$$

Исходя из формул (3.63) и (3.64), получаем для $p=1\div m$

$$Q_p\ldots Q_2Q_1(\bar{b}_1\bar{b}_2\ldots\bar{b}_m)=$$

$$=(\bar{r}_1\bar{r}_2\ldots r_p\bar{b}^{(p)}{}_{p+1}\ldots\bar{b}^{(p)}{}_m).$$

Отсюда

$$Q=Q_mQ_{m-1}\ldots Q_p\ldots Q_2Q_1. \tag{3.65}$$

Матрицы Q_p будем выбирать со следующей блочной структурой:

$$Q_1=\hat{Q}_{n\times n},\ Q_p=\begin{pmatrix} E_{p-1} & 0_{(p-1)\times(n-p+1)} \\ 0_{(n-p+1)\times(p-1)} & \hat{Q}_{n-p+1}\end{pmatrix},\ \text{для } p=2\div n, \tag{3.66}$$

где нижний индекс у каждого блока указывает размерность соответствующей ему матрицы. Введя обозначения $\hat{Q}_{n-p+1}=\hat{Q}_p$ $E_{n-p+1}=E_p$, определим матрицу $\hat{Q}_p$ как

$$\hat{Q}_p=E_p-k_p(\bar{q}_p,\ \bar{q}_p{}^{\mathrm{T}}), \tag{3.67}$$

где $\bar{q}_p$ — вектор-столбец:

$$\bar{q}_p=(q_{pp}b^{(p-1)}{}_{p+1,\,p}\ldots b^{(p-1)}{}_{p+k,\,p}\ldots b_{np}{}^{(p-1)})^{\mathrm{T}}.$$

Матрица $\hat{Q}_p$, получившая название *матрицы отражения*, является симметричной и ортогональной.

Представляя вектор $\bar{b}_p{}^{(p-1)}$ через подвекторы $\bar{b}_{p_1}{}^{(p-1)}$ размерности $p-1$ и $\bar{b}_{p_2}{}^{(p-1)}$ размерности $n-p+1$, с учетом (3.66) имеем

$$Q_p\bar{b}_p{}^{(p-1)}=\begin{pmatrix} \bar{b}_{p_1}{}^{(p-1)} \\ \hat{Q}_pb_{p_2}{}^{(p-1)}\end{pmatrix}.$$

Отсюда на основании формулы (3.63)

$$\bar{b}_{p_1}{}^{(p-1)}=\bar{r}_{p_1} \text{ и } \hat{Q}_p\bar{b}_{p_2}{}^{(p-1)}=\bar{r}_{p_2}. \tag{3.68}$$

Распишем вторую из этих формул с учетом (3.67)

$$\hat{Q}_p \bar{b}_{p_2}^{(p-1)} = \bar{b}_{p_2}^{(p-1)} - k_p(\bar{q}_p,\ \bar{q}_p^{\tau}) \Leftrightarrow \hat{Q}_p \bar{b}_{p_2}^{(p-1)} =$$
$$= \bar{b}_{p_2}^{(p-1)} - k_p(\bar{q}_p^{\tau},\ \bar{b}_{p_2}^{(p-1)})\,\bar{q}_p.$$

Полученное выражение в координатной форме записи примет вид

$$b_{pp}^{(p-1)} - k_p(\bar{q}_p^{\tau},\ \bar{b}_{p_2}^{(p-1)})\,q_{pp} = r_{pp},$$
$$b^{(p-1)}{}_{p+k,\,p} - k_p(\bar{q}_p^{\tau},\ \bar{b}_{p_2}^{(p-1)})\,b^{(p-1)}{}_{p+k,\,p} = 0 \text{ для } k = 1 \div (n-p). \tag{3.69}$$

Из этих уравнений находим параметры k_p и q_{pp}. Так, из второго уравнения имеем

$$k_p = (\bar{q}_p^{\tau},\ \bar{b}_{p_2}^{(p-1)})^{-1}, \text{ если } (\bar{q}_p^{\tau},\ \bar{b}_{p_2}^{(p-1)}) \neq 0. \tag{3.70}$$

После подстановки этого выражения для k_p в первое уравнение (3.69) находим

$$b_{pp}{}^{p-1} - q_{pp} = r_{pp}.$$

Если во втором выражении (3.68) перейти к евклидовой норме вектора, то

$$(\hat{Q}_p \bar{b}_{p_2}^{(p-1)},\ \hat{Q}_p \bar{b}_{p_2}^{(p-1)}) = \| \bar{b}_{p_2}^{(p-1)} \|^2_E \Leftrightarrow \| \bar{b}_{p_2}^{(p-1)} \|_E = \sigma_p r_{pp},$$
$$\sigma_p = \begin{cases} +1, \text{ если } r_{pp} > 0, \\ -1, \text{ если } r_{pp} < 0. \end{cases}$$

В результате вычисление q_{pp} можно осуществлять по формуле

$$q_{pp} = b_{pp}^{(p-1)} - \sigma_p \sqrt{\sum_{k=0}^{n-p} (b^{p-1}{}_{p+k;\,p})^2}, \tag{3.71}$$
$$\sigma_p = \begin{cases} -1, \text{ если } b_{pp}^{(p-1)} > 0, \\ +1, \text{ если } b_{pp}^{(p-1)} < 0. \end{cases}$$

Таким образом, для получения ортогональной и симметричной матрицы Q_p необходимо выполнить:

1) вычисление значения $t = \sqrt{\sum_{k=0}^{n-p} (b_{p+k,\,p}^{p-1})^2}$;

2) вычисление элемента q_{pp} по формуле $q_{pp} = b_{pp}^{(p-1)} - \sigma_p t$, где $\sigma = \begin{cases} -1, \text{ если } b_{pp}^{(p-1)} > 0, \\ 1, \text{ если } b_{pp}^{(p-1)} < 0; \end{cases}$

3) вычисление значения K_p по формуле (3.70) или $K_p = (t^2 + b_{pp}^{(p-1)} t)^{-1}$, если $(\bar{q}_p^{\tau},\ b_{p_2}^{(p-1)}) \neq 0$ и $K_p = 0$, если $(\bar{q}_p^{\tau},\ b_{p_2}^{(p-1)}) = 0$;

4) вычисление элементов матрицы отражения $\hat{Q}_p$ по формуле (3.67);

5) получение матрицы Q_p по формулам (3.66).

Матрица Q_p, вычисляемая по описанному выше алгоритму, как нетрудно проверить, удовлетворяет поставленным выше условиям (3.69) и (3.64) и является ортогональной и симметричной. После вычисления матриц Q_p для $p=1\div m$ может быть определена матрица Q по формуле (3.65) в результате перемножения матриц Q_p, что приводит к значительной арифметической работе, превышающей вычислительные затраты, непосредственно связанные с приведением матрицы B к треугольной форме по формуле (3.62).

Представленный выше алгоритм получения матрицы отражения Q_p и использования этих матриц для ортогонального разложения по формуле (3.62) можно применить и для сингулярного разложения матриц. Пусть исходная матрица B размера $[n\times m]$ имеет $\operatorname{Rg} B=k$ и первые k столбцов ее независимы, тогда описанным выше способом она может быть преобразована к виду $QB=R$, где матрица R имеет структуру, показанную на рис. 1, *а*. Применим описанный способ ортогонального разложения, но уже к матрице R^{T}, в результате чего поддиагональные элементы матрицы R^{T} будут аннулированы, т. е. получим

$$SR^{\mathrm{T}}=D^{\mathrm{T}}{}_k, \qquad S=S_nS_{n-1}\ldots S_l\ldots S_1,$$

где S_l — ортогональные матрицы, вычисляемые по приведенному выше алгоритму; $D_k{}^{\mathrm{T}}$ — матрица, у которой $d_{ii}\neq 0$ для $i=1\div k$, а все остальные элементы $d_{ij}=0$. Причем с помощью матриц перестановок строк можно так организовать вычислительный процесс, чтобы выполнялось условие $d_{ii}\geqslant d_{kk}$ при $i>k$. В конце процесса сингулярного разложения получают матрицу $D_r=(D_r{}^{\mathrm{T}})^{\mathrm{T}}$.

Ортогональное разложение матрицы (3.62) можно реализовать и по алгоритму, основанному на методе последовательных плоских вращений, каждое из которых аннулирует (заведомо приравнивает нулю) только один элемент преобразуемой матрицы. Если же требуется аннулировать, например в столбце $\bar{b}_p{}^{(p-1)}$ матрицы $B^{(p-1)}$, все элементы с номерами от $p+1$ до n, то для этого можно воспользоваться ортогональной матрицей S_p, удовлетворяющей, как и матрице Q_p в методе отражений, условиям (3.63) и (3.64). Однако в отличие от Q_p матрица S_p определяется через произведение матриц плоских вращений как

$$S_p=S_{pn}\ldots S_{pk}\ldots S_{p,p+1}, \tag{3.72}$$

где S_{pk} — ортогональная *матрица плоского вращения* размера $[n\times n]$, равная

$$S_{pk}=P_{p+1,\,k}\begin{pmatrix} E_{p-1} & & O \\ & \hat{S}_{p\,k} & \\ O & & E_{n-p-1} \end{pmatrix} P_{p+1,\,k} \text{ для } k=(p+1)\div n, \tag{3.73}$$

где $\hat{S}_{pk}$ — матрица элементарного вращения вида

$$\hat{S}_{pk}=\begin{pmatrix} s_{pp}^{(k)} & s_{pk} \\ s_{kp} & s_{kk} \end{pmatrix}, \tag{3.74}$$

в которой

$$s_{pp}^{(k)}=s_{kk}, \quad s_{pk}=-s_{kp} \text{ и } (s_{pp}^{(k)})^2+s^2{}_{kp}=1, \tag{3.75}$$

$P_{p+1,\,k}$ — матрица перестановки $(p+1)$-й и k-й строк.

Для определения значений s_{pp} и s_{pk} воспользуемся равенствами (3.75) и требованием, предъявляемым к матрице S_{pk}, аннулирования элемента с номером k в p-м столбце преобразуемой матрицы. Так, в столбце $\bar{b}_p^{(p-1)}$ матрицы $B^{(p-1)}$ требуется аннулировать элемент под номером $(p+1)$. Выполнив преобразование

$$S_{p,p+1}\bar{b}_p^{(p-1)}=\bar{b}_p^{(p-1)_1},$$

получим новый столбец, обозначенный $\bar{b}_p^{(p-1)_1}$, в котором элемент под номером $(p+1)$ будет равен нулю, т. е. $b_{p+1,\,p}^{(p-1)_1}=0$. Для данного случая

$$S_{p,\,p+1}\bar{b}_p^{(p-1)}=\begin{pmatrix} E_{p-1} & O \\ & \hat{S}_{p,\,p+1} \\ O & E_{n-p+1} \end{pmatrix}\begin{pmatrix} \bar{b}_{p_1}^{(p-1)} \\ b_{pp}^{(p-1)} \\ b_{p+1,\,p}^{(p-1)} \\ \bar{b}_{p_2}^{(p-1)} \end{pmatrix}\Leftrightarrow$$

$$\Leftrightarrow\begin{pmatrix} \bar{b}_{p_1}^{(p-1)} \\ s_{pp}^{(p+1)}b_{pp}^{(p-1)}+s_{p+1,\,p}b_{p+1,\,p}^{(p-1)} \\ s_{p+1,\,p}b_{pp}^{(p-1)}+s_{pp}^{(p+1)}b_{p+1,\,p}^{(p-1)} \\ \bar{b}_{p_2}^{(p-1)} \end{pmatrix}=\begin{pmatrix} \bar{b}_{p_1}^{(p-1)} \\ b_{pp}^{(p-1)_1} \\ 0 \\ b_{p+2,\,p}^{(p-1)} \\ b_{p_2}^{(p-1)_1} \end{pmatrix}.$$

Отсюда с учетом условия, что $\bar{b}_{p+1,\,p}^{(p-1)_1}=0$, имеем

$$-s_{p+1,\,p}b_{pp}^{(p-1)}+s_{pp}^{(p+1)}b_{p+1,\,p}^{p-1}=0,$$

$$s^2{}_{p+1,\,p}+(s_{pp}^{(p+1)})^2=1.$$

После решения этой системы получим

$$b_{pp}^{(p-1)_1}=\sqrt{(b_{pp}^{(p-1)})^2+(b_{p+1,\,p}^{(p-1)})^2},$$

$$s_{p+1,\,p}=b_{p+1,\,p}^{(p-1)}/b_{pp}^{(p-1)_1}, \quad s_{pp}^{(p+1)}=b_{pp}^{(p-1)}/b_{pp}^{(p-1)_1}. \tag{3.76}$$

Аннулируя теперь в столбце $\bar{b}_p^{(p-1)1}$ элемент под номером $(p+2)$ за счет преобразования

$$S_{p,\,p+2}\bar{b}_p^{(p-1)1}=\bar{b}_p^{(p-1)2}\Leftrightarrow$$

$$\Leftrightarrow\left(P_{p+1,\,p+2}\begin{pmatrix}E_{p-1} & O\\ & \hat{S}_{p,\,p+2}\\ O & & E_{n-p-1}\end{pmatrix}P_{p+1,\,p+2}\right)\begin{pmatrix}\bar{b}_{p_1}^{(p-1)}\\ b_{pp}^{(p-1)1}\\ 0\\ b_{p+2,\,p}^{(p-1)}\\ \bar{b}_{p_2}^{(p-1)1}\end{pmatrix}=\bar{b}_p^{(p-1)2}$$

из условия, что $b_{p+2,\,p}^{(p-1)2}=0$, получаем систему

$$-s_{p+2,\,p}b_{pp}^{(p-1)1}+s_{pp}^{(p+2)}b_{p+2,\,p}^{(p-1)}=0,$$

$$s^2{}_{p+2,\,p}+(s_{pp}^{(p+2)})^2=1.$$

Решением этой системы являются значения:

$$s_{p+2,\,p}=b_{p+2,\,p}^{(p-1)}/b_{pp}^{(p-1)2},\qquad s_{pp}^{(p+2)}=b_{pp}^{(p-1)1}/b_{pp}^{(p-1)2},$$

$$b_{pp}^{(p-1)2}=\sqrt{(b_{pp}^{(p-1)1})^2+(b_{p+2,\,p}^{(p-1)})^2}=\sqrt{\sum_{i=0}^{2}(b_{p+i,\,p})^2}\tag{3.77}$$

После k последовательных преобразований подобного типа для $k=1\div(n-p)$ получим

$$S_{p,p+k}=P_{p+1,p+k}\begin{pmatrix}E_{p-1} & & O\\ & \hat{S}_{p,p+k} & \\ O & & E_{n-p-1}\end{pmatrix}P_{p+1,p+k},\tag{3.78}$$

где элементы матрицы $\hat{S}_{p,\,p+k}$ определяются по формулам

$$s_{p+k,p}=b_{p+k,p}^{(p-1)}/b_{pp}^{(p-1)k},\ s_{pp}^{(p+k)}=b_{pp}^{(p-1)k-1}/b_{pp}^{(p-1)k},\tag{3.79}$$

$$b_{pp}^{(p-1)k}=\sqrt{\sum_{i=0}^{k}(b_{p+i,p}^{(p-1)})^2},\ b_{pp}^{(p-1)k-1}=\sqrt{\sum_{i=0}^{k-1}(b_{p+i,p}^{(p-1)})^2}.$$

В результате описанных преобразований на k-м шаге вектор-строка $\bar{b}_p^{(p-1)}$ будет приведен к виду

$$s_{p,p+k}S_{p,p+k-1}\cdots S_{p,p+2}S_{p,p+1}\bar{b}_p^{(p-1)}=$$

$$=(b_{p_1}^{(p-1)}\ b_{pp}^{(p-1)k}\underbrace{0\ldots0}_{k}\ b_{p+k+1,p}^{(p-1)}\ \bar{b}_{p_2}^{(p-1)k-1})^{\mathrm{T}}.\tag{3.80}$$

Здесь подвектор $\bar{b}_{p_2}^{(p-1)k-1}$ размерности $n-p-k-1$ и его элементы равны соответствующим по номерам элементам подвек-

тора $\bar{b}_p^{(p-1)}$. Таким образом, после $k=n-p$ описанных шагов получим

$$S_{pn}S_{p,n-1}\ldots S_{p,p+2}S_{p,p+1}\bar{b}_p^{(p-1)} = \\ = \left(\bar{b}_{p_1}^{(p-1)}b_{pp}^{(p-1)}{}_{n-p}\underbrace{0\ldots0}_{n-p}\right) \Leftrightarrow S_p\bar{b}_p^{(p-1)} = \bar{r}_p, \qquad (3.81)$$

где векторы $\bar{b}_p^{(p-1)}$ и $\bar{r}_p$ определяются, как и в методе отражения (3.63).

Для нахождения матрицы S_p требуется выполнить:

1) вычисление элементов матрицы $\hat{S}_{p,\,p+k}$ по формулам

$$t_{p,k} = \sqrt{\sum_{i=0}^{k}(b_{p+i,p}^{(p-1)})^2},$$

$$s_{p+k,p} = b_{p+k,p}^{(p-1)}/t_{p,k},\ s_{pp}^{(p+k)} = \sqrt{1-s_{p+k,p}^2};$$

2) вычисление элементов матрицы $S_{p,\,p+k}$ (этот процесс связан с присвоением новых номеров i, j для $s_{pp}{}^{(p+k)}$ и $s_{p+k,\,p}$);

3) вычисление элементов матрицы S_p по формуле (3.72).

Вычислив элементы матрицы S_p для $p=1\div m$, находим матрицу ортогонального преобразования Q по формуле, аналогичной (3.65):

$$Q = S_mS_{m-1}\ldots S_p\ldots S_2S_1. \qquad (3.82)$$

При реализации ортогональных преобразований отражения и вращения для решения конкретных задач на конкретных вычислительных средствах может возникнуть необходимость в различных модификациях описанных выше методов, вызванная такими соображениями, как увеличение скорости и точности вычислений, защищенность от машинных нулей и переполнения, сокращение требований к памяти ЭВМ, использование разреженности матриц, распараллеливание вычислительных процессов и др.

Рассмотрим один из модифицированных методов вращения, получивший название метода Джентльмена, который позволяет сократить число арифметических операций по сравнению с алгоритмом вращения, приведенным выше. Пусть требуется аннулировать элементы столбца $\bar{b}_p^{(p-1)}$ матрицы $B^{(p-1)}$ начиная с $(p+1)$ до n. Тогда при аннулировании $(p+1)$ элемента матрица $S_{p,\,p+1}$, согласно (3.76), примет вид

$$\hat{S}_{p,p+1} = \begin{pmatrix} s_{pp}^{(p+1)} & s_{p+1,p} \\ -s_{p+1,p} & s_{pp}^{(p+1)} \end{pmatrix}.$$

Ее можно представить следующим образом: $\hat{S}_{p,\,p+1} = D_{p,\,p+1}H_{p,\,p+1}$, где $D_{p,\,p+1}$ — диагональная матрица, т. е.

$$\hat{S}_{p,p+1}=E \text{ при } b_{p+1,p}^{(p-1)}=0,$$

$$\hat{S}_{p,p+1}=\begin{pmatrix} s_{pp}^{(p+1)} & 0 \\ 0 & s_{pp}^{(p+1)} \end{pmatrix}\begin{pmatrix} 1 & s_{p+1,p}/s_{pp}^{(p+1)} \\ -s_{p+1,p}/s_{pp}^{(p+1)} & 1 \end{pmatrix}$$

$$\text{при } |b_{p+1,p}^{(p-1)}| \leqslant |b_{pp}^{(p-1)}|, \qquad (3.83)$$

$$\hat{S}_{p,p+1}=\begin{pmatrix} s_{p+1,p} & 0 \\ 0 & s_{p+1,p} \end{pmatrix}\begin{pmatrix} s_{pp}^{(p+1)}/s_{p+1,p} & 1 \\ -1 & s_{pp}^{(p+1)}/s_{p+1,p} \end{pmatrix}$$

$$\text{при } |b_{pp}^{(p-1)}| < |b_{p+1,1}|.$$

Так как преобразование вращения $S_{p,\,p+k}$ изменяет только две строчки матрицы $B^{(p-1)}$, а именно строчки с номерами p и $p+k$, то достаточно рассмотреть, как изменяется подматрица

$$B_{p,p+1}^{(p-1)}=\begin{pmatrix} b_{pp}^{(p-1)} & b_{p,p+1}^{(p-1)} & \dots & b_{pm}^{(p-1)} \\ b_{p+1,p}^{(p-1)} & b_{p+1,p+1}^{(p-1)} & \dots & b_{p+1,m}^{(p-1)} \end{pmatrix}$$

матрицы $B^{(p-1)}$ после умножения ее слева на матрицу вращения $\hat{S}_{p,\,p+1}$. С учетом введенного представления (3.83) получим

$$\hat{S}_{p,p+1}B_{p,p+1}^{(p-1)}=D_{p,p+1}H_{p,p+1}B_{p,p+1}^{(p-1)} \Leftrightarrow D_{p,p+1}\tilde{B}_{p,p+1}^{(p-1)1}=B_{p,p+1}^{(p-1)1}, \qquad (3.84)$$

$$\tilde{B}_{p,p+1}^{(p-1)1}=H_{p,p+1}B_{p,p+1}^{(p-1)}.$$

Применяя описанный выше метод аннулирования элемента под номером $(p+1)$ в столбце $\bar{b}_p^{(p-1)}$ для аннулирования элемента под номером $(p+2)$ в столбце $\bar{b}_p^{(p-1)1}$ матрицы $B^{(p-1)1}$, получим

$$\hat{S}_{p,p+2}B_{p,p+2}^{(p-1)1}=B_{p,p+2}^{(p-1)1}.$$

Здесь подматрица $B_{p,\,p+2}^{(p-1)1}$ матрицы $B^{(p-1)1}$, являющаяся результатом первого шага преобразования $S_{p,\,p+1}B^{(p-1)}=B^{(p-1)1}$, примет вид

$$B_{p,p+2}^{(p-1)1}=\begin{pmatrix} b_{pp}^{(p-1)1} & b_{p,p+1}^{(p-1)1} & \dots & b_{pm}^{(p-1)1} \\ b_{p+2,p}^{(p-1)} & b_{p+2,p+1}^{(p-1)} & \dots & b_{p+2,m}^{(p-1)} \end{pmatrix}.$$

Заметим, что вторая строчка подматрицы $B_{p,\,p+2}^{(p-1)1}$ совпадает со второй строчкой подматрицы $B_{p,\,p+2}^{(p-1)}$, т. е. не изменяется от первого шага преобразования. Представляя матрицу $\hat{S}_{p,\,p+2}$ в виде

$$\hat{S}_{p,p+2}=D_{p,p+2}H_{p,p+2},$$

где $D_{p,\,p+2}$ — диагональная матрица, получим

$$\hat{S}_{p,p+2}=E \text{ при } b_{p+2,p}^{(p-1)}=0,$$

$$\hat{S}_{p,p+2}=\begin{pmatrix} s_{pp}^{(p+2)} & O \\ O & s_{pp}^{(p+2)} \end{pmatrix}\begin{pmatrix} 1 & s_{p+2,p}/s_{pp}^{(p+2)} \\ -s_{p+2,p}/s_{pp}^{(p+2)} & 1 \end{pmatrix}$$

$$\text{при } |b_{p+2,p}^{(p-1)}|>|b_{pp}^{(p-1)_1}|, \tag{3.85}$$

$$\hat{S}_{p,p+2}=\begin{pmatrix} s_{p+2,p} & O \\ O & s_{p+2,p} \end{pmatrix}\begin{pmatrix} s_{pp}^{(p+2)}/s_{p+2,p} & 1 \\ -1 & s_{pp}^{(p+2)}/s_{p+2,p} \end{pmatrix}$$

$$\text{при } |b_{p+2,p}^{(p-1)}|>|b_{pp}^{(p-1)_1}|,$$

где величины $s_{pp}^{(p+2)}$ и $s_{p+2,p}$ находятся по формулам (3.77). С учетом введенного представления (3.85) имеем

$$\tilde{B}_{p,p+2}^{(p-1)_2}=H_{p,p+2}B_{p,p+2}^{(p-1)_1},$$

$$D_{p,p+2}\tilde{B}_{p,p+2}^{(p-1)_2}=B_{p,p+2}^{(p-1)_2}. \tag{3.86}$$

В результате k шагов аналогичных преобразований для подматрицы $B_{p,\,p+k}^{(p-1)_{k-1}}$ матрицы $B^{(p-1)\,k-1}$ будем иметь

$$B_{p,p+k}^{(p-1)_{k-1}}=\begin{pmatrix} b_{pp}^{(p-1)_{k-1}} & b_{p,p+1}^{(p-1)_{k-1}} & \dots & b_{pm}^{(p-1)_{k-1}} \\ b_{p+k,p}^{(p-1)} & b_{p+k,p+1}^{(p-1)} & \dots & b_{p+k,m}^{(p-1)} \end{pmatrix},$$

$$\hat{S}_{p,p+k}B_{p,p+k}^{(p-1)_{k+1}}=B_{p,p+k}^{(p-1)_k}.$$

Представляя матрицу $\hat{S}_{p,\,p+k}$ в виде

$$\hat{S}_{p,p+k}=D_{p,p+k}H_{p,p+k},$$

где $D_{p,\,p+k}$ — диагональная матрица, получим

$$\hat{S}_{p,p+k}=E \text{ при } b_{p+k,p}^{(p-1)}=0,$$

$$\hat{S}_{p,p+k}=\begin{pmatrix} s_{pp}^{(p+k)} & O \\ O & s_{pp}^{(p+k)} \end{pmatrix}\begin{pmatrix} 1 & s_{p+k,p}/s_{pp}^{(p+k)} \\ -s_{p+k,p}/s_{pp}^{(p+k)} & 1 \end{pmatrix}$$

$$\text{при } |b_{p+k,p}^{(p-1)}|\leqslant|b_{pp}^{(p-1)_{k-1}}|, \tag{3.87}$$

$$\hat{S}_{p,p+k}=\begin{pmatrix} s_{p+k,p} & O \\ O & s_{p+k,p} \end{pmatrix}\begin{pmatrix} s_{pp}^{(p+k)}/s_{p+k,p} & 1 \\ -1 & s_{pp}^{(p+k)}/s_{p+k,p} \end{pmatrix}$$

$$\text{при } |b_{p+k,p}^{(p-1)}|>|b_{pp}^{(p-1)_{k-1}}|,$$

где $s_{pp}^{(p+k)}$ и $s_{p+k,\,p}$ находятся по формуле (3.79). С учетом данного представления имеем

$$\tilde{B}_{p,p+k}^{(p-1)_k}=H_{p,p+k}B_{p,p+k}^{(p-1)_{k-1}},$$

$$D_{p,p+k}\tilde{B}_{p,p+k}^{(p-1)_k}=B_{p,p+k}^{(p-1)_k}. \tag{3.88}$$

Сокращение числа арифметических операций в методе Джентльмена по сравнению с методом вращения, описываемым выше алгоритмом вычисления матрицы S_p, связано с присутствием двух единиц в матрице H. Это позволяет выполнить операцию перемножения матриц по первой формуле в (3.88), т. е. получить матрицу $\bar{B}$ за $2n$ сложений и $2n$ умножений, тогда как перемножение матрицы $S_{p,\,p+k}B_{p,\,p+k}^{(p-1)_k-1}$ выполняется за $2n$ операций сложений и $4n$ операций умножений. В методе Джентльмена в памяти ЭВМ вместо матрицы B последовательно хранятся матрицы $D_{p,\,p+k}$ и $B_{p,\,p+k}$. При этом, однако, явный вид матрицы S_k не получается, а следовательно, нельзя получить и матрицу Q.

В заключение отметим, что элементы матрицы $D_{p,\,p+k}$ уменьшаются с ростом k, а элементы матрицы $B_{p,\,p+k}$, наоборот, возрастают. Хотя эти изменения элементов и носят медленный характер, но для больших систем уравнений необходимо в алгоритме Джентльмена предусматривать процедуру, позволяющую контролировать величину этих изменений и при необходимости вводить масштабирование соответствующих строк в матрицах $B^{(p-1)_k}$.

Кроме метода отражения и вращения для ортогонального разложения широко применяется и метод ортогонализации Грама — Шмидта, алгоритм которого был рассмотрен в § 1.3. В табл. 3 приведены некоторые характеристики этих методов

Таблица 3

Способ разложения	Режим вычисления	$N_{\text{умн}}$	Точность разложения	Дополнительная память
1	2	3	4	5
Метод отражения	С накоплением	$2/3\ n^3$	$2{,}9\ n$	$3\ n$
Метод вращения (Джентльмена)	Обычный	$4/3\ n^3$	$2{,}9\ n$	n
Метод Грама — Шмидта	С накоплением	n^3	1,0	$n^2/2$

при использовании их для ортогонального разложения матрицы. Причем предполагается, что все матрицы квадратные и имеют один и тот же порядок, равный n. В графе $N_{\text{умн}}$ приведены только главные члены числа операций умножения и деления, причем операции извлечения квадратного корня в состав главных членов не входят. Алгебраические задачи, особенно с матрицами большого порядка, требуют для своего решения значительных ресурсов памяти ЭВМ. Один из путей экономии памяти — размещение информации о сомножителях, получаемых при ортогональном разложении на месте исходной матри-

цы (при условии, что исходная матрица не потребуется для дальнейших вычислений). Однако этого места не всегда бывает достаточно, и нужно выделять дополнительную память. В графе 5 приведены главные члены числа полных слов памяти ЭВМ, которые необходимо добавить для размещения сомножителей. Точность разложения — одна из важнейших характеристик разложения матрицы на множители. Для всех приведенных разложений матрицы A эквивалентное возмущение δA [см. формулу (2.53)] оценивается по формуле (2.79)

$$\varepsilon_A \leqslant f(n)\bar{\varepsilon}_{\text{окр}},$$

где $f(n)$ зависит только от n и от способа получения разложения. В графе 4 приведены главные члены функции $f(n)$. При этом следует иметь в виду, что в таблице оценивается только точность вычислений, а влияние точности исходных данных на разложение не учитывается. В графе 2 указывается, с какой точностью должны выполняться промежуточные арифметические операции для получения результата разложения матрицы A с той точностью, которая приведена в графе 4. Напомним, что в режиме накопления вычисления выполняются со словами двойной длины (для ЕС ЭВМ — 8-байтовыми), а в обычном режиме — просто со словами (для ЕС ЭВМ — 4-байтовыми).

Анализ данных, приведенных в табл. 3, показывает, что нельзя дать однозначного ответа на вопрос, какое из разложений лучше. Выбор метода зависит от задачи и конкретного вычислительного средства, с помощью которого эта задача решается. Так, если матрица A общего вида и имеет мало нулевых элементов, то метод отражения имеет преимущество перед методами вращения и ортогонализации с точки зрения скорости вычислений. Однако если матрица A имеет преобладающее число нулевых элементов (т. е. A — разреженная матрица), то метод отражения и метод ортогонализации Грама — Шмидта будут уступать и по скорости вычислений, и по объему необходимой дополнительной памяти методу вращений. Это связано с тем, что в методе вращения изменению подлежат только элементы двух строк преобразуемой матрицы при каждом элементарном вращении. Если проанализировать точность разложений, то наилучшие результаты следует ожидать от метода ортогонализации Грама — Шмидта. Но, с другой стороны, как показывает практика [1], в методе Грама — Шмидта плохо выполняются условия ортогональности, тогда как в методах отражения и вращения условия ортогональности выполняются с высокой точностью.

§ 3.7. РАЗРЕЖЕННЫЕ СИСТЕМЫ ЛИНЕЙНЫХ УРАВНЕНИЙ И АЛГОРИТМЫ ИХ РЕШЕНИЯ

Приведем еще одну модифицированную схему решения СЛАУ методом Гаусса. Как было показано в § 3.4, первый шаг прямого хода метода Гаусса связан с переходом от исходной системы $Ax=b$ к эквивалентной системе $A^{(1)}x=b^{(1)}$, в которой элементы первого столбца, начиная с a_{21} до a_{n1}, аннулированы. Опишем этот переход в виде невырожденного преобразования вида

$$L_1Ax=L_1b \Leftrightarrow A^{(1)}x=b^{(1)}.$$

Второй шаг исключения опишем тоже с помощью невырожденного преобразования вида

$$L_2A^{(1)}x=L_2b^{(1)} \Leftrightarrow A^{(2)}x=b^{(2)}.$$

Отсюда следует

$$L_2L_1Ax=L_2L_1b.$$

Продолжая этот процесс, на k-м шаге исключения будем иметь

$$L_k\dots L_2L_1Ax=L_k\dots L_2L_1b. \tag{3.89}$$

В этой системе все элементы, находящиеся под диагональными элементами $a_{11}, a_{22}^{(1)}, \dots, a_{kk}^{(k-1)}$, будут аннулированы. Спрашивается, какой вид должна иметь матрица L_k, чтобы k-й шаг прямого хода метода Гаусса был реализован. Такой матрицей является нижняя треугольная матрица, имеющая вид

$$L_k=(\bar{e}_1\bar{e}_2\dots\bar{\eta}_k\dots\bar{e}_n), \tag{3.90}$$

у которой $\bar{e}_i$ — единичный вектор-столбец с i-й компонентой, равной единице, а компоненты вектора-столбца $\bar{\eta}_k$ вычисляются по формулам

$$\eta_{ik}=0 \text{ при } i<k, \quad \eta_{kk}=1/a^{(k-1)}{}_{kk}, \tag{3.91}$$
$$\eta_{ik}=-a_{ik}{}^{(k-1)}/a_{kk}{}^{(k-1)} \text{ при } i>k,$$

где элементы $a_{ik}{}^{(k-1)}$ находятся по формулам (3.36). После описанных выше шагов получим

$$L_n\dots L_k\dots L_2L_1Ax=L_n\dots L_k\dots L_2L_1b. \tag{3.92}$$

Введя обозначения

$$L=L_n\dots L_k\dots L_2L_1 \text{ и } LA=R, \tag{3.93}$$

уравнение (3.92) запишем как

$$Rx=Lb, \tag{3.94}$$

где R — треугольная матрица с единичной диагональю, имеющая вид (3.34). Вычислительный процесс, связанный с полу-

чением зависимости (3.92), является прямым ходом в рассматриваемой модифицированной схеме метода Гаусса.

Для решения системы (3.94), как и в методе Гаусса, будем выполнять обратный ход, описываемый на первом шаге хода преобразованием вида

$$U_nRx = U_nLb \Leftrightarrow R_nx = U_nLb, \tag{3.95}$$

где U_n — верхняя треугольная матрица:

$$U_n = (\bar{e}_1\bar{e}_2 \ldots \bar{e}_{n-1}\bar{\xi}_n), \tag{3.96}$$

$\bar{e}_i$ — единичный вектор-столбец с i-й компонентой, равной единице, а $\bar{\xi}_n$ — вектор-столбец, компоненты которого определяются по формулам

$$\xi_{nn} = 1, \quad \xi_{in} = -a^{(i)}{}_{in} \text{ при } i = (n-1) \div 1. \tag{3.97}$$

Матрица R_n, входящая в уравнение (3.95), принимает вид

$$R_n = \begin{pmatrix} 1 & a_{12}^{(1)} & \ldots & a^{(1)}{}_{1,n-1} & 0 \\ & 1 & \ldots & a^{(2)}{}_{2,n-1} & 0 \\ & & \ldots & \ldots & \\ & 0 & & 1 & 0 \\ & & & & 1 \end{pmatrix}.$$

После второго шага обратного хода получаем

$$U_{n-1}R_nx = U_{n-1}U_nLb \Leftrightarrow R_{n-1}x = U_{n-1}U_nLb,$$

где U_{n-1} — верхняя треугольная матрица:

$$U_{n-1} = (\bar{e}_1\bar{e}_2 \ldots \bar{\xi}_{n-1}\bar{e}_n),$$

в которой компоненты вектора-столбца $\bar{\xi}_{n-1}$ определяются по формулам

$$\xi_{n-1,n-1} = 1, \quad \xi_{i,n-1} = -a^{(i)}{}_{i,n-1} \text{ для } i = (n-2) \div 1.$$

При этом матрица R_{n-1} имеет вид

$$R_{n-1} = \begin{pmatrix} 1 & a_{12}^{(1)} & \ldots & a^{(1)}{}_{1,n-2} & 0 & 0 \\ & 1 & \ldots & a^{(2)}{}_{2,n-2} & 0 & 0 \\ & & \ldots & \ldots & & \\ & & & 1 & 0 & 0 \\ & 0 & & & 1 & 0 \\ & & & & & 1 \end{pmatrix}.$$

После m шагов обратного хода получаем

$$R_mx = U_m \ldots U_{n-1}U_nLb, \tag{3.98}$$

где матрица U_m находится по формуле

$$U_m=(\bar{e}_1\bar{e}_2\ldots\bar{e}_{m-1}\bar{\xi}_m\bar{e}_{m+1}\ldots\bar{e}_n), \tag{3.99}$$

в которой компоненты вектора-столбца $\bar{\xi}_m$ вычисляются как

$$\bar{\xi}_{m,m}=1, \quad \xi_{im}=-a^{(i)}_{im} \text{ для } i=(m-1)\div 1. \tag{3.100}$$

Матрица R_m, входящая в формулу (3.98), будет иметь вид

$$R_m=\begin{pmatrix} 1 & a_{12}^{(1)} & \ldots & a^{(1)}_{1,m-1} & 0 & \ldots & 0 \\ & 1 & \ldots & a^{(2)}_{2,m-1} & 0 & \ldots & 0 \\ & & & \ldots & \ldots & & \\ & & & 1 & 0 & \ldots & 0 \\ & & O & & 1 & \ldots & 0 \\ & & & & & & 1 \end{pmatrix}. \tag{3.101}$$

В результате на $(n-1)$-м шаге обратного хода получим

$$R_2x=U_2\ldots U_m\ldots U_nLb, \tag{3.102}$$

где

$$U_2=(\bar{e}_1\bar{\xi}_2\bar{e}_3\ldots\bar{e}_n), \quad R_2=E.$$

Введя обозначение

$$U=U_2\ldots U_m\ldots U_n, \tag{3.103}$$

запишем уравнение (3.102) в виде

$$x=ULb. \tag{3.104}$$

Отсюда при условии, что матрица системы A невырожденная, получаем

$$A^{-1}=UL \Leftrightarrow A^{-1}=$$
$$=U_2\ldots U_m\ldots U_nL_n\ldots L_k\ldots L_1. \tag{3.105}$$

Приведенное представление обратной матрицы получило название *эллиминативной формы обратной матрицы*. Вычисление обратной матрицы через ее эллиминативную форму осуществляется по формуле

$$\bar{a}'_i=\bar{e}_iUL \text{ для } i=1\div n, \tag{3.106}$$

где $\bar{a}_i'$ — i-й вектор-строка обратной матрицы $A^{-1}=(\bar{a}_1'\ldots\bar{a}_i'\ldots\ldots\bar{a}_n')^{\mathrm{T}}$; $\bar{e}_i$ — единичный вектор-строка, у которого i-я компонента равна единице. После вычисления элементов обратной матрицы нахождение решения системы $Ax=b$ сводится к процедуре вычисления скалярных произведений

$$(\bar{a}'_i, \ b)=x_i \text{ для } i=1\div n. \tag{3.107}$$

Описанный выше процесс решения системы линейных уравнений с использованием эллиминативного представления обратной матрицы предполагает выбор в качестве ведущих элементов при реализации прямого хода (вычисление матрицы L) элементов $a_{kk}^{(k-1)}$. Однако выбор ведущего элемента в описанном методе, как, например, для метода Гаусса с выбором главного элемента, может быть таким, при котором за ведущий выбирается элемент с заранее неизвестным номером. В этом случае на каждом шаге прямого хода перед процессом исключения (аннулирования) элементов соответствующего столбца осуществляются процесс выбора главного элемента по заранее заданной стратегии и затем перестановка столбцов и строк в матрице $\hat{A}^{(k)}$ таким образом, чтобы выбранный главный элемент оказался на месте с номером (k, k). С учетом указанного, процедуры представления обратной матрицы в эллиминативной форме и на ее основе решения СЛАУ преобразуются в отличие от описанных выше следующим образом. На первом шаге прямого хода будем иметь

$$L_1(P_1AQ_1)Q_1x=L_1P_1b,$$

где P_1 и Q_1 — матрицы перестановок строк и столбцов. Введя обозначения

$$P_1AQ_1=\tilde{A}, \quad L_1\tilde{A}=A^{(1)} \text{ и } Q_1x=y_1,$$

получим

$$A^{(1)}y_1=L_1P_1b.$$

На втором шаге прямого хода будем иметь

$$A^{(2)}y_2=L_2P_2L_1P_1b,$$

где

$$P_2A^{(1)}Q_2=\hat{A}^{(1)}, \quad L_2\hat{A}^{(1)}=A^{(2)} \text{ и } Q_2Q_1x=y_2.$$

И в результате n шагов прямого хода получим

$$\begin{gathered}\hat{R}y_n=L_nP_n\ldots L_2P_2L_1P_1b,\\ y_n=Q_n\ldots Q_2Q_1x,\end{gathered} \tag{3.108}$$

где $\hat{R}$ — верхняя треугольная матрица с единичной диагональю, причем

$$\hat{R}=A^{(n)} \Leftrightarrow \hat{R}=L_nP_n\ldots L_2P_2L_1P_1AQ_1Q_2\ldots Q_n. \tag{3.109}$$

После ввода обозначений

$$\hat{L}=L_nP_n\ldots L_2P_2L_1P_1 \text{ и } Q=Q_1Q_2\ldots Q_n$$

приведенная выше формула (3.108) примет вид

$$\hat{R}=\hat{L}AQ, \quad y_n=Qx. \tag{3.110}$$

Используя теперь обратный ход для решения уравнения (3.108) на основе матричных преобразований U_m, где $m=n\div 2$, получим

$$y_n = U\hat{L}b.$$

С учетом того, что $x=Qy_n$ и $x=A^{-1}b$, имеем

$$A^{-1}=QU\hat{L}. \qquad (3.111)$$

Представление обратной матрицы в виде формулы (3.111) будем называть *обобщенной эллиминативной формой обратной матрицы.* Алгоритм нахождения обратной матрицы и решения СЛАУ с использованием обобщенной эллиминативной формы будет состоять в вычислении векторов-строк $\bar{a}_i'$ и решений x_i для $i=1\div n$ по формулам

$$\bar{a}'_i=\bar{e}_iQU\hat{L}, \quad (\bar{a}'_i,\ b)=x_i. \qquad (3.112)$$

В тех случаях, когда получение обратной матрицы необязательно, алгоритм решения СЛАУ несколько упрощается и состоит в следующих вычислениях:

$$y_n=U\hat{L}b \text{ и } x=Qy_n. \qquad (3.113)$$

Алгоритм Гаусса с применением эллиминативной формы обратной матрицы будем в дальнейшем называть LU-алгоритмом. Его реализация, как следует из формул (3.112) и (3.113), связана с процедурой скалярного умножения векторов, которая может быть выполнена, во-первых, без перемножения их нулевых компонент, во-вторых, в режиме накопления. При этом эффективность LU-алгоритма будет существенным образом зависеть от числа нулевых элементов в матрице $A^{(k)}$. Это прежде всего связано с возможностью при реализации LU-алгорима в памяти ЭВМ хранить только ненулевые компоненты векторов η_k и ξ_k вместе с информацией об их адресах (номерах). И следовательно, при наличии в матрицах $A^{(k)}$ большого числа нулевых элементов удается существенно сократить требования к числу необходимых ячеек памяти ЭВМ при реализации LU-алгоритма. Естественно, что если в исходной матрице системы число нулевых элементов незначительно, то организация хранения ненулевых компонент векторов ξ_k и η_k и их номеров будет неоправданной, так как приведет только к увеличению необходимой памяти ЭВМ для решения системы.

Введем в рассмотрение *число нулевых элементов*, входящих в матрицу A. Обозначим его как $N(0)$. Тогда для нахождения числа ненулевых элементов матрицы A получим

$$N(\bar{0})=mn-N(0).$$

Определение 3.1. Матрицу A будем называть матрицей с *разреженной структурой*, если число ее нулевых элементов более чем на порядок превосходит число ее ненулевых элементов.

Из данного определения следует, что для матриц с разреженной структурой должно выполняться условие

$$N(0) \geqslant 10kN(\overline{0}),$$

где $k>1$. Кроме того, эти матрицы, как правило, высокого порядка. Среди невырожденных матриц примерами элементарных матриц с разреженной структурой являются диагональные матрицы.

Системы линейных уравнений, матрицы которых являются матрицами с разреженной структурой, будем называть *разреженными системами линейных уравнений.*

Процесс решения СЛАУ, как следует из рассмотренных уже алгоритмов, сопровождается различными преобразованиями исходной системы. Так, большинство прямых методов решения СЛАУ связано с последовательным (пошаговым) аннулированием элементов матрицы исходной системы и приведением ее к треугольному виду (или диагональному, как, например, при сингулярном разложении). В результате на каждом из таких шагов преобразований вновь получаемая матрица может обладать существенно меньшей разреженностью, чем преобразуемая.

В качестве примера рассмотрим решение СЛАУ методом Гаусса, матрица которой имеет вид

$$A=\begin{pmatrix} * & * & * & \dots & * & \dots & * \\ * & * & & & & & \\ * & & * & & & O & \\ \vdots & & & \ddots & & & \\ * & & O & & * & & \\ \vdots & & & & & \ddots & \\ * & & & & & & * \end{pmatrix} \qquad (3.114)$$

Здесь звездочкой обозначены ненулевые элементы матрицы A. После первого шага исключения будет получена матрица

$$A^{(1)}=\begin{pmatrix} * & * & * & \dots & * & \dots & * \\ 0 & * & * & \dots & * & \dots & * \\ 0 & * & * & \dots & * & \dots & * \\ \dots & & \dots & & \dots & & \dots \\ 0 & * & * & \dots & * & \dots & * \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & * & * & \dots & * & \dots & * \end{pmatrix}$$

Таким образом, все $N(0)=(n-1)(n-2)$ — нулевые элементы исходной матрицы A — в результате выполненного преобразования стали в матрице $A^{(1)}$ ненулевыми.

Определение 3.2. Образование в преобразованной матрице $B=FA$ ненулевых элементов на тех местах, на которых до выполнения преобразования F в матрице A находились нулевые элементы, будем называть *локальным заполнением матрицы A для преобразования F.*

За меру локального заполнения может быть принято как общее число локально заполненных нулевых элементов, так и коэффициент локального заполнения, определяемый по формуле

$$\eta(A, F)=\begin{cases}\eta=1, & \text{если } N_A(0)=0,\\ \eta=\overline{N}_B(0)/N_A(0), & \text{если } N_A(0)\neq 0,\end{cases} \tag{3.115}$$

где $\overline{N}_B(0)$ равно общему числу локально заполненных нулевых элементов матрицы A за счет преобразования F. Следовательно, если в матрице B в результате преобразования F над матрицей A не имеется ни одного локально заполненного элемента, то $\eta(A, F)=0$, если же все нулевые элементы матрицы A стали локально заполненными, то $\eta(A, F)=1$.

Возникают вопросы, существует ли такое преобразование F_k на каждом k-м шаге исключения метода Гаусса, при котором достигается минимальное локальное заполнение матрицы $A^{(k)}$, и если существует, то как его определить. Введем в рассмотрение матрицу B_k, полученную путем замены ненулевых элементов в подматрице A_k, которая состоит из последних $n-k$ строк и $n-k$ столбцов матрицы $A^{(k)}$. Теперь составим матрицу G_k по формуле

$$G_k=B_k(\overline{B}_k)^{\mathrm{T}}B_k, \tag{3.116}$$

где $\overline{B}_k$ — матрица, получаемая из матрицы B_k путем замены всех ее нулевых элементов единицами, а всех элементов, равных единице, — нулями.

Теорема Тьюарсона. Локальное заполнение матрицы $A^{(k)}$ будет минимальным при $(k+1)$-м шаге гауссова исключения, если в качестве главного элемента выбрать элемент $a_{st}{}^{(k)}\neq 0$, где $s=k+\alpha-1$ и $t=k+\beta-1$, а номера α и β соответствуют номерам элемента $\hat{g}_{\alpha,\beta}{}^{(k)}$, являющегося наименьшим элементом матрицы G_k.

С доказательством этой теоремы можно познакомиться в [23]. Отметим следующее: элементы матрицы G_k являются целыми положительными числами и число элементов $\hat{g}^{(k)}{}_{\sigma p \beta q}$, принимающих наименьшее значение среди значений элементов матрицы G_k, может быть несколько. Кроме того, элементы $a_{st}{}^{(k)}$, выбираемые в качестве главных в матрице $A^{(k)}$, могут

оказаться весьма малыми по абсолютному значению, что в конечном счете может повлиять на устойчивость решения СЛАУ. Для обеспечения численной устойчивости решения системы на роль главных элементов выбираются такие элементы $a_{st}^{(k)}$, которые, кроме малого локального заполнения матрицы $A^{(k)}$ (необязательно минимального), удовлетворяют и условию $|a_{st}^{(k-1)}|>\varepsilon$, где число ε задается из соображений устойчивости решения системы.

Исходя из изложенного выше, алгоритм решения СЛАУ на основе использования обобщенной эллиминативной формы обратной матрицы и минимального локального заполнения, который будем называть LU-алгоритмом минимального заполнения, должен включать такие процедуры, как:

1) вычисление на каждом шаге прямого хода элементов матрицы G_k по формуле (3.116);

2) нахождение среди них наименьших по величине элементов $\hat{g}^{(k)}{}_{\alpha_p\beta_q}$;

3) выбор среди элементов матрицы A_k тех, номера которых соответствуют значениям $s_p=k+\alpha_p-1$ и $t_q=k+\beta_q-1$;

4) нахождение среди элементов $a^{(k)}{}_{s_p t_q}$ такого элемента $a_{st}^{(k-1)}$, который удовлетворяет условию $|a_{st}^{(k-1)}|=\max|a_{s_p t_q}^{(k-1)}|$;

5) получение матриц перестановок $P_k=P_{ks}$ и $Q_k=Q_{kt}$ и вычисление по формулам (3.91) и (3.97) векторов η_k и ξ_k, входящих в состав матриц L_k и U_k;

6) вычисление решения системы по формуле (3.113) и обратной матрицы по формуле (3.112).

В качестве примера выполним первый шаг исключения для системы, матрица которой имеет размер [5×5] и структуру вида (3.114). Сначала по формуле (3.116) найдем элементы матрицы G_1

$$G_1=\begin{pmatrix}1&1&1&1&1\\1&1&0&0&0\\1&0&1&0&0\\1&0&0&1&0\\1&0&0&0&1\end{pmatrix}\begin{pmatrix}0&0&0&0&0\\0&0&1&1&1\\0&1&0&1&1\\0&1&1&0&1\\0&1&1&1&0\end{pmatrix}\begin{pmatrix}1&1&1&1&1\\1&1&0&0&0\\1&0&1&0&0\\1&0&0&1&0\\1&0&0&0&1\end{pmatrix}=$$

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

В полученной матрице G_1 наименьшими по величине являются элементы $g_{22}=g_{33}=g_{44}=g_{55}=0$. Этим элементам в матрице A соответствуют элементы a_{22}, a_{33}, a_{44}, a_{55}. Пусть среди этих эле-

ментов наибольшим по модулю является a_{55}, тогда именно он и выбирается в качестве главного. Одновременно по значениям элементов g_{ij} можно определить, каково будет локальное заполнение матрицы $A^{(1)}$, если в качестве главного элемента будет выбран элемент a_{ij} с номером, соответствующим номеру элемента g_{ij}. Так, если в качестве главного элемента был бы выбран элемент a_{11}, то локальное заполнение было бы равно 12 и это соответствовало бы методу исключения по обычной схеме Гаусса. Для рассматриваемого же алгоритма, когда в качестве главного элемента выбран элемент a_{55}, локальное заполнение будет равно нулю. При этом $P_1 = P_{15}$ и $Q_1 = Q_{15}$.

Оценка производительности LU-алгоритма минимального заполнения свидетельствует о необходимости выполнения довольно значительной вычислительной работы, ведущей к минимальному локальному заполнению, которое к тому же не всегда приводит к *минимальному глобальному заполнению.* Поэтому во многих случаях более рационально использовать стратегию выбора главного элемента, предложенную Марковицем, приводящую к меньшим вычислительным затратам, хотя и порождающую заполнение, возможно, и не минимальное, но достаточно малое. Суть стратегии Марковица состоит в следующем. Пусть $N_i^{(k)}(\overline{0})$ обозначает число ненулевых элементов в i-й строке матрицы A_k, а $N_j^{(k)}(\overline{0})$ — число ненулевых элементов в j-м столбце A_k.

Определение 3.3. *Ценой Марковица* элемента $a_{ij}^{(k-1)}$ матрицы A_k называется число

$$M^{(k)}_{ij} = (N_i^{(k)}(\overline{0}) - 1)(N_j^{(k)}(\overline{0}) - 1), \quad i, j = k \div n. \tag{3.117}$$

Цена $M_{ij}^{(k)}$ показывает число элементов, изменяющих свое значение при переходе от $A^{(k)}$ к $A^{(k+1)}$, если в качестве главного элемента выбран элемент $a_{ij}^{(k-1)}$. Положим

$$M^{(k)}_{\alpha_p \beta_q} = \min_{i,j} \{M^{(k)}_{i,j},\ i, j = k \div n\}. \tag{3.118}$$

Далее по номерам α_p и β_q будем выбирать из A_k элементы $a_{\alpha_p \beta_q}^{(k-1)}$, а затем среди этих элементов — главный элемент $a_{st}^{(k-1)}$, удовлетворяющий условию

$$| a_{st}^{(k-1)} | = \max_{p,q} | a_{s_p t_q}^{(k-1)} | .$$

Таким образом, начиная с 3-го пункта, вычисления по LU-алгоритму минимального заполнения и LU-алгоритму с ценой Марковица производятся по однотипной схеме.

Рассмотрим выбор главного элемента на первом шаге исключения для приведенного выше примера (3.114). Цены Марковица можно объединить в виде матрицы $M^{(1)} = \{M_{ij}^{(1)}\}$. Тогда для данного примера матрица $M^{(1)}$ примет вид

$$M^{(1)}=\begin{pmatrix} 16 & 4 & 4 & 4 & 4 \\ 4 & 1 & * & * & * \\ 4 & * & 1 & * & * \\ 4 & * & * & 1 & * \\ 4 & * & * & * & 1 \end{pmatrix},$$

где звездочками обозначены ее невычисляемые элементы, положения которых соответствуют нахождению нулевых элементов в матрице A. Из анализа элементов $M_{ij}^{(1)}$ следует, что только a_{22}, a_{33}, a_{44}, a_{55} могут быть в числе главных элементов на первом шаге исключения. И при условии, что $|a_{55}|=\max\{|a_{22}|, |a_{33}|, |a_{44}|, |a_{55}|\}$, за главный элемент выбирается a_{55}. Кстати, этот результат совпадает с результатом, когда стратегия выбора главного элемента в том же примере осуществлялась за счет минимизации локального заполнения.

Таким образом, из рассмотренного примера видно, что только за счет оптимальной стратегии выбора главных элементов в схеме Гаусса можно существенно сократить по сравнению с традиционной схемой Гаусса локальное заполнение исходной матрицы системы.

Пусть теперь требуется решить систему $Ax\cong b$ по методу наименьших квадратов, причем структура матрицы размера $[n\times m]$ является разреженной и $n\geqslant m$. Используя традиционную схему перехода от исходного уравнения к нормальной системе уравнения при условии, что матрица Грама $G=E$, получим

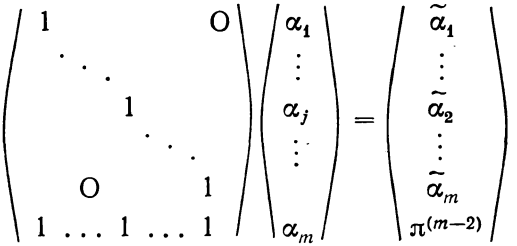
$$A^{\tau}Ax^*=A^{\tau}b \Leftrightarrow Tx^*=d, \tag{3.119}$$

где матрица $T=A^{\tau}A$ и имеет размер $[m\times m]$, а $d=A^{\tau}b$. В результате выполненного перехода разреженность структуры матрицы T может стать значительно меньшей, чем матрицы A, за счет локального заполнения последней. В подтверждение этого рассмотрим систему уравнений вида

$$\begin{pmatrix} a_{11} & & & & \mathrm{O} \\ & \ddots & & & \\ & & a_{jj} & & \\ & & & \ddots & \\ \mathrm{O} & & & & a_{mm} \\ a_{m+1,1} & \cdots & a_{m+1,j} & \cdots & a_{m+1,m} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_j \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_j \\ \vdots \\ b_m \\ b_{m+1} \end{pmatrix} \tag{3.120}$$

Кстати, заметим, что к системе вида (3.120) приводится следующая задача. Допустим, измерены внутренние углы выпуклого m-угольника $\alpha_1\ldots$, α_j, $\ldots$, α_m. Требуется, используя только одно уравнение связи $\alpha_1+\alpha_2+\ldots+\alpha_j+\ldots+\alpha_m=\pi(m-2)$, полу-

чить уравненные значения выполненных измерений. Системой уравнений поправок для этой задачи будет



Здесь $\tilde{\alpha}_i$ обозначает измеренное значение внутреннего угла α_j. Сравнение этой системы с системой (3.120) показывает, что они однотипные по своей структуре.

Если в результате традиционных преобразований (3.119) осуществить переход от системы (3.120) к соответствующей нормальной системе уравнений, то, как нетрудно убедиться в полученной нормальной системе уравнений, матрица T будет иметь полностью заполненную структуру. И это не позволит применить для данной нормальной системы уравнений описанные выше эффективные вычислительные алгоритмы, основанные на использовании разреженности структуры матрицы системы. Поэтому поставим задачу построения такого алгоритма решения СЛАУ по методу наименьших квадратов, при котором переход от исходной к нормальной системе уравнений сохранял бы высокую разреженность системы. В том случае, когда подобный переход выполнен, решение полученной нормальной системы уравнений целесообразно осуществлять по алгоритмам решения разреженных систем уравнений.

Для реализации поставленной задачи от исходной системы $Ax \cong b$, которая может быть и несовместной, перейдем за счет введения дополнительных неизвестных x_{m+i} к совместной системе уравнений вида

$$(\bar{a}_i, x) - x_{m+i} = b_i \text{ для } i = 1 \div n. \tag{3.121}$$

Здесь $\bar{a}_i$ — i-я строка матрицы A. Напомним, что мы приняли $n \geqslant m$. В системе (3.121) число неизвестных равно $(n+m)$, а число уравнений — n. Решение системы (3.121) будем выполнять при соблюдении условия

$$\min \sum_{i=1}^{n} x^2_{m+i}. \tag{3.122}$$

Используя для этого метод неопределенных множителей Лагранжа, получим выражение для определения условного минимума

$$\min\{\sum_{i=1}^{n} x^2_{m+i} + \sum_{k=1}^{n} \lambda_k[(\bar{a}_k, x) - x_{m+k} - b_k]\}.$$

Из этого выражения на основании необходимых условий существования экстремума функции нескольких переменных для $k=1\div n$ имеем

$$(\bar{a}_k, \hat{x}) - 0{,}5\lambda_k = b_k,$$
$$(\bar{a'}_k, \bar{\lambda}) = 0,$$

где $\hat{x}$ — вектор решения уравнения (3.121), удовлетворяющий условию (3.122), $\bar{\lambda}$ — вектор-столбец, компонентами которого являются неопределенные множители Лагранжа; a_l' — l-я строка матрицы A^{T}. Полученную систему уравнений можно записать в блочном виде

$$\begin{pmatrix} A & E_n \\ O_m & A^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \hat{x} \\ \bar{\lambda}^* \end{pmatrix} = \begin{pmatrix} b \\ \bar{o} \end{pmatrix}, \qquad (3.123)$$

где $\bar{\lambda}^*$ — вектор-столбец, определяемый как $\bar{\lambda}^* = 0{,}5\bar{\lambda}$; E_n — единичная матрица размера $[n\times n]$; O_m — нулевая матрица размера $[m\times m]$; $\bar{o}$ — нулевой подвектор размера $[m\times 1]$. Введя в рассмотрение вектор невязки $v=Ax-b$, из системы (3.123) непосредственно вытекает, что подвектор ее решения $\hat{x}$ совпадает с псевдорешением системы $Ax\cong b$, т. е. $\hat{x}=x^*$. Кроме того, так как $A^{\mathrm{T}}v=0$, то $v=\bar{\lambda}^*$. Отсюда следует, что систему (3.123) можно представить в виде

$$\begin{pmatrix} A & E_n \\ O_m & A^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} x^* \\ v \end{pmatrix} = \begin{pmatrix} b \\ \bar{o} \end{pmatrix}. \qquad (3.124)$$

Система (3.124) получила название *расширенной системы* уравнений по отношению к системе $Ax\cong b$. Таким образом, для вычисления псевдорешения можно использовать не только нормальную, но и расширенную систему уравнений, которая, как нетрудно заметить, полностью сохраняет разреженность исходной системы уравнений.

В качестве примера найдем псевдорешение системы (3.120), используя для этого расширенную систему уравнений и LU-алгоритм с ценой Марковица. Система (3.124) в рассматриваемом примере будет иметь вид

$$\begin{pmatrix} D & E_m & \bar{o} \\ \bar{a}_{m+1} & \bar{o}^{\mathrm{T}} & 1 \\ O_m & D & \bar{a}^{\mathrm{T}}_{m+1} \end{pmatrix} \begin{pmatrix} x^* \\ v \end{pmatrix} = \begin{pmatrix} b \\ \bar{o} \end{pmatrix},$$

где D — диагональная матрица с элементами a_{jj} $(j=1\div m)$, определяемыми из (3.120); $\bar{a}^{\mathrm{T}}_{m+1}$ — вектор-столбец, транспонированный к вектору-строке $\bar{a}_{m+1}$ матрицы системы (3.120). Можно убедиться, что минимальные цены Марковица (и минимальное заполнение на каждом шаге прямого хода) будут обеспечиваться, если в качестве главных элементов при прямом ходе будут выбраны следующие элементы: $a_{kk}{}^{(k-1)}=a_{kk}$ при $k=1\div m$, $a^{(l-1)}{}_{l+1,\,l}$ при $l=(m+1)\div 2m$ и $a^{(2m)}{}_{2m+1,\,2m+1}$. В результате после прямого хода приведенная выше система преобразуется к виду

$$\begin{pmatrix} E_m & D^{-1} & \bar{O} \\ O_m & E_m & D^{-1}\bar{a}^{\mathrm{T}}_{m+1} \\ \bar{O}^{\mathrm{T}} & \bar{O}^{\mathrm{T}} & 1 \end{pmatrix} \begin{pmatrix} x^* \\ v \end{pmatrix} = \begin{pmatrix} D^{-1}\bar{b}_1 \\ \bar{O} \\ b^*{}_{m+1} \end{pmatrix},$$

где значение $b^*{}_{m+1}$ вычисляется по формуле

$$b^*{}_{m+1}=\left(b_{m+1}-\sum_{i=1}^{m}\frac{b_i a_{m+1,i}}{a_{ii}}\right)\left(1+\sum_{i=1}^{m}\left(\frac{a_{m+1,i}}{a_{ii}}\right)^2\right)^{-1}.$$

Так как в LU-алгоритмах умножение на нулевые элементы исключается, то из полученных зависимостей легко подсчитать общее число операций умножения и деления, необходимых для нахождения псевдорешения системы (3.120) по изложенному алгоритму

$$N^{(\text{об})}{}_{\text{оп}}=N_{\text{пр.х}}+N_{\text{обр.х}}\cong(4m+1)+2m\cong 6m.$$

При этом объем памяти ЭВМ, требуемый для хранения как исходной информации, так и информации о результатах промежуточных вычислений, должен отводиться только под хранение $3n$ чисел.

Если теперь сравнить решение системы (3.120) по традиционной схеме метода наименьших квадратов, т. е. по схеме (3.119) и по вычислительной схеме, основанной на решении системы (3.124) по LU-алгоритму с минимизацией локального заполнения, то получим, что вторая схема позволяет сократить в $(m/3)^2$ раза количество операций умножения и в $(m/2)$ раза объем информации, необходимый для хранения в ЭВМ. Причем эти оценки практически неулучшаемы. Сравнение указанных вычислительных схем для других по структуре разреженности систем уравнений можно характеризовать коэффициентами выигрыша по быстродействию $\eta_{\text{оп}}=N_{\text{оп}}{}^{\text{тр}}/N_{\text{оп}}{}^{\text{об}}$ и сокращения необходимого объема памяти $\eta_{\text{яч}}=N_{\text{яч}}{}^{\text{тр}}/N_{\text{яч}}{}^{\text{об}}$, значения которых, как уже указывалось, в общем случае будут ниже, чем для системы (3.120), и целиком зависеть от структуры матрицы A, т. е.

$$(m/3)^2 \geqslant \eta_{оп} \geqslant 1, \quad m/2 \geqslant \eta_{яч} \geqslant 1. \qquad (3.125)$$

Причем $\eta_{оп}=1$ и $\eta_{яч}=1$ в том случае, когда матрица A полностью заполнена, т. е. $N(O)=0$.

§ 3.8. ОРГАНИЗАЦИЯ ХРАНЕНИЯ ЧИСЛОВОЙ ИНФОРМАЦИИ В ЭВМ ПРИ РАБОТЕ С РАЗРЕЖЕННЫМИ СИСТЕМАМИ УРАВНЕНИЙ

С точки зрения выделения ресурсных вычислительных средств под решение СЛАУ с матрице разреженной структуры самым важным параметром является возможность экономного хранения в ЭВМ исходных данных и результатов промежуточных вычислений, так как в конечном счете для больших систем уравнений от этого параметра в значительной степени будет зависеть и такой важный параметр любого вычислительного процесса, как быстродействие решения задачи, например за счет сокращения наиболее длинных операций ЭВМ — числа обращений к внешним запоминающим устройствам. Для ряда систем, матрицы которых имеют специальную разреженную структуру, например (3.120), удается обеспечить хранение в ЭВМ только ненулевых элементов исходной матрицы и матриц, получаемых в процессе промежуточных вычислений по LU-алгоритму, без организации хранения номеров элементов этих матриц. Это возможно в тех случаях, когда расположение запоминаемых элементов можно описать в виде аналитических выражений. Для матриц с разреженной структурой общего вида хранение в ЭВМ ее ненулевых элементов всегда должно сопровождаться запоминанием номеров, однозначно определяющих положение этих элементов в матрице. Такой вид хранения в ЭВМ элементов матриц получил название хранения в упакованной форме.

Одним из распространенных в вычислительной практике, и в частности при решении СЛАУ, способов хранения в памяти ЭВМ элементов разреженной матрицы в упакованной форме является использование связных списков. Опишем суть этого способа хранения для матрицы размера $[n \times m]$, содержащей τ ненулевых элементов a_{ij}, предусмотрев возможность экономной записи в памяти ЭВМ ненулевых элементов, вновь образуемых при преобразовании исходной матрицы. Под хранение всей матрицы выделим две части памяти: память для начальных адресов столбцов (П. Н. А) и память для записи ненулевых элементов матрицы (П. Э. З.). Первая часть памяти П. Н. А. содержит m последовательно расположенных ячеек памяти, меньший адрес которых соответствует и меньшему номеру столбца матрицы. В ячейках П. Н. А. записываются адреса (условные номера) первых ненулевых элементов соответству-

ющих столбцов матрицы. Например, если адресом (условным номером) первого ненулевого элемента a_{sj} j-го столбца матрицы A является s_j, то в j-й ячейке в П. Н. А. будет записана положительная константа целого типа s_j, соответствующая номеру s. Вторая часть памяти П. Э. З. состоит из *записей*, связанных с ненулевыми элементами матрицы A. Здесь под записью элемента a_{ij} будем понимать упорядоченную тройку значений i, a, p, где i — номер строки, a — значение элемента a_{ij} и p — адрес следующего ненулевого элемента j-го столбца. Если же запись соответствует последнему ненулевому элементу столбца j, то значение p принимается равным нулю. Будем предполагать, что записи хранятся по столбцам и под каждое из трех значений одной записи отводится одна ячейка памяти. Таким образом, на одну запись приходятся три ячейки памяти, П. Э. З. состоит из 3τ ячеек памяти, а для хранения матрицы A с использованием связанных списков требуется объем памяти в $n+3\tau$ ячеек.

Главное преимущество описанной схемы хранения заключается в том, что новые ненулевые элементы, образующиеся в столбцах в процессе вычислений, могут быть достаточно просто размещены в П. Э. З. Более того, все записи в П. Э. З. не обязательно должны быть сосредоточены в одной области памяти, а могут быть разбросаны по всей доступной памяти ЭВМ (но, конечно, группами из трех последовательных ячеек).

Покажем на примере, каким образом появление нового ненулевого элемента влияет на информацию, записываемую в ячейки памяти П. Н. А и П. Э. З. Пусть вначале в 3-м столбце матрицы первым ненулевым элементом являлся элемент $a_{23}=0{,}5$, а последним — $a_{43}=1{,}5$. Предположим, что начальная ячейка П. Н. А. имеет адрес 101. Тогда адрес ячейки П. Н. А., соответствующей третьему столбцу матрицы, будет равен 103 (табл. 4). Пусть начальные адреса записей для a_{23} и a_{43} соответственно равны 200 и 203. Далее предположим, что в результате процесса исключения при прямом ходе алгоритма Гаусса произошло локальное заполнение элемента a_{33} (т. е. нулевое значение a_{33} перешло, допустим, в значение 2,5) и запись для a_{33} размещается в ячейках, первая из которых имеет адрес 300 (вариант 1).

Таким образом, включение нового ненулевого элемента потребовало лишь изменения содержимого 202-й ячейки памяти.

Рассмотрим теперь вариант 2 изменения исходной матрицы, когда вместо элемента $a_{33}^{(k)}$ ненулевым стал бы элемент $a_{13}^{(k)}$ (например, $a_{13}^{(k)}=3{,}5$). Соответствующая ему запись хранится (как и для варианта 1) начиная с 300-й ячейки. Как видно из табл. 4, и в этом случае, для того чтобы включить новый ненулевой элемент, в исходном связном списке необходимо изменить только содержимое одной ячейки 103.

Таблица 4

Адреса ячеек памяти ЭВМ	103	200	201	202	203	204	300	301	302
Текущее содержание ячеек (исходная информация)	200	2	0,5	203	4	1,5	—	—	—
Новое содержание ячеек при $a_{33}^{(k)} \neq 0$ (вариант 1)	200	2	0,5	300	4	1,5	3	2,5	203
Новое содержание ячеек при $a_{13}^{(k)} \neq 0$ (вариант 2)	300	2	0,5	203	4	1,5	1	3,5	200

Если в процессе вычислений (например, при прямом ходе) некоторые ненулевые элементы становятся равными нулю, то ячейки памяти, занятые записями, соответствующими этим элементам, освобождаются и могут быть использованы для хранения записей новых ненулевых элементов. С этой целью можно организовать новые связные списки для хранения начальных адресов таких освободившихся записей. Реализуется это следующим образом. Под хранение начального адреса первой свободной записи в ЭВМ отводится отдельная ячейка памяти. Третья ячейка каждой свободной записи, включая и первую запись, должна содержать адрес следующей свободной записи, если же данная свободная запись является последней в списке свободных записей, то третья ячейка должна содержать нуль. Первая и вторая ячейки каждой свободной записи при этом остаются свободными.

Для иллюстрации описанного способа образования списков освободившихся записей рассмотрим такой пример. Пусть свободными записями являются записи с начальными адресами 101 и 201 и требуется добавить к списку еще одну освободившуюся запись с начальным адресом 301. Если ячейка 50 предназначена для хранения в ней адреса первой свободной записи, то требуемые изменения в содержимом ячеек памяти могут быть представлены результатами, приведенными в табл. 5.

Кроме связных списков существуют и другие виды упаковок матриц, которые в отдельных случаях позволяют обеспечить более эффективное хранение информации с точки зрения необходимого объема памяти ЭВМ или с точки зрения выбора

Таблица 5

Адреса ячеек памяти ЭВМ	50	101	102	103	201	202	203	301	302	303
Текущее содержание ячеек	101	—	—	201	—	—	0	—	—	—
Новое содержание ячеек	301	—	—	201	—	—	0	—	—	101

времени (считывания) этой информации из памяти (внешней, оперативной) ЭВМ. Более подробно с этими вопросами можно ознакомиться в [25].

Глава 4

ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ СИСТЕМ УРАВНЕНИЙ

Рассмотренные в главе 3 прямые методы решения СЛАУ при условии исключения ошибок в исходной информации (ошибок измерения) и ошибок вычислений (ошибок округлений) приводят к точному решению. В отличие от прямых методов в итерационных методах получают последовательность приближенных решений и достаточно близкий к пределу член этой последовательности принимается за решение системы. Таким образом, итерационные методы решения систем имеют некоторую ошибку даже при условии, что ошибки измерения и округления отсутствуют. Однако из этого факта, безусловно, нельзя делать вывод о том, что итерационные методы решения являются менее точными, чем прямые. Так, если рассматривать *сходящиеся итерационные методы решения систем*, то в ряде случаев ошибка решения в них может быть значительно меньше, чем ошибка, вызванная ошибками округления при получении решения прямым методом. Причем в процессе решения системы итерационным методом удается достаточно простыми способами выполнять и оценку точности вычисленного решения. Исходя из отмеченных выше преимуществ итерационных методов решения систем, их часто используют совместно с прямыми методами для уточнения решения последних и оценки точности полученного решения.

§ 4.1. АЛГОРИТМ ИТЕРАЦИОННОГО УТОЧНЕНИЯ ПРИБЛИЖЕННОГО РЕШЕНИЯ

Изложим один из наиболее распространенных методов итерационного уточнения приближенного решения СЛАУ, где в качестве приближенного выступает решение, полученное прямым

методом. Этот метод опишем только применительно к прямым методам, основанным на LU-разложениях матрицы системы, хотя этот метод может быть использован в комбинации с любым другим прямым методом решения СЛАУ.

Обозначим вектор, полученный после решения системы $Ax=b$ прямым методом, через $x^{(0)}$. Вычислим соответствующую невязку $v^{(0)}=b-Ax^{(0)}$. Тогда вектор ошибки решения равен $d^{(0)}=x^{\tau}-x^{0}$, где $x^{(\tau)}$ — точное решение, и удовлетворяет уравнению

$$Ad^{(0)}=v^{(0)}. \tag{4.1}$$

Для обобщенного LU-алгоритма эта формула преобразуется к виду

$$x^{(0)}=QU^{-1}L^{-1}Pb, \quad d^{(0)}=QU^{-1}L^{-1}Pv^{(0)},$$

где Q и L — матрицы перестановок. При этом за первое уточненное значение решения системы берется

$$x^{(1)}=x^{(0)}+d^{(0)}.$$

Выполняя аналогичные по виду вычисления с $x^{(1)}$, вместо $x^{(0)}$ получим второе уточненное значение

$$v^{(1)}=b-Ax^{(1)},$$
$$d^{(1)}=QU^{-1}L^{-1}Pv^{(1)},$$
$$x^{(2)}=x^{(1)}+d^{(1)}.$$

Получение q-го уточненного значения решения СЛАУ связано с выполнением таких операций для $i=0\div(q-1)$:

$$\begin{aligned} v^{(i)}&=b-Ax^{(i)},\\ d^{(i)}&=QU^{-1}L^{-1}Pv^{(i)},\\ x^{(i+1)}&=x^{(i)}+d^{(i)}. \end{aligned} \tag{4.2}$$

При этом значение $x^{(q)}$ принимается за окончательное решение системы, если для него выполняются условия

$$\| x^{(q)}-x^{(q-1)}\| < \bar{\varepsilon}_{\text{окр}} \| x^{(q)}\| \tag{4.3}$$

либо начиная с $q=2$

$$\| d^{(q)}\| > \| d^{(q-1)}\|, \tag{4.4}$$

либо

$$q=MAXIT. \tag{4.5}$$

Здесь $\bar{\varepsilon}_{\text{окр}}=p^{-t+1}$ — относительная ошибка округления, вызванная вычислениями на ЭВМ со словами, под мантиссу которых отведено t разрядов, а $MAXIT$ — максимальное число итераций, задаваемое или вычисляемое до решения СЛАУ.

Найдем условие, при выполнении которого алгоритм итерационного уточнения сходится. С этой целью преобразуем выражение для $d^{(i)}$, входящее в формулы (4.2), за счет подстановки в него значения $v^{(i)}$ из первой формулы (4.2):

$$d^{(i)} = QU^{-1}L^{-1}Pb + QU^{-1}L^{-1}PAx^{(i)}.$$

Если предположить, что при реализации LU-алгоритма возникают суммарные ошибки округления, описываемые матрицей возмущения G, т. е.

$$A = PLUQ - G, \tag{4.6}$$

то приведенное выше выражение для $d^{(i)}$ примет вид

$$d^{(i)} = x^{(0)} - x^{(i)} + QU^{-1}L^{-1}PGx^{(i)}.$$

Обозначая матрицу $QU^{-1}L^{-1}PG$ через F, получаем

$$F = QU^{-1}L^{-1}PG, \quad d^{(i)} = x^{(0)} - x^{(i)} + Fx^{(i)}. \tag{4.7}$$

После подстановки полученного значения $d^{(i)}$ в выражение для $x^{(i+1)}$ в формуле (4.2) будем иметь

$$x^{(i+1)} = x^{(0)} + Fx^{(i)}. \tag{4.8}$$

Выполняя последовательную подстановку в эту формулу значений $x^{(i)}$ для $i = 0 \div (q-1)$, получим

$$x^{(1)} = x^{(0)} + Fx^{(0)}, \quad x^{(2)} = x^{(0)} + Fx^{(1)} \Rightarrow$$
$$\Rightarrow x^{(2)} = x^{(0)} + Fx^{(0)} + F^2x^{(0)} \ldots, x^{(q)} =$$
$$= (E_0 + F + F^2 + \ldots + F^q)x^{(0)}. \tag{4.9}$$

На основании формулы (4.6) можно записать

$$Ax = PLUQx - Gx \Leftrightarrow QU^{-1}L^{-1}Pb = x^{(\tau)} - Fx^{(\tau)} \Leftrightarrow$$
$$\Leftrightarrow x^{(0)} = x^{(\tau)} - Fx^{(\tau)}.$$

Подставив это значение $x^{(0)}$ в формулы (4.9), получим

$$x^{(\tau)} - x^{(q)} = (E + F + \ldots + F^q)Fx^{(\tau)} -$$
$$- (E + F + \ldots + F^{q-1})Fx^{(\tau)} \Leftrightarrow x^{(\tau)} - x^{(q)} = F^q x^{(\tau)}. \tag{4.10}$$

Теперь значение $x^{(q)}$, определяемое по формуле (4.9), используем для нахождения $d^{(q)}$ по формуле (4.7)

$$d^{(q)} = x^{(0)} - x^{(q)} + Fx^{(q)} \Rightarrow d^{(q)} =$$
$$= x^{(0)} - \sum_0^q F^i x^{(0)} + \sum_0^q F^{(i+1)} x^{(0)}.$$

Отсюда имеем

$$d^{(q)} = F^{q+1}x^{(0)} \Leftrightarrow d^{(q)} = F^q d^{(0)}. \tag{4.11}$$

На основе полученных равенств (4.10) и (4.11) достаточно просто находятся признаки сходимости алгоритма итерационного уточнения (4.2).

Теорема 4.1. Предположим, что все вычисления по алгоритму (4.2) выполняются без ошибок округления. Тогда для сходимости алгоритма итерационного уточнения (4.2), т. е. для того, чтобы $x^{(q)} \to x^{\tau}$ при $q \to \infty$, необходимо и достаточно, чтобы наибольший модуль собственных значений матрицы F был строго меньше единицы, или достаточно, чтобы для согласованной нормы матрицы F выполнялось условие $\|F\| < 1$.

Доказательство. Из определения сходимости последовательности векторов сходимость последовательности (4.9) будет обеспечена, если

$$\lim_{g \to \infty} \| x^{(\tau)} - x^{(q)} \| = 0.$$

Это выражение с учетом формулы (4.10) преобразуется к виду

$$\lim_{g \to \infty} \| F^q x^{(\tau)} \| = 0.$$

После перехода к согласованной норме матрицы получим следующее выражение, определяющее сходимость последовательности (4.9) к точному решению $x^{(\tau)}$:

$$\lim_{g \to \infty} \| F^q \| = 0. \tag{4.12}$$

Достаточным признаком выполнения данного условия, как следует из следствия 1 теоремы 1.29, является требование того, чтобы

$$\| F \| < 1, \tag{4.13}$$

а необходимым и достаточным признаком выполнения условия (4.12), как следует из следствия 2 теоремы 1.29, является неравенство

$$| \lambda_1(F) | < 1. \tag{4.14}$$

Здесь $|\lambda_1(F)| = \max |\lambda_i(F)|$, где $\lambda_i(F)$ — собственные значения матрицы F.

Следствие. Если выполнены условия теоремы 4.1, то

$$\lim_{g \to \infty} d^{(q)} = 0. \tag{4.15}$$

Переходя к согласованной норме матрицы, в формуле (4.11), получаем

$$\| d^{(q)} \| \leqslant \| F^i \| \, \| d^{(0)} \|.$$

Отсюда при условии, что $\|d^{(0)}\| \neq 0$, вытекает доказательство следствия.

Остановимся теперь на некоторых особенностях алгоритма итерационного уточнения (4.2)—(4.5). Вначале поясним, чем вызвана необходимость применения в этом алгоритме трех критериев остановки (4.3), (4.4) и (4.5). Так, при $|\lambda_1(F)| \ll 1$ итерационный процесс (4.2) будет быстро сходиться и, как правило, будет остановлен условием (4.3). Если $|\lambda_1(F)| < 1$, но значение $|\lambda_1(F)|$ близко к единице, то сходимость процесса (4.2) может быть очень медленной и даже в отдельных случаях из-за ошибок округлений при вычислениях по формулам (4.2) не обеспечивать выполнения условия (4.3). В этих случаях итерации будут остановлены условием (4.5). Если же $|\lambda_1(F)| \geqslant 1$, то процесс (4.2) расходится, в результате чего соотношение (4.15) не будет выполняться, а это будет обнаружено при проверке условия (4.4).

Как было подчеркнуто выше, теорема 4.1 справедлива только при предположении, что алгоритм итерационного уточнения выполняется без ошибок округления. Однако при реализации алгоритма (4.2) ошибки вычисления невязок $v^{(i)}$ сравнимы по величине с самими невязками. Но если правая часть выражения для вычисления $d^{(i)}$ по алгоритму (4.2) не имеет верных знаков, то алгоритм итерационного уточнения не может привести к уточнению приближенного решения $x^{(i)}$. Исходя из этого, невязки v_i обычно вычисляют с увеличенной разрядностью, например с двойной точностью на ЕС ЭВМ.

Будучи сходящимся, алгоритм итерационного уточнения по сравнению с прямым методом (например, LU-алгоритмом) дает более точное решение и правдоподобную оценку ошибки. Но это приводит к необходимости использовать дополнительную память для хранения матрицы A и к увеличению времени для реализации вычислений по формулам (4.2)—(4.5). Укажем требования к памяти и времени счета (через число арифметических операций для LU-алгоритма и алгоритма итерационного уточнения).

LU-алгоритм

Память	n^2
Время	$1/3\, n^3$

Алгоритм итерационного уточнения

Память	$2n^2$
Время	$1/3 n^3 + 2qn^2$

Дополнительные вычислительные ресурсы, которые необходимы для реализации алгоритма итерационного уточнения, весьма значительны, особенно если идет речь о решении больших систем. К тому же прямые методы в ряде случаев дают и приемлемую точность решения.

Поэтому при обработке больших систем с матрицами разреженной структуры применение итерационного алгоритма име-

ет несколько другое предназначение, связанное в основном с ограничением локального заполнения при каждом шаге исключения прямого метода. Это может достигаться как за счет естественного обнуления на каждом шаге вновь образуемого элемента, если абсолютная величина его не превосходит произведения единичной ошибки округления $\varepsilon_{\text{окр}}$ на модуль исключаемого элемента, так и за счет отбрасывания малых по модулю элементов, появляющихся в ходе исключения. Данный процесс обнуления элементов в преобразованной матрице осуществляется по величине вводимого порога T, и каждый раз, как процесс исключения порождает новый элемент, меньший T по абсолютной величине, этот элемент заменяется нулем.

Если порог $T_0=\varepsilon_{\text{окр}}a_{ij}^{(k)}$, то это соответствует процессу естественного обнуления. Как правило, значение порога T выбирается существенно большим значения T_0, в результате чего в преобразованной матрице появляется и значительно большее число нулевых элементов, но, с другой стороны, это и приводит к ухудшению точности решения СЛАУ прямым методом. Поэтому для восстановления утерянной точности весьма эффективен рассмотренный выше процесс итерационного уточнения. При этом вычисления, связанные с реализацией алгоритма (4.2)—(4.5), требуют немного времени по сравнению с LU-разложением матрицы системы. Таким образом, если при введении порога T удается сохранить высокую разреженность исходной СЛАУ, то ее решение с использованием алгоритма итерационного уточнения по сравнению с чисто прямыми методами решения позволяет уменьшить необходимый объем памяти и сократить время счета на ЭВМ за счет выполнения арифметических операций только над элементами, имеющими ненулевые значения.

В работе [25] приведен пример решения нормальной системы уравнения, матрица которой размера $[10^3\times10^3]$ имеет следующую структуру:

$$A=\begin{pmatrix}
* & * & 0 & 0 & * & 0 & . & . & . & 0\\
* & * & * & 0 & 0 & * & 0 & & & .\\
0 & * & * & * & 0 & 0 & * & . & & .\\
0 & 0 & * & * & * & 0 & 0 & . & . & .\\
* & 0 & 0 & * & * & * & 0 & . & . & 0\\
0 & * & 0 & 0 & * & * & * & . & . & *\\
. & 0 & * & 0 & 0 & * & * & * & * & 0\\
. & . & . & . & . & . & . & . & . & .\\
0 & & & . & * & . & . & * & * & *\\
0 & . & . & . & 0 & * & 0 & 0 & * & *
\end{pmatrix}$$

Здесь звездочками отмечены ненулевые элементы матрицы A.

Таблица 6

Алгоритм	Память ячеек	Время счета, с	Ошибка $\|x-x\|$
LU-алгоритм	45 850	182.31	$2{,}02\cdot10^{-1}$
IR-алгоритм	14 082	8.50	$1{,}83\cdot10^{-6}$

В табл. 6 даны результаты решения этой системы прямым методом (*LU*-алгоритмом) и методом итерационного уточнения совместно с *LU*-алгоритмом, который в таблице обозначен как *IR*-алгоритм. При реализации *IR*-алгоритма был задан барьер $T=0{,}01$ и при вычислениях как по *LU*-, так и по *IR*-алгоритму умножения на нулевые элементы не производились. Кроме того, организация хранения ненулевых элементов осуществлялась в этих алгоритмах с использованием связных списков.

При реализации *IR*-алгоритма потребовалось 16 итераций, что составило всего 10% общего времени решения системы.

Теоремой 4.1 были установлены признаки сходимости алгоритма итерационного уточнения, одним из которых является $\|F\|<1$. Используя этот признак сходимости и формулу (4.10), относительную оценку точности решения СЛАУ по *IR*-алгоритму можно определять как

$$\varepsilon_{x_q}=\frac{\|x^{(\tau)}-x^{(q)}\|}{\|x^{(\tau)}\|}\leqslant\|F\|^{q+1}\ \text{для}\ q>1.$$

С учетом выражений (4.7) эта формула преобразуется к виду

$$\varepsilon_{x_q}\leqslant(\|A^{-1}\|\,\|G\|)^{q+1}\Leftrightarrow\varepsilon_{x_q}\leqslant\left(K^{-}(A)\frac{\|G\|}{\|A\|}\right)^{q+1}\ \text{для}\ q>1. \quad (4.16)$$

В этой формуле для оценки отношения $\|G\|/\|A\|$ можно использовать формулы (2.79). Отметим также, что формула (4.16) получена при условии, что ошибки округления, связанные с вычислениями по формулам (4.2), не учитываются.

Анализ формулы (4.16) показывает, что, чем лучше обусловлена матрица системы (т. е. меньше $K^{-}(A)$), тем процесс итерационного уточнения сходится быстрее. Кроме того, эта формула может служить и для выбора порога T при решении систем с матрицами разреженной структуры, а именно: необходимо выбирать порог T таким образом, чтобы не нарушалось условие

$$K^{-}(A)\frac{\|G(T)\|}{\|A\|}<1. \quad (4.17)$$

В этой формуле $G(T)$ — возмущение *LU*-алгоритма за счет введения порога T.

Алгоритм итерационного уточнения можно применить и для решения систем линейных уравнений общего вида по методу наименьших квадратов. В этом случае наиболее эффективное решение получается тогда, когда от исходной системы $Ax=b$, где A — матрица размера $[n\times m]$, а $v=b-Ax$, переходят к расширенной системе (3.123), которую можно записать в виде, при котором матрица системы имеет симметричную структуру

$$\begin{pmatrix} E & A \\ A^{\mathrm{T}} & O \end{pmatrix}\begin{pmatrix} v \\ x \end{pmatrix}=\begin{pmatrix} b \\ \overline{O} \end{pmatrix}. \tag{4.18}$$

После перехода к системе вида (4.18) она решается по алгоритму итерационного уточнения совместно, например, с LU-алгоритмом и учетом разреженной структуры матрицы расширенной системы.

§ 4.2. МЕТОД ПРОСТОЙ ИТЕРАЦИИ

Основное применение итерационные методы нашли при решении систем уравнений большой размерности. Это вызвано тем, что решение, например, системы $Ax=b$, где A — невырожденная матрица размера $[n\times n]$, нельзя выполнить за меньшее чем $N\cong n^3/3$ число операций умножения и деления. В то же время итерационные методы позволяют в ряде случаев проводить решение тех же систем за существенно меньшее число операций, которое непосредственно зависит от скорости сходимости итерационного процесса. Поэтому скорость сходимости, как, впрочем, и точность приближенного решения, является основной характеристикой любого итерационного процесса.

Пусть задана система уравнений вида $F(x)=0$, где F — непрерывное отображение конечномерного нормированного пространства $L_n \to L_n$, от исходного уравнения всегда можно перейти к эквивалентной системе вида

$$x=\Psi(x). \tag{4.19}$$

Будем находить приближенное решение этого уравнения по следующей формуле:

$$x^{(k+1)}=\Psi(x^{(k)},\ x^{(k-1)},\ldots,x^{(k-i)}) \text{ для } i=0\div k, \tag{4.20}$$

считая, что *начальное приближение*, которое будем обозначать $x^{(0)}$, имеется или каким-либо образом может быть задано. Такой процесс будем называть *итерационным* i-го порядка. Если приближение $x^{(k+1)}$ зависит явно только от k, т. е. $i=0$, то (4.20) будем называть итерационным процессом первого порядка. Итерационный процесс будем считать линейным, если отображение Ψ не зависит от $x^{(k)}$, и стационарным, если отображение Ψ не зависит от номера итерации. Применительно к ре-

шению СЛАУ наиболее распространены линейные стационарные итерационные процессы первого порядка, имеющие вид

$$x^{(k+1)}=Cx^{(k)}+d. \tag{4.21}$$

В дальнейшем вычислительный процесс, основанный на решении систем по формуле (4.21), будем называть *методом простой итерации.*

Таким образом, если требуется найти приближенные решения системы $Ax=b$ методом простой итерации, то сначала исходную систему приводим к эквивалентной системе вида $x=Cx+d$. Затем по заданному значению $x^{(0)}$ вычисляем первое приближенное значение x по формуле $x^{(1)}=Cx^{(0)}+d$, потом — второе приближенное значение x по формуле $x^{(2)}=Cx^{(1)}+d$ и т. д.

Определение 4.1. Пусть имеется последовательность приближенных решений $\{x^{(k)}\}_{k\geqslant 1}$ системы уравнений. Тогда итерационный процесс будем называть сходящимся, если для любого начального приближения $x^{(0)}$ последовательность $\{x^{(k)}\}$ сходится к решению x, т. е.

$$\lim_{k\to\infty} x^{(k)}=x. \tag{4.22}$$

При вводе вектора ошибки $\Delta^{(k)}=x^{(k)}-x$ и его оценки через норму вектора

$$\delta^{(k)}=\|\Delta^{(k)}\|=\|x^{(k)}-x\| \tag{4.23}$$

условие сходимости (4.22) записывается в виде

$$\lim_{k\to\infty}\|x^{(k)}-x\|=0. \tag{4.24}$$

Теорема 4.2. Для сходимости метода простой итерации необходимо и достаточно, чтобы наибольшее по модулю собственное значение матрицы C было меньше единицы.

Доказательство. Будем последовательно вычитать из уравнения $x=Cx+d$ приближенные решения $x^{(k+1)}$, определяемые по формуле (4.21) для $k=0, 1, 2, \ldots$. В результате получим систему вида

$$x-x^{(k+1)}=C(x-x^{(k)}), \quad k=0, 1, 2, \ldots. \tag{4.25}$$

Подставляя первое уравнение этой системы во второе, затем преобразованное второе уравнение в третье и т. д., имеем

$$x-x^{(k+1)}=C^{k+1}(x-x^{(0)}), \quad k=0, 1, 2, \ldots. \tag{4.26}$$

Переходя к пределу этой последовательности и исходя из формулы (4.24), получаем

$$\lim_{k\to\infty}\|C^{k+1}(x-x^{(0)})\|=0.$$

Это выражение для согласованной нормы матрицы при условии, что $x \neq x^{(0)}$, преобразуется к виду

$$\lim_{k\to\infty} \| C^{k+1} \| = 0. \tag{4.27}$$

Отсюда, исходя из следствия 2 теоремы 1.29:

$$\lim_{k\to\infty} x^{(k)} = x \Leftrightarrow | \lambda_1(C) | < 1.$$

Кроме того, из формулы (4.27) и следствия 1 теоремы 1.29 получаем и достаточный признак сходимости по методу простой итерации:

$$\| C \| < 1. \tag{4.28}$$

Как отмечалось выше, важной характеристикой итерационного процесса является скорость его сходимости. Для оценки скорости сходимости итерационного процесса будем использовать показатель η_k, характеризующий, во сколько раз уменьшается норма ошибки после k итераций. Данная оценка получила название *средней скорости сходимости.* Для процесса (4.21) на основании (4.26) имеем

$$\| x - x^{(k+1)} \| \leqslant \| C^{k+1} \| \| x - x^{(0)} \|.$$

Исходя из этой зависимости, за показатель η_k может быть выбрана следующая характеристика:

$$\eta_k(C) = -k^{-1} \ln \| C^{k+1} \|. \tag{4.29}$$

Так как из условий $|\lambda(C)| < 1$ следует, что

$$\lim_{k\to\infty} (\| C^k \|)^{1/k} = \lambda_1(C),$$

то, переходя в неравенстве (4.29) к пределу при $k \to \infty$, получим

$$\eta_\infty(C) = \lim_{k\to\infty} \eta_k(C) = -\ln \lambda_1(C), \tag{4.30}$$

где $\lambda_1(C) = \max |\lambda_i(C)|$ — спектральный радиус матрицы C. Коэффициент $\eta_\infty(C)$ получил название *асимптотической скорости сходимости.* Отметим, что $\eta_\infty(C)$ не зависит от выбора нормы матрицы, тогда как $\eta_k(C)$ от выбора нормы матрицы зависит. Оценку быстродействия итерационных методов можно осуществлять также и по числу арифметических операций $N_{оп}$. Особенно эта оценка удобна, если производится сравнение эффективностей итерационного метода и прямого. Как и для прямого метода, значение $N_{оп}$ будем в основном определять через число операций умножения. Так, в случае оценки быстродействия метода простой итерации после k итераций получим

$$N_{оп} \simeq kn^2 + N_c, \tag{4.31}$$

где N_c — число арифметических операций, приходящихся на переход к эквивалентной системе $x=Cx+d$, у которой $\|C\|<1$. В частности, из этой формулы видно, что предельный выигрыш по быстродействию при использовании этого метода вместо прямого метода решения СЛАУ составляет $n/3$ раза при условии, что $N_c<n^2$ и $k=1$. Как правило, значение $k\ll n/3$ и основная задача связана с уменьшением N_c.

Весьма важной характеристикой итерационного процесса является оценка ошибки приближенного решения, за которую принимается или формула (4.23), или формула

$$\varepsilon^{(k)}=\frac{\|x^{(k)}-x\|}{\|x\|}=\frac{\delta^{(k)}}{\|x\|}. \tag{4.32}$$

В итерационных алгоритмах по величине $\delta^{(k)}$ и $\varepsilon^{(k)}$ определяется условие выхода из циклического процесса. Для получения формулы, удобной для вычисления значения $\delta^{(k)}$, преобразуем уравнение (4.25), предварительно прибавив к левой и правой частям его вектор $x^{(k)}$ и решая это уравнение относительно x:

$$x=(E-C)^{-1}[(x^{(k+1)}-x^{(k)})+(E-C)x^{(k)}].$$

После подстановки полученного по этой формуле значения x в правую часть уравнения (4.25) будем иметь

$$x-x^{(k+1)}=C(E-C)^{-1}(x^{(k+1)}-x^{(k)}).$$

Перейдя к согласованным нормам и учитывая неравенство (2.81), при условии, что $\|C\|<1$, получаем

$$\|x-x^{(k+1)}\|\leqslant\frac{\|C\|\,\|E\|}{1-\|C\|}\|x^{(k+1)}-x^{(k)}\|. \tag{4.33}$$

При реализации итерационных алгоритмов решения СЛАУ на ЭВМ оценка вычислительной точности приближенного решения, выполняемая по формуле (4.33), очень удобна, так как требует незначительной по объему дополнительной памяти только под хранение предыдущего приближенного решения и всего около nk операций умножения.

В некоторых случаях формулу (4.33) целесообразно представить в модифицированном виде

$$\|x-x^{(k+1)}\|\leqslant\frac{\|C^{k+1}\|\,\|E\|}{1-\|C\|}\|x^{(1)}-x^{(0)}\|. \tag{4.34}$$

Из данной формулы легко определить число необходимых итераций, обеспечивающих заданную точность решения системы. Так, если $k_{\bar{\delta}}$ — число итераций, соответствующих точности $\bar{\delta}$, то на основании формулы (4.34) получим

$$k_{\bar{\delta}}=\frac{\ln\bar{\delta}-\ln\delta^{(1)}}{\ln\|C\|}. \tag{4.35}$$

Анализ этой формулы показывает, что с уменьшением нормы матрицы $\|C\|$ или спектрального радиуса матрицы C сходимость итерационного процесса (4.21) убыстряется, так как при этом происходит уменьшение значения k_δ.

Таким образом, параметры, характеризующие качество метода простой итерации, в основном зависят от вида эквивалентной системы

$$x = Cx + d. \tag{4.36}$$

Обычно переход от исходной системы общего вида $Ax = b$ к эквивалентной системе (4.36), обеспечивающей хорошую сходимость метода простой итерации, является наиболее сложной задачей при решении СЛАУ этим методом. Пусть матрица исходной системы невырожденная, имеет размер $[n \times n]$. Тогда всегда найдется такая невырожденная матрица G, для которой

$$C = E - G^{-1}\mathrm{A}, \quad d = G^{-1}b. \tag{4.37}$$

Матрица G, входящая в эти формулы, получила название *матрицы расщепления*. В зависимости от способа выбора матрицы G имеется ряд модификационных методов, в конечном счете сводящихся к методу простой итерации. Большое применение в вычислительной практике нашли методы Ричардсона, Якоби, Гаусса — Зейделя, верхней релаксации.

Блок-схема алгоритма решения системы методом простой итерации приведена на рис. 10.

§ 4.3. ИТЕРАЦИОННЫЕ МЕТОДЫ РИЧАРДСОНА И ЯКОБИ

Для рассмотрения модифицированных методов, сводящихся к методу простой итерации, введем ряд понятий.

Определение 4.2. Метод простой итерации будем называть *симметризуемым*, если для некоторой невырожденной матрицы W матрица $W(E-C)W^{-1}$ является симметричной и положительно-определенной. При этом матрица W называется *матрицей симметризации.*

Из этого определения, в частности, следует, что если матрицы исходной системы и расщепления симметрические и положительно-определенные, то метод простой итерации будет симметризуемым. В этом случае, например, матрицей симметризации может быть любая матрица W, для которой матрица расщепления $G = W^{\mathrm{T}}W$.

Определение 4.3. Метод простой итерации будем называть *экстраполированным методом*, если имеет место равенство

$$x^{(k+1)} = C_\gamma x^{(k)} + \gamma d, \tag{4.38}$$

где матрица C_γ определяется по формуле

$$C_\gamma = \gamma C + (1 - \gamma)E. \tag{4.39}$$

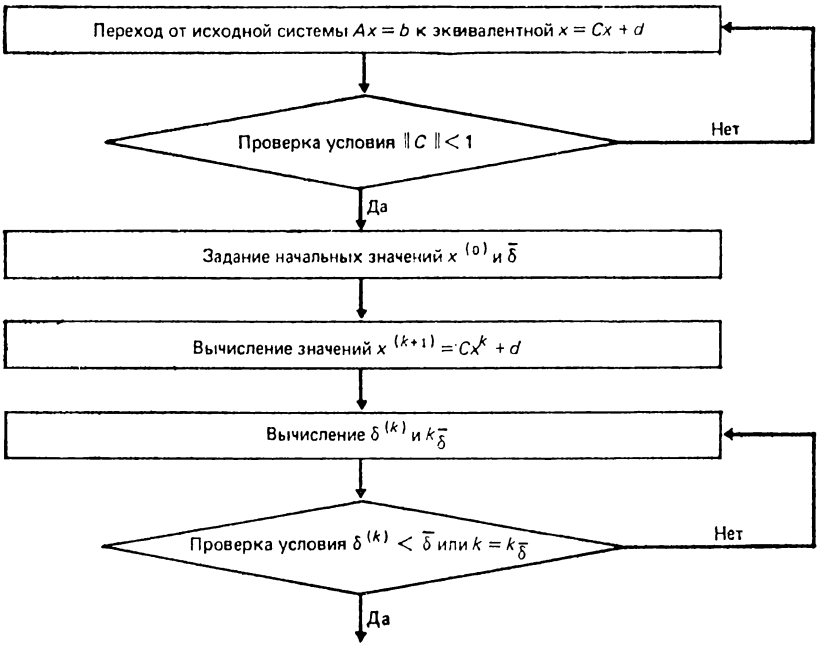


Рис. 10. Блок-схема алгоритма метода простой итерации

Теорема 4.3. Если для решения нормальной системы уравнения используется симметризуемый метод простой итерации, то всегда найдется такой соответствующий ему экстраполированный метод, который является сходящимся.

Для доказательства теоремы в качестве параметра γ в экстраполированном методе возьмем значение $\bar{\gamma}$:

$$\bar{\gamma} = 2(2 - \lambda_1(C) - \lambda_n(C))^{-1}. \tag{4.40}$$

Тогда из (4.39) получим

$$C_{\bar{\gamma}} = 2(2 - \lambda_1(C) - \lambda_n(C))^{-1}(C - \bar{\lambda}E). \tag{4.41}$$

Здесь значение $\bar{\lambda} = (\lambda_1 + \lambda_n)/2$, где λ_1 и λ_n — соответственно наибольшее и наименьшее (ненулевое) значения матрицы.

Для нахождения спектрального радиуса $\lambda_1(C_{\bar{\gamma}})$ матрицы $C_{\bar{\gamma}}$ воспользуемся отношением Релея (1.160). Тогда

$$\lambda_1(C_{\bar{\gamma}}) = \max \frac{(C_{\bar{\gamma}}x, x)}{(x, x)}.$$

Отсюда с учетом (4.41)

$$\lambda_1(C_{\bar{\gamma}}) = \frac{2}{2 - \lambda_1(C) - \lambda_n(C)} \max \left[\frac{(Cx, x)}{(x, x)} - \bar{\lambda} \right].$$

Так как из свойства отношения Релея $\lambda_1(C)=\max\dfrac{(Cx,\,x)}{(x,\,x)}$, то предыдущее выражение преобразуется к виду

$$\lambda_1(C_{\bar\gamma})=\frac{2}{2-\lambda_1(C)-\lambda_n(C)}\,(\lambda_1(C)-\bar\lambda).$$

И далее после элементарных преобразований получаем

$$\lambda_1(C_{\bar\gamma})=\frac{\lambda_1(C)-\lambda_n(C)}{2-\lambda_1(C)-\lambda_n(C)}<1. \tag{4.42}$$

Теорема доказана.

Из теоремы 4.3 и формулы (4.30) следует, что экстраполированный метод со значением $\bar\gamma$, определяемым по формуле (4.40), обладает асимптотической скоростью сходимости. Такой метод получил название *оптимального экстраполированного метода* решения систем линейных уравнений.

Пусть задана нормальная система уравнений $Ax=b$. Применим для ее решения метод Ричардсона, который описывается следующей формулой:

$$x^{(k+1)}=(E-A)x^{(k)}+b. \tag{4.43}$$

Данная формула в результате элементарных преобразований приводится к виду (4.21). В этом случае $C=E-A$, а матрицей расщепления будет $G=E$. Так как матрица A положительно-определенная и симметрическая, то любое собственное значение матрицы C будет определяться через соответствующее собственное значение матрицы A как $\lambda(C)=1-\lambda(A)$. Отсюда на основании теоремы 4.2 получаем условие сходимости метода (4.43) для нормальной системы уравнений:

$$\lambda_1(A)<2. \tag{4.44}$$

Метод Ричардсона можно представить и в более обобщенной форме

$$x^{(k+1)}=(E-\bar\gamma A)x^{(k)}+\bar\gamma b, \tag{4.45}$$

где

$$\bar\gamma=2(2-\lambda_1(C)-\lambda_n(C))^{-1}=2(\lambda_1(A)+\lambda_n(C))^{-1}.$$

Здесь матрицей C в формуле (4.21) является матрица

$$C_{\bar\gamma}=E-\bar\gamma A,$$

спектральный радиус которой

$$\lambda_1(C_{\bar\gamma})=\frac{\lambda_1(A)-\lambda_n(A)}{\lambda_1(A)+\lambda_n(A)}=\frac{K^+(A)-1}{K^+(A)+1}, \tag{4.46}$$

где $K^+(A)$ — спектральное число обусловленности матрицы A. Напомним, что $K^+(B)=\|B\|_2\|B^+\|_2\Rightarrow K^+(B)=\rho_1(B)/\rho_n(B)$, где

ρ_1 — наибольшее, а ρ_n — наименьшее ненулевое сингулярное число матрицы B. Причем если B — невырожденная матрица, то $K^+(B)=K_2^-(B)$. Используя теорему 4.3, получаем, что обобщенный метод Ричардсона (4.45) для нормальной системы уравнений всегда сходится с асимптотической скоростью

$$\eta_\infty(C_\gamma^-)=-\ln\frac{K^+(A)-1}{K^+(A)+1}\simeq\frac{2}{K^+(A)}. \tag{4.47}$$

Из данной формулы следует, что, чем хуже обусловлена матрица A, тем медленнее сходимость итерационного метода Ричардсона.

Предположим теперь, что имеет место следующее блочное разбиение системы $Ax=b$, у которой матрица A размера $[n\times n]$:

$$\begin{pmatrix} A_{11} & A_{12} & \dots & A_{1q} \\ A_{21} & A_{22} & \dots & A_{2q} \\ \dots & & \dots & \\ A_{q1} & A_{q2} & \dots & A_{qq} \end{pmatrix}\begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \dots \\ \bar{x}_q \end{pmatrix}=\begin{pmatrix} \bar{b}_1 \\ \bar{b}_2 \\ \dots \\ \bar{b}_q \end{pmatrix}. \tag{4.48}$$

При этом разбиение системы на блоки осуществляется по тем же правилам, которые изложены в § 3.3:

клетки (блоки) A_{ij} и A_{kl} не имеют общих элементов матрицы A, если $i\neq k$ или $j\neq l$;

элементы, входящие в каждую клетку, соответствуют элементам матрицы A при сохранении последовательности их взаимного расположения, т. е. если $a_{i_s j_t}$ и $a_{i_p j_r}$ — элементы клетки A_{ij}, соответствующие элементам a_{gk} и a_{ml} матрицы A, то $s-p=g-m$ и $t-r=k-l$;

клетки A_{ii} — квадратные матрицы.

Представим матрицу A в виде

$$A=A_1+A_2+A_3, \tag{4.49}$$

где

$$A_1=\begin{pmatrix} A_{11} & & & O \\ & A_{22} & & \\ & & & \\ O & & & A_{qq} \end{pmatrix},\quad A_2=\begin{pmatrix} 0 & A_{12} & \dots & A_{1q} \\ & 0 & \dots & A_{2q} \\ & & \ddots & \\ O & & & 0 \end{pmatrix},$$

$$A_3=\begin{pmatrix} 0 & & & O \\ A_{21} & 0 & & \\ \vdots & & \ddots & \\ A_{q1} & A_{q2} & \dots & 0 \end{pmatrix}.$$

В принятых обозначениях метод Якоби (4.48) записывается как

$$A_1x^{(k+1)}=-(A_2+A_3)x^{(k)}+b. \tag{4.50}$$

При условии, что A_1 — невырожденная матрица, от итерационного процесса (4.50) нетрудно перейти к методу простой итерации (4.21), в котором матрица C и вектор d будут определяться по формулам

$$C = -A_1^{-1}(A_2+A_3) = E - A_1^{-1}A,$$
$$d = (A_{11}^{-1}\overline{b}_1 \ldots A_{qq}^{-1}\overline{b}_q). \qquad (4.51)$$

Если в исходной системе матрица A симметрическая и положительно-определенная, то матрица A_1 также будет симметрической и положительно-определенной, а следовательно, представление (4.51) возможно.

Частным случаем представления (4.49) является такой, при котором клетки A_{ii} являются матрицами размера $[1\times 1]$. Соответствующий этому случаю метод Якоби (4.50) получил название *точечного* в отличие от всех остальных случаев, для которых метод Якоби будем называть *блочным*.

Пусть задана система $Ax=b$, для элементов матрицы которой выполняется свойство

$$\text{либо } |a_{ii}| > \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|,$$
$$\text{либо } |a_{jj}| > \sum_{i=1}^{j-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|. \qquad (4.52)$$

Матрицы, обладающие отмеченными выше свойствами, называются *матрицами с диагональным преобладанием*. Если для решения такой системы, обладающей первым свойством (4.52), применить точечный метод Якоби, то такой итерационный процесс будет сходящимся. Действительно, матрица C при этом имеет вид

$$C = -\begin{pmatrix} 0 & (a_{12}a^{-1}{}_{11}) & \ldots & (a_{1n}a^{-1}{}_{11}) \\ (a_{21}a^{-1}{}_{22}) & 0 & \ldots & (a_{2n}a^{-1}{}_{22}) \\ \ldots & \ldots & \ldots & \ldots \\ (a_{n_1}a^{-1}{}_{nn}) & (a_{n_2}a^{-1}{}_{nn}) & \ldots & 0 \end{pmatrix}.$$

Вычисляя для этой матрицы l-норму, получим

$$\|C\|_l = \Big(\sum_{j=1}^{s-1} |a_{sj}| + \sum_{j=s+1}^{n} |a_{sj}|\Big) |a_{ss}|^{-1} < 1.$$

Это и есть один из достаточных признаков сходимости метода простой итерации. Если для исходной системы выполняется второе условие (4.52), то, переходя к ее решению по точечному методу Якоби, будем иметь

$$A_1 x^{(k+1)} = -(A_2+A_3)x^{(k)} + b.$$

Примем следующее обозначение:

$$A_1 x^{(m)} = y^{(m)} \text{ для } m = 1, 2, \ldots, k, k+1, \ldots \tag{4.53}$$

Тогда после несложных преобразований точечный метод Якоби преобразуется к методу простой итерации относительно векторов $y^{(m)}$:

$$y^{(k+1)} = Cy^{(k)} + b, \tag{4.54}$$

где

$$C = E - AA_1^{-1}.$$

Этот метод итерации будет сходящимся, так как для c-нормы матрицы C получим

$$\| C \|_c = \left(\sum_{i=1}^{t-1} | a_{it} | + \sum_{i=t+1}^{n} | a_{it} | \right) | a_{tt} |^{-1} < 1.$$

После вычисления приближенного значения $y^{(k+1)}$ приближенное значение $x^{(k+1)}$ находим из выражения (4.53) по формуле

$$x^{(k+1)} = A_1^{-1} y^{(k+1)}.$$

Еще раз отметим, что сходимость описанного выше точечного метода Якоби при решении СЛАУ обеспечивается для любых линейных систем, удовлетворяющих только условиям (4.52), причем исходная система может и не быть нормальной системой уравнений.

Определение 4.4. Блочная матрица A системы (4.48) обладает свойством разбиения $\mathrm{BL}(A)$, при условии, что имеются такие два непересекающихся непустых подмножества S_R и S_Q множества $S = \{1, 2, \ldots, q\}$, что, во-первых, $S_R \cup S_Q = S$, а во-вторых, если клетки A_{ij} не являются нулевыми матрицами при $i \neq j$, то либо $i \in S_R$ и $j \in S_Q$, либо $i \in S_Q$ и $j \in S_R$.

В качестве примера рассмотрим блочную матрицу A, не имеющую ни одного нулевого блока. В этом случае можно выбрать такие подмножества S_R и S_Q, что $S_R = \{1\}$ и $S_Q = \{2, 3, \ldots, q\}$. Нетрудно убедиться, что такое разбиение матрицы A на блоки обладает свойством $\mathrm{BL}(A)$. Другим примером разбиения матрицы A на блоки, обладающего свойством $\mathrm{BL}(A)$, является блочно-трехдиагональная матрица вида

$$A = \begin{pmatrix} A_{11} & A_{12} & & & & O \\ A_{21} & A_{22} & A_{23} & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & & \cdot & A_{q-1,q} \\ O & & & & A_{q,q-1} & A_{qq} \end{pmatrix} \tag{4.55}$$

Так, предполагая для удобства, что q — четное, в качестве S_R и S_Q для матрицы A, представленной в виде (4.55), возьмем $S_R=\{1,3,\ldots,q-1\}$ и $S_Q=\{2,4,\ldots,q\}$, при которых будет удовлетворяться определение 4.4.

Теперь можно получить признак сходимости блочного метода Якоби (4.50) исходя из следующего свойства. Если симметричная и положительно-определенная матрица A обладает свойством $\mathrm{BL}(A)$, то спектральный радиус матрицы $C=E-A_1^{-1}A$ удовлетворяет неравенству $\lambda_1(C)<1$, кроме того, $\lambda_n(C)=-\lambda_1(C)$. Если теперь перейти к оптимальному экстраполированному методу Якоби, то он на основании (4.38) примет вид

$$x^{(k+1)}=C_{\bar{\gamma}}x^{(k)}+\bar{\gamma}d,$$

где $C_{\bar{\gamma}}$ с учетом (4.39) и (4.51) будет иметь вид

$$C_{\bar{\gamma}}=E-\bar{\gamma}A_1^{-1}A,$$

причем для случая, когда $\lambda_n(C_{\bar{\gamma}})=-\lambda_1(C_{\bar{\gamma}})$, имеем $\bar{\gamma}=1$. Отсюда следует, что $C_{\bar{\gamma}}=C_{\bar{\gamma}}$, т. е. метод Якоби совпадает с оптимальным экстраполяционным методом Якоби. Следовательно, при сделанных предложениях метод Якоби всегда сходится с асимптотической скоростью.

Таким образом, если задана нормальная система уравнений, у которой матрица A невырожденная и имеет блочно-трехдиагональную структуру вида (4.55), то для ее решения можно применить блочный метод Якоби, который будет сходиться с асимптотической скоростью

$$\eta_\infty=-\ln\lambda_1(C).$$

Как было показано выше, для метода Якоби могут быть применены две различные по виду эквивалентные схемы, первая из которых описывается формулой (4.50), а вторая — формулой (4.21), где матрица C и вектор d вычисляются по формулам (4.51). Встает вопрос, какая из этих схем наиболее рациональна с точки зрения объема вычислений. Так как скорость сходимости $\eta_\infty(C)$ для обеих схем одинакова, то сравнительную оценку объема вычислений целесообразно осуществлять с использованием формулы (4.31). Дальнейший сравнительный анализ будем осуществлять в предположении, что матрица исходной системы симметрическая, положительно-определенная и имеет блочное представление (4.48), при котором все клетки одного размера $[n/q\times n/q]$.

Реализация метода Якоби по первой схеме (4.50) выполняется в два этапа. Сначала вычисляется значение

$$y^{(k+1)}=-(A_2+A_3)x^{(k)}+b \text{ для } k=0,1,2,\ldots,$$

а затем в результате решения уравнения

$$A_1 x^{(k+1)} = y^{(k+1)} \tag{4.56}$$

находится приближенное решение $x^{(k+1)}$ исходной системы уравнений. Так как матрица A_1 системы (4.56) симметрическая и положительно-определенная, то для ее решения можно применить метод квадратного корня (или снова метод итерации). При этом разложение каждой клетки A_{ii} на произведение двух транспонированных треугольных матриц $A_{ii} = T'_{ii} T_{ii}$ надо выполнить всего один раз для всех шагов итераций и только обратный ход придется выполнять на каждом итерационном шаге. Следовательно, число операций умножения и деления, необходимых для обеспечения решения СЛАУ методом Якоби при использовании первой схемы, будет при $n/q \geqslant 4$ порядка

$$N^{(1)}{}_{\text{оп}} \simeq kn^2 + \left[\left(\frac{n}{q}\right)^3 \frac{q}{6} + \left(\frac{n}{q}\right)^2 qk\right].$$

После элементарных преобразований эта формула преобразуется к виду

$$N^{(1)}{}_{\text{оп}} \simeq \left(\frac{n}{q}\right)^2 \left(kq^2 + kq + \frac{n}{6}\right). \tag{4.57}$$

Если метод Якоби реализовать по второй схеме, то он будет включать, во-первых, процедуру вычисления элементов матрицы $C = -A_1^{-1}(A_2 + A_3)$, которая выполняется всего один раз для всех шагов итерации, и, во-вторых, процедуру вычисления приближенного решения $x^{(k+1)}$ по методу простой итерации. Число операций умножения и деления, требуемое для реализации этой схемы при $n/q \geqslant 4$, будет соответствовать значению

$$N^{(2)}{}_{\text{оп}} \simeq n^2 k + \left[\frac{1}{2}\left(\frac{n}{q}\right)^3 q + \left(\frac{n}{q}\right)^3 q(q-1)\right].$$

Эта формула преобразуется к виду

$$N_{\text{оп}}^{(2)} \simeq \left(\frac{n}{q}\right)^2 \left(kq^2 + nq - \frac{n}{2}\right). \tag{4.58}$$

Из отношения

$$\gamma_{12} = \frac{N^{(1)}{}_{\text{оп}}}{N^{(2)}{}_{\text{оп}}} \simeq \frac{kq^2 + kq + n/6}{kq^2 + nq - n/2}$$

можно определить условия, при которых применение одной из схем в методе Якоби является более эффективным. Так, при $\gamma_{12} > 1$ эффективнее вторая схема, а при $\gamma_{12} < 1$ — первая. Отсюда получаем следующее условие для оценки эффективности второй схемы по сравнению с первой:

$$\begin{cases} 3kq - 3nq + 2n > 0, \\ 2 \leqslant q \leqslant n/4. \end{cases} \tag{4.59}$$

Из анализа этих зависимостей получаем, что при $k > n - n/q$ и $2 \leqslant q \leqslant n/4$, а также при $q \geqslant n/4$ и любом значении k вторая схема решения будет эффективнее первой. В остальных случаях первая схема эффективнее второй. В частности, если $q=2$, т. е. матрица A разбита на четыре клетки, то при $k < n$ первая схема эффективнее второй, а при точечном разбиении, т. е. $q=n$, наоборот.

§ 4.4. ИТЕРАЦИОННЫЕ МЕТОДЫ ВЕРХНЕЙ РЕЛАКСАЦИИ. МЕТОД ГАУССА — ЗЕЙДЕЛЯ

Одним из наиболее эффективных и широко используемых итерационных методов для решения нормальной системы уравнений полного ранга является метод последовательной верхней релаксации. Изложим суть данного метода. Пусть исходная система уравнений $Ax=b$ разбита на блоки в виде (4.48) и, кроме того, матрица A представлена, как и в разложении (4.49). Будем для данной системы рассматривать итерационный процесс, описываемый следующим образом:

$$A_1 x^{(k+1)} = -\omega(A_2 x^{(k+1)} + A_3 x^{(k)} - b) + \\ + (1-\omega) A_1 x^{(k)}. \tag{4.60}$$

При $\omega > 1$ приведенный процесс называется *методом верхней релаксации*, при $\omega < 1$ — *методом нижней релаксации*, а при $\omega = 1$ — *методом Гаусса — Зейделя*. Параметр ω, входящий в формулу (4.60), называется *коэффициентом релаксации*. В дальнейшем будет рассматриваться метод верхней релаксации. От выражения (4.60) перейдем к выражению вида

$$x^{(k+1)} = \omega C_1 x^{(k+1)} + (\omega C_2 + (1-\omega)E) x^{(k)} + \omega d, \tag{4.61}$$

предположив, что A_1 — невырожденная матрица, будем иметь

$$C_1 = -A_1^{-1} A_2, \quad C_2 = -A_1^{-1} A_3, \quad d = A_1^{-1} b.$$

После несложных преобразований выражение (4.61) для метода релаксации примет вид, эквивалентный методу простой итерации:

$$x^{(k+1)} = C_\omega x^{(k)} + d_\omega,$$

где

$$C_\omega = (E - \omega C_1)^{-1}(\omega C_2 + (1-\omega)E), \\ d_\omega = \omega(E - \omega C_1)^{-1} d. \tag{4.62}$$

В тех случаях, когда для блочного представления системы (4.48) имеем $q=n$, метод релаксации называется *точечным*, а при $2 \leqslant q < n$ — *блочным*.

На основании теоремы 4.1 необходимым и достаточным условием сходимости метода верхней релаксации является $\lambda_1(C_\omega)<1$, где λ_1 — спектральный радиус матрицы C_ω. Естественно, что проверка этого условия сходимости — весьма сложная вычислительная задача.

Имеются более приемлемые для практического использования признаки сходимости метода верхней релаксации. Один из них основан на следующем свойстве, которое приведем без доказательства. Для спектрального радиуса матрицы C_ω выполняется свойство

$$\lambda_1(C_\omega) \geqslant |\omega - 1|, \tag{4.63}$$

где равенство имеет место тогда и только тогда, когда все собственные значения матрицы C_ω равны $|\omega-1|$. Отсюда вытекает необходимое условие сходимости метода релаксации

$$0 < \omega < 2, \tag{4.64}$$

которое для метода верхней релаксации принимает вид $1<\omega<2$. В том случае, когда матрица исходной системы симметрическая и положительно-определенная, условие (4.64) является необходимым и достаточным признаком сходимости метода верхней релаксации.

В общем случае (даже для нормальных систем уравнений) метод верхней релаксации является несимметризуемым, а следовательно, для него нельзя применить экстраполяцию как способ ускорения сходимости основного метода. В связи с этим возникает вопрос, а при каком же значении $\omega \in (0; 2)$ обеспечивается максимальное значение асимптотической скорости сходимости, что естественно равносильно достижению минимума $\lambda_1(C_\omega)$. В общем случае на данный вопрос не удается найти ответ. Однако если матрица A, представленная в блочном виде, является согласованно упорядоченной, то такое значение ω_0 можно найти. Это значение ω_0 называется оптимальным значением коэффициента релаксации.

Определение 4.5. Матрица A системы $Ax=b$, представленной в блочном виде (4.48), является *согласованно упорядоченной*, если для некоторого значения $t \in S$, где множество $S=\{1, 2, \ldots, q\}$:

1) существуют непустые подмножества $S_i \subset S$ при $i=1 \div t$, такие, что

$$S_i \cap S_j = \varnothing, \text{ если } i \neq j,$$

$$\mathrm{U}^t_{i-1} S_k = S;$$

2) для клеток A_{ij} матрицы A выполняется условие: если $A_{ij}\neq 0$ при $i\neq j$ и $i\in S_k$, где $S_k\subset S$, то при $j>i$ имеем $j\in S_{k+1}$, где $S_{k+1}\subset S$, а при $j<i$: $j\in S_{k-1}$, где $S_{k-1}\subset S$.

Отметим, что свойство согласованной упорядоченности и свойство BL(A) определяются относительно заданного разбиения матрицы A на блоки. При этом если матрица A обладает свойством BL(A) относительно заданного разбиения на блоки, то с помощью соответствующей перестановки строк и столбцов матрицы A можно получить согласованно упорядоченную матрицу. Если же положить $S_R=\{S_i\colon\ i$ — нечетно$\}$ и $S_Q=\{S_i\colon\ i$ — четно$\}$, где S_R и S_Q — подмножества S, взятые из определения 4.4, то любая согласованно упорядоченная матрица обладает свойством BL(A).

Отсюда следует, что любая матрица с разбиением на блоки при $q=2$ является согласованно упорядоченной и обладает свойством BL(A).

Примером согласованной и упорядоченной матрицы при $q>2$ является блочно-трехдиагональная матрица вида (4.55). Приведем пример еще одной блочной матрицы

$$A=\begin{pmatrix} A_{11} & A_{12} & O & A_{14} \\ A_{21} & A_{22} & A_{23} & O \\ O & A_{23} & A_{33} & A_{34} \\ A_{41} & O & A_{43} & A_{44} \end{pmatrix}.$$

Эта матрица, как нетрудно проверить, обладает свойством BL(A), но не обладает свойством согласованной упорядоченности. Однако согласованное упорядочение может быть достигнуто за счет перестановки местами, например, 1-й и 2-й блочной строки и соответственно 1-го и 2-го блочного столбца, т. е. матрица $\tilde{A}$ будет согласованно упорядоченной, если

$$\tilde{A}=Q_{12}AQ_{12}=\begin{pmatrix} A_{22} & A_{21} & A_{23} & O \\ A_{12} & A_{11} & O & A_{14} \\ A_{32} & O & A_{33} & A_{34} \\ O & A_{41} & A_{43} & A_{44} \end{pmatrix}.$$

В работе [24] показано, что, если матрица исходной системы согласованно упорядоченная и положительно-определенная, то для значения ω_0 имеем

$$\omega_0=\frac{2}{1+\sqrt{1-\lambda^2_1(B)}}, \qquad (4.65)$$

а спектральный радиус матрицы перехода C_ω в методе релаксации будет определяться как

$$\lambda_1(C_\omega)=\begin{cases}\left(\dfrac{\omega\lambda_1(B)+\sqrt{\omega^2\lambda_1(B)-4(\omega-1)}}{2}\right)^2 & \text{если } 0<\omega\leqslant\omega_0,\\ \omega-1, \quad \text{если } \omega_0\leqslant\omega<2.\end{cases} \tag{4.66}$$

где $\lambda_1(B)$ — спектральный радиус матрицы перехода в методе Якоби, определяемой по формуле (4.51) как $B=E-A_1^{-1}A$. Напомним, что $\lambda_1(B)<1$. Из формулы (4.65) следует, что $1<\omega_0<2$. С помощью формул (4.66) и (4.65) легко находится спектральный радиус для матрицы C_{ω_0}:

$$\lambda_1(C_{\omega_0})=\frac{1-\sqrt{1-\lambda^2_1(B)}}{1+\sqrt{1-\lambda^2_1(B)}}. \tag{4.67}$$

Из сравнения значений $\lambda_1(C_\omega)$ и $\lambda_1(C_{\omega_0})$ следует, что

$$\lambda_1(C_\omega)>\lambda_1(C_{\omega_0}) \text{ для } \omega\neq\omega_0, \tag{4.68}$$

а потому существует только одно значение $\omega=\omega_0$, при котором минимизируется значение $\lambda_1(C_\omega)$, а следовательно, обеспечивается максимальная сходимость метода релаксации.

Если теперь сравнить значение $\lambda_1(C_{\omega_0})$ и $\lambda_1(B)$, то при условии, что $\lambda_1(B)<1$ и $\lambda_1(B)\neq0$:

$$\lambda_1(C_{\omega_0})<\lambda_1(B). \tag{4.69}$$

Таким образом, метод релаксации при оптимальном выборе коэффициента ω_0 обеспечивает более быструю сходимость по сравнению с методом Якоби даже в том случае, когда в методе Якоби используется оптимальная экстраполяция.

Метод релаксации при $\omega=1$, как отмечалось выше, получил название метода Гаусса — Зейделя. Исходя из формулы (4.60), он представляется в виде

$$A_1x^{(k+1)}=-A_2x^{(k+1)}-A_3x^{(k)}+b \tag{4.70}$$

или после несложных преобразований

$$x^{(k+1)}=C_1x^{(k+1)}+C_2x^{(k)}+d, \tag{4.71}$$

где $C_1+C_2=C$ — матрица в эквивалентной системе $x=Cx+d$, причем $C_1=-A_1^{-1}A_2$, $C_2=-A_1^{-1}A_3$, $d=A_1^{-1}b$. Выражение (4.71) в координатной записи представляется как

$$x_1^{(k+1)} = c_{11}x_1^{(k)} + c_{12}x_2^{(k)} + \ldots + c_{1i}x_i^{(k)} + \ldots + c_{1n}x_n^{(k)},$$

$$x_2^{(k+1)} = c_{21}x_1^{(k+1)} + c_{22}x_2^{(k)} + \ldots + c_{2i}x_i^{(k)} + \ldots + c_{2n}x_n^{(k)},$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \tag{4.72}$$

$$x_i^{(k+1)} = c_{i1}x_1^{(k+1)} + c_{i2}x_2^{(k+1)} + \ldots + c_{i,i-1}x_{i-1}^{(k+1)} + c_{i,i}x_i^{(k)} + \ldots + c_{in}x_n^{(k)},$$

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

$$x_n^{(k+1)} = c_{n1}x_1^{(k+1)} + c_{n2}x_2^{(k+1)} + \ldots + c_{ni}x_i^{(k+1)} + \ldots + c_{n,n-1}x_{n-1}^{(k+1)} + c_{nn}x^{(k)}.$$

Так как в методе Гаусса — Зейделя коэффициент релаксации $\omega = 1$ и удовлетворяет условию (4.64), то, следовательно, метод Гаусса — Зейделя (4.70) всегда сходится для нормальных систем уравнений полного ранга. При этом спектральный радиус сходимости для матрицы перехода в методе Гаусса — Зейделя будет

$$\lambda_1(M) = (\lambda_1(B) + \sqrt{\lambda_1(B)})/2, \tag{4.73}$$

где матрица перехода в методе Гаусса — Зейделя определяется формулой

$$M = -(E + A_1^{-1}A_2)^{-1}A_1^{-1}A_3. \tag{4.74}$$

Метод релаксации (а следовательно, и метод Гаусса — Зейделя) аналогично методу Якоби может быть реализован по схеме (4.60) или (4.61). При этом по первой схеме (4.60) сначала вычисляется значение

$$y^{(k+1)} = -\omega(A_2x^{(k+1)} + A_3x^{(k)} - b) + (1 - \omega)A_1x^{(k)}, \tag{4.75}$$

а затем — значение приближенного решения

$$x^{(k+1)} = A_1y^{(k+1)}.$$

Для ответа на вопрос, какая же из этих схем эффективнее при решении конкретной СЛАУ методом верхней релаксации (или Гаусса — Зейделя), можно воспользоваться оценками, которые были получены при выборе аналогичных по виду схем, реализующих метод Якоби. Напомним, что при $k < n$ и $q = 2$ эффективнее первая схема по сравнению со второй, а для точечного метода более эффективна вторая схема.

Глава 5

ИНТЕРПОЛИРОВАНИЕ ФУНКЦИЙ

§ 5.1. ПОСТАНОВКА ЗАДАЧИ

На практике не всегда удается описывать изучаемые функциональные зависимости аналитически. Часто приходится иметь дело с функциями $y=f(x)$, заданными лишь таблично:

$$\begin{array}{cccccc} x_0 & x_1 & x_2 & \dots & x_n, \\ y_0 & y_1 & y_2 & \dots & y_n, \end{array} \tag{5.1}$$

где $y_0=f(x_0)$, $y_1=f(x_1), \dots, y_n=f(x_n)$ и $x_0<x_1<\dots<x_n$. При этом известно, что областью определения $f(x)$ служит некоторый промежуток $[a, b]$, включающий в себя отрезок $[x_0, x_n]$. Поэтому возникают вопросы о значениях функции в промежуточных точках $x\in[a, b]$, не фигурирующих среди абсцисс (5.1). Подобные вопросы составляют задачи интерполирования функций (при этом говорят об экстраполировании, если $x\in[a, b]$, но $x\notin[x_0, x_n]$. Совокупность точек $x_0, \dots, x_n$ называют *сеткой*, а сами эти точки — *узлами сетки* или *узлами интерполяции*.

Примеры интерполирования функций в геодезической практике чрезвычайно разнообразны. Достаточно вспомнить интерполяцию отметок в тахеометрической съемке, прогнозы осадок реперов и деформаций инженерных сооружений, приближенное оценивание значений аномалии силы тяжести между гравиметрическими пунктами, интерполяцию компонент уклонений отвеса и высот квазигеоида в физической геодезии, работы с аэроснимками в фотограмметрии и т. д.

При этом наиболее часто возникают задачи интерполирования функций двух переменных. Тем не менее все характерные черты проблемы удобно сначала просмотреть для одномерной задачи, т. е. для случая интерполирования функции одной переменной.

Итак, используя числа $y_i=f(x_i)$, $i=1, \dots, n$ и предполагая, что функция f обладает той или иной гладкостью (например, непрерывна на $[a, b]$ вместе с первыми двумя производными), требуется решить одну из следующих задач:

1) локальной интерполяции, когда необходимо указать способ, с помощью которого можно было бы приписать значению f в заданной между узлами промежуточной точке $x'\in\in(x_i, x_{i+1})$ некоторое «разумное» приближенное число, точность которого можно оценить;

2) глобальной интерполяции, когда требуется восстановить функцию f, т. е. найти такую аналитически заданную функцию φ, называемую интерполяционной, которая приближала бы f

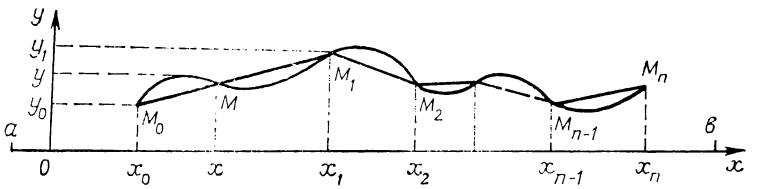


Рис. 11. Геометрическая интерпретация линейной интерполяции

на $[a, b]$, а ее значения в узлах интерполяции совпадали бы с заданными значениями функции f

$$\varphi(x_i)=f(x_i)=y_i. \tag{5.2}$$

Заметим, что вторая задача включает в себя первую и ее решение менее надежно. Поэтому на практике стараются ограничиваться решением локальных задач интерполяции.

В простейших случаях обычно достаточно воспользоваться линейной интерполяцией, которую можно выразить формулой

$$\varphi(x)=y_{i-1}+\frac{y_i-y_{i-1}}{x_i-x_{i-1}}(x-x_{i-1}),\ i=1, 2, \ldots, n. \tag{5.3}$$

Здесь имеется в виду, что $x\in[x_{i-1}, x_i]$. Именно линейная интерполяция (5.3) обычно используется при нахождении промежуточных значений функции в таблицах логарифмов, тригонометрических функций и т. п. С геометрической точки зрения формуле (5.3) соответствует прямолинейное звено ломаной (рис. 11) между двумя точками плоскости: $M_i(x_i, y_i)$ и $M_{i+1}(x_{i+1}, y_{i+1})$. При этом для вычисления $\varphi(x)$ с точным значением абсциссы x требуются лишь два узла интерполяции: ближайший слева и ближайший справа.

Такой простейший подход, использующий для каждой точки из всей исходной информации (5.1) лишь два узла, далеко не всегда обеспечивает требуемую точность интерполяции. Другими словами, ошибка интерполяции $|f(x)-\varphi(x)|$ может быть слишком велика. Поэтому приходится прибегать к интерполяционной функции более сложной, чем ломаная с прямолинейными звеньями.

Чаще всего $\varphi(x)$ выбирается из класса полиномов (многочленов). В этом случае интерполяция называется *полиномиальной* или *параболической* (линейная интерполяция есть частный случай параболической, т. е. $\varphi(x)$ есть многочлен 1-й степени).

При изучении полиномиальной интерполяции полезно иметь в виду известную из математического анализа теорему Вейерш-

трасса о том, что если $f(x)$ непрерывна на отрезке $[a, b]$, то для всякого $\varepsilon>0$ найдется такой полином $\varphi_k(x)$ степени k, что $|f(x)-\varphi_k(x)|<\varepsilon$ для любого $x\in(a, b)$. Отметим, что из теоремы не следует, что нужный полином φ_k является именно интерполяционным. Поэтому упомянутая теорема не является обоснованием выбора интерполяционной функции в виде полинома при решении задач глобальной интерполяции. Однако простота полиномиальной структуры и достаточная эффективность полиномов при решении локальных задач сделали полиномиальную интерполяцию одним из классических численных методов.

При изучении полиномиальной интерполяции таблично заданных функций предстоит ответить на следующие вопросы:

1) как определить наименьшую возможную степень k интерполяционного полинома $\varphi(x)$;

2) как построить этот полином и как им пользоваться;

3) как произвести оценивание точности интерполирования.

§ 5.2. ТАБЛИЧНЫЕ РАЗНОСТИ

Ответ на вопрос об определении наименьшей возможной степени k интерполяционного полинома $\varphi(x)$ зависит от гладкости интерполируемой функции: большая гладкость позволяет ограничиться меньшей степенью интерполяционного полинома. Но как оценить гладкость таблично заданной функции (5.1), когда известные средства дифференциального исчисления непосредственно применить не удается?

Предположим сначала, что исходная функция (5.1) имеет равноотстоящие узлы, т. е. $x_{i+1}-x_i=\Delta x_i\equiv h$, где $i=0, 1, \ldots, \ldots, n-1$. Положительная константа h называется *шагом таблицы*. Вычислим приращения функции:

$$\Delta y_0=y_1-y_0=f(x_0+h)-f(x_0),$$

$$\Delta y_1=y_2-y_1=f(x_0+2h)-f(x_0+h),$$

. . . . . .

$$\Delta y_{n-1}=y_n-y_{n-1}=f(x_0+nh)-f(x_0+(n-1)h).$$

Они называются конечными разностями 1-го порядка. Очевидно, что если узлов интерполяции $n+1$, то конечных разностей 1-го порядка будет n. По ним можно составить $n-1$ конечных разностей второго порядка:

$$\Delta^2 y_0=\Delta y_1-\Delta y_0,$$

$$\Delta^2 y_1 = \Delta y_2 - \Delta y_1,$$

. . . . . .

$$\Delta^2 y_{n-2} = \Delta y_{n-1} - \Delta y_{n-2}.$$

Вообще конечная разность порядка $k \leqslant n$ определяется через разности предыдущего порядка:

$$\Delta^k y_i = \Delta^{k-1} y_{i+1} - \Delta^{k-1} y_i.$$

Таких разностей имеется $n+1-k$. Наивысший порядок возможной разности есть n.

Отметим связь между конечными разностями и производными функции, интересную с точки зрения связи дискретного и непрерывного анализа. Так как

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

то

$$f'(x) \approx \frac{\Delta y}{h}. \tag{5.4}$$

Легко показать, что и вообще

$$f^n(x) \approx \frac{\Delta^n y}{h^n}, \tag{5.5}$$

хотя погрешность формулы (5.5) быстро растет с увеличением n.

Если некий полином k-й степени задан аналитически, то можно найти его производную k-го порядка и убедиться, что она обязательно представляет собой константу. Аналогичное утверждение справедливо и для таблично заданного полинома, но только роль производных играют конечные разности.

Теорема 5.1. Если $y = f(x)$ — полином степени k со старшим членом $a_0 x^k$, то конечная разность k-го порядка есть величина постоянная, равная $\Delta^k y = k! a_0 h^k$, а все разности более высокого порядка, чем k, равны нулю. Для практики особенно важно утверждение обратного смысла.

Теорема 5.2. Если при любом h конечные разности k-го порядка постоянны, то рассматриваемая функция $f(x)$ есть полином степени k. Доказательство можно найти в работе [4].

Приведенные теоремы обосновывают следующее практическое правило: степень интерполяционного полинома выбирается так, чтобы она совпадала с порядком практически постоянных (соответственно точности, с которой заданы значения $f(x)$ в узлах интерполяции) конечных разностей.

Итак, хотя мы пока не располагаем методикой построения интерполяционного полинома, но простой критерий для выяснения его степени получен.

Таблица 7

x, с	y, дм	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
1/30	1,19				
		+38			
2/30	1,57		+11		
		+49		+1	
3/30	2,06		+12		−4
		+61		−3	
4/30	2,67		+ 9		+6
		+70		+3	
5/30	3,37		+12		−5
		+82		−2	
6/30	4,19		+10		+4
		+92		+2	
7/30	5,11		+12		
		+104			
8/30	6,15				
Контроль:		+496	+66	+1	+1

В этой главе предполагается, что все цифры в исходных значениях $f(x)$ верные. Но и в этом случае погрешность y_0, y_1, ..., y_n достигает половины единицы младшего десятичного разряда. Значит, погрешность в разностях 1-го порядка может достигать уже единицы последнего разряда, в разностях 2-го порядка — двух единиц, в разностях 3-го порядка — четырех единиц последнего разряда и т. д.

Для гладких функций конечные разности убывают с увеличением порядка и можно найти такой порядок разностей, при котором они отличаются друг от друга на величины, соизмеримые с их погрешностями. Причем с дальнейшим увеличением порядка разности начинают расти, имея беспорядочные знаки.

Пример 5.1. Даны результаты измерения y высоты падения тела за время x (первые два столбца табл. 7). Определить порядок практически постоянных разностей.

Решение приводим в табл. 7, причем разности удобно писать целыми числами в единицах последнего знака. Относительно стабильно ведут себя разности 2-го порядка, а затем картина ухудшается. Следовательно, $\varphi(x)$ целесообразно искать в виде полинома именно 2-й степени. (В данном случае результат можно было предвидеть, так как $y=y_0+v_0x+gx^2/2$, где v_0 — начальная скорость, g — ускорение свободного падения.)

Пусть теперь узлы интерполяции x_0, x_1, ..., x_n не обязательно равноотстоящие. В этом случае аналогичную конечным разностям роль играют так называемые разделенные разности или «подъемы» функции.

Разделенными разностями 1-го порядка называются величины, имеющие смысл средних скоростей изменения функции:

$$[x_0, x_1] = \frac{y_1 - y_0}{x_1 - x_0}; \quad [x_1, x_2] = \frac{y_2 - y_1}{x_2 - x_1},$$

и вообще

$$[x_i, x_{i+1}] = \frac{y_{i+1} - y_i}{x_{i+1} - x_i}.$$

Разделенными разностями 2-го порядка называются величины

$$[x_0, x_1, x_2] = \frac{[x_1, x_2] - [x_0, x_1]}{x_2 - x_0};$$

$$[x_1, x_2, x_3] = \frac{[x_2, x_3] - [x_1, x_2]}{x_3 - x_1}.$$

Они связаны с изменением средней скорости изменения функции при переходе от предыдущего интервала (x_{i-1}, x_i) к следующему (x_i, x_{i+1}).

Разделенные разности k-го порядка определяются через разделенные разности $(k-1)$-го порядка с помощью рекуррентного соотношения

$$[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{[x_{i+1}, \ldots, x_{i+k}] - [x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}. \quad (5.6)$$

Для выявления связи дискретного и непрерывного анализа интересно иметь в виду следующее. В интервале (x_0, x_k) существует такая точка ξ, что

$$[x_0, x_1, \ldots, x_k] = \frac{f^{(k)}(\xi)}{k!}, \quad (5.7)$$

поэтому

$$\lim_{\substack{x_1 \to x_0 \\ x_2 \to x_0 \\ \cdots \\ x_k \to x_0}} [x_0, x_1, \ldots, x_k] = \frac{f^{(k)}(x_0)}{k!}. \quad (5.8)$$

Для приближенных «прикидок» иногда полезно соотношение

$$f^{(k)}(x) \approx k! [x_0, x_1, \ldots, x_k]. \quad (5.9)$$

Для разделенных разностей имеют место утверждения, аналогичные теоремам 5.1 и 5.2.

Теорема 5.3. Если $f(x)$ — полином степени k, то разделенные разности k-го порядка есть константы, не зависящие от узлов $x_0, x_1, \ldots, x_k$ и равные коэффициенту при старшей степе-

ни x в полиноме $f(x)$. Все разделенные разности большего, чем k, порядка равны нулю.

Эта теорема лежит в основе практического правила, позволяющего назначать степень интерполяционного полинома в случае неравноотстоящих узлов так, чтобы она совпадала с порядком практически постоянных разделенных разностей.

В дальнейшем нам полезно выражать значения функции $f(x_k)$ в произвольном узле интерполяции x_k через значение функции $f(x_0)$ в начальном узле x_0 и начальные значения разделенных разностей $[x_0, x_1]$, $[x_0, x_1, x_2]$, $[x_0, x_2, x_3]$ .... Это выражение можно вывести по индукции самостоятельно. Оно имеет вид

$$\begin{aligned} f(x_k) = f(x_0) + (x_k - x_0)[x_0, x_1] + \\ + (x_k - x_0)(x_k - x_1)[x_0, x_1, x_2] + \\ + \ldots + (x_k - x_0)(x_k - x_1)\ldots(x_k - \\ - x_{k-1})[x_0, x_1, \ldots, x_k]. \end{aligned} \tag{5.10}$$

Мы рассмотрели два вида табличных разностей: конечные и разделенные. Связь между ними для таблицы с постоянным шагом h легко установить по индукции

$$[x_0, x_0+h, \ldots, x_0+kh] = \frac{\Delta^k y_0}{k!h^k}. \tag{5.11}$$

Практически всегда можно пользоваться только разделенными табличными разностями.

§ 5.3. ИНТЕРПОЛЯЦИОННЫЙ ПОЛИНОМ

Рассмотрим вопрос о построении полинома $\varphi(x)$ и о его использовании.

Пусть по-прежнему некоторая функция $f(x)$ задана таблицей своих значений (5.1) в узлах интерполяции x_0, x_1, ..., x_n, причем они могут быть не равноотстоящими, но обязательно различными. С помощью табличных разностей, как описано в предыдущем параграфе, установлено, что подходящая степень интерполяционного полинома $\varphi(x)$ равна $k \leqslant n$. Требуется составить сам полином $\varphi(x)$. При этом надо учитывать следующую теорему.

Теорема 5.4. Существует один полином $\varphi(x)$ степени не выше n со свойствами (5.2) или, что все равно, через $n+1$ точку плоскости всегда можно провести параболу n-го порядка, и притом только одну.

Пусть искомый полином имеет вид

$$\varphi(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n,$$

где коэффициенты a_0, a_1, ..., a_n подлежат определению.

Начальные условия (5.2) фактически задают систему $n+1$ линейных уравнений с $n+1$ неизвестными a_0, a_1, ..., a_{n-1}, a_n:

$$a_0x_i^n+a_1x_i^{n-1}+\ldots+a_{n-1}x_i+a_n=y_i, \tag{5.12}$$

где $i=0, 1, \ldots, n$.

Определитель системы (5.12) есть определитель Вандермонда, и, так как x_i различны между собой, он отличен от нуля. Поэтому данная система всегда имеет единственное решение, что и требовалось доказать.

Из теоремы следует, что для решения поставленной в начале параграфа задачи не обязательно следует использовать все $n+1$ узлов интерполяции. Поэтому при $k<n$ возникает задача о выборе исходных узлов. Решается она в зависимости от назначения конструируемого полинома. Предположим, что $\varphi(x)$ предназначается для работы в той части области определения $f(x)$, которая представляется первыми $k+1$ узлами x_0, x_1, ..., ..., x_k.

Теорема 5.5. Искомый интерполяционный полином может быть записан в виде

$$\varphi(x)=\sum_{i=0}^{k} y_i\Lambda_i(x), \tag{5.13}$$

где

$$\Lambda_i(x)=\frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_k)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_k)}=$$

$$\prod_{j=0}^{k}(x-x_j)(x_i-x_j)^{-1}, \quad i\neq j. \tag{5.14}$$

$\Lambda_i(x)$ называются *полиномами влияния узла* x_i.

Обратим внимание на структуру записи (5.14): в числителе отсутствует сомножитель вида $(x-x_i)$, а в знаменателе нет сомножителя (x_i-x_i). Ясно, что $\Lambda_i(x)$ есть полином степени k для всякого $i=0, 1, 2, \ldots, k$, причем

$$\Lambda_i(x_j)=\begin{cases} 1, & \text{если } j=i,\\ 0, & \text{если } j\neq i.\end{cases} \tag{5.15}$$

Поэтому и полином (5.13) имеет степень не выше k и удовлетворяет условиям интерполяции (5.2), поскольку

$$\varphi(x_j)=y_0\cdot 0+y_1\cdot 0+\ldots+y_j\cdot 1+$$
$$+y_{j+1}\cdot 0+\ldots+y_k\cdot 0=y_j,$$

где j принимают значения 0, 1, ..., k.

Теорема доказана, так как других полиномов с теми же свойствами не может быть согласно предыдущей теореме.

Замечание 5.1.

$$\sum_{i=0}^{k}\Lambda_i(x)\equiv 1. \tag{5.16}$$

В самом деле, если предположить, что $y_i=1$, то $k=0$, так как табличные разности 0-го порядка (т. е. сами ординаты) постоянны. Поэтому $\varphi(x)\equiv 1$ и (5.16) сразу следует из (5.13). Свойством (5.16) удобно пользоваться для контроля составления полиномов влияния.

Замечание 5.2. Полиномы влияния зависят только от узлов интерполяции и не зависят от интерполируемой функции.

Замечание 5.3. Полином k-й степени (5.13) наиболее целесообразно использовать в той части области определения $f(x)$, которая расположена примерно посредине конфигурации именно первых $k+1$ узлов из всех заданных $n+1$, причем $k\leqslant n$. Если точка x', в которой надо оценить значение $f(x)$, не удовлетворяет этому условию, то предварительно из $n+1$ заданных узлов надо выбрать $k+1$ узлов, ближайших к x', и затем перенумеровать их от 0 до k в порядке возрастания. Допустимо и упорядочивание этих узлов по степени близости к x', т. е. по признаку $|x_i-x'|\geqslant|x_j-x'|$, если $i>j$.

Форму записи (5.13) интерполяционного полинома принято называть формой Лагранжа.

Пример 5.2.

x	0	1	3	4
y	−4	0,5	0,5	8

А. Решить глобальную задачу с помощью интерполяционного полинома в форме Лагранжа не выше 3-й степени.

Решение. Согласно (5.14):

$$\Lambda_0(x)=\frac{(x-1)(x-3)(x-4)}{(0-1)(0-3)(0-4)}=-\frac{1}{12}x^3+\frac{2}{3}x^2-\frac{19}{12}x+1,$$

$$\Lambda_1(x)=\frac{(x-0)(x-3)(x-4)}{(1-0)(1-3)(1-4)}=\frac{1}{6}x^3-\frac{7}{6}x^2+2x,$$

$$\Lambda_2(x)=\frac{(x-0)(x-1)(x-4)}{(3-0)(3-1)(3-4)}=-\frac{1}{6}x^3+\frac{5}{6}x^2-\frac{2}{3}x,$$

$$\Lambda_3(x)=\frac{(x-0)(x-1)(x-3)}{(4-0)(4-1)(4-3)}=\frac{1}{12}x^3-\frac{1}{3}x^2+\frac{1}{4}x.$$

Контроль (5.16) выполняется. Согласно (5.13), после приведения подобных членов

$$\varphi(x)=x^3-5{,}5x^2+9x-4.$$

Б. Решить локальную задачу, т. е. приближенно оценить значение $f(x)$ в промежуточной точке $x'=2$.

Решение. Интерполяционный полином и полиномы влияния находить нет необходимости. Целесообразно сразу в числителях $\Lambda_i(x)$ заменить x на 2. Получим: $\Lambda_0(2)=-1/6$, $\Lambda_1(2)=2/3$, $\Lambda_2(2)=2/3$, $\Lambda_3(2)=-1/6$.

Контроль $\sum_{i=0}^{3}\Lambda_i(2)=1$ выполняется. Подставляя числа $\Lambda_i(2)$ в (5.13), получаем $f(2)\approx\varphi(2)=0$.

Интерполяционный полином в форме Лагранжа удобно применять в тех случаях, когда ведется многократное интерполирование по одним и тем же узлам (пусть даже различных функций). Дело в том, что в этом случае можно заранее составить полиномы влияния. Вычисления особенно упрощаются при интерполировании многих функций в одной и той же промежуточной точке x', так как при этом полиномы влияния представляют собой постоянные числа. Разумеется, для разных узлов и разных точек x' они различны.

Уясним существенный недостаток интерполяционного полинома в форме Лагранжа. Пусть по заданной табличным способом функции $f(x)$ вычислено ее интерполированное значение в некоторой промежуточной точке x' с помощью полинома k-й степени:

$$\varphi_k(x') \approx f(x').$$

После того как вычисления закончены, представим что возникла необходимость снова получить $f(x')$, но с привлечением одного дополнительного узла (и, следовательно, с помощью полинома $k+1$ степени $f_{k+1}(x)$). Объясняется это дополнительно проведенными измерениями, желанием проконтролировать и уточнить уже имеющийся результат и другими вариантами. В этом случае в формуле (5.13) не только добавится одно дополнительное слагаемое, но и все коэффициенты (5.14) придется перевычислять заново, так как они зависят от всех $k+1$ узлов сразу и в них появятся дополнительные множители. Этот недостаток можно избежать, если представить интерполяционный полином в иной форме. Обратимся для этого к формуле (5.10). В силу условий (5.2), интерполяционный полином $\varphi(x)$ имеет те же разделенные разности, что и интерполируемая функция $f(x)$. Поэтому равенства не изменятся, если f в (5.10) заменить на φ вида (5.13). При этом равенства становятся тождеством на отрезке $[x_0, x_n]$, поскольку $\varphi(x)$, являясь полиномом k-й степени, имеет все разделенные разности большего, чем k, порядка, равные 0 (см. теорему 5.3). Таким образом, интерполяционный полином можно записать в виде

$$\varphi(x) = y_0 + \sum_{i=1}^{k} ([x_0, x_1, \ldots, x_i] \prod_{j=0}^{i-1} (x - x_j)), \qquad (5.17)$$

который называется формой Ньютона. При $k=1$ он совпадает с линейной функцией (5.3), где $i=1$.

Формулы (5.13) и (5.17) представляют собой лишь различные формы записи одного и того же интерполяционного полинома. Однако форма Ньютона более удобна тем, что при добавлении к узлам $x_0, x_1, \ldots, x_k$ нового узла x_{k+1} все ранее найденные члены остаются без изменения и в полиноме добавляется лишь один дополнительный член вида

$$[x_0, x_1, \ldots, x_k, x_{k+1}](x-x_0)(x-x_1)\ldots(x-x_k)=$$
$$=c_k\prod_{j=0}^{k}(x-x_j).$$

Это позволяет, последовательно добавляя по одному дополнительному узлу, постепенно увеличивать точность результата интерполяции.

Так как табличные разности обычно быстро убывают с увеличением порядка, то ближайшие к данной точке x' узлы интерполяции окажут основное влияние на искомую величину, а остальные будут давать лишь небольшие поправки. Поэтому решение локальной задачи интерполяции для заданной точки x' надо организовать таким образом, чтобы очередность привлечения каждого нового узла x_i зависела от близости его к x' (см. замечание 5.3). Принятая нами упорядоченность исходных узлов по мере их возрастания служит лишь для определенности. Для счета порядок нумерации узлов безразличен. Рекомендуется оставлять только те члены, которые больше допустимой погрешности. Тем самым достигается возможность автоматически определять достаточную степень интерполяционного полинома и общее количество узлов, необходимое для решения локальной задачи.

Благодаря перечисленным свойствам, интерполяционный полином в форме Ньютона наиболее удобен для расчетов на ЭВМ. Аналогичным свойством характеризуется и так называемая интерполяционная схема Эйткена [4].

На рис. 12 приведена блок-схема одного из простейших алгоритмов решения локальной задачи интерполяции с точностью до $\varepsilon>0$ с помощью полинома в форме Ньютона. Форма Лагранжа применяется в основном в теоретических вопросах.

Известны и другие формы записи интерполяционного полинома, которые для некоторых частных случаев более эффективны, чем форма Ньютона. Однако соответствующие преимущества обычно несущественны при использовании ЭВМ, и на них можно не останавливаться.

§ 5.4. ТОЧНОСТЬ ПОЛИНОМИАЛЬНОЙ ИНТЕРПОЛЯЦИИ

Пусть для функции $f(x)$, заданной в виде (5.1) с $n+1$ узлом интерполяции $x_0, x_1, \ldots, x_n$, построен интерполяционный полином n-й степени $\varphi(x)$. Значения $\varphi(x)$ совпадают со значениями $f(x)$ только в узлах интерполяции. Поэтому естественно возникает вопрос о величине отличия $\varphi_n(x')$ от $f(x')$ в промежуточных точках $x'\in[x_0, x_n]$:

$$R_n(x')=f(x')-\varphi_n(x'). \tag{5.18}$$

От чего зависит погрешность $R_n(x')$? Прежде всего, от структуры $f(x)$. Если, например, $f(x)$ есть полином степени не вы-

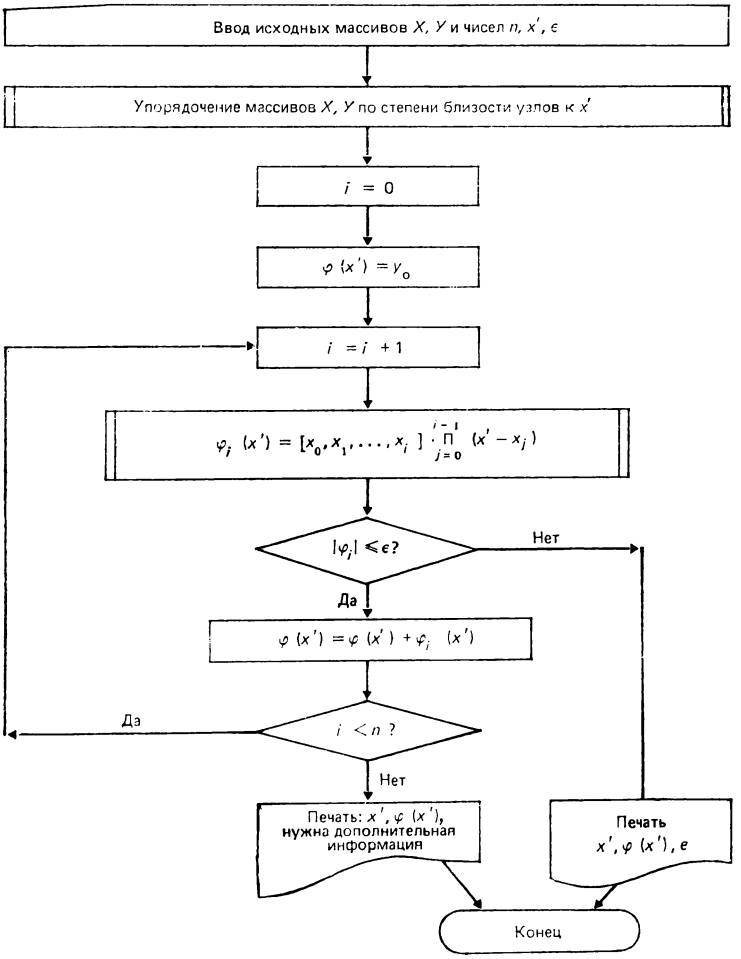


Рис. 12. Блок-схема алгоритма интерполяции с помощью полинома Ньютона

ше n, то $R_n(x')=0$. Если же о свойствах $f(x)$ ничего не известно, то оценить величину погрешности невозможно. Справедливо следующее утверждение.

Теорема 5.6.

Если исходные данные в таблице $f(x)$ безошибочны, ошибки округления при вычислениях $\varphi(x')$ отсутствуют и $x'\in[x_0, x_n]$, а функция $f(x)$ на отрезке $[x_0, x_n]$ обладает непрерывными производными до $(n+1)$-го порядка включительно, то $\exists\xi$:

$$R_n(x')=\frac{f^{(n+1)}(\xi)}{(n+1)!}(x'-x_0)(x'-x_1)\ldots(x'-x_n), \qquad (5.19)$$

где ξ зависит от x' и $\xi \in (x_0, x_n)$. Доказательство приведено в [4].

Полной ясности формула (5.19) не дает, так как положение точки ξ в интервале (x_0, x_n) не определено. Поэтому на практике обычно пытаются оценивать верхнюю границу для $[f^{n+1}(x)]$ при $x \in [x_0, x_n]$.

Введем обозначение

$$M_{n+1} = \max | f^{(n+1)}(x) | \text{ при } x \in [x_0, x_n].$$

Тогда (5.19) дает следующую оценку погрешности интерполяции:

$$| R_n(x') | \leqslant \frac{M_{n+1}}{(n+1)!} \prod_{j=0}^{n} | x' - x_j |. \qquad (5.20)$$

На практике редко имеется достаточно информации для вычисления величины M_{n+1}. Приближенное представление о ней можно получить с помощью формул (5.5) и (5.9).

С точки зрения связи между анализом дискретным и непрерывным интересно отметить аналогию представления функции $f(x)$ в виде

$$f(x) = \varphi_n(x) + R_n(x),$$

где правая часть определяется формулами (5.17) и (5.20) или (5.13) и (5.20), с известной формулой Тейлора. Произведение разностей $(x-x_0)(x-x_1)\ldots(x-x_n)$ является обобщением степени бинома, а разделенные разности являются как бы обобщенными производными. Если узлы $x_0, x_1, \ldots, x_n$ стягиваются в одну точку, например в x_0, то, согласно (5.8), формула Ньютона обращается в формулу Тейлора. Таким образом, формулу Ньютона можно рассматривать как обобщение формулы Тейлора на случай дискретного анализа.

Ранее было рекомендовано выбирать такое минимальное количество $i+1$ ближайших к x' узлов интерполяции, при котором выполняется неравенство $[\varphi_{i+1}] < \varepsilon$, где $\varepsilon > 0$ — заданная точность интерполяции (см. блок-схему на рис. 12). Теперь видно, что данное решение приближенно, поскольку это эмпирическое правило не может обеспечить выполнение строгого неравенства $|R_i(x'| < \varepsilon$, что следует из подстановки приближенного соотношения (5.9) в формулу (5.20). Для оценивания величины $|R_i(x')|$ целесообразно пользоваться разделенной разностью i-го порядка, максимальной по модулю.

Пример 5.3. Оценить точность ε линейной интерполяции (5.3) в случае равноотстоящих узлов с шагом h.

Решение. Так как интерполяция линейна, то имеются лишь два ее узла x_0, x_1 и потому $n=1$, $x_1 = x_0 + h$. Согласно (5.20), имеем

$$|R_1(x)| \leqslant \frac{M_2}{2} |x-x_0| |x-x_1| = \varepsilon.$$

Легко убедиться, что $|x-x_0||x-x_1| \equiv |x-x_0||x-x_0-h|$ достигает максимума, равного $h^2/4$, при $x=(1/2)(x_0+x_1)$. Поэтому для всякого $x \in [x_0, x_1]$

$$|R_1(x)| \leqslant \frac{1}{8} M_2 h^2 \approx \frac{1}{8} |\Delta^2 y| \approx \varepsilon,$$

где $M_2 \approx |\Delta^2 y|/h^2$ согласно (5.5).

Если величина M_{n+1} в правой части неравенства (5.20) является характеристикой только интерполируемой функции, то величина $\Pi|x'-x_j|$ зависит от расположения узлов сетки, которыми можно распоряжаться. Минимум этой величины достигается в том случае, если узлы совместить с корнями полинома Чебышева (см. [9]). Однако выигрыш в точности не оправдывает заметное усложнение численных расчетов.

Отметим, что в любом случае

$$\prod_{j=0}^{n} |x' - x_j| < \max_j |x' - x_j|^{n+1} \leqslant (nh)^{n+1},$$

где h — шаг сетки (для неравномерной — максимальное расстояние между соседними узлами). Если число узлов фиксировано, а $h \to 0$, то полиномиальная интерполяция функции с ограниченной $(n+1)$-й производной имеет погрешность порядка $O(h^{n+1})$. В таких случаях говорят, что обеспечивается $(n+1)$-й порядок малости относительно шага. Порядок малости является важной сравнительной характеристикой сеточных методов. Он позволяет также использовать часто эффективный так называемый двойной пересчет по схеме Рунге (см. § 7.5).

Если же h сетки фиксирован, то увеличивать степень интерполяционного полинома не всегда целесообразно. Если о гладкости интерполируемой функции ничего не известно, то вполне может случиться, что величины M_n (в 5.20) с увеличением n растут быстрее, чем $n!$, либо вовсе не существуют, начиная с некоторого порядка. В таких ситуациях использование новых дополнительных узлов только ухудшит реальную точность интерполяции.

Приближение произвольной функции интерполяционным полиномом теоретически приемлемо (и потому надежно) лишь в рамках локальных задач. Если при этом 4—5 узлов не обеспечивают заданную точность интерполяции, то целесообразнее не увеличивать количество узлов, а уменьшать шаг таблицы. Если интерполируемая функция имеет по меньшей мере вторую производную ограниченной, то даже линейная интерполяция может обеспечить высокую точность — надо только достаточно измельчить расстояние между узлами.

Глобальное описание функции не всегда оправдано. В связи с этим вспомним указанную выше аналогию между интерполя-

ционным полиномом в форме Ньютона и конечным рядом Тейлора: последний тоже отражает поведение функции лишь в окрестности точки разложения.

Решение интерполяционных задач особенно усложняется, когда исходные сеточные значения функции имеют разного рода погрешности (это особенно часто имеет место при решении геодезических задач). В таких ситуациях лишь табличные разности первых двух порядков бывают не слишком далеки от истины, но относительные погрешности велики уже и у них. Поэтому табличные разности высоких порядков практически носят случайный характер и, разумеется, непригодны для интерполирования.

§ 5.5. ИНТЕРПОЛЯЦИЯ СПЛАЙНАМИ

Основные определения и свойства. В предыдущем параграфе отмечались недостатки глобального описания таблично заданной функции (5.1) с помощью полиномиальной интерполяции. Однако для локального описания полиномы являются почти идеальным средством благодаря своей простоте. В связи с этим возникла мысль приближать интерполируемую функцию на каждом отрезке между двумя соседними узлами сетки $[x_{i-1}, x_i]$, $i=1, \ldots, n$ с помощью персонального полинома той или иной степени, а затем «склеить» эти полиномы в узлах так, чтобы обеспечить в целом гладкость интерполяционного агрегата (интерполянта) вплоть до нужной производной. Получающиеся в результате кусочно-полиномиальные функции, обладающие определенной гладкостью, называются *сплайнами* и играют важнейшую роль в современной вычислительной математике. Сплайн p-й степени «склеивается» из кусочных полиномов p-й степени так, что первые $p-1$ производные его были непрерывны на отрезке $[x_0, x_n]$, а p-я производная интегрируема с квадратом на этом отрезке.

Простейшим примером может служить сплайн 1-й степени (линейный сплайн), графиком которого служит ломаная (см. рис. 11).

Можно доказать [12], что сплайны степени p обладают следующим важным свойством: их $(p-1)$-я производная минимальна на отрезке интерполяции в том смысле, что

$$\int_{x_0}^{x_n} \left(\frac{d^{p-1}\varphi}{dx^{p-1}} \right)^2 dx = \min \tag{5.21}$$

по сравнению с любыми другими интерполяционными функциями.

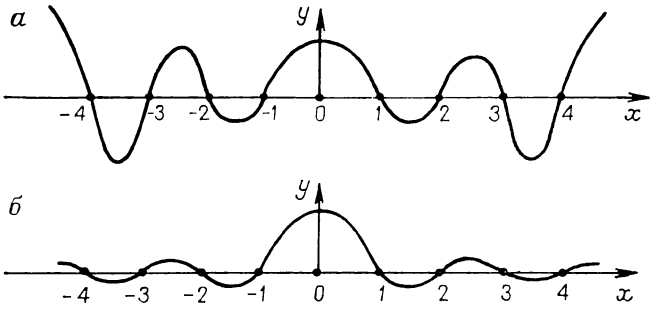


Рис. 13. Графики:
а — интерполяционного полинома; *б* — кубического сплайна

Замечание 5.4. Свойство (5.21) является характеристическим (т. е. справедливо только для сплайнов степени p) и потому может быть положено в основу определения сплайна p-й степени.

Если $p=3$, то под знаком интеграла в (5.21) стоит квадрат второй производной, определяющей кривизну соответствующей кривой. Поэтому сплайн 3-й степени (кубический) является «самым неискривленным» по сравнению с любыми другими интерполянтами. Преимущества интерполянта в виде кубического сплайна по сравнению с интерполянтом в виде обычного полинома видны из следующего.

Требуется решить глобальную задачу интерполяции для некоторой функции, заданной на сетке с постоянным единичным шагом. Сетка состоит из девяти узлов, симметрично расположенных относительно нуля. Исходные значения функции равны нулю во всех узлах, кроме центрального, где функция равна единице. На рис. 13 показаны графики интерполяционного полинома и кубического сплайна, разрешающих поставленную задачу.

У сплайна нет присущей полиному осцилляции на концах.

Предположим теперь, что рассматриваемую задачу надо решить графически. Возьмем для этого гибкую металлическую линейку. Поставив линейку на ребро, будем изгибать ее, придерживая пальцами одновременно в нескольких местах так, чтобы ребро проходило сразу через все точки. Можно доказать, что прочерченная кривая совпадает с графиком кубического сплайна (поскольку уравнением свободного равновесия упругого бруска служит равенство нулю четвертой производной). Этому обстоятельству сплайны обязаны своим названием (по-английски *spline* — рейка, упругий брусок). Уникальное свойство минимальной кривизны сделало кубические сплайны наиболее распространенными в задачах сплайновой интерполяции.

Построение кубического сплайна. Рассмотрим построение кубического сплайна на примере решения задачи глобальной интерполяции для сеточной функции (5.1).

На отрезке $[x_0, x_n]$ требуется найти функцию $\varphi(x)$, удовлетворяющую условиям:

а) $\varphi(x)$ непрерывна вместе со своими производными до 2-го порядка включительно;

б) на каждом i-м отрезке $[x_{i-1}, x_i]$ сетки $\varphi(x)$ является кубическим полиномом вида

$$\varphi_i(x) = \begin{cases} a_{0i} + a_{1i}(x - x_{i-1}) + a_{2i}(x - x_{i-1})^2 + \\ + a_{3i}(x - x_{i-1})^3 \text{ при } x \in [x_{i-1}, x_i]; \\ 0 \text{ при } x \notin [x_{i-1}, x_i], \ i = 1, \ldots, n; \end{cases}$$

$$\varphi(x) = \sum_{i=1}^{n} \varphi_i(x) \text{ при } x \in [x_0, x_n];$$

в) в узлах сетки выполняются условия интерполяции

$$\varphi(x_i) = f(x_i) \equiv y_i, \ i = 0, 1, \ldots, n.$$

Так как функции $\varphi_i(x)$ представляют собой различные кубические полиномы на каждом i-м отрезке $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, \ldots, n$, то неизвестных коэффициентов в данной задаче $4n$. Однако независимых неизвестных (в силу условий непрерывности) оказывается только $n+1$. В качестве них удобно выбрать значения m_i вторых производных искомого сплайна в узлах. Числа $m_i = \varphi''(x_i)$, $i = 0, 1, \ldots, n$ будем называть *моментами сплайна.*

Функция $\varphi_i(x)$ по условию есть полином 3-й степени на каждом i-м частичном отрезке. Следовательно, ее вторая производная — линейная функция на этом отрезке. Составленная из них вторая производная $\varphi''(x)$ сплайна должна быть непрерывной на всем отрезке $[x_0, x_n]$, графиком ее служит ломаная с ординатами m_i (см. рис. 11). Поэтому можно записать

$$\varphi''_i(x) = m_{i-1} \frac{x_i - x}{h_i} + m_i \frac{x - x_{i-1}}{h_i}, \ x \in [x_{i-1}, x_i], \qquad (5.22)$$

где $h_i = x_i - x_{i-1}$. Эта функция линейна относительно x, и ее значения на концах отрезка $\varphi_i''(x_{i-1}) = m_{i-1}$, $\varphi_i''(x_i) = m_i$ обеспечивают непрерывное «склеивание» φ_i'' в $\varphi''(x)$.

Проинтегрируем теперь дважды обе части равенства (5.22). Результат удобно записать в виде

$$\varphi_i(x) = m_{i-1} \frac{(x_i - x)^3}{6h_i} + m_i \frac{(x - x_{i-1})^3}{6h_i} + C_{1i} \frac{x_i - x}{h_i} + C_{2i} \frac{x - x_{i-1}}{h_i},$$

где C_{1i} и C_{2i} — произвольные постоянные (правильность легко проверить обратным двойным дифференцированием и получе-

нием (5.22)). Значения C_{1i} и C_{2i} определяются краевыми условиями

$$\varphi_i(x_{i-1})=y_{i-1} \Rightarrow \frac{1}{6} m_{i-1}h_i^2+C_{1i}=y_{i-1},$$

$$\varphi_i(x_i)=y_i \Rightarrow \frac{1}{6} m_i h_i^2+C_{2i}=y_i.$$

Окончательно имеем ($i=1, 2, \ldots, n$ — номер частичного отрезка $[x_{i-1}, x_i]$):

$$\varphi_i(x)=m_{i-1}\frac{(x_i-x)^3}{6h_i}+m_i\frac{(x-x_{i-1})^3}{6h_i}+\left(y_{i-1}-\frac{1}{6}m_{i-1}h^2{}_i\right)\frac{x_i-x}{h_i}+$$

$$+\left(y_i-\frac{1}{6}m_i h^2{}_i\right)\frac{x-x_{i-1}}{h_i}, \qquad (5.23)$$

$$\varphi'{}_i(x)=-m_{i-1}\frac{(x_i-x)^2}{2h_i}+m_i\frac{(x-x_{i-1})^2}{2h_i}+$$

$$+\frac{1}{h_i}(y_i-y_{i-1})-\frac{h_i}{6}(m_i-m_{i-1}). \qquad (5.24)$$

Из выражения (5.23) следует, что все коэффициенты искомого сплайна выражаются через моменты. Основу для вычисления этих моментов дает условие непрерывности $\varphi'(x)$. Оно состоит в том, что для каждого узла x_i с номером $i=1, \ldots, n-1$ (кроме крайних) значения производной сплайна, вычисленные с помощью φ'_i по формуле (5.24) и φ'_{i-1} по формуле*

$$\varphi'{}_{i+1}(x)=-m_i\frac{(x_{i+1}-x)^2}{2h_{i+1}}+m_{i+1}\frac{(x-x_i)^2}{2h_{i+1}}+$$

$$+\frac{1}{h_{i+1}}(y_{i+1}-y_i)-\frac{h_{i+1}}{6}(m_{i+1}-m_i), \qquad (5.25)$$

должны совпадать. Составив уравнение $\varphi'_i(x_i)=\varphi'_{i+1}(x_i)$, имеем

$$\frac{1}{6}h_i m_{i-1}+\frac{1}{3}(h_i+h_{i+1})m_i+\frac{1}{6}h_{i+1}m_{i+1}=$$

$$=\frac{y_{i+1}-y_i}{h_{i+1}}-\frac{y_i-y_{i-1}}{h_i}. \qquad (5.26)$$

Таких уравнений будет $n-1$ с $n+1$ неизвестными. Чтобы получить единственное решение, надо указать еще два условия для сплайна (обычно в крайних узлах x_0 и x_n). Это можно сделать, если имеется дополнительная информация об интерполируемой функции. На практике удобно дополнительно потребовать нулевую кривизну сплайна на концах, т. е.

---

* Формула (5.25) получена из (5.24) формальной заменой i на $i+1$.

$$m_0 = m_n = 0. \tag{5.27}$$

В рассмотренной аналогии кубического сплайна с гибкой металлической линейкой условие (5.27) означает, что концы линейки свободно опущены. Виды других краевых условий и рекомендации по их выбору можно найти в [8]. Мы ограничимся условиями (5.27).

Совокупность уравнений (5.26), (5.27) дает следующую систему линейных алгебраических уравнений для определения моментов $m_1, \ldots, m_{n-1}$ сплайна:

$$Am = By, \tag{5.28}$$

где A — квадратная симметричная трехдиагональная матрица размера $[(n-1)\times(n-1)]$, зависящая только от взаимного расположения (конфигурации) узлов. Ее главную диагональ составляют числа $2(h_1+h_2)$, $2(h_2+h_3)$, ..., $2(h_{n-1}+h_n)$. Соседние верхняя и нижняя диагонали одинаковы и состоят из чисел $h_2, h_3, \ldots, h_{n-1}$. Все остальные элементы — нули. Столбец m длиной $n-1$ содержит искомые моменты сплайна $m_1, \ldots, m_{n-1}$. Прямоугольная матрица B имеет размерность $[(n-1)\times(n+1)]$. Ее первую строчку составляют числа $6/h_1$, $6[(-1/h_1)-(1/h_2)]$, $6/h_2$, далее нули. Вторую строчку составляют числа 0, $6/h_2$, $6[(-1/h_2)-(1/h_3)]$, $6/h_3$ и нули. Третья строчка: 0, 0, $6/h_3$, $6[(-1/h_3)-(1/h_4)]$, $6/h_4$, затем нули. Аналогично заполняются остальные строчки. В последней строчке отличными от нуля являются только последние три числа: $6/h_{n-1}$, $6[(-1/h_{n-1})-(1/h_n)]$, $6/h_n$.

Столбец y длиной $n+1$ содержит исходные ординаты $y_0, \ldots, y_n$. Матрица A имеет строгое диагональное преобладание в том смысле, что

$$|a_{ii}| - \sum_{j \neq i} |a_{ij}| = r_i > 0, \tag{5.29}$$

где $i, j = 1, 2, \ldots, n-1$.

Можно доказать, что матрица A положительно определена и потому система (5.28) имеет единственное решение. Следовательно, сплайн $\varphi(x)$ также однозначно определяется по формулам (5.23).

Поскольку матрица A имеет специальную структуру (трехдиагональную), то объем вычислений при решении системы (5.28) возрастает с ростом n значительно медленнее, чем в случае плотно заполненной матрицы коэффициентов. Метод прогонки позволяет эффективно решать системы типа (5.28) даже при очень большом n.

Диагональное преобладание матрицы A позволяет легко оценить число обусловленности $\mu(A)$. Можно доказать [9], что $\mu(A) \leqslant \|A\| \cdot \max_i (1/r_i)$, где r_i определяются формулой (5.29).

Поэтому если величина $\max\limits_i |a_{ii}| (\min\limits_i r_i)^{-1}$ невелика, то система (5.28) хорошо обусловлена.

Отметим частный случай, когда исходная сетка узлов имеет постоянный шаг $h \equiv h_i$. Введем новую переменную $t = (x-x_0)/h$, имеющую стандартные узлы $t_i = i$, где $i = 0, 1, \ldots, \ldots, n$. Шаг равен единице. На главной диагонали матрицы A стоят одинаковые числа 4, а соседние верхнюю и нижнюю диагонали составляют одинаковые числа 1. Столбец By в (5.28) длиной $n-1$ заполняют увеличенные в 6 раз вторые конечные разности $\Delta^2 y_{i-1}$, $i = 1, \ldots, n-1$. Другими словами, каждому узлу i (кроме крайних $i=0$ и $i=$п) соответствует уравнение

$$m_{i-1} + 4m_i + m_{i+1} = 6(y_{i+1} - 2y_i + y_{i-1}) = 6\Delta^2 y_{i-1}. \qquad (5.30)$$

Дополнив решение системы таких уравнений моментами (5.27), получаем по формуле (5.23)

$$\varphi_i(t) = \frac{h^2}{6} m_{i-1}(i-t)^3 + \frac{h^2}{6} m_i (t-i+1)^3 +$$

$$+ \left(y_{i-1} - \frac{h^2}{6} m_{i-1}\right)(i-t) + \left(y_i - \frac{h^2}{6} m_i\right)(t-i+1) \qquad (5.31)$$

и t в (5.31) заменяем на $(1/h)(x-x_0)$, чтобы прийти к старой переменной. При решении локальных задач часто удобнее пользоваться непосредственно формулой (5.31), подставляя туда для каждой промежуточной точки x' соответствующее $t' = (1/h)(x'-x_0)$.

Пример 5.4.

x	1	3	5	7
y	2	5	6	3
i	0	1	2	3

С помощью кубического сплайна решить локальную задачу интерполяции в точке $x' = 4$.

Решение. $n=3$, $h=2$, $t = (1/2)(x-1)$. Система (5.28) согласно (5.30), имеет вид

$$\left.\begin{array}{l} 4m_1 + m_2 = -12 \\ m_1 + 4m_2 = -24 \end{array}\right\} \Rightarrow \quad m_1 = -1{,}6; \quad m_2 = -5{,}6; \quad m_0 = m_3 = 0.$$

Интересующая нас точка $t' = (1/2)(4-1) = 1{,}5$ принадлежит отрезку $[1, 2]$, имеющему номер $i=2$. По формуле (5.31) имеем

$$\varphi_2(t) = \frac{4,\,1;\,6}{6}(2-t)^3 - \frac{4,\,5;\,6}{6}(t-1)^3 +$$

$$+ \left(5 + \frac{4,\,1;\,6}{6}\right)(2-t) + \left(6 + \frac{4,\,5;\,6}{6}\right)(t-1),$$

$$f(4) \approx \varphi_2(1{,}5) = 7{,}30.$$

С точки зрения экономии вычислений целесообразно формулу (5.23) записать в виде [9]

$$\varphi_i(x) = y_{i-1} + \frac{x - x_{i-1}}{h_i}\left\{(y_i - y_{i-1}) - (x_i - x)\times\right.$$
$$\left.\times\left[\frac{1}{6} m_{i-1}(x_i - x + h_i) + \frac{1}{6} m_i (x + h_i - x_{i-1})\right]\right\},$$

требующем выполнения лишь шестнадцати арифметических операций. Если же сплайном приходится пользоваться многократно, то полезно вычислить и хранить в памяти ЭВМ коэффициенты a_{0i}, a_{1i}, a_{2i}, a_{3i} в записи кубического полинома по степеням $(x - x_{i-1})$. Это можно сделать с помощью формул $a_{0i} = y_{i-1}$; $a_{1i} = [(y_i - y_{i-1})/h_i] - (m_i + 2m_{i-1}) \times h_i/6$; $a_{2i} = m_{i-1}/2$; $a_{3i} = (m_i - m_{i-1})/6h_i$, $i = 1, 2, \ldots, n$.

Далее удобно воспользоваться схемой Горнера:

$$\varphi_i(x) = y_{i-1} + u_i(a_{1i} + u_i(a_{2i} + u_i a_{3i})), \quad u_i = x - x_{i-1},$$

требующей выполнения лишь шести арифметических операций.

Кубические сплайны обладают хорошими аппроксимирующими свойствами. Можно доказать, что если интерполируемая функция f непрерывна на отрезке $[x_0, x_n]$ вместе со своими первыми l производными ($l = 0, 1, 2, 3$), то для погрешности справедливо неравенство

$$\max_{x_0 \leqslant x \leqslant x_n} |f^{(p)}(x) - \varphi^{(p)}(x)| \leqslant Ch^{l-p} = O(h^{l-p}), \quad p \leqslant l, \qquad (5.32)$$

где C — неотрицательная константа, не зависящая от сетки; h — шаг сетки (для неравномерной — максимальное расстояние между соседними узлами). Одно из следствий — возможность использования схемы Рунге (см. § 7.5). Известны и более точные оценки [3], [9].

§ 5.6. ИНТЕРПОЛЯЦИЯ И КОЛЛОКАЦИЯ НА ПЛОСКОСТИ

В § 5.1 этой главы уже отмечалось, что в геодезической практике часто используется интерполяция функций именно двух переменных. Постановка задач аналогична уже разобранной. Только теперь интерполируемой функцией служит функция $z = f(x, y)$, которую можно трактовать как плоское скалярное поле. Предполагается, что эта функция обладает определенной гладкостью в заданной ограниченной замкнутой плоскости D, но ее значения $z_1, \ldots, z_n$ известны лишь в отдельных точках $P_1, \ldots, P_n$ этой области, называемых узлами сетки. Совокупность узлов определяет собой сетку той или иной конфигурации.

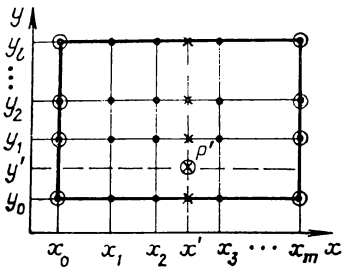


Рис. 14. Регулярная прямоугольная сетка

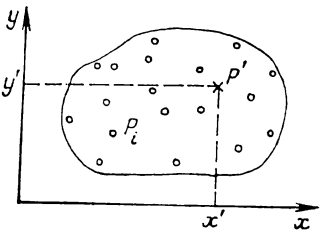


Рис. 15. Хаотичная сетка

Сложность решения задач интерполяции на плоскости существенно зависит от конфигурации узлов сетки. Мы рассмотрим две стандартные ситуации: а) в прямоугольной области задана регулярная прямоугольная сетка (рис. 14); б) в односвязной области D задана хаотичная сетка (рис. 15).

Случай регулярной сетки. В случае регулярной сетки уже известная техника интерполяции функций одной переменной является обычно достаточной для решения локальных задач интерполяции и функций двух переменных. Так, на прямоугольной сетке удобна последовательная интерполяция. Пусть заданы $z_{ij}=f(x_i, y_i)$, где $i=0, 1, \ldots, m$ и $j=0, 1, \ldots, n$, и требуется оценить значение $f(x', y')$. Сначала выполним одномерную интерполяцию для каждой j-й строки и вычислим значения $\varphi(x', y_0)$, $\varphi(x', y_1)$, ..., $\varphi(x', y_l)$. Соответствующие точки на рис. 14 отмечены крестиками. Далее эти числа используются для интерполяции по столбцу x' для вычисления нужного числа $\varphi(x', y')$. По каждой переменной можно брать свое число узлов и пользоваться либо полиномиальной интерполяцией, либо сплайновой. Некоторым недостатком последовательной интерполяции является завышение степени результирующего интерполяционного полинома двух переменных. Например, если по каждой переменной выполнять линейную интерполяцию, т. е. пользоваться полиномом 1-й степени, то результирующий интерполяционный полином двух переменных представит собой билинейную форму вида $a_0+a_1x+a_2y+a_3xy$, т. е. будет уже полиномом 2-й степени.

Можно построить и полином минимальной степени k, если использовать треугольную конфигурацию ближайших к точке P' узлов в количестве $(k+1)(k+2)/2$. Так, при $k=1$ интерполяционный полином двух переменных в форме Ньютона имеет для узлов (x_0, y_0), (x_0, y_1), (x_1, y_0) следующий вид (рис. 16):

$$\varphi_1(x, y)=z_{00}+[x_0, x_1; y_0](x-x_0)+ \\ +[x_0; y_0, y_1](y-y_0), \qquad (5.33)$$

где разделенные разности представляют собой

$$[x_0, x_1; y_0] = (z_{10} - z_{00})/(x_1 - x_0),$$
$$[x_0; y_0, y_1] = (z_{01} - z_{00})/(y_1 - y_0). \qquad (5.34)$$

Если привлечь еще три узла (x_2, y_0), (x_1, y_1), (x_0, y_2), то $k=2$ и соответствующий квадратичный полином $\varphi_2(x, y)$ получается добавлением к $\varphi_1(x, y)$ еще трех слагаемых вида

$$[x_0; y_0, y_1, y_2](y - y_0)(y - y_1) +$$
$$+ [x_0, x_1; y_0, y_1](x - x_0)(y - y_0) +$$
$$+ [x_0, x_1, x_2; y_0](x - x_0)(x - x_1). \qquad (5.35)$$

Хорошие результаты получаются с помощью последовательной интерполяции кубическими сплайнами. Построение результирующей кусочно-бикубической функции подробно описано в [12]. В результате получается кусочная функция $\varphi(x, y)$, представляющая собой в каждой ячейке бикубический полином двух переменных и удовлетворяющая на границе области краевому условию: вторая производная по внешней нормали равна нулю. Необходимый алгоритм подробно описан в работах [3], [12].

Случай хаотической сетки. Коллокация как обобщение интерполяции. Более сложными, но зато и более распространенными на практике являются задачи интерполяции с хаотичной сетки. Обобщим рассмотренные выше задачи интерполяции. Более абстрактный подход к решению позволит получать решения более широкого, чем интерполяция, класса задач и приведет к методу обработки разнородных измерений, чрезвычайно эффективному с точки зрения геодезических приложений.

Пусть по-прежнему изучаемая функция $f(P)$ принадлежит некоторому множеству H функций, обладающих в области D определенной гладкостью. На H заданы $n+1$ линейных линейно независимых функционалов $L_1, L_2, \ldots, L_n, F$ и известны значения первых n из этих функционалов на f, т. е. известны числа $l_i = L_i f$, где $i = 1, \ldots, n$. Рассмотрим следующие задачи:

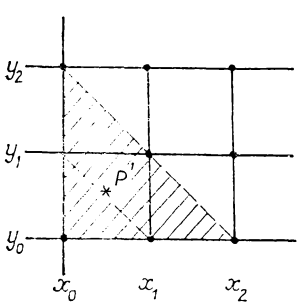


Рис. 16. Треугольная конфигурация узлов для построения интерполяционного полинома минимальной степени

1) локальной коллокации, где надо указать способ, с помощью которого можно было бы приписать значению Ff некоторое «разумное» приближенное число;

2) глобальной коллокации, в которой требуется восстановить функцию f, т. е. найти такую аналитически заданную функцию φ, которая

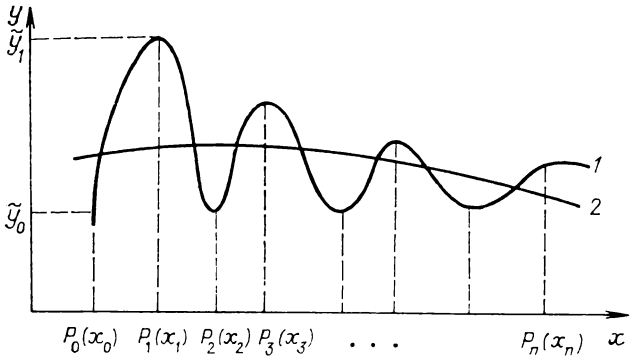


Рис. 17. Интерполяционная (*1*) и аппроксимирующая (*2*) кривые

приближала бы f на D, а значения функционалов L_i на φ совпадали бы с заданными числами l_i, т. е.

$$L_i\varphi = L_i f \equiv l_i;\ i=1, \ldots, n. \qquad (5.36)$$

Задание на H функционала L означает задание определенного правила, по которому каждой функции f из H можно поставить в соответствие единственное действительное число Lf, например $\int\limits_D f(P)dD$ или $f(P_0)$, где P_0 — фиксированная точка в D. Функционал последнего типа называется *дельта-функционалом* и обозначается $\delta_{P_0}(f)$. Понятия линейности и линейной независимости определяются стандартным образом [22].

Замечание 5.5. Если участвующие в постановке задач функционалы $L_1, \ldots, L_n, F$ есть дельта-функционалы $\delta_{P_1}, \ldots, \ldots, \delta_{P_n}, \delta_{P'}$, соответствующие узлам сетки $P_1, \ldots, P_n$ и промежуточной точке P' (рис. 17), то задачи коллокации называются задачами интерполяции. В частности, эти задачи совпадают с теми, которые рассмотрены в предыдущих параграфах главы (при $D=[a,b]$). Таким образом, интерполяция есть частный случай коллокации.

Пример 5.5. Пусть H — множество всех функций, непрерывных на заданном отрезке $D=[0,1]$. О функции $f(x)\in H$ известно, что $f(0)=1$, $f'(1/2)=5$, $f''(1)=6$.

Требуется: 1) приближенно оценить значение $\int\limits_0^1 f(x)dx$; 2) восстановить функцию $f(x)$ на $[0,1]$.

Пример 5.6. Пусть D — заданный район земной поверхности, который допустимо считать частью плоскости. Даны точки P_1, P_2, P_3, $P'\in D$ и результаты измерений силы тяжести в точках P_1, P_2 и астрономической широты в точке P_3. Требуется: 1) вычислить высоту квазигеоида в точке P'; 2) восстановить в области D возмущающий потенциал.

Числа l_1, l_2, l_3 получаются вычитанием из результатов измерений соответствующих модельных значений, обусловленных нормальным гравитационным полем Земли, H — сужение на D множества функций, гармонических вне Земли и регулярных на бесконечности. Степень гладкости функций из

H обычно известна из предварительной статистической обработки различных характеристик реального гравитационного поля. Вид функционалов на геопотенциале, определяющих понятия силы тяжести, астрономической широты, высоты квазигеоида представлен в [14].

Решение глобальной задачи коллокации сводится к решению системы n уравнений (5.36), которую запишем в виде

$$Lf=l, \tag{5.37}$$

где L — столбец n исходных функционалов L_i, l — столбец n заданных чисел l_i.

Система уравнений (5.37) имеет тот же смысл, что и система параметрических уравнений связи в ТМОГИ. Однако имеется и существенное отличие. В классической теории уравнивания в роли неизвестного всегда выступает некоторый конечномерный вектор (например, набор координат пунктов или поправок к приближенным координатам). Теперь же неизвестной является функция, представляющая собой в общем случае элемент бесконечномерного пространства. Разумеется, если дополнительно известно, что H представляет собой множество полиномов, степень которых не превышает заданную m, то H конечномерно и всякий элемент из него описывается не более чем $m+1$ числами — коэффициентами полинома.

Упражнение 5.1. Решить задачи, сформулированные в примере 5.5 при условии, что H — множество полиномов, степень которых не превышает 2.

Ответ: решение единственное, $\varphi(x)=f(x)=1+2x+3x^2$; $F(f)=3$.

Однако произвольную функцию нельзя описать конечным набором параметров. Каким бы большим, но конечным ни было число n, количество уравнений в (5.37) будет меньше, чем число неизвестных. Поэтому система (5.37) имеет бесчисленное множество решений. Обозначим это множество через $L^{-}l$. Всякая функция из него решает задачи коллокации, и для выделения какого-то одного решения необходима дополнительная информация. В предыдущих параграфах эта информация состояла в указании конкретного класса функций, из которого выбирались решения (полином заданной степени, кубический сплайн). Теперь мы распорядимся имеющейся неопределенностью решения более целенаправленно. На основании физического смысла задачи коллокации обычно удается назвать некоторый функционал $\Phi\varphi$, $\varphi\in H$, характеризующий меру качества коллокации: чем меньше значение этого функционала на элементе $\varphi\in L^{-1}l$, тем этот элемент предпочтительнее. Естественно поэтому из всевозможных функций φ, удовлетворяющих условиям коллокации (5.36), выбрать такую $\hat{\varphi}$, которая обеспечивает минимум функционалу качества

$$\hat{\varphi}(P)=\arg\min_{\varphi\in L^{-}l}\Phi\varphi\approx f(P). \tag{5.38}$$

Требования (5.36), (5.38) определяют общий принцип решения глобальных задач коллокации, в частности интерполяции (см. замечание 5.5): надо найти в H такую функцию $\hat{\varphi}(P)$, на которой значения исходных функционалов $L_1, \ldots, L_n$ совпадают с заданными числами $l_1, \ldots, l_n$, а значение функционала качества минимально, т. е.

$$L_i\hat{\varphi} = L_i f \equiv l_i, \quad i = 1, \ldots, n, \ \Phi\hat{\varphi} = \min. \tag{5.39}$$

Функция $\hat{\varphi}(P)$, удовлетворяющая условиям (5.39), называется *коллокационным (интерполяционным) сплайном.*

Если глобальная задача решена, то решение локальной задачи получается совсем просто:

$$Ff \approx F\hat{\varphi}. \tag{5.40}$$

Вид функционала качества должен назначаться из инженерных соображений, основанных на понимании прикладного значения решаемой задачи. Однако вид Φ играет важную роль и с математической точки зрения, поскольку решение должно быть единственным. Укажем два вида наиболее простых функционалов качества, для которых доказаны существование и единственность решения задач коллокации и которые пригодны для решения широкого круга двумерных геодезических задач:

$$\Phi\varphi = \iint_D \left[\left(\frac{\partial^2\varphi}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2\varphi}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2\varphi}{\partial y^2}\right)^2\right] dD, \tag{5.41}$$

$$\Phi\varphi = \iint_D \left[(\nu\varphi)^2 + \left(\frac{\partial\varphi}{\partial x}\right)^2 + \left(\frac{\partial\varphi}{\partial y}\right)^2\right] dD, \tag{5.42}$$

где $\nu > 0$ — числовой параметр.

Наиболее полно и просто задачи коллокации формулируются в терминах теории гильбертовых пространств. (Для понимания требуются лишь простейшие понятия этой теории, с которыми можно познакомиться в работах [14] и [22].)

Пусть H_1, H_2 — два гильбертовых пространства, $T : H_2 \to H_2$ — линейный ограниченный оператор, действующий из H_1 в H_2. Над пространством H_1 задана линейно независимая система линейных ограниченных функционалов L_i $(i = 1, \ldots, n)$. В качестве решения глобальной задачи коллокации (интерполяции, см. замечание 5.5) выбирается такой элемент $\hat{\varphi} \in H_1$, который удовлетворяет двум условиям:

$$L_i\hat{\varphi} \equiv (L_i^*, \hat{\varphi})_{H_1} = L_i f \equiv l_i, \ i = 1, \ldots, n;$$

$$\Phi\hat{\varphi} = (T\hat{\varphi}, T\hat{\varphi})_{H_2} = \| T\hat{\varphi} \|^2_{H_2} = \min, \tag{5.43}$$

где символы $(L_i^*, \hat{\varphi})_{H_1}$, $(T\hat{\varphi}, T\hat{\varphi})_{H_2}$ обозначают скалярные произведения в пространствах H_1, H_2 соответственно; l_i — наперед заданные числа, $i = 1, \ldots, n$; L_i^* — элементы пространства H_1, которые являются представителями функ-

ционала L_i в смысле теоремы Рисса о представлении линейного функционала в гильбертовом пространстве. Функция $\hat{\varphi}$, удовлетворяющая условиям (5.43), называется коллокационным (интерполяционным) сплайном.

В общем случае оператор T может иметь ненулевое ядро. Это означает, что величина $||T\varphi||_{H_2}$ является полунормой в пространстве H_1. Значит, $\hat{\varphi}$ — такое решение уравнения (5.37), которому соответствует минимальная полунорма. Решения подобного рода называют *T-нормальными* [20]. Формулировки и доказательства теорем о существовании и единственности коллокационных сплайнов в зависимости от H_1, H_2, T приведены в [3]. В качестве H_1 и H_2 обычно используются пространства Соболева $W_2^q(D)$. Они являются гильбертовыми и состоят из функций пространства $L_2(D)$, которые имеют в D суммируемые с квадратом обобщенные производные до порядка q включительно, причем $W_2^0=L_2$. Скалярное произведение в таком пространстве определяется формулой

$$(u, v)_{W_2^q} = \int_D \sum_{|\alpha|=0}^{q} \sum_{(\alpha)} \frac{\partial^{|\alpha|} u}{\partial x^{\alpha_1} \partial y^{\alpha_2}} \frac{\partial^{|\alpha|} v}{\partial x^{\alpha_1} \partial y^{\alpha_2}} dD, \qquad (5.44)$$

где $|\alpha|=\alpha_1+\alpha_2$, $\alpha_1 \geqslant 0$, $\alpha_2 \geqslant 0$, а $\sum_{|\alpha|}$ обозначает суммирование по всем производным порядка $|\alpha|$.

В качестве оператора T обычно используется дифференциальный p-го порядка оператор вида

$$Tu = \nabla^p u = \left\{ \sqrt{\frac{p!}{\alpha_1!\,\alpha_2!}} \frac{\partial^p u}{\partial x^{\alpha_1} \partial y^{\alpha_2}} \right\}, \qquad (5.45)$$

$\alpha_1+\alpha_2=|\alpha|=p$, осуществляющий отображение $H_1=W_2^q$, $q \geqslant p$ в пространство обобщенных градиентов $H_2=\prod_{i=1}^{\beta} L_2(D)$, где β — количество различных пар, $\alpha_1 \geqslant 0$, $\alpha_2 \geqslant 0$, для которых $\alpha_1+\alpha_2=p$. При этом

$$\|Tu\|^2_{H_2} = \int_D \frac{p!}{\alpha_1!\,\alpha_2!} \left(\frac{\partial^p u}{\partial x^{\alpha_1} \partial y^{\alpha_2}} \right)^2 dD. \qquad (5.46)$$

Если $p=0$, то $T=E$ — единичный оператор, $H_1=H_2=W_2^q$.

В частности, функционал (5.41) получается по формулам (5.43), (5.45), (5.46) при $p=q=2$. Прообразом оператора T служат элементы $W_2^2(D)$, а образом является тройка функций

$$\partial^2\varphi/\partial x^2,\ \sqrt{2}\partial^2 u/\partial x \partial y,\ \partial^2 u/\partial y^2.$$

Функционал (5.46) с $\nu=1$ получается по тем же правилам при $p=0$, $q=1$, $T=E$, $H_1=H_2=W_2'$. Это справедливо и для функций с количеством переменных, отличным от двух. Так, для функций одной переменной $f(x)$ с $D=[x_0, x_n]$ задача коллокации (5.47) с $L_i f=f(x_i)$, $i=1, \ldots, n$ и оператором p-кратного дифференцирования T, действующего из W_2^3 в L_2, совпадает с интерполяционной задачей построения полиномиального сплайна $(p+1)$-й степени под условием (5.21) (см. замечания 5.4, 5.5). При $p=2$ решением задачи служит кубический сплайн.

Теорема 5.7. Для существования и единственности решения двумерной (т. е. на плоскости) задачи глобальной коллокации (5.39), (5.41), (5.42) необходимо и достаточно, чтобы среди точек $p_1, \ldots, p_n$ нашлись три, не лежащие на одной прямой. Доказательство приведено в [3].

Коллокационный (интерполяционный) сплайн минимальной нормы. Обсудим вопросы построения коллокационного (интер-

поляционного) сплайна. Сначала рассмотрим задачу (5.39), (5.42). Вид функционала (5.42) связан с тем множеством H, среди функций которого предполагается искать решение. В данном случае множество H удобно трактовать как гильбертово пространство $W_2^1(D)$ [см. выражение (5.44)] с $q=1$. $W_2^1(D)$ достаточно понимать как бесконечномерное евклидово пространство дифференцируемых в D функций двух переменных со скалярным произведением

$$(u, v)=\int_D \left(\nu^2 uv+\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\frac{\partial v}{\partial y}\right)dD,\ \nu\neq 0. \qquad (5.47)$$

При этом функционал (5.42) является квадратом нормы $\|\varphi\|^2$ в этом пространстве. Следовательно, коллокационный сплайн $\hat{\varphi}$, удовлетворяющий условиям (5.39), является решением системы (5.37), имеющим в пространстве $H=W_2^1(D)$ минимальную норму. Решение минимальной нормы называется *нормальным решением*. Таким образом, коллокационным сплайном, разрешающим в H глобальную задачу коллокации, является нормальное (по метрике H) решение системы уравнений (5.37).

Линейная система уравнений (5.37), как уже отмечалось, имеет бесчисленное множество решений. Это равносильно тому, что соответствующая приведенная (однородная) система $L\varphi=0$ имеет не только нулевые решения, так же как в случае системы линейных алгебраических уравнений. Множество всех таких решений называется *ядром оператора* L и обозначается $\mathrm{Ker}\,L$. Линейные функционалы $L_1, \ldots, L_n$, составляющие оператор L, не чувствительны к функциям из ядра, т. е. добавление к какому-либо решению системы (5.37) любой функции из $\mathrm{Ker}\,L$ не отражается на исходных числах $l_1, \ldots, l_n$. Нормальное решение принадлежит так называемому ортогональному дополнению к $\mathrm{Ker}\,L$, которое обозначается $\mathrm{Ker}^{\perp}L$ и заполнено такими функциями из H, что каждая из них ортогональна любой функции из $\mathrm{Ker}\,L$. Множество $\mathrm{Ker}^{\perp}L$ представляет собой n-мерное подпространство пространства H. Базисные элементы этого подпространства будем обозначать L_i^* и называть их *представителями исходных функционалов* L_i, $i=1, \ldots, n$.

Строгое доказательство указанных фактов приведено в работах [14], [16]. Всякому линейному ограниченному на H функционалу F можно поставить в соответствие единственную функцию $F^*(P)$, $P\in D$ из H такую, что значение F на любой функции φ из H можно вычислить как скалярное произведение в H функций $\varphi(P)$ и $F^*(P)$, т. е.

$$F\varphi=(F^*, \varphi). \qquad (5.48)$$

Такая функция $F^*(P)$ и называется представителем функционала F. Ниже будет указан простой способ нахождения представителя по виду функционала.

Итак, нет необходимости искать нормальное решение во всем бесконечномерном пространстве H, достаточно найти его в n-мерном подпространстве $\mathrm{Ker}^{\perp}L$. Существуют такие числа $a_1, \ldots, a_n$, что

$$\hat{\varphi}[P]=a_1L_1^*(P)+\ldots+a_nL_n^*(P), \tag{5.49}$$

и задача сводится к нахождению этих чисел, являющихся координатами коллокационного (интерполяционного) сплайна наименьшей нормы в базисе ортогонального дополнения к ядру исходных функционалов.

Подействуем на обе части равенства (5.49) функционалом L_i, $i=1, \ldots, n$ и, учитывая (5.36), получим

$$a_1L_iL_1^*+a_2L_iL_2^*+\ldots+a_nL_iL_n^*=l_i. \tag{5.50}$$

Если предположить, что представители функционалов известны, то числа $L_iL_j^*$, $i, j=1, \ldots, n$ легко вычисляются. Поэтому совокупность уравнений (5.50) представляет собой систему n линейных алгебраических уравнений с n неизвестными $a_1, \ldots, \ldots, a_n$.

Запишем эту систему в матричном виде

$$\underset{n\times n}{G}\,\underset{n\times 1}{a}=\underset{n\times 1}{l}, \tag{5.51}$$

где $G=g_{ij}$, а $g_{ij}=L_iL_j^*$, $i=1, \ldots, n$, $j=1, \ldots, n$, $a^{\mathrm{T}}=(a_1, \ldots, a_n)$. Можно доказать, что матрица G симметрична и положительно определена. Поэтому система (5.50) имеет единственное решение. Для нахождения его в общем случае удобно пользоваться методом квадратного корня. Подставляя a в (5.49), получаем окончательное решение глобальной задачи коллокации

$$\begin{aligned}\hat{\varphi}(P)&=\sum_{i=1}^{n}a_iL_i^*(P)=\\ &=\underset{1\times n}{L^*(P)}\,\underset{n\times n}{G^{-1}}\,\underset{n\times 1}{l}=\underset{n\times 1}{\Lambda(\mathrm{P})l}=\\ &=\Lambda(P)Lf=\Pi f,\end{aligned} \tag{5.52}$$

где представители $L_i^*(P)$, $i=1, \ldots, n$ являются некоторыми заранее определенными функциями двух переменных, $L^*(P)$ — строка представителей, $\Lambda(P)=L^*(P)\,G^{-1}$.

Функции $\Lambda_1(P), \ldots, \Lambda_n(P)$, составляющие строку $\Lambda(P)$, являются новым базисом в $\mathrm{Ker}^{\perp}L\subset H$. Его спецификой являются следующие равенства:

$$L_i\Lambda_j(P)=\begin{cases}1, & \text{если } i=j;\\ 0, & \text{если } i\neq j,\end{cases} \tag{5.53}$$

так что базисный элемент $\Lambda_i(P)$ — функция влияния функционала L_i. Базис со свойствами (5.53) называется *базисом Лагранжа*, соответствующим функционалам $L_1, \ldots, L_n$ на множе-

стве H. Мы уже имели дело с подобным базисом в условиях, когда все исходные функционалы есть дельта-функционалы (частный случай коллокации — интерполяция, см. замечание 5.5), а H — множество полиномов, степень которых не превышает заданную [см. формулы (5.13)—(5.15)].

Если L^- в (5.38) трактовать как оператор $E_n \to W_2^q$, псевдообратный к оператору L системы (5.37), то $\Lambda(P)$ является главным псевдообратным оператором L^+, поскольку он переводит исходный числовой столбец l в решение $\hat{\varphi}$ минимальной нормы в пространстве $H=W_2^q : \hat{\varphi}=L^+l$.

Оператор $\Pi=\Lambda(P)\ L=L^+L$ является проектором из пространства W_2^q в его подпространство $\mathrm{Ker}^\perp L$. Легко проверить непосредственно, что $\Pi\times\Pi=\Pi^2=\Pi$. Это означает, что на $\mathrm{Ker}^\perp L$ оператор Π является тождественным: если изучаемая функция f принадлежит $\mathrm{Ker}^\perp L$, то нормальное (главное псевдообратное) решение $\hat{\varphi}$ в точности совпадает с истинным решением f. При этом решение локальной задачи окажется безошибочным, т. е. $\mathrm{Ker}(F-\hat{F})=\mathrm{Ker}^\perp L$ для $\forall F\in H^*$.

Итак, решение глобальной задачи коллокации (интерполяции) рекомендуется выполнять в следующем порядке.

1. Составить функции $L_1^*(P),\ldots, L_n^*(P)$, являющиеся представителями исходных функционалов в избранном пространстве H.

2. Заполнить квадратную матрицу $\underset{n\times n}{G}$ элементами $g_{ij}=L_iL_j^*(P)$, $i=1, \ldots, n$, $j=1, \ldots, n$.

3. Решить систему (5.51) методом квадратного корня.

4. Решение глобальной задачи записать в виде первого равенства (5.52).

5. Решение локальной задачи имеет вид (5.40).

Следующие две теоремы раскрывают оптимальные свойства рекомендованного глобального решения $\hat{\varphi}$ минимальной нормы с точки зрения решения локальной задачи (5.40).

Теорема 5.8. [7]. Выберем какое-нибудь положительное число $C\geqslant\|\hat{\varphi}\|$ и на каждом решении $\varphi\in L^-l$, удовлетворяющем условию $\|\varphi\|\leqslant C$, вычислим значение любого линейного ограниченного функционала F. Множество всех полученных чисел образует отрезок на числовой оси. Длина этого отрезка зависит от вида F и величины выбранной константы C, но серединой отрезка всегда служит значение F именно на $\hat{\varphi}$.

Прежде чем сформулировать следующую теорему, вспомним, что линейное пространство может быть заполнено различными математическими элементами. Пусть H^* — множество всех линейных функционалов, которые на любой функции φ из H ограничены. Легко убедиться, что такое множество является линейным пространством. Ограниченность функционала F на H означает существование такой константы $C>0$, что $F\varphi\leqslant C\cdot\|\varphi\|$ для всякой функции φ из H. Наименьшая такая константа называется *нормой функционала* и обозначается $\|F\|$. Из определения

представителя (5.48) и известного неравенства Коши — Буняковского следует, что

$$\| F \|_{H^*} = \| F^* \|_H. \tag{5.54}$$

Нижний индекс указывает, в каком пространстве считается норма. Таким образом, H^* — нормированное пространство и мера близости двух функционалов F_1 и F_2 из H^* определяется числом $\|F_1 - F_2\|$. Норма (5.54) позволяет ввести скалярное произведение в H^* по правилу

$$(F_1, F_2)_{H^*} = (F_1^*, F_2^*)_H. \tag{5.55}$$

Теорема 5.9. Определим приближенный функционал $\hat{F} \approx F$, разрешающий локальную задачу коллокации (5.44), равенством $\hat{F}(f) = F(\hat{\varphi})$.
Тогда

$$\| F - \hat{F} \| = \min \tag{5.56}$$

по сравнению с любыми другими функционалами F из n-мерного подпространства в H^*, натянутого на исходные линейно независимые функционалы $L_1, \ldots, L_n$. Другими словами, $\hat{F}$ является элементом наилучшего линейного приближения [22] к F по метрике H^* из заданного подпространства.

Теорема 5.10. Обозначим $F(P) - \hat{\varphi}(P) = v(P) \in \operatorname{Ker} L$ в силу (5.39). Справедливо равенство (обобщенная теорема Пифагора)

$$\| \hat{\varphi} \|^2 + \| v \|^2 = \| f \|^2.$$

при этом

$$\| v \| = \| f - \hat{\varphi} \| = \min \tag{5.57}$$

по сравнению с любыми другими функциями из n-мерного подпространства $\operatorname{Ker}^{\perp} L \subset H$. Таким образом, $\hat{\varphi}$ является элементом наилучшего линейного приближения [22] к f по метрике H из заданного подпространства (см. § 6.1).

Заметим, что все другие функции из $\operatorname{Ker}^{\perp} L$, отличные от $\hat{\varphi}$, не удовлетворяют условиям коллокации (5.39). Доказательство можно найти в работе [14].

Построение нормального решения задач коллокации (интерполяции) с помощью воспроизводящего ядра. Переходим теперь к вопросам практической реализации схемы вычислений, которая указана выше.

Основная трудность заключается в выполнении 1-го пункта. Однако все упрощается, если для пространства H, которому принадлежит изучаемая функция $f(P)$, $P \in D$, известно так называемое воспроизводящее ядро. *Воспроизводящим ядром* называется такая симметричная функция $Я(P, Q)$ двух точек $P, Q \in D$, которая обладает двумя свойствами:

а) если Q — любая фиксированная точка D, то $Я(P, Q)$ как функция точки P принадлежит H;

б) значение любой функции $\varphi \in H$ в произвольной точке Q имеет вид

$$\varphi(Q) = (\varphi(P),\ Я(P, Q)), \tag{5.58}$$

что выражает воспроизводящее свойство ядра.

Здесь при вычислении скалярного произведения по формуле (5.47) текущей точкой интегрирования служит P.

Пример 5.7. Пусть $D=[-1,1]$, а H — множество всех полиномов на D, степень которых не превышает единицу. Введем на H скалярное произведение по правилу $(u, \tilde{v}) = \int_{-1}^{1} u(x)v(x)dx$. Таким образом, H — двумерное евклидово пространство, а воспроизводящим ядром является $Я(x, \xi) = 1/2 + (3/2)x\xi$, где $x, \xi \in [-1,1]$. Проверьте это по определению.

Пример 5.8. Пусть $D=[-\pi, \pi]$, а H — множество всех функций, дифференцируемых на D. Введем скалярное произведение по правилу $(u, v) = \int_{-\pi}^{\pi} (\nu^2 u(x) \times v(x) + u'(x)v'(x))dx$, где $\nu \neq 0$. Таким образом, H — бесконечномерное евклидово пространство $W_2^1([-\pi, \pi])$. Воспроизводящее ядро существует и равно [7]

$$Я(x, \xi) = \begin{cases} \dfrac{\operatorname{ch}\nu(\pi - \xi)\operatorname{ch}\nu(\pi + x)}{\nu \operatorname{sh}(2\pi\nu)} & \text{при } -\pi \leqslant x \leqslant \xi \leqslant \pi; \\ \dfrac{\operatorname{ch}\nu(\pi + \xi)\operatorname{ch}\nu(\pi - x)}{\nu \operatorname{sh}(2\pi\nu)} & \text{при } -\pi \leqslant \xi \leqslant x \leqslant \pi, \end{cases}$$

где $x, \xi \in [-\pi, \pi]$.

Пример 5.9. Пусть D — квадратная область $-\pi \leqslant x \leqslant \pi$, $-\pi \leqslant y \leqslant \pi$, $H = W_2'(D)$ со скалярным произведением (5.47). Воспроизводящее ядро существует и равно [7]

$$Я(P, Q) = Я(x, y; \xi, \eta) = \frac{1}{4\pi^2\nu^2} + \frac{1}{2\pi^2} \times$$

$$\times \sum_{k=1}^{\infty} \frac{\cos\frac{1}{2}k(x - \pi)\cos\frac{1}{2}k(\xi - \pi)}{\nu^2 + k^2/4} +$$

$$+ \frac{1}{2\pi^2} \sum_{m=1}^{\infty} \frac{\cos\frac{1}{2}m(y - \pi)\cos\frac{1}{2}m(\eta - \pi)}{\nu^2 + m^2/4} +$$

$$+ \frac{1}{\pi^2} \sum_{k=1}^{\infty}\sum_{m=1}^{\infty} \frac{\cos\frac{1}{2}k(x - \pi)\cos\frac{1}{2}k(\xi - \pi)\cos\frac{1}{2}m(y - \pi)\cos\frac{1}{2}m(\eta - \pi)}{\nu^2 + k^2/4 + m^2/4}.$$

Замечание 5.6. С помощью масштабного множителя перед переменными отрезок $[-\pi, \pi]$ в примере 5.8 и квадрат $[-\pi, \pi; -\pi, \pi]$ в примере 5.9 легко преобразовать к любому дру-

гому отрезку числовой оси и соответственно к любому другому прямоугольнику, стороны которого параллельны координатным осям плоскости x, y.

Следующие теоремы раскрывают основные свойства воспроизводящего ядра и его роль в решении задач коллокации.

Теорема 5.11. Для того чтобы гильбертово пространство H имело воспроизводящее ядро, необходимо и достаточно, чтобы $\varphi(P)$ было ограничено для всякой функции φ из H и для всякой точки P из D.

Теорема 5.12. Если в H существует воспроизводящее ядро, то оно единственно и связано с элементами $e_i(P)$, $i=1, 2, \ldots$ ортонормированного базиса пространства H формулой

$$Я(P, Q)=\sum_{i=1}^{\infty} e_i(P)e_i(Q),$$

где $P, Q \in D$.

Упражнение 5.2. Пользуясь последней формулой, проверьте результат примера 5.7.

Теорема 5.13. Представителем $L_i^*(P)$ линейного ограниченного функционала L_i, $i=1, \ldots, n$ в пространстве H с воспроизводящим ядром $Я(P, Q)$ служит функция $L_i Я(P, Q)$, полученная действием функционала L_i на ядро по одной из точек (результат действия, например, по точке Q удобно обозначать $Я(P, L_i)$). Другими словами, $L_i^*(P)=Я(P, L_i)$. В силу симметрии ядра, функции $Я(L_i, Q)$ и $Я(P, L_i)$ — одно и то же. Доказательство приведенных теорем и другие подробности использования воспроизводящих ядер можно найти в работах [14], [16].

Пример 5.10. Зафиксируем на отрезке $[-1,1]$ точку $\tilde{x}=1/2$ и рассмотрим дельта-функционал $\delta_{1/2}\varphi=\varphi(1/2)$ в условиях примера 5.7. Его представителем служит полином $\delta_{1/2}Я(x, \xi)=Я(x, 1/2)=1/2+3x/4=\delta^*_{1/2}(x)$.

Упражнение 5.3. По данным примера 5.7 найти представителя функционала $L\tilde{u}=\int_{-1}^{1} u(x)dx$.

Ответ: $L^*(x)\equiv 1$.

Итак, если воспроизводящее ядро известно, то теорема 5.13 полностью решает вопрос о нахождении в H представителей $L_i^*(P)$ исходных функционалов L_i, $i=1, \ldots, n$. Следует отметить, что если скалярное произведение (5.51) задано, то найти воспроизводящее ядро в случае области D произвольной формы довольно затруднительно. Поскольку скалярное произведение (5.51) задано априори, используют стандартную область в виде прямоугольника (см. результаты примера 5.9 с учетом замечания 5.6). На практике часто поступают наоборот: сначала задаются удобным для вычислений воспроизводящим ядром. Примером служат следующие положительно определенные функции

двух точек P и Q плоскости, зависящие только от расстояния r между этими точками:

$$Я(r)=D\left(1+\frac{r}{\mu}-\frac{r^2}{2\mu^2}\right)\exp\left(-\frac{r}{\mu}\right), \tag{5.59}$$

$$Я(r)=D(1+r^2/\mu^2)^{-m}. \tag{5.60}$$

В этом выражении D и μ — свободные параметры, причем $D>0$, $m=1/2{,}1$ или $3/2$.

Положительная определенность ядра обязательна и состоит в том, что соответствующее двумерное преобразование Фурье должно быть неотрицательным. Ядра, зависящие только от расстояния между точками, называются *изотропными*. Для них упомянутое преобразование имеет вид:

$$\overline{Я}(\omega)=\int_0^\infty Я(r)I_0(\omega r)r\,dr,\ \omega\geqslant 0, \tag{5.61}$$

и называется *преобразованием Ганкеля*. Здесь $I_0(x)$ — функция Бесселя нулевого порядка.

Ядро (5.60) удобно для решения коллокационных задач физической геодезии (см. пример 5.6), поскольку при $m=1/2$ или $3/2$ оно легко продолжается в верхнее полупространство $z>0$ функциями соответственно

$$Я(P,Q)=\frac{D\mu}{(r^2+a^2)^{1/2}}$$

или

$$Я(P,Q)=\frac{Da\mu^2}{(r^2+a^2)^{3/2}},$$

гармоническими по каждой точке. Здесь $a=z(P)+z(Q)+\mu$, $z(P)$ — аппликата точки P.

Если определить H как пространство, для которого выбранная положительно определенная симметричная функция двух точек служит воспроизводящим ядром, то вид соответствующего скалярного произведения остается, как правило, неизвестным*. Но для вычислений он и не нужен. В самом деле, согласно схеме вычислений и теореме 5.13, решение глобальной задачи коллокации имеет вид:

$$\underset{1\times 1}{f(P)}\approx\underset{1\times 1}{\hat{\varphi}(P)}=\underset{1\times n}{Я(P,L)}\,\underset{n\times n}{Я^{-1}(L,\ L)}\,\underset{n\times 1}{l}. \tag{5.62}$$

Для решения локальной задачи нет необходимости получать

* Величина q в (5.44) устанавливается в зависимости от скорости убывания функции типа (5.61) с увеличением ω.

$\hat{\varphi}(P)$ в явном виде и только затем пользоваться формулой (5.40). Удобнее сразу воспользоваться формулой

$$Ff \approx F\hat{\varphi} = \underset{1\times n}{Я(F,\ L)}\ \underset{n\times n}{Я^{-1}(L,\ L)}\ \underset{n\times 1}{l}. \tag{5.63}$$

Здесь $Я(P, L) = L^*(P)$ — строка n функций $Я(P, L_i)$, полученных действием исходного функционала L_i, $i=1, \ldots, n$ на ядро $Я(P, Q)$ по точке Q; $Я(F, L)$ — строка n чисел, полученных действием функционала F на функции $Я(P, L_i)$; $Я(L, L) = G$ — квадратная размера $[n\times n]$ матрица, (i, j)-позицию в которой занимает число $g_{ij} = Я(L_i, L_j)$, полученное действием на ядро функционалом L_i по точке P и функционалом L_j по точке Q.

Упражнение 5.4. Пусть изучаемая функция $y=f(x)$ принадлежит пространству $H=W_2^1[-\pi, \pi]$, описанному в примере 5.8. Исходными числами служат пять значений $l_i=L_if=f(x_i)$, где $x_i=-\pi+i(\pi/2)$, $i=0, 1, 2, 3, 4$. Пользуясь теоремой 5.13, найдите представителей $L_1^*(x)$, $L_2^*(x)$ функционалов $L_1=f(-\pi/2)$, $L_2=f(0)$ и убедитесь, что на пересечении второй строки и третьего столбца матрицы G стоит число $g_{2,3}=(L_1, L_2)^*=Я(L_1, L_2)=Я(-\pi/2, 0)=\mathrm{ch}(\nu\pi/2)\cdot\mathrm{ch}(\nu\pi)/(\nu\cdot\mathrm{sh}(2\nu\pi))$. Действуя аналогично, сформируйте матрицу $G=Я(L, L)$ размера [5×5] и убедитесь, что ее обратная матрица $G^{-1}=Я^{-1}(L, L)$ будет трехдиагональной. Позиции 1.1 и 5.5 ее главной диагонали занимают числа $\mathrm{ch}(\nu\pi/2)$, остальные элементы главной диагонали в 2 раза больше. Все элементы двух соседних диагоналей равны -1.

Упражнение 5.5. По данным упражнения 5.4 оцените значение функционала $Ff=\int_{-\pi}^{\pi} f(x)dx$, пользуясь алгоритмом (5.63).

Ответ: $Я(F, L)$ представляет собой строку из пяти одинаковых чисел ν^{-2}; $\hat{F}f=2\nu^{-1}\mathrm{th}(\nu\pi/4)\times[(l_0/2)+l_1+l_2+l_3+(l_4/2)]$.

Точность можно оценить только в терминах воспроизводящего ядра. Необходимая для этого формула имеет вид

$$\|F-\hat{F}\|^2 = \underset{1\times 1}{Я(F, F)} - \underset{1\times n}{Я^{\mathrm{T}}(L,\ F)}\ \underset{n\times n}{Я^{-1}(L,\ L)}\ \underset{n\times 1}{Я(F,\ L)}. \tag{5.64}$$

Левая часть этого выражения — стандартная мера близости приближенного функционала $\hat{F}$, определенного равенством в теореме 5.9 и разрешающего локальную задачу коллокации относительно участвующего в ней функционала F; $Я(F, F)$ обозначает число, полученное действием функционала F на ядро $Я(P, Q)$ дважды (один раз по точке P и один раз по точке Q).

Полезно отметить, что $Я(L_i, L_j)$ является скалярным произведением функционалов L_i и L_j в пространстве H^*:

$$(L_i,\ L_j) = Я(L_i,\ L_j), \tag{5.65}$$

Это следует из (5.55), определения представителя (5.48) и воспроизводящего свойства ядра (5.58). Поэтому G в (5.51) есть

матрица Грама для исходных функционалов в H^*. Следствием положительной определенности воспроизводящего ядра является положительная определенность квадратичной формы, вычитаемой в правой части равенства (5.64). Поэтому число $\|F-\hat{F}\|$, являющееся мерой точности решения локальной задачи в соответствии с метрикой избранного пространства H, обязательно уменьшается с увеличением n.

Функционал $\hat{F}$ позволяет приписать искомому числу $F(f)$ разумное приближение $\hat{F}f=F\hat{\varphi}$. Число $\Delta=(F-\hat{F})f=F(f-\varphi)$ представляет собой истинную ошибку аппроксимации (если $F=L_i$, $i=1,\ldots,n$, то $\Delta=0$). Следующее неравенство устанавливает предельную ошибку такой аппроксимации:

$$|\Delta| \leqslant \|F-\hat{F}\|\,(\|f\|^2-\|\hat{\varphi}\|^2)^{1/2}, \tag{5.66}$$

где первый сомножитель правой части определяется формулой (5.64), норма нормального решения глобальной задачи коллокации определяется формулой

$$\underset{1\times n}{\|\hat{\varphi}\|^2}=\underset{1\times n}{l^{\mathrm{T}}}\underset{n\times n}{Я^{-1}(L,\ L)}\underset{n\times 1}{l}, \tag{5.67}$$

а норма интерполируемой функции должна оцениваться по дополнительной информации. Второй сомножитель правой части неравенства (5.66) является характеристикой точности решения глобальной задачи.

Обратите внимание на условия минимума (5.56) и (5.57) для обоих упомянутых сомножителей. Доказательства можно найти в работе [14].

Статистически оптимальная коллокация (интерполяция). Все характеристики точности решения задач коллокации представляют собой меру близости в терминах нормы того или иного пространства. Можно ли придать им привычный для геодезистов смысл средних квадратических ошибок и дисперсий? Можно, если в качестве воспроизводящего ядра $Я(P, Q)$ взять ковариационную функцию $K(P, Q)$, характеризующую статистическую структуру изучаемой детерминированной функции $f(P)$:

$$Я(P,\ Q) \equiv K(P,\ Q). \tag{5.68}$$

В основе такого подхода лежит ковариационный анализ неслучайных функций. Математическое ожидание M при этом определяется как усреднение не по ансамблю реализаций, как в теории случайных полей, а по рассматриваемой области D задания изучаемой функции

$$Mf=\frac{1}{S_D}\iint\limits_D f(x, y)\,dxdy, \tag{5.69}$$

где S_D — площадь D.

Если $Mf=0$, то говорят, что скалярное поле, определяемое в D функцией $f(P)$, $P\in D$, центрировано. В этом случае дисперсия Df поля определяется соотношением

$$Df=M(f^2). \tag{5.70}$$

Ковариационный момент, соответствующий точкам $P, Q\in D$, вводится как функция только расстояния r между этими точками по правилу

$$K(r)=M(f(P)f(Q)), \tag{5.71}$$

причем усреднение предполагается для всевозможных пар точек области D, удаленных на фиксированное расстояние r.

Практически обычно задаются последовательностью значений $r_i=i\Delta r$, $i=0, 1, 2,\ldots$ с некоторым шагом $\Delta r>0$. Для каждого r_i определяется эмпирический ковариационный момент $\widetilde{K}(r_i)$ как среднее произведение всех доступных значений изучаемой функции, удаленных друг от друга на расстояние $(i-1/2)\Delta r\leqslant r_i<(i+1/2)\Delta r$. Полученные эмпирические моменты сглаживаются подходящей положительно определенной функцией $K(r)$. Использование для этого функций вида (5.59), (5.60) состоит в подборе таких значений свободных параметров D и μ, при которых функция $K(r)$ наилучшим образом согласовывается с эмпирическими моментами $\widetilde{K}(r_i)$. Параметр D играет роль дисперсии изучаемого поля (функции $f(P)$), а μ однозначно связан с радиусом корреляции. Чем больше D и меньше μ, тем сложнее структура изучаемого поля с точки зрения задач коллокации (интерполяции).

Зависимость ковариации только от расстояния постулирует изотропность изучаемого поля. Известны и более сложные методы, позволяющие отыскивать ковариацию как функцию не только расстояния, но и азимута направления $\overrightarrow{PQ}$ [14], [18].

Полученная тем или иным методом ковариационная функция и выбирается в качестве воспроизводящего ядра. Можно доказать, что в условиях (5.68) скалярное произведение (5.55), (5.65) в H^* двух функционалов L_i, L_j приобретает смысл ковариации $\operatorname{cov}(l_i, l_j)$ между значениями l_i, l_j этих функционалов на изучаемой функции $f(P)$, т. е.

$$\begin{aligned}(L_i, L_j)&=\text{Я}(L_i, L_j)=\\&=\operatorname{cov}(l_i, l_j)=M(l_i, l_j).\end{aligned} \tag{5.72}$$

Поэтому, в частности,

$$\|L_i\|^2=\text{Я}(L_i, L_i)=\sigma_i^2=M(l_i^2), \tag{5.73}$$

где σ_i^2 — дисперсия величины l_i и матрица $\text{Я}(L, L)=G$ в (5.51) — ковариационная матрица исходных значений $l_1,\ldots, l_n$.

Мера точности $\|F-\hat{F}\|^2$, определяемая формулой (5.64), представляет собой дисперсию ошибки Δ, а $\|F-\hat{F}\|$ есть среднее квадратическое значение этой ошибки. Это делает особенно ценным указанное выше свойство оптимальности (5.56). Таким образом, описанный метод решения задач коллокации (интерполяции) в условиях (5.66) является статистически оптимальным и совпадает с известным в ковариационной теории случайных функций методом оптимального линейного прогноза [19]. Нельзя забывать, что смысл использованных понятий «дисперсия» и «средняя квадратическая ошибка» полностью определяется смыслом математического ожидания (5.69) и отличается от этих терминов в традиционной теории вероятностей применительно к случайным ошибкам измерений. В рассматриваемой ситуации эти ошибки отсутствуют, а дисперсия ошибки $\Delta(P)$ аппроксимации заданного дельта-функционала $\delta_P f \approx \hat{\varphi}(P)$ представляет собой средний квадрат функции ошибок $\Delta(P)$ по исходной области определения [14].

Решение задач коллокации с воспроизводящим ядром, которое отражает структуру ковариационной функции лишь приближенно, напоминает уравнивание с неточными весами — окончательные оценки остаются несмещенными, но дисперсия их возрастает.

Подведем итоги решения задач коллокации (интерполяции) в заданном пространстве H с помощью сплайна минимальной нормы.

Сплайн минимальной нормы есть нормальное решение системы функциональных уравнений (5.37).

Дана общая схема вычислений решения задач коллокации.

Для практической реализации этой схемы необходимо воспроизводящее ядро пространства H.

Если ядро известно, то решение глобальной задачи дает выражение (5.62), решение локальной задачи имеет вид (5.63), оценивание точности может быть выполнено по формулам (5.64), (5.66), (5.67).

В условиях (5.68) решения задач коллокации становятся статистически оптимальными.

Возможность оценить точность решения в терминах средних квадратических ошибок является одним из важнейших достоинств средней квадратической коллокации с точки зрения геодезических приложений.

Наиболее трудоемкой численной операцией является решение системы n линейных алгебраических уравнений (5.51), где n — количество исходных данных. Недостатком метода является то, что матрица коэффициентов в общем случае полностью заполнена.

Если n велико, то решение выполнимо лишь в том случае, когда тем или иным способом удается упростить структуру матрицы коэффициентов системы (5.51), например сделать ее ленточной. Один из возможных путей в этом направлении состоит в таком преобразовании изучаемой функции, после которого соответствующее воспроизводящее ядро $Я(P, Q)$ быстро убывает с увеличением расстояния между точками P и Q. Это приводит к большому количеству нулей в матрице коэффициентов и заметно упрощает решение соответствующей системы. Другим систематическим приемом разрежения матрицы системы, подлежащей решению, является метод конечных элементов, изложенный в § 6.6.

С увеличением n обусловленность матрицы коэффициентов системы (5.51) заметно ухудшается. Это приводит к неустойчивости решения задач коллокации в том смысле, что незначительные изменения элементов матрицы коэффициентов или правой части системы (5.51) вызывают недопустимо большие изменения конечных результатов решения. То же происходит и в случае сравнительно небольшого n, но при наличии среди исходных функционалов L_i, близких к линейной зависимости друг от друга. Примером может служить интерполяция в плоской области с явно неравномерно расположенными в ней узлами (все узлы близки к одной и той же прямой и т. п.). Возможная мера регуляризации (т. е. обеспечения устойчивого решения) таких задач указана в § 6.4.

Дифференциальный функционал качества. Рассмотрим теперь задачу интерполяции как частный случай коллокации (см. замечание 5.5). Ее решение с использованием дифференциального функционала качества известно в аналитическом виде.

Исходными данными служат значения изучаемой функции $l_i = f(P_i)$, $i = 1, \ldots, n$ в узлах P_i, хаотически расположенных в плоской области D. Это означает, что исходными функционалами L_i являются дельта-функционалы $\delta_{P_i} \equiv f(P_i)$.

Теорема 5.14. Решение (5.38) глобальной задачи интерполяции (5.39) в плоской области D с исходными числами $l_i = L_i f = \delta_{P_i} f$, $i = 1, \ldots, n$ и функционалом качества (5.41) имеет вид

$$\hat{f}(x, y) = \frac{1}{2} \sum_{i=1}^{n} a_i r_i^2(x, y) \ln r_i^2(x, y) + \\ + \tau_1 + \tau_2 x + \tau_3 y, \tag{5.74}$$

где

$$r_i^2(x, y) = (x - x_i)^2 + (y - y_i)^2 = |\overrightarrow{P_i P}|^2,$$

$P(x, y)$ — текущая точка области D. Коэффициенты $a_1, \ldots, a_n$ составляют столбец a, коэффициенты τ_1, τ_2, τ_3 составляют стол-

бец τ корней следующей системы линейных алгебраических уравнений:

$$\left(\begin{array}{c|c} G & M \\ \hline M^{\mathrm{T}} & \Theta \end{array}\right)\left(\begin{array}{c} a \\ \hline \tau \end{array}\right)=\left(\begin{array}{c} l \\ \hline \upsilon \end{array}\right), \tag{5.75}$$

где $G=(g_{ij})$ — симметричная матрица размера $[n\times n]$ с элементами $g_{ij}=|\overrightarrow{P_iP_j}|^2\ln|\overrightarrow{P_iP_j}|$ при $i\neq j$ и $g_{ii}=0$; M — прямоугольная матрица размера $[n\times 3]$, строчки которой заполняют числа 1, x_i, y_i; выполнение условий теоремы 5.7 обеспечивает этой матрице полный ранг; Θ — нуль-матрица размера $[3\times 3]$; l — столбец n исходных значений функции f в узлах $P_i(x_i, y_i)$; ϑ — нуль-столбец с тремя строчками. Доказательство можно найти в работе [3]. Там же отмечается, что при решении системы (5.75) возникает следующее. Если значения координат точек P заданы в реальных единицах измерения (метрах, километрах и т. д.), то элементы матрицы коэффициентов могут сильно колебаться по величине. Это может привести к значительным ошибкам округления при решении системы методом Гаусса, так как аддитивные операции с величинами разных порядков всегда опасны для реализации на ЭВМ с ограниченной мантиссой. Поэтому рекомендуется ввести новые переменные

$$\begin{aligned} x' &= (x-x_{\min})/r, \\ y' &= (y-y_{\min})/r, \end{aligned} \tag{5.76}$$

где $x_{\min}$, $x_{\max}$, $y_{\min}$, $y_{\max}$ — экстремальные значения координат исходных точек P_i, r — наибольшее число из $x_{\max}-x_{\min}$, $y_{\max}-y_{\min}$.

Отметим также, что $\lim r\ln r=0$ при $r\to 0$. Поэтому особенность, возникающая при использовании выражения (5.74) для решения локальной задачи интерполяции в точке P', близкой к какому-нибудь узлу сетки P_i, устранима: соответствующее слагаемое надо заменить нулем. Интерполяционный сплайн (5.74) непрерывен и имеет непрерывные первые производные. Порядок точности относительно шага сетки более 3/2 [3].

Пример 5.11 [3]. Известны значения некоторой функции $z=f(x, y)$ в пяти узлах P_i плоскости

P_i	P_1	P_2	P_3	P_4	P_5
x_i	—2/3	0,25	—0,40	—0,00	0,70
y_i	—1/3	—0,50	0,70	0,00	—0,40
z_i	—1	1	—1	1	—1
x_i'	0,0000	0,6707	0,1951	0,4878	1,0000
y_i'	0,2439	0,1220	1,0000	0,4878	0,1951

Требуется решить следующие задачи интерполяции: а) локальную для точки $Q(-1/2, 1/2)$, т. е. оценить значение $f(Q)$; б) глобальную, в которой требуется восстановить функцию.

Координаты узлов x_i, y_i преобразуются по формулам: $x_i'=(x_i-u)r$, $y_i'=(y_i-u)/r$, $r=v-u$, $u=-2/3$ и $v=0{,}70$ — соответственно наименьшая и наибольшая координаты узлов. Преобразованные узлы принадлежат единичному квадрату D.

Составляя и решая систему (5.79) с $l_i=z_i$, получим $f(x', y')\approx\hat{\varphi}(x', y')$ в виде (5.78) с $a_1=-1{,}4284$, $a_2=2{,}4651$, $a_3=-0{,}2967$, $a_4=1{,}6704$, $a_5=-2{,}4104$, $\tau_1=0{,}5189$, $\tau_2=-0{,}0949$, $\tau_3=-0{,}5019$, $f(Q)\approx\hat{\varphi}(Q)=-0{,}8654$.

Пример 5.12. Информация о рельефе в некоторой местности D представляет собой значения отметок z_i в хаотически расположенных точках $P_i(x_i, y_i)$, $i=1, ..., n$. По имеющимся данным требуется построить горизонтали. Можно рекомендовать следующую схему вычислений: 1) составить и решить систему (5.75); 2) пересчитать значения интерполяционного сплайна (5.74) в узлы регулярной прямоугольной сетки с удобными шагами.

Наиболее трудоемкой численной операцией является решение системы $n+3$ уравнений (5.75), где n — количество исходных данных. Матрица коэффициентов в общем случае оказывается плотной, что является недостатком метода. Систематический прием разрежения матрицы методом конечных элементов описан в § 6.6.

§ 5.7. ОСНОВНЫЕ НАПРАВЛЕНИЯ ИСПОЛЬЗОВАНИЯ ИНТЕРПОЛЯЦИИ И КОЛЛОКАЦИИ

Интерполирование функций играет значительную роль в современной вычислительной математике и в численных методах геодезии. Рассмотренные задачи имеют самостоятельное значение для приближенного представления и вычисления функций, например, различных трансформант гравитационного поля, а также возникают в качестве вспомогательных при решении более сложных задач. Перечислим основные направления использования интерполяции и коллокации.

1. Численное решение нелинейных уравнений и систем.

В основе лежит приближенная замена искомых корней нелинейного уравнения $f(x)=0$ корнями уравнения $\varphi(x)=0$, где $\varphi(x)$ — интерполяционная функция для $f(x)$. Взяв, например, за x_{i+1} корень интерполяционного полинома 1-й степени, построенного по значениям $f(x_i)$ и $f'(x_i)$ в узле x_i или по значениям $f(x_{i-1})$, $f(x_i)$ в узлах x_{i-1}, x_i, приходят соответственно к методам касательных (Ньютона) и секущих: $x_{i+1}=x_i-f(x_i)/f'(x_i)$, $x_{i+1}=x_i-f(x_i)/[x_{i-1}, x_i]$, где $[x_{i-1}, x_i]$ — разделенная разность (см. § 5.2).

Другой подход основан на интерполировании обратной функции $x=g(y)$, т. е. нахождении x для нужного y, если задана таблица $y_i=f(x_i)$. Для монотонных функций нет разницы между прямым и обратным интерполированием и исходную таблицу можно считать как задание $x_i=g(y_i)$.

2. Численное дифференцирование и интегрирование.

Дифференцируемая или интегрируемая функция на всей области или на ее составных частях заменяется построенной интерполяционной функцией. Основная задача при этом состоит в правильном оценивании точности. Подобные вопросы часто встречаются в физической геодезии: вычисление уклонений отвеса с помощью дифференцирования высот квазигеоида или интегрирования аномалии силы тяжести по формуле Венинга-Мейнеса и т. п.

3. Численное решение дифференциальных уравнений.

Особенно часто возникает в космической геодезии. Производные искомых функций заменяются интерполяционными формулами численного дифференцирования. Другие функции, входящие в уравнение, также часто заменяются интерполяционными.

4. Численное решение интегральных уравнений.

Возникает в физической геодезии. Искомая функция $f(x)$ заменяется какой-либо интерполяционной с выбранными узлами x_i. Приближенные значения $f(x_i)$ находятся из системы, полученной после подстановки вместо независимой переменной узлов интерполяции x_i.

Глава 6

АППРОКСИМАЦИЯ ФУНКЦИЙ

§ 6.1. ПОСТАНОВКА ЗАДАЧИ. ЭЛЕМЕНТ НАИЛУЧШЕГО ПРИБЛИЖЕНИЯ

Слово «аппроксимация» происходит от латинского *approximo* — «приближаюсь» и означает приближенную замену одних математических объектов (например, чисел, функций) другими, более простыми и в том или ином смысле близкими к исходным. Мера близости оговаривается специально и имеет принципиальное значение.

Примером задач аппроксимации могут служить все задачи локальной и глобальной интерполяции и коллокации, рассмотренные в предыдущей главе. Характеристическое требование задач интерполяции и коллокации, выделяющее их из более общего класса задач аппроксимации, состоит в том, что значения восстановленной (приближенно) функции в узлах интерполяции (значения исходных функционалов на приближенно восстановленной функции) обязаны в точности совпадать с заданными числами.

Пусть, например, требуется решить глобальную задачу интерполяции для функции одной переменной $y=f(x)$, $x\in D=[a, b]$, заданной таблично

$$\begin{array}{cccccc} x & x_0 & x_1 & \dots & x_n, \\ y & \tilde{y}_0 & \tilde{y}_1 & \dots & \tilde{y}_n. \end{array} \tag{6.1}$$

Мы научились находить различные решения $\varphi(x)$ этой задачи при соблюдении условий

$$\varphi(x_i) = \tilde{f}(x_i) \equiv \tilde{y}_i, \quad i = 0, 1, \dots, n. \tag{6.2}$$

Эти условия естественны, когда ошибками исходных данных можно пренебречь.

Например, для ковариационной функции глобального гравитационного поля Земли известны аналитические выражения [14], [16], но они слишком громоздки для многократного использования на ЭВМ (особенно при необходимости постоянно убирать низкочастотные составляющие). Рекомендуется составить для них функцию, заданную с помощью (6.1), а вычисления для промежуточных точек выполнять более простыми средствами, например с помощью кубических сплайнов.

Однако во многих случаях нет необходимости в точности удовлетворять условиям (6.2). В частности, при решении геодезических задач значения $\tilde{y}_0, \tilde{y}_1, \dots, \tilde{y}_n$ из (6.1) часто получены из измерений и имеют погрешности. Если эти погрешности значительны, то выполнение условий (6.2) приводит к незакономерным «выбросам» случайного характера в структуре $\varphi(x)$. Поэтому необязательно выполнять условия (6.2), которые приводят, с одной стороны, к сравнительно большому объему вычислений, а с другой — к формальному копированию погрешностей исходных данных. Таким образом, возникает новая задача: требуется по данным (6.1) подобрать такую аналитическую функцию $\varphi(x)$, $x \in [a, b]$, которая имела бы простейшую структуру, сглаживала особенности заданной экспериментальной таблицы и наилучшим образом отражала общий ход изменения $f(x)$ в среднем. Такая задача называется *задачей аппроксимации (приближения) функций.* Она тесно связана с интерполяцией и в данной постановке является ее обобщением. Упоминавшаяся функция $\varphi(x)$ называется теперь *аппроксимирующей функцией* (см. рис. 17).

В поставленной задаче необходимо уточнить, как наилучшим образом отразить общий ход изменения $f(x)$ и из какого класса функций выбирать $\varphi(x)$. Эти вопросы зависят от физического смысла задачи и характера распределения вероятностей ошибок измерений исходных ординат (абсциссы по-прежнему будем считать безошибочными).

Что касается класса функций, из которого выбирают аппроксимирующие функции, мы будем поступать по той же схеме, что и в предыдущей главе: начнем с простейших — полиномов одной переменной и закончим сплайнами на плоскости с двумя видами функционала качества. Будем рассматривать функции

$f(P)$, $\varphi(P)$, $g(P)$,... одной, двух или трех переменных, где P — текущая точка области D соответственно одномерной (числовая ось или часть ее), двумерной (плоскость или часть ее), трехмерной (пространственная область).

К критериям близости $\varphi(P)$ к $f(P)$ тоже можно подойти поразному. В общем случае удобно пользоваться терминологией линейных евклидовых пространств H, где близость между функциями φ, $f \in H$ измеряется расстоянием r между ними по соответствующей метрике пространства H

$$r(\varphi, f) = \| f - \varphi \|_H. \tag{6.3}$$

Если исходная функция задана таблично, т. е. на множестве отдельных точек $P_i \in D$, то аппроксимация называется *точечной;* если исходная функция задана на всем множестве D, то аппроксимация называется *интегральной.*

Запас аппроксимирующих функций часто составляет некоторое подпространство $H_1 \in H$ с известным базисом $g_0(P)$, $g_1(P)$,..., $g_k(P)$, так что

$$\varphi(P) = a_0 g_0(P) + a_1 g_1(P) + \ldots + a_k g_k(P), \tag{6.4}$$

и коэффициенты a_0, a_1,..., a_k находятся под условием минимума расстояния между f и φ:

$$r^2(\varphi, f) = \| v \|^2_H = \| f - \sum_{j=0}^{k} a_j g_j \|^2_H = \min. \tag{6.5}$$

Функция из H_1, обеспечивающая минимум (6.5), называется *функцией наилучшего приближения* в смысле метрики пространства H и обозначается $\hat{\varphi}$. Ее легко найти стандартными способами. В самом деле

$$\| v \|^2_H = (f - \sum_j a_j g_j,\ f - \sum_j a_j g_j)_H =$$
$$= \| f \|^2_H - 2a^{\tau} b + a^{\tau} G a,$$

где a — столбец $k+1$ неизвестных коэффициентов a_0, a_1,..., a_k; b — столбец скалярных произведений (f, g_i), $i = 0,1,\ldots, k$; G — матрица Грама, (i, j)-позицию в которой занимает скалярное произведение $(g_i, g_j)_H$, $i, j = 0,1\ldots, k$. Дифференцируя r^2 по a_l и приравнивая производные к нулю, получаем невырожденную систему $k+1$ линейных алгебраических уравнений относительно $k+1$ искомых коэффициентов a_i:

$$Ga = b. \tag{6.6}$$

Таким образом, функция наилучшего приближения имеет вид (6.4), где коэффициенты a_0,..., a_k находятся из решения системы (6.6).

Функция $\hat{v}(P) = f(P) - \hat{\varphi}(P)$ представляет собой уклонение наилучшей аппроксимирующей функции от аппроксимируемой,

а ее норма служит критерием точности аппроксимации. Легко проверить, что $\hat{v}(P)$ ортогональна подпространству H_1 и потому справедлива теорема Пифагора:

$$\hat{v} \perp \hat{\varphi}, \quad \|f\|^2_H = \|\hat{\varphi}\|^2_H + \|\hat{v}\|^2_H. \tag{6.7}$$

Мы разъяснили решение задачи о наилучшей аппроксимации из подпространства для случая, когда пространство H заполнено функциями, определенными в области D. Все сказанное остается в силе и при заполнении H объектами любой другой математической природы.

Рассмотрим геометрическую интерпретацию вышеизложенного. Пусть H — трехмерное пространство геометрических векторов с базисом e_1, e_2, e_3; H_1 — двумерная плоскость с базисом g_0, g_1; элемент φ в плоскости подбирается так, чтобы длина вектора уклонений v была минимальной. Это автоматически приводит к перпендикулярности v к плоскости H_1 (рис. 18).

Пример 6.1. Вспомним параметрический способ уравнивания геодезической сети методом наименьших квадратов. Пусть для определения k неизвестных (например, координат пунктов) измерены n значений углов и расстояний и составлена линеаризованная система параметрических уравнений связи. Столбцы матрицы коэффициентов этой системы играют роль элементов g_j и образуют k-мерное подпространство H_1 в стандартном n-мерном евклидовом пространстве H. Аппроксимируемым элементом f является столбец правой части системы (измеренный вектор). Аппроксимирующий элемент φ (уравненный вектор) удовлетворяет геометрическим условиям сети, т. е. должен быть точной линейной комбинацией элементов g_j. Это равносильно принадлежности к H_1. Таким образом, задание проекта сети равносильно заданию подпространства H_1. Измеренный вектор не принадлежит H_1 в силу неизбежных ошибок измерений. Процесс уравнивания состоит в подборе вектора из H_1, ближайшего к измеренному вектору в смысле (6.5). Корнями системы нормальных уравнений (6.6) служат координаты уравненного вектора φ в базисе подпространства H_1. Квадрат длины вектора уклонений характеризует точность уравнивания, т. е. $\|v\|^2/(n-k)$ есть дисперсия единицы веса. Соотношение (6.7) используется, в частности, для контроля вычислений.

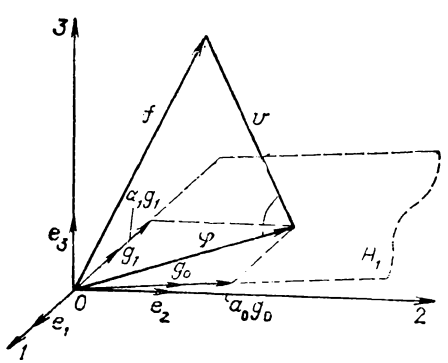


Рис. 18. Геометрическая интерпретация задачи отыскания элемента наилучшего приближения

Пример 6.2. Решение локальной задачи коллокации, изложенное в § 5.6 предыдущей главы, есть обычное решение задачи о наилучшей аппроксимации в пространстве функционалов со скалярным произведением (5.55), (5.65). Исходные функционалы образуют подпространство H_1 (см. теорему 5.9 в предыдущей главе).

Упражнение 6.1. Пользуясь изложенной методикой построения элемента наилучшей аппроксимации, докажите теорему 5.9 из предыдущей главы.

Точечная аппроксимация функций называется *квадратичной*, если скалярное произведение в H определяется по следующему правилу:

$$(f_1(P),\ f_2(P))_H \equiv \sum_{i=0}^{n} f_1(P_i) f_2(P_i),\ P \in D. \tag{6.8}$$

Условие (6.5) имеет вид

$$\| v \|^2 = \sum_{i=0}^{n} v_i^2 = \sum_{i=0}^{n} (f(P_i) - \varphi(P_i))^2 = \min \tag{6.9}$$

и обеспечивается методом наименьших квадратов. Пользуясь сведениями из курса теории математической обработки геодезических измерений легко доказать, что если узлы $x_0, x_1, \ldots, x_n$ фиксированы, а погрешности ординат $\tilde{y}_0, \tilde{y}_1, \ldots, \tilde{y}_n$ подчиняются нормальному закону распределения вероятностей, то требование (6.9) совпадает с условиями метода максимального правдоподобия и в этом смысле из всех возможных является наиболее обоснованным. Если исходные значения функции неравноточны, то в определении скалярного произведения (6.8) каждое слагаемое надо умножить на вес ρ_i соответствующего узла P_i.

Иногда важно так подобрать $\varphi(x)$, чтобы избежать больших уклонений $\varphi(x)$ от y_i. В таком случае все внимание уделяется величине $\Delta = \max |\varphi(x_i) - f(x_i)|$, $i = 0, 1, \ldots, n$, которая и принимается в качестве меры близости. При таком подходе $\varphi(x)$ находится под условием

$$\Delta \equiv \max |\varphi(x_i) - f(x_i)| = \min,\ i = 0, 1, \ldots, n. \tag{6.10}$$

Этот процесс называется *точечной равномерной аппроксимацией.*

Условие (6.10) является более жестким, чем (6.9) и его практическая реализация вызывает серьезные затруднения. Поэтому в дальнейшем будем использовать условие (6.9). Оно приводит к сравнительно простым вычислениям, достаточным с точки зрения общей теории обработки измерений.

§ 6.2. ТОЧЕЧНАЯ КВАДРАТИЧНАЯ АППРОКСИМАЦИЯ ПОЛИНОМАМИ

Пусть некоторая функция $f(x)$ задана в виде (6.1), причем узлы $x_0, x_1, \ldots, x_n$ — фиксированные точные числа, а ординаты $\tilde{y}_0, \tilde{y}_1, \ldots, \tilde{y}_n$ — результаты измерений. Пусть с помощью табличных разностей установили, что $f(x)$ может быть аппроксимирована полиномом $\varphi_k(x)$ степени $k < n$. Запишем искомый полином в виде

$$\varphi_k(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_k x^k.$$

Роль базисных функций $(k+1)$-мерного подпространства H_1 играют степенные функции $g_j(x) = x^j$, где $j = 0, 1, \ldots, k$. Необходимо определить коэффициенты $a_0, a_1, \ldots, a_k$.

Если бы $f(x)$ действительно представляла собой полином k-й степени, а все y_i были бы безошибочны, то очевидно

$$\left.\begin{aligned} a_0+a_1x_0+a_2x^2{}_0+\ldots+a_kx_0{}^k-\tilde{y}_0=0;\\ a_0+a_1x_1+a_2x_1{}^2+\ldots+a_kx_1{}^k-\tilde{y}_1=0;\\ \ldots\ldots\ldots\ldots\ldots\\ a_0+a_1x_n+a_2x^2{}_n+\ldots+a_kx^k{}_n-\tilde{y}_n=0. \end{aligned}\right\} \qquad (6.11)$$

Это означает, что измеренный вектор $\tilde{y}=(\tilde{y}_0, \tilde{y}_1, \ldots, \tilde{y}_n)^{\tau}$ принадлежит подпространству H_1. Но поскольку предположения о полиномиальной структуре $f(x)$ и безошибочности ординат $\tilde{y}$ практически не выполняются, то система (6.11) оказывается несовместной. Поэтому каким бы мы ни взяли вектор a, в правой части этой системы всегда будет находиться некоторый вектор уклонений $v_{n+1,1}$, отличный от нуль-вектора $O_{n+1,1}$:

$$Xa-\tilde{y}=v, \qquad (6.12)$$

где X — матрица размера $[(n+1)\times(k+1)]$ коэффициентов системы (6.11), а v — столбец $n+1$ уклонений.

Условие (6.9) приводит к необходимости решать систему уравнений (6.6). Согласно (6.8) (i, j)-позицию матрицы G занимает число $(g_i, g_j)=\sum\limits_{s=0}^{n} x^i{}_s\cdot x^j{}_s=x_0{}^{i+j}+x_1{}^{i+j}+\ldots+x_n{}^{i+j}$, где $i=0, 1, \ldots, k$ и $j=0, 1, \ldots, k$. В матричном виде система (6.6) имеет вид

$$X^{\tau}Xa=X^{\tau}\tilde{y} \qquad (6.13)$$

и называется системой нормальных уравнений. Решая ее, например, методом квадратного корня, получим вектор коэффициентов a и, следовательно, составим искомый полином $\varphi(x)$.

Оценивание точности полученных коэффициентов $a_0, a_1, \ldots, a_k$ осуществляется точно так же, как в схеме косвенных измерений метода наименьших квадратов. В простейших случаях о качестве аппроксимации можно судить по значениям компонент вектора v из (6.12).

Заметим, что если $k=n$, то аппроксимирующий полином $\varphi(x)$ совпадает с интерполяционным полиномом Лагранжа для системы точек $x_0, x_1, \ldots, x_n$, причем $\|v\|^2=0$. В этом смысле аппроксимация функций представляет собой более общий процесс, чем интерполяция.

Интересно отметить, что полином наилучшего квадратичного приближения обладает тем же свойством, что и интерполяционная функция: разность $f(x)-\varphi_k(x)$ на промежутке $[a, b]$ имеет не менее $k+1$ нулей [9].

Пример 6.3. Пусть требуется аппроксимировать аналитическую зависимость высоты y падения тела от времени x по данным первых двух

Таблица 8

x	1	t_i	$t^2{}_i$	$t^3{}_i$	$t^4{}_i$	$\tilde{y}_i$	$x_i\tilde{y}_i$	$x^2{}_i\tilde{y}_i$	$v_i=y_i-\varphi(t_i)$
1/30	1	0	0	0	0	1,19	0	0	0
2/30	1	1	1	1	1	1,57	1,57	1,57	$-4\cdot10^{-3}$
3/30	1	2	4	8	16	2,06	4,12	8,24	$-6\cdot10^{-3}$
4/30	1	3	9	27	81	2,67	8,01	24,03	$+4\cdot10^{-3}$
5/30	1	4	16	64	256	3,37	13,48	53,92	$+4\cdot10^{-3}$
6/30	1	5	25	125	625	4,19	20,95	104,75	0
7/30	1	6	36	216	1296	5,11	30,66	183,96	$-4\cdot10^{-3}$
8/30	1	7	49	343	2401	6,15	43,05	301,35	$+4\cdot10^{-3}$
Σ	8	28	140	784	4676	26,31	121,84	677,82	$[v^2]=116\cdot10^{-6}$

столбцов табл. 7. Кроме того, желательно оценить величину ускорения g. Результаты измерения величин y обозначаем $\tilde{y}$.

Решение.

1. Прежде всего, необходимо назначить степень k аппроксимирующего полинома. В данном случае, учитывая физический смысл задачи (а в общем случае — результаты примера 5.1), можно принять $k=2$. Таким образом

$$\varphi(x)=a_0+a_1x+a_2x^2,$$

где a_0 — путь, пройденный к началу отсчета; a_1 — скорость в момент отсчета; a_2 — половина ускорения свободного падения.

2. Поскольку имеем равноотстоящие абсциссы с постоянным шагом $h=1/30$ с, то целесообразно ввести новую переменную

$$t=\frac{x-x_0}{h}=\frac{x-1/30}{1/30}=30x-1. \qquad (6.14)$$

Заполняем первый и третий столбцы табл. 8.

3. Составление нормальных уравнений (6.13) удобно делать в той же таблице. Здесь матрицу уравнений погрешностей $X_{8,3}$ составляют данные второго, третьего, четвертого столбцов.

Таким образом, система нормальных уравнений такова:

$$8a_0'+28a_1'+140a_2'=26{,}31;$$

$$28a_0'+140a_1'+784a_2'=121{,}84;$$

$$140a_0'+784a_1'+4676a_2'=677{,}82.$$

Корни этой системы соответствуют новой переменной t и потому обозначены a'.

4. Решение системы дает: $a_0'=1{,}186$, $a_1'=0{,}3302$, $a_2'=0{,}0540$. Поэтому $\hat{\varphi}(t)=1{,}19+0{,}330t+0{,}054t^2$.

5. Возвращаемся к старой переменной x согласно (6.14), $\hat{\varphi}(x)=1{,}19+0{,}330(30x-1)+0{,}054(30x-1)^2$. После преобразований окончательно имеем $\hat{\varphi}(x)=0{,}910+6{,}66x+48{,}6x^2$. Таким образом, $g=972$ см/с.

6. В десятом столбце подсчитываем уклонения v_i согласно (6.12) (в тысячных долях), характеризующие качество аппроксимации. Полученная сумма $[v^2]=116\cdot10^{-6}$ может быть использована для подсчета средних квадратических ошибок коэффициентов полинома (в частности, полученного g)

по обычной схеме косвенных измерений метода наименьших квадратов. Заметим лишь, что информативность подобных средних квадратических ошибок меньше, чем в схеме уравнительных вычислений геодезических сетей, поскольку неизвестно поведение изучаемой функции между заданными узлами.

Описанная методика наименьших квадратов принципиально проста. Однако с увеличением k она становится довольно громоздкой, и определитель матрицы нормальных уравнений уменьшается, что сильно затрудняет решение системы (6.13). Есть и более существенный недостаток, заключающийся в следующем. Пусть для заданной функции (6.1) подобран полином $\varphi_k(x)$ степени k, т. е. по вышеописанной методике найдены коэффициенты $a_0, a_1, \ldots, a_k$. Но после вычисления вектора v выяснилось, что точность получилась недостаточной и потому следует произвести аппроксимацию полиномом степени $k+1$. На практике удобно постепенно наращивать степень аппроксимирующего полинома, пока не будут достигнуты заданные допуски на вектор v. При переходе от степени k к степени $k+1$ вся предыдущая работа оказывается бесполезной, так как у полинома $(k+1)$-й степени не только добавится новый коэффициент a_{k+1}, но и все остальные коэффициенты $a_0, a_1, \ldots, a_k$ тоже изменятся. Необходимо иметь такую схему вычисления коэффициентов полиномов последовательных степеней, при которой найденные один раз коэффициенты не будут изменяться при переходе от полинома k-й степени к полиному $(k+1)$-й степени, а необходимо только выполнить добавочные вычисления для нахождения a_{k+1}. Это можно сделать, если аппроксимирующую функцию (6.4) искать в виде линейной комбинации специально подобранных базисных функций, обладающих на заданной сетке свойством ортогональности [4].

§ 6.3. АППРОКСИМАЦИЯ КУБИЧЕСКИМИ СПЛАЙНАМИ НА ОТРЕЗКЕ

Вернемся к задаче аппроксимации данных (6.1). При изучении интерполяции в предыдущей главе отмечалась оптимальность условия (5.21) с $p=2$. Поэтому и в качестве решения задачи аппроксимации рекомендуется выбирать среди всех дважды дифференцируемых на отрезке $[x_0, x_n]$ функций $\varphi(x)$ ($H=W_2^2([a,b])$) такую, для которой интеграл (5.21) имеет минимальное значение. Это требование надо разумно сочетать с задачей сглаживания ошибок в исходных ординатах. Выходом является отыскание аппроксимирующей функции $\hat{\varphi}(x)$ под комбинированным условием минимума следующего функционала:

$$\Phi_p\varphi = \int_{x_0}^{x_n} [\varphi'(x)]^2 dx + \sum_{i=0}^{n} p_i v_i^2, \qquad (6.15)$$

где $p_0, p_1, \ldots, p_n$ — заданные положительные числа, веса исходных ординат; $v_i = \tilde{y}_i - \varphi(x_i)$. В отличие от интерполяции в общем случае $v_i \neq 0$. Чем больше вес p_i, тем большее значение придается выполнению интерполяционного условия в i-м узле. При малых весах условия интерполяции становятся менее значимыми и особое внимание уделяется обеспечению малой кривизны $\varphi(x)$.

Теорема 6.1. Решением задачи аппроксимации данных (6.1) под условием минимума функционала (6.15) является кубический сплайн, т. е. функция, удовлетворяющая условиям а) и б) § 5.5 и условию (5.27).

Доказательство. Обозначим через $\hat{\varphi}_p(x)$ экстремаль функционала (6.15), на которой он обращается в минимум (нижний индекс обозначает зависимость от весов), и пусть $\hat{y}_i = \hat{\varphi}_p(x_i)$. Обозначим через $\varphi(x)$ кубический сплайн, который является интерполяционным для табличных данных $(x_i, \hat{y}_i)$, $i = 0, 1, \ldots, n$, т. е. $\varphi(x_i) = \hat{y}_i$. Так как $v_i = \tilde{y}_i - \hat{\varphi}_p(x_i) = \tilde{y}_i - \varphi(x_i) = \tilde{y}_i - \hat{y}_i$, то второе слагаемое в правой части (6.15) одинаково для функций $\hat{\varphi}_p(x)$ и $\varphi(x)$. Поэтому

$$\int_{x_0}^{x_n} (\hat{\varphi}''_p)^2 dx \leqslant \int_{x_0}^{x_n} (\varphi'')^2 dx. \tag{6.16}$$

Но в § 5.5 показано, что интерполяционный кубический сплайн $\varphi(x)$ — единственная функция, обеспечивающая минимум интегралу (6.16). Поэтому $\hat{\varphi}_p(x) \equiv \varphi(x)$, что и требовалось доказать.

В связи с этим результат аппроксимации под условием минимума функционала (6.15) будем называть *аппроксимирующим (сглаживающим) сплайном.*

Замечание 6.1. Аппроксимирующий сплайн $\hat{\varphi}_p(x)$ для данных $(x_i, \tilde{y}_i)$ является интерполяционным сплайном $\varphi(x)$ для данных $(x_i, \hat{y}_i)$, где $\hat{y}_i = \hat{\varphi}_p(x)$, $i = 0, 1, \ldots, n$. Если ошибками исходных ординат можно пренебречь $(\tilde{y}_i = y)$, то аппроксимирующий сплайн совпадает с интерполяционным.

Рассмотрим построение аппроксимирующего сплайна. Обозначим через H_1 множество всех интерполяционных кубических сплайнов, соответствующих заданным (6.1) абсциссам $x_0, x_1, \ldots, x_n$ и всевозможным ординатам $y_0, y_1, \ldots, y_n$. Оно называется *пространством интерполяционных сплайнов.* Согласно теореме 6.1, нужный аппроксимирующий сплайн нет необходимости искать во всем множестве H дважды дифференцируемых на отрезке функций. Достаточно искать его в $H_1 \subset H$ — пространстве интерполяционных сплайнов. Возьмем из H_1 произвольный сплайн и составим для него функционал (6.15). Учитывая (5.22), получим

$$\Phi(\varphi, p)=\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left(m_{i-1}\frac{x_i-x}{h_i}+m_i\frac{x-x_{i-1}}{h_i}\right)^2 dx+$$

$$+\sum_{i=0}^{n}p_i(y_i-\tilde{y}_i)^2=\sum_{i=1}^{n}m_i\left(\frac{1}{6}h_i m_{i-1}+\frac{1}{3}(h_i+h_{i+1})m_i+\right.$$

$$\left.+\frac{1}{6}h_{i+1}m_{i+1}\right)+\sum_{i=0}^{n}p_i(y_i-\tilde{y}_i)^2=(Am,m)+\sum_{i=0}^{n}p_i(y_i-\tilde{y}_i)^2,$$

где m — столбец $n-1$ моментов $m_i=\varphi_p''(x_i)$, $i=1, 2, \ldots, n$; $m_0=m_n=0$; A — матрица коэффициентов, стоящая в левой части системы уравнений (5.28). Таким образом, функционал зависит от $n+1$ переменных $y_0, y_1, \ldots, y_n$. Для нахождения минимума продифференцируем функционал по y и приравняем производную к нулю. Получим

$$\frac{\partial\Phi}{\partial y_i}=\frac{\partial}{\partial y_i}(Am, m)+2p_i(y_i-\tilde{y}_i)=2\left(\frac{\partial(Am)}{\partial y_i}, m\right)+$$

$$+2p_i y_i-2p_i\tilde{y}_i=0. \qquad (6.17)$$

Но, согласно (5.28), $Am=By$, поэтому

$$\left(\frac{\partial(Am)}{\partial y_i}, m\right)=\left(\frac{\partial(By)}{\partial y_i}, m\right)=\left(\frac{\partial y}{\partial y_i}, B^{\mathrm{T}}m\right)=(B^{\mathrm{T}}m)$$

и условие минимума (6.17) в матричной форме имеет вид

$$B^{\mathrm{T}}m+Py=P\tilde{y}, \qquad (6.18)$$

где P — диагональная матрица весов. Умножая слева на BP^{-1} и учитывая (5.28), получим окончательно

$$(A+BP^{-1}B^{\mathrm{T}})m=B\tilde{y}. \qquad (6.19)$$

Матрица коэффициентов этой системы пятидиагональна, симметрична и положительно определена. Оптимальный метод решения таких систем описан в [8]. Определив моменты m аппроксимирующего сплайна, легко вычислить на основании (6.18) его значения в узлах сетки, т. е.

$$\hat{y}=\tilde{y}-P^{-1}B^{\mathrm{T}}m, \qquad (6.20)$$

и составить окончательное выражение (5.23). Столбец $v=\tilde{y}-\hat{y}=P^{-1}B^{\mathrm{T}}m$ позволяет оценить точность аппроксимации.

Таким образом, рекомендуется следующая схема вычислений аппроксимации таблично заданной функции (6.1) кубическими сплайнами.

1. Формирование матриц A и B по правилам, описанным в § 5.5.

2. Составление системы n—1 линейных алгебраических уравнений (6.19) с n—1 неизвестными моментами $m_1, \ldots, m_{n-1}$ аппроксимирующего сплайна.

3. Решение системы (6.19) и определение моментов $m_1, \ldots, m_{n-1}$ (моменты m_0 и m_n, соответствующие крайним узлам, полагаются равными нулю).

4. Вычисление сеточных значений аппроксимирующего сплайна по формуле (6.20).

5. Выражение аппроксимирующего сплайна для каждого частичного промежутка сетки имеет вид (5.23), только надо y_i заменить на $\hat{y}_i$. По-прежнему полезны рекомендации по экономизации вычислений, приведенные в § 5.5.

§ 6.4. АППРОКСИМАЦИЯ НА ПЛОСКОСТИ СПЛАЙНОМ МИНИМАЛЬНОЙ НОРМЫ. СРЕДНЯЯ КВАДРАТИЧЕСКАЯ КОЛЛОКАЦИЯ

Пусть теперь изучаемая функция $z=f(P)$ есть функция двух переменных x, y, непрерывная в некоторой плоской области D и отнесенная по инженерным соображениям к некоторому евклидову (гильбертову W_2^2) пространству H. Информация об изучаемой функции состоит из n чисел $\tilde{l}_1, \tilde{l}_2, \ldots, \tilde{l}_n$, представляющих собой в общем случае измеренные на f значения n известных линейных ограниченных функционалов $L_1, L_2, \ldots, L_n$. Требуется решать задачи коллокации, сформулированные в § 5.6, но учесть при этом, что исходные числа $\tilde{l}_i$ содержат ошибки, которыми нельзя пренебречь. На это указывает волнистая черта над l ($\Delta_i=\tilde{l}_i-l_i$ — ошибка измерения). Рекомендуется использовать результат § 5.6.

Решение задач «чистой» коллокации выполнялось на основании условий (5.39) с функционалами качества (5.41), (5.42), для более общего случая — на основании условий (5.43). В предыдущей главе подробно рассмотрено нормальное решение в виде сплайна минимальной нормы в пространстве H, в частности в виде экстремали функционала (5.42), и отмечена его эффективность для геодезических приложений. Будем использовать те же методы. Отличие состоит в том, что теперь условия коллокации (в частности, интерполяции, см. замечание 5.5) не следует выполнять точно, так как ошибки измерений, содержащиеся в числах $\tilde{l}_i$, желательно не «копировать» а по возможности сгладить. Поэтому скомбинируем норму $\|\varphi\|_H$ в выбранном пространстве H с нормой невязки уравнения (5.37), т. е. с величиной $\|v\|^2_{E_n} = \sum_{i=1}^{n} p_i v_i^2$, где уклонение $v_i=L_i\varphi-\tilde{l}_i$, p_i — вес результата $\tilde{l}_i$, а φ — любая функция из H, которая может быть решением глобальной задачи коллокации со сглаживанием

$$\widetilde{\Phi}_\alpha\varphi=\|\tilde{l}-L\varphi\|^2_{E_n}+\alpha\|\varphi\|^2_H. \qquad (6.21)$$

Такой функционал называется *сглаживающим*. Здесь α — неотрицательный числовой параметр, соизмеряющий влияние слагаемых на величину сглаживающего функционала. В дальнейшем мы объясним, из каких соображений рекомендуется назначать величину параметра α. Пока под α будем понимать фиксированную константу.

Функция $\hat{\varphi}_\alpha \in H$, обеспечивающая минимум сглаживающему функционалу (6.21), называется аппроксимирующим (сглаживающим) сплайном минимальной нормы:

$$\hat{\varphi}(P) = \arg\min_{\varphi \in H} \tilde{\Phi}_\alpha \varphi \approx f(P). \tag{6.22}$$

В геодезической литературе решение задач аппроксимации с помощью такого сплайна принято называть *средней квадратической коллокацией.*

Прежде чем рассматривать соответствующую схему вычислений, установим связь между сплайнами коллокационным (5.38) и аппроксимирующим (6.22).

Теорема 6.2. Аппроксимирующий сплайн $\hat{\varphi}_\alpha$, построенный по исходным данным D, H, L, $\tilde{l}$, совпадает с коллокационным сплайном $\hat{\varphi}$, построенным по исходным данным D, H, L, $\tilde{l}_\alpha$, где $\tilde{l}_\alpha = L\hat{\varphi}_\alpha$ совпадает с $\tilde{l}$ только при $\alpha = 0$.

Доказательство выполняется по той же схеме, по которой сделано доказательство теоремы 6.1. Его можно найти в работах [3], [12].

Следствие. Аналогично (5.49), аппроксимирующий сплайн $\hat{\varphi}_\alpha$ при любом параметре $\alpha \geqslant 0$ есть линейная комбинация представителей $L_i^*(P)$ исходных функционалов L_i:

$$\hat{\varphi}_\alpha(P) = a_1 L_1^*(P) + \ \ldots\ + a_n L_n^*(P). \tag{6.23}$$

Коэффициенты a_i определяются под условием минимума сглаживающего функционала (6.21). Поэтому далее поступаем стандартно: подставим $\varphi = \Sigma a_i L_i^*$ в (6.21), продифференцируем $\tilde{\Phi}_\alpha(a_1, \ldots, a_n)$ по a_i и полученные n производных приравняем к нулю.

Имеем

$$\tilde{\Phi}_\alpha(a_1, \ \ldots, \ a_n) = V^\tau P V + \alpha a^\tau G a, \tag{6.24}$$

где V — столбец n уклонений $v_i = \sum_{j=1}^{n} g_{ij} a_i - \tilde{l}_i$, $g_{ij} = L_i L_j^*$, P — диагональная матрица весов, G — квадратная матрица размера $[n \times n]$ с элементами g_{ij}, a — столбец n переменных коэффициентов. В матричном виде

$$V = Ga - \tilde{l}, \tag{6.25}$$

$$\frac{\partial\widetilde{\Phi}_\alpha}{\partial a}\equiv 2PV\frac{\partial V}{\partial a}+2\alpha Ga\equiv 2GP(Ga-\tilde{l})+2\alpha Ga=0,$$

$$(G+\alpha P^{-1})a=\tilde{l},\ \text{или}\ G_\alpha a=\tilde{l}. \qquad (6.26)$$

Окончательно система n уравнений (6.26) с n неизвестными a_i получена умножением обеих частей предшествующего ей равенства на $P^{-1}G^{-1}$. Матрица коэффициентов системы (6.26) симметрична и положительно определена (нормальна). Ее можно сравнить с системой (5.51). Решение системы (6.26) зависимо от α и потому обозначается a_α. Окончательный вид аппроксимирующего сплайна запишем в виде, аналогичном (5.52):

$$\hat{\varphi}_\alpha(P)=\sum_{i=1}^{n}a_{i\alpha}L_i^*(P)=L^*(P)a_\alpha=L^*(P)G_\alpha^{-1}\tilde{l}. \qquad (6.27)$$

Все сказанное в § 5.6 относительно определения функций $L_i^*(P)$ остается без изменений. По-прежнему необходимо знать воспроизводящее ядро или ковариационную функцию (см. теорему 5.13).

Что касается назначения в сглаживающем функционале (6.21) величины параметра α, то здесь существует множество различных подходов. Отметим, что описанная методика аппроксимации совпала с известным в математике вариационным методом регуляризации некорректных задач [20]. С геодезическими приложениями этого метода можно познакомиться по работам [2], [16], [21]. Мы не будет останавливаться на методах регуляризации. Заметим, что выбор параметра α оказывается особенно важным в тех случаях, когда исходная задача [см. систему уравнений (5.37)] оказывается слабоустойчивой. Это происходит, например, в том случае, когда среди исходных функционалов L_i имеются два дельта-функционала δ_{P_1} и δ_{P_2}, относящиеся к близким друг к другу точкам P_1, P_2 плоскости D.

Укажем два наиболее известных способа назначения параметра α, обеспечивающего устойчивое решение задачи и потому называемого *параметром регуляризации.*

Первый способ наиболее прост и состоит в следующем: в качестве α выбирается дисперсия единицы веса σ^2, характеризующая точность имеющихся результатов измерений $\tilde{l}$. Это наиболее обосновано в тех случаях, когда в качестве воспроизводящего ядра используется ковариационная функция аппроксимируемого поля $z=f(P)$. В самом деле, если $Я(P,Q)=K(P,Q)$ и $\alpha=\sigma^2$, то описанная схема построения аппроксимирующего сплайна совпадает с известным в теории случайных функций оптимальным линейным прогнозом и фильтрацией по Колмогорову — Винеру [19].

Второй способ называется *назначением параметра α по невязке.* В его основу положено то, что величины уклонений $v_{i\alpha}=L_i\hat{\varphi}_\alpha-\tilde{l}_i$ вычисленных значений l_i от измеренных $\tilde{l}_i$ должны

соответствовать точности измерений. Если точность выполненных измерений невысока, то нет смысла добиваться малых уклонений. Для каждого фиксированного значения α уклонения вычисляются по формуле (6.25)

$$V_\alpha = Ga_\alpha - \tilde{l}. \tag{6.28}$$

Многократно решая систему уравнений (6.26) с различными значениями α, надо остановиться на таком значении α, при котором приближенно (порядка 10%) соблюдается равенство

$$\sum_{i=1}^{n} P_i v^2_{i\alpha} = v^{\mathrm{T}}_\alpha v_\alpha \approx n\sigma^2. \tag{6.29}$$

Замечание 6.2. Обозначим через H_σ множество таких функций из H, которые обеспечивают уравнениям $L_i\varphi = \tilde{l}_i$, $i = 1, \ldots, n$ невязку $V^{\mathrm{T}}V \leqslant n\sigma^2$, где $V = L_i\varphi - l_i$. Можно доказать, что аппроксимирующий сплайн (6.22) с выбором α из условия (6.29) имеет минимальную норму по сравнению со всеми функциями из H_σ. Этот факт придает конкретный смысл последним двум словам в термине «аппроксимирующий сплайн минимальной нормы».

Рекомендуется следующая схема вычислений средней квадратической коллокации с выбором параметра α по невязке.

1. Определить воспроизводящее ядро того пространства H, в котором предполагается решать задачу.
2. Сформировать матрицу $G = Я(L, L)$ (см. § 5.6).
3. Задаться начальным значением α, например положить $\alpha = \sigma^2$.
4. Составить систему (6.26) с выбранным значением α.
5. Найти решение a_α системы (6.26), например, методом квадратного корня.
6. Вычислить уклонения по формуле (6.28).
7. Проверить выполнение условия (6.29).
8. Если условие (6.29) выполнено, то аппроксимирующий сплайн минимальной нормы имеет вид (6.27), где функции $L_i^*(P)$ определяются в соответствии с теоремой 5.13. В противном случае следует изменить параметр α и перейти к п. 4.

Пример 6.4. Пусть в условиях примера 5.7 даны числа $l_1 = L_1 f = f(0) = 1$ и $l_2 = L_2 f = f(s) = 1 + s$, где $0 < s < 1$ — фиксированная константа. Требуется оценить значение функционала $Ff = f'(0)$.

Поскольку исходные числа l предполагаются безошибочными, то воспользуемся интерполяционным сплайном минимальной нормы $\hat{\varphi}(x)$. Вычисление по формуле (5.63) приводит к формуле $\hat{Ff} = (l_2 - l_1)/s$, которая совпадает с точной формулой дифференцирования $f'(0) = (f(s) - f(0))/s$ для линейной функции и потому дает точный ответ, равный единице при любом $s > 0$, несмотря на то что $\det G = 3s^2/4 \to 0$ при $s \to 0$. Предположим теперь, что вместо точных чисел l_1, l_2 мы имеем только результаты измерения $\tilde{l}_1$, $\tilde{l}_2$, ошибки которых $\Delta_1 = \tilde{l}_1 - l_1$, $\Delta_2 = \tilde{l}_2 - l_2$ соизмеримы с величиной s. Получить

приемлемое решение по формуле (5.63) уже не представляется возможным. Так, при $s=0{,}001$, $\Delta_1=-0{,}05$, $\Delta_2=0{,}05$ получается $Ff=100$.

Решим ту же задачу с помощью аппроксимирующего сплайна по формуле (6.27). Веса обоих измерений полагаем одинаковыми и равными единице, а в качестве σ^2 возьмем число $(\Delta_1^2+\Delta_2^2)/2$, $\sigma=0{,}05$. Реализация указанной выше схемы вычислений с выбором параметра α по невязке, т. е. из условия (6.29), дала значение искомой производной $f'(0)=F\hat{\varphi}_\alpha=1$, что совпадает с его точным значением. При этом $\alpha=7{,}51\cdot10^{-5}$, что в 3 раза меньше $\sigma^2=2{,}5\cdot10^{-5}$.

§ 6.5. АППРОКСИМАЦИЯ НА ПЛОСКОСТИ ДИФФЕРЕНЦИАЛЬНЫМ СПЛАЙНОМ

Обобщим содержание § 5.6 на случай, когда исходные значения $\tilde{l}_i=f(P_i)+\Delta_i$ содержат непренебрегаемые ошибки Δ_i, $i=1, 2, \ldots, n$. Так же как и в § 6.3 и 6.4, в точности выполнять условия интерполяции $\varphi(P_i)=\tilde{l}_i$ не следует. Поэтому введем сглаживающий функционал

$$\tilde{\Phi}_\alpha\varphi=\sum_{i=1}^{n}p_iv_i^2+\alpha\Phi\varphi, \qquad (6.30)$$

который складывается из меры близости функции φ к изучаемой функции (p_i — вес i-го числа $\tilde{l}_i$, $v_i=\varphi(P_i)-\tilde{l}_i$) и меры «искривленности» функции φ (второе слагаемое здесь определяется формулой (5.42); α — неотрицательный числовой параметр, имеющий тот же смысл, что и в предыдущем параграфе.

Дважды дифференцируемая функция $\hat{\varphi}_\alpha$, обеспечивающая минимум сглаживающему функционалу (6.30), называется аппроксимирующим (сглаживающим) сплайном с дифференциальным 2-го порядка функционалом качества или *аппроксимирующим дифференциальным 2-го порядка сплайном.*

Можно доказать, что, аналогично теоремам 6.1 и 6.2, аппроксимирующий дифференциальный сплайн φ_α представляет собой интерполяционный дифференциальный сплайн (5.74), построенный по исходным данным $\tilde{l}_{i\alpha}=\hat{\varphi}_\alpha(P_i)$. Коэффициенты a и τ в выражении (5.74) находятся из решения системы $n+3$ уравнений, имеющей ту же структуру, что и система (5.75). Надо только к блоку G добавить матрицу αP^{-1}, где P — диагональная матрица весов, а на место столбца l в правой части поставить столбец имеющихся исходных значений функции $\tilde{l}$. Параметр α подбирается методом, указанным в § 6.4.

Пусть у функции $f(P)$, $P\in D$ известны значения $\tilde{l}_i$ только в узлах $P_i(x, y)$ хаотичной сетки области D, $i=1, \ldots, n$. Значения $\tilde{l}_i$ получены с весом p_i и средней квадратической ошибкой единицы веса σ. Общая схема вычислений аппроксимации $f(P)$ дифференциальным 2-го порядка сплайном с выбором параметра α по невязке следующая [3].

1. Сформировать матрицы G и M так, как это описано в § 5.6.

2. Задаться начальным значением α, например положить $\alpha=\sigma^2$.

3. Составить систему линейных уравнений вида

$$\begin{pmatrix} G+\alpha P^{-1} & M \\ M^{\mathrm{T}} & \Theta \end{pmatrix}\begin{pmatrix} a \\ \tau \end{pmatrix}=\begin{pmatrix} \tilde{l} \\ \upsilon \end{pmatrix}$$

с выбранным значением α.

4. Найти решение a_α, τ_α этой системы с учетом рекомендаций (5.76).

5. Составить выражение (5.74)

$$\hat{\varphi}_\alpha(x, y)=\frac{1}{2}\sum_{i=1}^{n} a_{i\alpha} r_i^2(x, y)\ln r_i^2(x, y)+\tau_{1\alpha}+\tau_{2\alpha}x+\tau_{3\alpha}y$$

и вычислить уклонение $v_{i\alpha}=\hat{\varphi}_\alpha(P_i)-\tilde{l}_i$, $i=1, \ldots, n$.

6. Проверить выполнение условия (6.29) с точностью порядка 10%.

7. Если условие (6.29) выполняется, то ответом служит выражение п. 5, с помощью которого можно легко вычислить приближенные значения изучаемой функции в узлах регулярной сетки, удобной для дальнейшего использования (см. пример 5.11). Если условие (6.29) не выполнено, то следует изменить параметр α и перейти к п. 3 схемы вычислений.

Можно доказать, что при $\sigma^2\to 0$ и, следовательно, $\alpha\to 0$ невязка $V_\alpha^{\mathrm{T}}V_\alpha$ стремится к нулю не медленнее $O(\sqrt{\alpha})$, а аппроксимирующий сплайн сходится к коллокационному сплайну по метрике пространства H.

Пример 6.5. Решение задач примера 5.11 с $\alpha=0{,}2$ дает следующие результаты [3]: $a_1=-0{,}9889$, $a_2=1{,}5901$, $a_3=-0{,}2938$, $a_4=1{,}3703$, $a_5=-1{,}6776$, $\tau_1=0{,}2925$, $\tau_2=0{,}1356$, $\tau_3=-0{,}4877$, $\hat{\varphi}_{0,2}(Q)=-0{,}8125$, $\hat{\varphi}_{0,2}(P_1)=-0{,}8022\neq\hat{\varphi}(P_1)=-1$. Если полагать, что исходные аппликаты измерены равноточно со средней квадратической ошибкой $\sigma=0{,}1$ и выбирать параметр α по невязке, то $\alpha\approx 0{,}006$.

§ 6.6. МЕТОД КОНЕЧНЫХ ЭЛЕМЕНТОВ

Общие сведения. С точки зрения вычислений основным недостатком изложенных выше методов интерполяции, коллокации и аппроксимации на плоскости с помощью сплайнов является необходимость решать системы линейных алгебраических уравнений с полными матрицами коэффициентов. Количество уравнений в таких системах не меньше количества n исходных функционалов L_i на изучаемой функции f. С увеличением n объем вычислений систем с полными матрицами возрастает очень быстро. В то же время в геодезической практике нередко воз-

никает необходимость совместной обработки очень большого количества результатов $\tilde{l}$ измерений функционалов L, например на геопотенциале. Достаточно вспомнить стандартные задачи физической геодезии об определении в заданной точке земной поверхности высоты квазигеоида или компонент уклонения отвесной линии. Для их решения нужна информация о значениях аномалии силы тяжести по всей поверхности Земли. Разумное использование такого громадного количества информации изложено в работах [16], [18]. Рассмотрим систематический прием, позволяющий существенно разредить матрицу коэффициентов системы, подлежащей решению при аппроксимации сплайнами, и тем самым значительно снизить объем необходимых вычислений. В основе лежит широко используемый в настоящее время в вычислительной математике метод конечных элементов. Чтобы уяснить смысл этого метода, вспомним общую схему решения задач аппроксимации с помощью сплайнов. Прежде всего обратим внимание на то, что эти задачи являются, по существу, задачами вариационными. Это означает, что решение $\hat{\varphi}$ отыскивается путем минимизации некоторого функционала сглаживания (6.15), (6.21), (6.30), а также (5.38), (5.43), (6.22). Экстремаль $\hat{\varphi}$ функционала сглаживания выбирается среди бесконечного множества различных функций, составляющих определенное пространство H. И хотя в общем случае это пространство бесконечномерно, нам всегда удавалось из теоретических соображений выделить в H некоторое конечномерное подпространство, среди элементов которого и имеет смысл искать решение $\hat{\varphi}$ [см. формулы (5.49), (6.23)]. Подставляя произвольную функцию φ этого подпространства в сглаживающий функционал, мы получали обычную функцию n переменных a_i координат φ в базисе n-мерного подпространства (6.24). Приравнивание производной этой функции по каждой переменной a_i к нулю давало систему уравнений (6.26), подлежащую решению. Структура (заполненность или разреженность) матрицы коэффициентов системы (6.26) полностью определяется базисными элементами того подпространства, в котором ищется решение.

Использование метода конечных элементов при аппроксимации сплайнами отличается от описанной выше схемы одним, но важным положением: решение предлагается искать приближенно в конечномерном подпространстве со специальным базисом, заведомо обеспечивающим разреженность матрицы коэффициентов системы уравнений, подлежащей решению. Отличие приближенного решения $\hat{\hat{\varphi}}$ от точного $\hat{\varphi}$ зависит от размерности N используемого подпространства и стремится к нулю при $N \to \infty$.

Средняя квадратическая коллокация на отрезке. Рассмотрим применение метода конечных элементов для решения следующей одномерной задачи средней квадратической коллокации.

Известно, что изучаемая функция $y=f(x)$, $x\in D=[a,b]$ принадлежит пространству дифференцируемых (не обязательно непрерывно) функций $H=W_2^1([a,b])$. Известны результаты $\tilde{l}_i$ измерений с весом p_i n значений l_i линейных ограниченных функционалов $L_i f$ на этой функции, $i=1,\ldots,n$. Требуется восстановить функцию f.

Наилучшей аппроксимацией будем считать такую функцию $\hat{\varphi}(x)$ из H, которая обращает в минимум сглаживающий функционал (6.21):

$$\Phi\varphi=\|L\varphi-\tilde{l}\|^2_{E_n}+\|\varphi\|^2_H=\sum_{i=1}^{n}p_i v^2{}_i+\int_a^b\left[(\nu\varphi)^2+\left(\frac{d\varphi}{dx}\right)^2\right]dx. \qquad (6.31)$$

Параметр α принят равным единице, $\nu>0$ — фиксированная константа; $L\varphi$ — столбец n чисел $l_i=L_i\varphi$, $i=1,\ldots,n$; $v_i=\varphi(x_i)-\tilde{l}_i$.

Итак

$$\hat{\varphi}(x)=\arg\min\Phi\varphi,\ \varphi\in H. \qquad (6.32)$$

Разобьем отрезок $[a,b]$ на равные части и пусть $\bar{x}_1, \bar{x}_2,\ldots,\bar{x}_N$, где $N\approx n$, обозначают концы этих частичных промежутков длиной h. Систему координат выберем так, чтобы $\bar{x}_1=h$. Узлы $\bar{x}_2,\ldots,\bar{x}_{N-1}$ будем называть *внутренними*, а узлы $\bar{x}_1=a$, $\bar{x}_N=b$ — *граничными*. Каждому внутреннему узлу с номером i поставим в соответствие кусочно-линейную функцию

$$\omega_i(x)=\begin{cases}1+(x-\bar{x}_i)/h \text{ при } \bar{x}_{i-1}\leqslant x\leqslant\bar{x}_i;\\ 1-(x-\bar{x}_i)/h \text{ при } \bar{x}_i\leqslant x\leqslant\bar{x}_{i+1};\\ 0 \text{ при } x<\bar{x}_{i-1} \text{ или } x>\bar{x}_{i+1},\ i=2,\ldots,N-1.\end{cases} \qquad (6.33)$$

Областью определения такой функции служит вся числовая ось, но от нуля она отличается только на конечном промежутке $(\bar{x}_{i-1}, \bar{x}_{i+1})$. Подобные функции называются *финитными*. Носитель финитной функции (та часть ее бесконечной области определения, где она отлична от нуля) представляет собой конечный элемент. Граничным узлам поставим в соответствие кусочно-линейные функции (рис. 19)

$$\omega_1(x)=\begin{cases}2-x/h \text{ при } h\leqslant x\leqslant 2h;\\ 0 \text{ при } x<h \text{ или } x>2h,\end{cases} \qquad (6.34)$$

$$\omega_N(x)=\begin{cases}-(N-1)+x/h \text{ при } (N-1)h\leqslant x\leqslant Nh;\\ 0 \text{ при } x<(N-1)h \text{ или } x>Nh.\end{cases}$$

Зададимся теперь какими-нибудь числами $z_1,\ldots,z_N$ и составим функцию вида

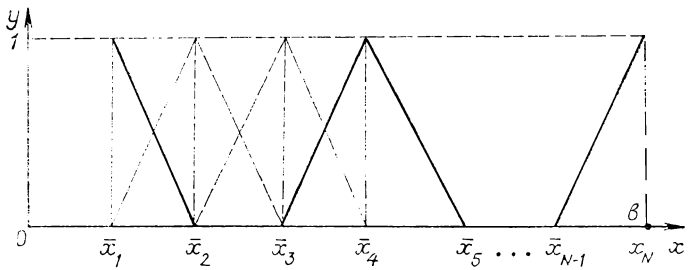


Рис. 19. Графики кусочно-линейных функций

$$\bar{\varphi}(x)=\sum_{i=1}^{N} z_i\omega_i(x). \tag{6.35}$$

Графиком этого кусочно-линейного интерполянта служит ломаная линия (рис. 20).

Множество всех функций вида (6.35), соответствующих всевозможным ординатам $z_1,\ldots,z_N$, образует N-мерное подпространство H^h пространства $H=W_2^1([a,b])$. Финитные функции $\omega_i(x)$, $i=1,\ldots,N$ составляют базис такого подпространства, который принято называть *базисом типа конечных элементов*. Любую функцию $f(x)$ из H можно в определенном смысле приблизить функцией $\bar{\varphi}(x)$ из H^h. Для этого достаточно взять в качестве z_i в (6.35) ординаты $f(\bar{x}_i)$, поскольку $\omega_i(x)$ — функции влияния i-го узла, удовлетворяющие условию (5.15). Геометрически это означает, что нужно вписать ломаную в график заданной функции (см. рис. 20). Точность аппроксимации увеличивается с измельчением шага h и может быть как угодно высокой. Поэтому решения (6.32) данной задачи будем выбирать не из всего бесконечномерного пространства H, а из N-мерного подпространства H^h типа конечных элементов. Это означает, что искомая аппроксимирующая функция ищется в виде (6.35) и определению подлежат только N коэффициентов z_i.

Далее все стандартно: подставляем (6.35) в (6.31), дифференцируем по z_i и приравниваем полученные производные к нулю. Получим:

$$\Phi\bar{\varphi}=\left(\sum_{i=1}^{N} z_iL\omega_i-\tilde{l},\sum_{j=1}^{N} z_jL\omega_j-\tilde{l}\right)_{E_n}+\left(\sum_{i=1}^{N} z_i\omega_i,\sum_{j=1}^{N} z_j\omega_j\right)_H=$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N} z_iz_j\,(L\omega_i,L\omega_j)_{E_n}-2\sum_{i=1}^{N} z_i\,(L\omega_i,\tilde{l})_{E_n}+(\tilde{l},\tilde{l})_{E_n}+$$

$$+\sum_{i=1}^{N}\sum_{j=1}^{N} z_iz_j\,(\omega_i\omega_j)_H=Z^{\mathrm{T}}\bar{G}Z-2Z^{\mathrm{T}}\bar{l}+(\tilde{l},\tilde{l})_{E_n}+Z^{\mathrm{T}}\bar{B}Z,$$

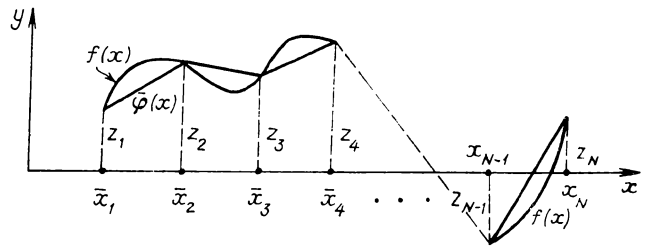


Рис. 20. График кусочно-линейного интерполянта

где $Z^{\mathrm{T}}=(z_1,\ldots,z_N)$, $\bar{G}$ и $\bar{B}$ — квадратные матрицы размера $[N\times N]$ с элементами

$$\bar{g}_{ij}=(L\omega_i,\ L\omega_j)_{E_n},$$
$$\bar{b}_{ij}=(\omega_i,\ \omega_j)_H; \tag{6.36}$$

$\bar{l}$ — столбец N чисел вида

$$\bar{l}_i=(L\omega_i,\ \tilde{l})_{E_n}; \tag{6.37}$$

$$\partial\Phi/\partial Z\equiv 2\bar{G}Z-2\bar{l}+2\bar{B}Z=0,$$

или

$$(\bar{G}+\bar{B})Z=\bar{l}. \tag{6.38}$$

Решая систему N линейных уравнений (6.38), получаем N нужных коэффициентов z_i для $\hat{\varphi}(x)$ типа (6.35). Чем больше взять N, тем точнее найденная функция $\hat{\varphi}(x)$ описывает истинное в смысле (6.32) решение $\hat{\varphi}(x)$, но тем большего порядка приходится решать систему уравнений (6.38). В чем же преимущество метода конечных элементов? В разреженности матрицы коэффициентов системы (6.38). В самом деле, функции $\omega_1,\ldots,\omega_N$ составляют базис типа конечных элементов и потому $\bar{b}_{ij}=0$ для всех i и j, для которых носители функций ω_i и ω_j не пересекаются. Другими словами, $\bar{b}_{ij}=0$ при $|i-j|>1$ и потому как бы велико ни было число N, в любой строчке матрицы $\bar{B}$ имеется не более трех ненулевых элементов. Если исходные функционалы связаны с фиксированными точками отрезка $[a,b]$, то это же можно сказать и о структуре матрицы $\bar{G}$. Учитывают также конкретный вид исходных функционалов L_i.

Упражнение 6.2. Пусть все исходные функционалы L_i в обсуждаемой задаче представляют собой значения изучаемой функции $f(x)$ в заданных точках $x_1,\ldots,x_n$ (не обязательно равноотстоящих) отрезка $[a,b]$, $\tilde{l}_i\approx l_i=f(x_i)$, $i=1,\ldots,n$. Полагая в (6.31) $\nu=1$, составьте формулы (6.36), (6.37) для формирования системы (6.38) и убедитесь, что все строчки матрицы коэффициентов этой системы содержат не более трех чисел, отличных от нуля.

Указание: $\bar{g}_{ij}=\sum\limits_{m=1}^{n}p_m\omega_i(x_m)\omega_j(x_m)$, $\bar{l}_i=\sum\limits_{m=1}^{n}p_m\tilde{l}_m\omega_i(x_m)$.

Следствием большой разреженности матрицы коэффициентов системы (6.38) является возможность решать подобные системы очень большого порядка при разумных требованиях к оперативной памяти используемой ЭВМ. Наиболее удобный способ решения при этом — способ итерации. Отметим также, что при программировании удобно пользоваться безразмерными абсциссами, выраженными в единицах шага h. Тогда $\bar{x}_i = i$, формулы (6.33), (6.34) упрощаются и отпадает необходимость в многократном делении на h при вычислениях $\bar{g}_{ij}$ и $\bar{l}_i$.

Средняя квадратическая коллокация на плоскости. На практике одномерные задачи средней квадратической коллокации редко имеют очень большое количество n исходных данных и поэтому возможно получать точные решения по схеме, описанной в § 6.4. В самом деле, если n не превышает одну-две сотни, то плотность матрицы коэффициентов (6.26) не имеет решающего значения и задача решается без упрощения. Однако в двумерных задачах (на плоскости) очень часто приходится иметь дело с большим n.

Пример 6.6. По значениям высот геоида, полученным из спутниковой альтиметрии в большом количестве хаотично расположенных точек океанической поверхности, требуется: а) восстановить возмущающий потенциал; б) пересчитать исходную информацию, заданную на хаотичной сетке, в узлы регулярной сетки; в) в заданных точках вычислить значения аномалии силы тяжести; г) получить средние интегральные значения высот геоида по заданным ячейкам ΔD.

Рассмотрим возможности метода конечных элементов для приближенного решения двумерных задач средней квадратической коллокации с большим количеством n исходных данных.

Постановка задачи подробно описана в § 6.4. Уточним только область определения D изучаемой функции $z = f(x, y)$. Пусть D — прямоугольная область $[a, b; c, d]$, а $P(x_i, y_i)$ — произвольные, но фиксированные точки этой области, к которым относятся исходные функционалы $L_i f$, $i = 1, \ldots, n$. Будем называть эти точки узлами коллокации и предполагать, что в общем случае они расположены в области D хаотично.

Аппроксимирующий сплайн минимальной нормы (6.22) будем искать приближенно в виде разложения $\bar{\varphi}$ по базису типа конечных элементов. Чтобы построить такой базис, надо конкретизировать пространство $H = W_2^q(D)$, к которому можно отнести изучаемую функцию f. Будем полагать для простоты, что $H = W_2^1(D)$, т. е. f непрерывна на D и имеет (не обязательно непрерывные) частные производные по x и по y. Скалярное произведение в таком пространстве задается с помощью формулы (5.47). Для построения пространства H^h типа конечных элементов, служащего подпространством в H, зададимся некоторым шагом $h > 0$ и введем в D квадратную сетку и прямоугольную

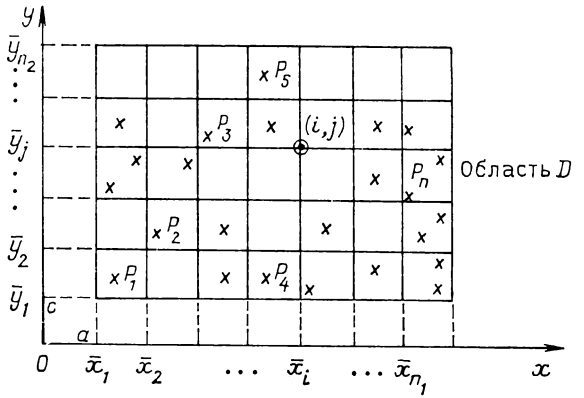


Рис. 21. Регулярная квадратная сетка для построения пространства H типа конечных элементов

систему координат так, как показано на рис. 21 (считаем, что b—a и d—c кратны h).

Узел регулярной сетки с координатами $\bar{x}_i = ih$, $\bar{y}_j = jh$ будем обозначать (i, j), где $i = 1, \ldots, n_1$ и $j = 1, \ldots, n_2$. Всего имеется $N = n_1 n_2$ таких узлов. Обычно $N \approx n$. Каждому узлу (i, j) поставим в соответствие кусочно-билинейную функцию

$$\omega_{ij}(x, y) = \omega_i(x)\omega_j(y), \tag{6.39}$$

где кусочно-линейные функции $\omega_i(x)$ определяются формулами (6.34), (6.35). График базисной функции $z = \omega_{ij}(x, y)$ изображен на рис. 22.

Функция $z = \omega_{ij}(x, y)$ определена на всей координатной плоскости xoy, но от нуля отлична только на конечном элементе области — квадрате $[\bar{x}_{i-1}, \bar{x}_{i+1}; \bar{y}_{j-1}, \bar{y}_{j+1}]$. Это финитная функция.

Формулы (6.39), (6.33), (6.34) позволяют легко составить аналитическое выражение для функции $\omega_{ij}(x, y)$, но в этом нет необходимости. Укажем лишь, что в каждом из четырех квадратов носителя функция $\omega_{ij}(x, y)$ имеет структуру квадратичного полинома вида $a_1 + a_2 x + a_3 y + a_4 xy$.

Итак, пространство H^h типа конечных элементов построено. Его размерность $N = n_1 n_2$, базисом служат функции $\omega_{ij}(x, y)$, где $i = 1, \ldots, n_1$ и $j = 1, \ldots, n_2$. Любая функция из H^h имеет вид

$$\bar{\varphi}(x, y) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} z_{ij}\omega_{ij}(x, y), \tag{6.40}$$

где z_{ij} — произвольные действительные числа.

Будем аппроксимирующий сплайн минимальной нормы (6.22), разрешающий глобальную задачу средней квадратической коллокации, приближенно выражать с помощью функций

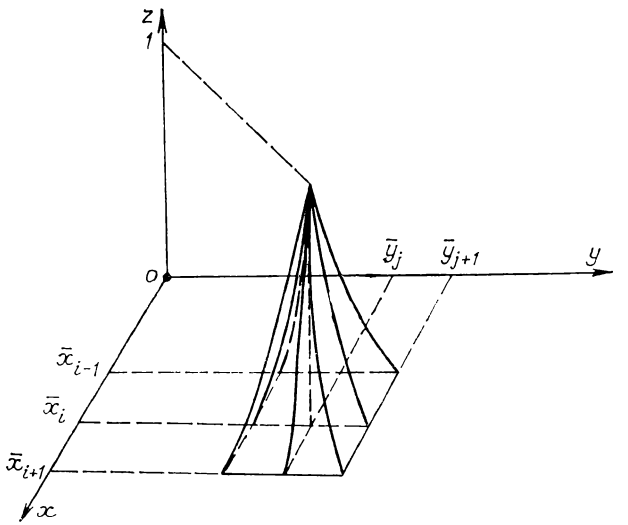


Рис. 22. График базисной функции

вида (6.40). Для этого поступают так же, как в случае одномерной задачи. В результате нужное решение $\hat{\bar{\varphi}}(x, y) \approx \hat{\varphi}(x, y)$ имеет вид (6.40), где коэффициенты z_{ij} — корни системы N линейных уравнений (6.38). С геометрической точки зрения корни z_{ij} представляют собой аппликаты функции $\hat{\bar{\varphi}}(x, y)$ в узлах (i, j).

Решать систему (6.38) рекомендуется методом итерации. Поэтому перепишем ее в виде

$$Z_{s+1} = D^{-1}[\bar{l} - (\bar{B}' + \bar{G}')Z_s, \qquad (6.41)$$

где $\bar{B} + \bar{G} = D + (\bar{B}' + \bar{G}')$; D — диагональная матрица, имеющая те же элементы, что и диагональ матрицы $\bar{B} + \bar{G}$; s — номер итерации.

Специфика обсуждаемой задачи обусловлена тем, что N может быть очень большим. Хранить в памяти ЭВМ матрицы $\bar{B}'$ и $\bar{G}'$, содержащие по N^2 чисел, невозможно. Поэтому желательно научиться эффективно умножать матрицы $\bar{B}'$ и $\bar{G}'$ на Z без существенных затрат на это оперативной памяти ЭВМ. Необходимые подробности изложены в [3], [17].

Итерации продолжаются до тех пор, пока очередной результат не будет отличаться от предыдущего по нужной метрике меньше желаемого значения, обусловленного точностью исходных данных.

Заключительные замечания. Одним из возможных приложений рассмотренной общей методики может служить задача а)

из примера 6.6. При решении задачи б) из того же примера все исходные функционалы $\tilde{L}_i$ следует полагать обычными дельта-функционалами. При этом $L\omega_{ij}$ представляет собой n-мерный столбец значений базисной функции ω_{ij} в узлах $P_1, \ldots, P_n$ исходной хаотичной сетки. Аналогично могут быть решены и другие многочисленные задачи геодезии и геофизики по пересчету скалярных полей с имеющей большое количество узлов хаотичной сетки на сетку регулярную. При этом все возможные последующие задачи, связанные с численным дифференцированием, интегрированием, автоматической рисовкой изолиний, значительно упрощаются.

Если известно, что решаемая задача неустойчива (например, задача в) из примера 6.6), то параметр α в сглаживающем функционале (6.21), (6.31) следует выбирать так, чтобы обеспечить решению регуляризирующие свойства (см. § 6.4).

Приближенное решение локальной задачи, состоящей в оценивании значения заданного функционала F на изучаемой функции f, имеет вид

$$Ff \approx \hat{F\varphi} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} z_{ij}\omega_{ij}(F), \tag{6.42}$$

где z_{ij} — решение системы (6.38), (6.41), а $\omega_{ij}(F)$ обозначает число, которое получается действием заданного функционала F на базисную функцию (6.39). Так, в задаче в) из примера 6.6 функционал F представляет собой совокупность действий, которые надо выполнить над функцией высоты геоида на плоскости для вычисления в заданной точке x_0, y_0 значения аномалии силы тяжести с точностью плоской аппроксимации [14].

Мы подробно рассмотрели приближенное решение задач средней квадратической коллокации методом конечных элементов. При этом базис составляется из простейших финитных функций (6.39) — непрерывных, но недифференцируемых. Следствием этого является непрерывность, но недифференцируемость и решения вида (6.40), что заметно сужает класс функционалов (6.42), которые можно вычислить на этом решении. Однако идейной основой метода конечных элементов является не гладкость базисных функций подпространства H^h, а их финитность. Поэтому описанная схема вычислений принципиально не меняется, если выбрать более гладкие базисные функции $\omega_{ij}(x, y)$ [8], [12].

Конечные элементы не обязательно должны иметь квадратную либо прямоугольную форму. Широко распространены конечные элементы в форме треугольников, в частности прямоугольных.

Наконец, размер носителей финитных функций, составляющих базис типа конечных элементов, может быть поставлен в зависимость от плотности исходных данных в различных частях

области D и от формы границы этой области. Регулярность сетки при этом теряется, но зато оптимизируется использование исходной информации в случае существенной неравномерности расположения узлов $P_1, \ldots, P_n$ по области D. Необходимые сведения для самостоятельного выполнения указанных обобщений можно найти, например, в работах [8], [12]. Готовые рекомендации относительно приближенного решения задачи § 6.5 на подпространстве типа конечных элементов с базисом (6.39) имеются в работе [3].

Можно доказать, что при вполне приемлемых условиях с повышением плотности исходных данных $\hat{\varphi}_\alpha \to \hat{\varphi}_\alpha$ по метрике соответствующего пространства, а $\hat{\varphi}_\alpha \to f$ с повышением точности исходных значений. Недостатком описанного в этом параграфе приложения метода конечных элементов является отсутствие достаточно простых разработок для оценивания точности аппроксимации.

Глава 7

ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

Рассмотрим основные методы приближенного вычисления определенных интегралов от функций одной переменной и двойных интегралов со стандартными областями интегрирования. Из несобственных интегралов изложены лишь интегралы 2-го рода со степенной особенностью как наиболее часто встречающиеся в геодезических приложениях.

§ 7.1. ИНТЕРПОЛЯЦИОННЫЙ ПОДХОД К ПРИБЛИЖЕННОМУ ВЫЧИСЛЕНИЮ ОПРЕДЕЛЕННЫХ ИНТЕГРАЛОВ В СОБСТВЕННОМ СМЫСЛЕ

Важнейшая в математическом анализе формула Ньютона — Лейбница

$$\int_a^b f(x)dx = F(b) - F(a),$$

где $F(x)$ — первообразная функция для $f(x)$, редко используется на практике, так как класс функций $f(x)$, для которых $F(x)$ выражается через элементарные функции, весьма узок. Замена интеграла I интегральной суммой $S_n = \sum_{i=1}^{n} f(x_i)\Delta x_i$ также неприменима ввиду медленной сходимости S_n к I при $n \to \infty$ и невозможности оценить точность при фиксированном n.

Кроме того, на практике $f(x)$ часто задана на сетке, т. е. ее значения известны лишь в отдельных точках $x_0, x_1, \ldots, x_n$ от-

резка интегрирования, а о поведении функции между этими точками имеется информация только качественного характера (степень гладкости и т. п.).

Обычно $f(x)$ приближенно заменяется другой функцией $\varphi(x)$, аппроксимирующей $f(x)$ на отрезке $[a, b]$ в том или ином смысле и легко интегрируемой.

Далее полагают, что

$$I = \int_a^b f(x)dx \approx \mathcal{I} = \int_a^b \varphi(x)dx. \tag{7.1}$$

Для аппроксимации используются методы, изложенные в гл. 5 и 6. Конкретный выбор их зависит от многих обстоятельств. Важнейшими среди них являются следующие: сведения о структуре исходной подынтегральной функции и точности ее значений в узлах сетки; возможность распоряжаться узлами сетки (их количеством и конфигурацией); требуемой точностью результата интегрирования, а также возможностью ее реально оценить. Наиболее изучена замена $f(x)$ интерполяционным полиномом $\varphi_n(x)$, которая и будет рассматриваться в дальнейшем.

Предположим, что подынтегральная функция задана на отрезке интегрирования таблично (5.1), узлы $x_0, \ldots, x_n$ фиксированы, а значения $y_0 = f(x_0), \ldots, y_n = f(x_n)$ безошибочны.

Пусть

$$f(x) = \varphi_n(x) + R_n(x) = \sum_{i=0}^{n} y_i \Lambda_i(x) + R_n(x), \tag{7.2}$$

где $\varphi_n(x)$ — интерполяционный полином Лагранжа n-й степени (5.13), $\Lambda_i(x)$ — полином влияния i-го узла (5.14), $R_n(x) = f(x) - \varphi_n(x)$ — погрешность аппроксимации (5.19).

Тогда очевидно, что

$$\int_a^b f(x)dx = \sum_{i=0}^{n} c_i y_i + r_n, \tag{7.3}$$

где

$$c_i = \int_a^b \Lambda_i(x)dx, \quad r_n = \int_a^b R_n(x)dx. \tag{7.4}$$

Приближенная формула

$$I = \int_a^b f(x)dx \approx \sum_{i=0}^{n} c_i f(x_i) = \mathcal{I} \tag{7.5}$$

называется *квадратурной формулой интерполяционного типа.* Числа c_i называются *весовыми коэффициентами* этой формулы, а исходные абсциссы $x_0, \ldots, x_n$ — *узлами интегрирования.* Если все узлы равноотстоящие, то расстояние h между ними называется *шагом интегрирования.* Квадратурная формула называет-

ся *замкнутого* или *открытого типа* в зависимости от того, входят или не входят концы a, b отрезка интегрирования в состав узлов.

Замечание 7.1. Весовые коэффициенты квадратурной формулы не зависят от подынтегральной функции, а зависят только от конфигурации узлов интегрирования.

Погрешность формулы (7.5) определяется величиной остаточного члена r_n. Вычислить ее непосредственно невозможно, поскольку точка ξ в формуле (5.19) зависит от x. Поэтому нас будет интересовать в дальнейшем возможность оценить величину $|r_n|$ сверху. С ростом n точность интегрирования увеличивается, но возрастают и затруднения, о которых говорилось в § 5.4. Чтобы не иметь дело с большими n, обычно отрезок $[a, b]$ делят на N частей (не обязательно равных), на каждом частичном отрезке интегрируют с помощью малого значения n и, пользуясь аддитивностью интеграла, результаты складывают. Получаются так называемые *составные квадратурные формулы*.

Говорят, что данная квадратурная формула имеет *алгебраическую точность порядка* k, если она точна для любого полинома, степень которого не превышает k. Это означает, что $r_n=0$ в формуле (7.3) и, следовательно, приближенное равенство (7.5) имеет вид точного равенства при условии, что $f(x)$ — любой полином степени не выше k. Чем выше порядок точности, тем больше оснований ожидать, что квадратурная формула даст хороший результат и в том случае, когда $f(x)$ — произвольная непрерывная функция (см. теорему Вейерштрасса, § 5.1).

Замечание 7.2. Квадратурная формула интерполяционного типа с $n+1$ узлами имеет алгебраический порядок точности не меньше n, поскольку если $f(x)$ — полином степени не выше n, то интерполяционный полином $\varphi_n(x) \equiv f(x)$ и, следовательно, $r_n=0$.

При вычислении интегралов от периодических функций естественно характеризовать тригонометрический порядок точности используемой квадратурной формулы. Говорят, что данная квадратурная формула имеет *тригонометрический порядок точности* k, если она точна для любого тригонометрического многочлена $a_0+\sum\limits_{i=1}^{m}(a_i\cos ix+b_i\sin ix)$ порядка m не выше k. В этом определении предполагается, что $f(x)$ имеет стандартный период 2π. Функции с любым другим периодом сводятся к этой стандартной ситуации известной линейной заменой переменной.

Теорема 7.1. Среди всех квадратурных формул с $n+1$ узлами наивысший тригонометрический порядок точности, равный n, имеет формула с равномерным расположением узлов на отрезке интегрирования $[0, 2\pi]$ и равными весовыми коэффициентами.

Другой важной характеристикой качества квадратной формулы является *порядок малости m соответствующего остаточного члена относительно шага интегрирования.* Наличие такой характеристики предполагает, что при равноотстоящих узлах с шагом h остаточный член имеет следующую структуру:

$$r=r(h)=h^mC(h), \tag{7.6}$$

где $C(h)$ — некоторая функция шага, такая, что $C(h)\to C\neq 0$ при $h\to 0$. Другими словами, при достаточно малом h

$$r=r(h)=Ch^m+o(h^m). \tag{7.7}$$

Значение величины m для заданной квадратурной формулы позволяет эффективно оценивать главную часть погрешности интегрирования по схеме Рунге (см. § 7.5).

В заключение этого параграфа отметим, что, хотя мы и договорились полагать исходные ординаты интегрируемой функции безошибочными, на практике они известны только приближенно. Поэтому естественно стремиться к использованию квадратурных формул, для которых сумма модулей весовых коэффициентов минимальна. В противном случае даже незначительные ошибки исходных значений функции могут привести к существенным погрешностям результата интегрирования.

Переходим к изложению основных квадратурных формул интерполяционного типа. Теоретически они отличаются друг от друга лишь степенью полинома Лагранжа, используемого для их вывода по схеме (7.2)—(7.5).

§ 7.2. ФОРМУЛА СРЕДНИХ ПРЯМОУГОЛЬНИКОВ

Пусть в формулах (7.2)—(7.5) значение $n=0$, т. е. $\varphi_0(x)\equiv \equiv f(x_0)=y_0$, а единственный узел x_0 находится в середине $\bar{x}$ отрезка $[a, b]$. Интерполяционный полином Лагранжа (5.13) представляет собой в данном случае постоянную функцию $\varphi_0(x)=f(\bar{x})\Lambda_0(x)$, где $\Lambda_0(x)\equiv 1$. Весовой коэффициент $C_0=\int\limits_a^b \Lambda_0(x)\,dx=b-a$ и, следовательно,

$$I\approx \bar{I}=\int_a^b f(x)dx=f(x)(b-a), \tag{7.8}$$

что и называется *квадратурной формулой среднего прямоугольника.* Геометрический смысл: площадь I криволинейной трапеции заменяется площадью $\bar{I}$ прямоугольника с длиной вертикальной стороны, равной $f(\bar{x})$ (рис. 23).

Иногда для аппроксимации I используют прямоугольник с тем же нижним основанием, но с вертикальной стороной длины $f(a)$ или $f(b)$. Получаются так называемые формулы левого или

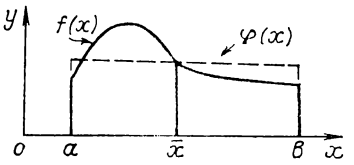


Рис. 23. Геометрическая интерпретация формулы средних прямоугольников

соответственно правого прямоугольника. Они обладают меньшей точностью, чем формула среднего прямоугольника, и в дальнейшем не рассматриваются.

Чтобы оценить погрешность формулы (7.8), разложим $f(x)$ в ряд Тейлора в окрестности точки $\bar{x}$:

$$f(x)=f(\bar{x})+f'(\bar{x})(x-\bar{x})+\frac{1}{2}f''(\xi)(x-\bar{x})^2,$$

где $\xi\in(\bar{x}, x)$ и зависит от x. Остаточный член (7.4) имеет вид

$$r_0=\int_a^b f(x)\,dx-(b-a)f(\bar{x})=f(\bar{x})(b-a)+\frac{1}{2}f'(\bar{x})(x-\bar{x})^2+$$

$$+\frac{1}{2\cdot 3}f''(\eta)(x-\bar{x})^3\Big|_a^b-(b-a)f(\bar{x}),$$

где $\eta\in[a, b]$. Второе слагаемое в правой части равно нулю. При вычислении 3-го слагаемого мы воспользовались известной теоремой о среднем значении интеграла: на отрезке интегрирования существует такая точка η, что $\int_a^b f''(\xi)(x-\bar{x})^2dx=$ $=f''(\eta)\int_a^b(x-\bar{x})^2dx$. Сокращая подобные члены и интегрируя, имеем

$$r_0=\frac{1}{6}f''(\eta)\left[\left(b-\frac{1}{2}(a+b)\right)^3-\left(a-\frac{1}{2}(a+b)\right)^3\right]$$

или после элементарных упрощений

$$|r_0|=\left|\frac{1}{24}f''(\eta)(b-a)^3\right|\leqslant\frac{M_2}{24}(b-a)^3, \qquad (7.9)$$

где $M_2=\max|f''(x)|$ при $x\in[a, b]$.

Замечание 7.3. Анализ полученной формулы интегрирования показывает, что формула (7.8) точна не только на полиномах 0-й степени (т. е. на постоянных функциях), как это можно было ожидать на основании рис. 23 и замечания 7.2, но и на любом полиноме 1-й степени (линейной функции), так как для

него $M_2=0$. Это является следствием симметричного выбора узла интегрирования x_0.

Для повышения точности получим *составную формулу средних прямоугольников*. Для этого отрезок $[a, b]$ делим на N частичных промежутков точками $a=X_0<X_1<\ldots<X_{N-1}<X_N=b$. Будем называть эти частичные промежутки шагами составной формулы. Для каждого шага применяем формулу (7.8) и результаты складываем:

$$\tilde{I}=\sum_{i=1}^{N}(X_i-X_{i-1})f(\bar{x}_i), \tag{7.10}$$

где $\bar{x}_i=\dfrac{X_{i-1}+X_i}{2}$. Это и есть составная формула средних прямоугольников. Она является квадратурной формулой открытого типа, имеет N узлов интегрирования $\bar{x}_i$ и характеризуется погрешностью

$$|r_0|\leqslant\frac{M_2}{24}\sum_{i=1}^{N}(X_i-X_{i-1})^3. \tag{7.11}$$

На равномерной сетке $X_i-X_{i-1}=(b-a)/N=h$ и потому

$$\tilde{I}=h\sum_{i=1}^{N}f(\bar{x}_i),\ |r_0|\leqslant\frac{Nh^3}{24}M_2=h^2\frac{b-a}{24}M_2. \tag{7.12}$$

Пример 7.1. Вычислить $I=\int_0^1 e^x dx$ по формуле средних прямоугольников с шагом 0,5 и оценить предельную погрешность.

Решение: $h=0{,}5$; $N=2$; $\bar{x}_1=0{,}25$; $\bar{x}_2=0{,}75$; $\tilde{I}=0{,}5\cdot(1{,}2840+2{,}1170)=1{,}7005\approx1{,}70$, $|r_0|\leqslant(0{,}5)^2\cdot2{,}718/24=0{,}028$. Точное значение $I=1{,}7183$, и потому $\tilde{I}-I\approx-0{,}02$.

Если подынтегральная функция задана таблично и никакой информации о величине M_2 нет, то величину h^2M_2 в формуле предельной погрешности (7.12) можно заменить модулем среднего значения соответствующих табличных разностей 2-го порядка $\overline{\Delta^2 y}$. В самом деле, для составной формулы имеем

$$r_0=\frac{h^3}{24}\sum_{i=1}^{N}f''(\eta_i)=\frac{h^3N}{24}\sum_{i=1}^{N}f''(\eta_i)/N$$

или, учитывая формулу (5.5),

$$r_0=h^2\frac{b-a}{24}\sum_{i=1}^{N}f''(\eta_i)/N\approx\frac{b-a}{24}\overline{\Delta^2 y}. \tag{7.13}$$

Шаг h составной формулы равен шагу интегрирования $h=\bar{x}_i-\bar{x}_{i-1}$. Поэтому ясно, что порядок малости остаточного чле-

на формулы прямоугольников относительно шага интегрирования равен двум, если $M_2<\infty$. Алгебраический порядок точности этой простейшей квадратурной формулы равен всего лишь единице, но зато на равномерной сетке она имеет наивысший из возможных тригонометрический порядок точности, равный $N-1$. Это следует из теоремы 7.1, поскольку все весовые коэффициенты в формуле (7.12) равны h.

Наличие у подынтегральной функции производных порядка, большего чем два, не улучшает точность интегрирования. Если $f(x)$ имеет ограниченной лишь первую производную, то погрешность интегрирования характеризуется оценкой, отличной от (7.12), а именно [9]

$$|r| \leqslant h \frac{b-a}{4} M_1. \tag{7.14}$$

§ 7.3. ФОРМУЛА ТРАПЕЦИЙ

Пусть теперь имеются два узла интегрирования x_0 и x_1, совпадающие соответственно с левым a и правым b концами отрезка интегрирования. Тогда в формулах (7.2)—(7.5) надо положить $n=1$, а полином Лагранжа $\varphi_1(x)=f(a)\Lambda_0(x)+f(b)\Lambda_1(x)$, где, согласно (5.14), $\Lambda_0(x)=(x-b)/(a-b)$, $\Lambda_1(x)=(x-a)/(b-a)$. Поэтому весовые коэффициенты нужной квадратурной формулы, в силу (7.4), имеют вид

$$C_0=\int_a^b \frac{x-b}{a-b}\,dx=\frac{1}{2}(b-a),\ C_1=C_0.$$

Соответствующая квадратура (7.5) запишется как

$$I \approx \bar{I}=\frac{1}{2}(b-a)[f(a)+f(b)] \tag{7.15}$$

и называется *формулой трапеции*. Такое название объясняется геометрической интерпретацией: кривая $f(x)$ заменяется стягивающей хордой $\varphi_1(x)$ и, следовательно, площадь I криволинейной трапеции аппроксимируется площадью $\bar{I}$ трапеции прямолинейной (рис. 24).

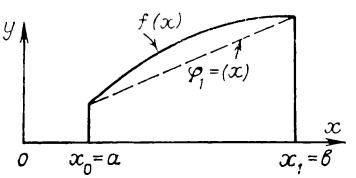


Рис. 24. Геометрическая интерпретация формулы трапеций

Погрешность формулы трапеции, согласно (7.4) и (5.19), имеет вид $r_1=\frac{1}{2}\int_a^b f''(\xi)\,(x-a)\,(x-b)\,dx$. Полагая $f''(x)$ непрерывной на $[a, b]$ и учитывая знакопостоянство $(x-a)\,(x-b)\leqslant 0$ при $x\in[a, b]$, применим теорему о среднем: существует на отрезке интегрирования такая точка η, что

$$r_1=\frac{1}{2}f''(\eta)\int_a^b(x-a)\,(x-b)\,dx=\frac{1}{2}f''(\eta)\int_a^b(x-a)\times$$

$$\times[(x-a)-(b-a)]\,dx.$$

Интегрируя это выражение, получим, что

$$|r_1|=\left|-\frac{1}{12}f''(\eta)\,(b-a)^3\right|\leqslant\frac{1}{12}M_2\,(b-a)^3, \qquad (7.16)$$

где $M_2=\max|f''(x)|$ при $x\in[a, b]$.

Для повышения точности найдем составную формулу трапеций. Для этого отрезок $[a, b]$ делим на N частичных промежутков (шаги составной формулы) точками $a=X_0<X_1<\ldots<X_N=b$. Для каждого шага применяем формулу (7.13) и результаты складываем:

$$\tilde{I}=\frac{1}{2}\sum_{i=1}^{N}(X_i-X_{i-1})(y_{i-1}+y_i). \qquad (7.17)$$

Это и есть составная формула трапеций. Она является квадратурной формулой замкнутого типа, имеет $N+1$ узлов интегрирования X_i, $i=0, 1, \ldots, N$ и характеризуется погрешностью

$$|r_1|\leqslant\frac{M_2}{12}\sum_{i=1}^{N}(X_i-X_{i-1})^3. \qquad (7.18)$$

На равномерной сетке $X_i-X_{i-1}=(b-a)/N=h$, и потому

$$\tilde{I}=\frac{h}{2}\left(\sum_{i=1}^{N}y_{i-1}+\sum_{i=1}^{N}y_i\right)=\frac{h}{2}\left(y_0+2\sum_{i=2}^{N}y_{i-1}+y_N\right)$$

или окончательно

$$\tilde{I}=h\left(\frac{y_0}{2}+y_1+\ldots-y_{N-1}+\frac{y_N}{2}\right),$$

$$|r_1|\leqslant\frac{M_2}{12}Nh^3=h^2\frac{b-a}{12}M_2. \qquad (7.19)$$

Пример 7.2. Решим задачу из примера 7.1 с помощью формулы трапеций: $h=0{,}5$; $N=2$; $x_0=0$; $x_1=0{,}5$; $x_2=1$; $\tilde{I}=0{,}5[(1/2)+1{,}6487+(2{,}7183/2)]=1{,}7539\approx 1{,}75$; $|r_1|\leqslant(0{,}5^2\cdot 2{,}718/12)=0{,}056$. Истинная ошибка $\tilde{I}-I\approx 1{,}75-1{,}72=0{,}03$.

Упражнение 7.1. Проделайте те же вычисления с шагом $h=0{,}25$.

Ответ: $\tilde{I}=1{,}7272\approx 1{,}73$, $|r_1|\leqslant 0{,}014$, $\tilde{I}-I\approx 0{,}01$.

Для таблично заданной подынтегральной функции справедлива формула, аналогичная (7.13), но со знаком «минус» и со знаменателем, в 2 раза меньшим.

Алгебраический порядок точности квадратурной формулы трапеций равен единице; порядок малости ее остаточного члена относительно шага интегрирования равен двум, если $M_2<\infty$.

Сопоставляя формулы (7.12) и (7.19), видим, что погрешность формулы средних прямоугольников примерно в 2 раза меньше, чем погрешность формулы трапеций. Поэтому если есть возможность выбирать между получением двух значений интегрируемой функции на концах отрезка интегрирования (шага составной формулы) или одного значения в середине этого промежутка, то предпочтение следует отдать второму варианту и воспользоваться формулой средних прямоугольников.

Однако лучше всего вычислить приближенное значение нужного интеграла и по формуле прямоугольников $\bar{I}_0$, и по формуле трапеций $\bar{I}_1$. Дело в том, что знаки главных частей погрешностей этих формул разные. Поэтому точное значение интеграла находится обычно между результатами счета по этим формулам. Если отложить $\bar{I}_0$ и $\bar{I}_1$ на числовой оси и поделить отрезок $[\bar{I}_0, \bar{I}_1]$ двумя точками на три равные части, то ближайшая к $\bar{I}_0$ точка деления дает уточненный результат, соответствующий рассмотренной в следующем параграфе формуле парабол (Симпсона).

Пример 7.3. Вычислим $I=\int_0^1 e^x dx$, используя результаты примеров 7.1 и 7.2. Имеем: $\bar{I}=1{,}7005+(1{,}7539-1{,}7005)/3=1{,}7183$. Из сопоставления с точным результатом видно, что приведенные знаки верны.

Все сказанное относительно точности справедливо в условиях, когда интегрируемая функция имеет ограниченную производную не менее 2-го порядка. Если же ограничена лишь 1-я производная, то погрешность интегрирования характеризуется оценкой (7.14).

Пример 7.4. Вычислим $I=\int_0^1 x^{3/2}dx$ по формуле трапеций с шагом $h=0{,}25$, $\bar{I}=0{,}25(0+0{,}125+0{,}3536+0{,}6495+0{,}5)=0{,}4070$. Вторая производная $3/(4\sqrt{x})$ подынтегральной функции не ограничена на отрезке интегрирования. Поэтому предельную ошибку результата приходится оценивать по формуле (7.14): $|r_1|\leqslant 0{,}25\cdot 0{,}25\cdot 1{,}5=0{,}09$. В данном случае интегрирование легко выполнить непосредственно $I=0{,}4$. Сравнение истинной ошибки $\bar{I}-I=0{,}007$ с оценкой предельной ошибки показывает, что последняя значительно больше.

Упражнение 7.2. Проделайте те же вычисления для $I=\int_0^1 \sqrt{x}dx$, где подынтегральная функция имеет еще меньшую гладкость.

Ответ: $\bar{I}=0{,}6433$, $\bar{I}-I\approx -0{,}02$.

§ 7.4. ФОРМУЛА ПАРАБОЛ (СИМПСОНА)

Аналогично выводится квадратурная формула при наличии трех узлов интегрирования ($n=2$); составляется по формуле (5.13) полином Лагранжа $\varphi_2(x)=y_0\Lambda_0(x)+y\Lambda_1(x)+y_2\Lambda_2(x)$ и вычисляются весовые коэффициенты (7.4). Если при этом узлы выбраны так, что $x_0=a$, $x_1=(a+b)/2$, $x_2=b$, то получается квадратурная формула вида

$$I\approx \tilde{I}=\frac{1}{6}(b-a)(y_0+4y_1+y_2). \tag{7.20}$$

Она называется *формулой* параболы или *формулой Симпсона.*

Геометрическая интерпретация: кривая $f(x)$ заменяется параболой, проходящей через точки (x_0, y_0), (x_1, y_1), (x_2, y_2) (рис. 25).

Вывод формулы для соответствующей погрешности интегрирования можно опустить. Окончательный результат следующий:

$$|r_2|=\left|-\frac{1}{2880}f^{(\mathrm{IV})}(\eta)(b-a)^5\right|\leqslant\frac{h^5}{90}M_4, \tag{7.21}$$

где η — некоторая точка отрезка интегрирования; $M_4=\max|f^{(\mathrm{IV})}(x)|$, $x\in[a,b]$; $h=(b-a)/2$.

Для повышения точности разделим отрезок интегрирования на N равных частей (шагов составной формулы) точками $a=X_0<X_1<\ldots<X_N=b$ и для каждой такой части применим формулу (7.20) с тремя узлами интегрирования — два узла по краям и один посредине. Полученные узлы интегрирования обозначим: $x_0, x_1, \ldots, x_n$, где $n=2N$, а расстояние между ними обозначим буквой $h=(b-a)/(2N)$ (рис. 26). Складывая результаты приближенного интегрирования по каждому из N шагов, получим составную формулу Симпсона (замкнутого типа):

$$\tilde{I}=\frac{h}{3}[y_0+4(y_1+y_3+\ldots+y_{n-1})+2(y_2+y_4+\ldots+y_{n-2})+y_n]. \tag{7.22}$$

Количество шагов N составной формулы может быть любым, а количество узлов интегрирования $n+1=2N+1$ всегда нечет-

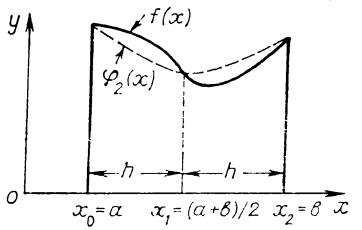


Рис. 25. Геометрическая интерпретация формулы парабол

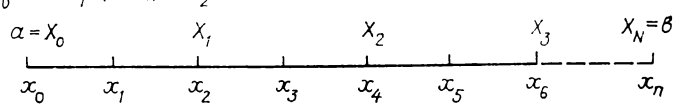


Рис. 26. Узлы составной формулы интегрирования

но. Шаг интегрирования постоянен и укладывается на отрезке интегрирования обязательно четное число раз $n=2N$. Все коэффициенты, стоящие в формуле (7.22) перед ординатами, соответствующими узлам с нечетными номерами, равны четырем, с четными номерами — двум, крайним узлам соответствуют единичные весовые коэффициенты. Для погрешности формулы (7.22) справедлива оценка

$$|r_2| \leqslant \frac{M_4}{90} Nh^5 = h^4 \frac{b-a}{180} M_4. \qquad (7.23)$$

Для таблично заданной функции по аналогии с выводом формулы (7.13) можно получить

$$r_2 \approx -\frac{b-a}{180} \overline{\Delta^4 y}, \qquad (7.24)$$

где $\overline{\Delta^4 y}$ — среднее значение табличной разности 4-го порядка для интегрируемой функции.

Пример 7.5. Вычислить $I=\int_0^{0,8} f(x)dx$ по формуле Симпсона и оценить точность, если подынтегральная функция задана таблично

x	0	0,1	0,2	0,3	0,4
y	1,0000	0,9950	0,9801	0,9553	0,9211
n	0	1	2	3	4

x	0,5	0,6	0,7	0,8
y	0,8776	0,8253	0,7648	0,6967
n	5	6	7	8

Решение: $h=0,1$; $N=4$; $n=2N=8$, по формуле (7.22) получаем $I \approx \bar{I} = (1/30)(1,0000+4\cdot 3,5927+2\cdot 2,7268+0,6967)=0,71735$.

Для оценки погрешности используем конечные разности в единицах десятитысячных долей (табл. 9).

Таблица 9

Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
	—99		
—50		0	
—149	—99		5
—248		5	
	—94		—4
—342		1	
	—93		4
—435		5	
	—88		1
—523		6	
	—82		0
—605		6	
	—76		
—681			

Воспользуемся формулой (7.24):

$$\overline{\Delta^4 y} \approx 2 \cdot 10^{-4},$$

$$|r_2| \approx \frac{0{,}8 \cdot 2 \cdot 10^{-4}}{180} \approx 10^{-6}.$$

В полученном результате все знаки можно считать верными.

Упражнение 7.3. Решите задачу из примера 7.1 с помощью формулы Симпсона.

Ответ: $I = 1{,}7189$; $|r_2| \leqslant 9 \cdot 10^{-4}$.

Алгебраический порядок точности формулы Симпсона равен трем ($M_4 = 0$ для всякого полинома, степень которого $\leqslant 3$), хотя из геометрической интерпретации (см. рис. 25) следует 2-й порядок точности. Повышение порядка является следствием того, что узлы интегрирования симметричны. Порядок малости остаточного члена относительно шага интегрирования равен четырем, при условии, что $M_4 < \infty$. Если же у подынтегральной функции ограничена лишь 3-я производная, то [9]

$$|r_2| \leqslant h^3 \frac{b-a}{72} M_3. \tag{7.25}$$

При меньшей гладкости точность формулы Симпсона становится даже хуже точности формулы средних прямоугольников [см. неравенства (7.12), (7.14)].

Общий вывод таков: если интегрируемая функция непрерывно дифференцируема по крайней мере трижды, то целесообразно по возможности пользоваться формулой Симпсона. В противном случае лучшие результаты дает формула средних прямоугольников.

§ 7.5. ВЫДЕЛЕНИЕ ГЛАВНОЙ ЧАСТИ ПОГРЕШНОСТИ АППРОКСИМАЦИИ НА СЕТКЕ МЕТОДОМ РУНГЕ

В гл. 5, 6 и 7 общим является то, что изучаемая функция (будучи непрерывной по своей физической природе) задана лишь дискретно — на некоторой сетке. Точность решения задач аппроксимации (интерполяции, коллокации, интегрирования) зависит от плотности исходных данных, т. е. от шага сетки (мы предполагаем для простоты, что сетка регулярная и имеет постоянный шаг h). Важной характеристикой качества той или иной аппроксимирующей формулы является порядок малости m соответствующего остаточного члена относительно шага сетки. Существование определенного m означает, что погрешность аппроксимации $r = r(h)$ имеет структуру (7.6), (7.7). Первое слагаемое правой части равенства (7.7) естественно называть главной частью погрешности аппроксимации, поскольку второе слагаемое имеет более высокий порядок малости.

Предположим теперь, что одна и та же задача локальной аппроксимации (интерполяции, коллокации, интегрирования) решена дважды: на сетке с шагом h получен результат $\bar{F}(h)$, а на сетке с шагом Kh получен результат $\bar{F}(Kh)$. Тогда если m для данного метода известен, то можно оценить главную часть погрешности аппроксимации. В самом деле, согласно (7.7)

$$\begin{aligned} r(h) &= F - \bar{F}(h) = Ch^m + o(h^m), \\ r(Kh) &= F - \bar{F}(Kh) = CK^m h^m + \\ &+ K \cdot o(h^m) = CK^m h^m + o(h^m), \end{aligned} \tag{7.26}$$

где F — точное значение искомой величины. Вычитая первое равенство из второго, получим

$$\bar{F}(h) - \bar{F}(Kh) = Ch^m(K^m - 1) + o(h^m).$$

Отсюда

$$r(h) \approx \tilde{r}(h) = Ch^m \approx \frac{\bar{F}(h) - \bar{F}(Kh)}{K^m - 1}. \tag{7.27}$$

Таким образом, расчет на двух сетках позволяет приближенно (с точностью до членов более высокого порядка) оценить точность аппроксимации на сетке с шагом h.

Найденную погрешность, хотя и приближенно, можно исключить из равенства (7.26) и получить результат с более высокой точностью:

$$\tilde{\bar{F}}(h) = \bar{F}(h) + \tilde{r}(h). \tag{7.28}$$

Указанная методика оценивания погрешности аппроксимации и повышения ее точности носит имя Рунге. Интересно отметить, что $\tilde{\bar{F}} \notin [\bar{F}(h), \bar{F}(Kh)]$, так что формула (7.28) представляет собой экстраполяцию.

Обычно производят так называемый двойной пересчет, полагая $K = 2$. Тогда формула (7.28) имеет вид:

$$F \approx \tilde{\bar{F}}(h) = \bar{F}(h) + [\bar{F}(h) - F(2h)]/(2^m - 1). \tag{7.29}$$

Величины порядка малости погрешностей различных методов указаны в §§ 5.4, 5.5, 7.2, 7.3, 7.4.

Упражнение 7.4. Уточните результат упражнения 7.1 методом Рунге, используя формулу (7.29) с $m=2$ и результат примера 7.2.

Ответ: $\tilde{\bar{I}} = 1{,}7272 + (1{,}7539 - 1{,}7272)/3 = 1{,}7183$.

Пример 7.6. Оценим точность результата I примера 7.5 методом Рунге. Для этого вычислим сначала тот же интеграл по формуле Симпсона (7.22) с удвоенным шагом $h=0{,}2$, принимая во внимание только следующую таблицу значений $y=f(x)$:

x	0	0,2	0,4	0,6	0,8
y	1,0000	0,9801	0,9211	0,8253	0,6967
n	0	1	2	3	4

$I=(2/30)(1{,}0000+4\cdot 1{,}8054+2\cdot 0{,}9211+0{,}6967)=0{,}71737$. Далее воспользуемся формулой (7.27), полагая $K=2$ и $m=4$; $r\approx|0{,}71735-0{,}71737|/15=2\cdot 10^{-5}/15\approx\approx 10^{-6}$.

Иногда порядок малости для данной аппроксимации существует, но неизвестен. Тогда его можно найти по *схеме Эйткена*. Для этого надо оценить искомую величину F на трех сетках с шагами h, Kh и K^2h. Обозначим соответствующие результаты $\bar{F}_1$, $\bar{F}_2$ и $\bar{F}_3$. Можно доказать [9], что $K^m\approx\approx(\bar{F}_3-\bar{F}_2)/(\bar{F}_2-\bar{F}_1)$ и потому

$$m\approx\tilde{m}=(\ln K)^{-1}\ln[(\bar{F}_3-\bar{F}_2)/(\bar{F}_2-\bar{F}_1)]. \qquad (7.30)$$

Точность этой формулы увеличивается при $h\to 0$. Можно m не вычислять, а сразу получать уточненный результат по формуле

$$\tilde{\bar{F}}=\bar{F}_1+(\bar{F}_2-\bar{F}_1)^2/(2\bar{F}_2-\bar{F}_1-\bar{F}_3). \qquad (7.31)$$

Пример 7.7. При вычислении интеграла в примере 7.4 отмечалось, что вторая производная подынтегральной функции не ограничена. Следовательно, для формулы трапеций $m\neq 2$. Оценим эту величину по схеме Эйткена. Для этого обозначим результат 0,4070, полученный в примере 7.4, через $\bar{I}_1$ и вычислим тот же интеграл по формуле трапеций еще дважды — с шагом 0,5 и с шагом 1. Получим соответственно $\bar{I}_2=0{,}5\ (0+0{,}3536+\\+0{,}5)=0{,}4268$ и $\bar{I}_3=1\,(0+0{,}5)=0{,}5$. Вычисления по формуле (7.29) дают $\tilde{m}=1{,}89$. Уточненное значение (7.31) оказывается равным $\tilde{\bar{I}}=0{,}3997$. Истинная ошибка $|\tilde{\bar{I}}-I|=3\cdot 10^{-4}$ более чем в 20 раз меньше истинной ошибки результата $\bar{I}_1$ из примера 7.4.

Упражнение 7.5. Пользуясь схемой Эйткена, оцените порядок малости погрешности интегрирования относительно шага сетки для квадратурной формулы трапеций при вычислении интегралов из упражнений 7.1 и 7.2 с шагами $h=0{,}25$, $2h=0{,}5$, $4h=1$. Убедитесь, что в силу различной гладкости подынтегральных функций результаты отличаются от теоретического значения $m=2$ по-разному.

Ответ: $m=1{,}98$ для $f(x)=e^x$ и $\tilde{m}=1{,}38$ для $f(x)=\sqrt{x}$.

Аналогично метод Рунге и схема Эйткена могут быть использованы для оценивания точности и последующего уточнения результата при решении других задач аппроксимации на сетке.

Для оценки погрешности интегрирования используется также следующий способ: интеграл вычисляется по двум различным квадратурным формулам и полученные результаты сравниваются. Совпадающие цифры старших разрядов считаются верными, а погрешность интегрирования — не превосходящей единицы последнего из совпадающих разрядов.

Если с ошибкой задан какой-нибудь узел интегрирования, то значение интеграла изменяется примерно на величину этой ошибки, умноженную на производную подынтегральной функции и весовой коэффициент в этом узле.

§ 7.6. ВЫБОР ШАГА ИНТЕГРИРОВАНИЯ

Ограничимся квадратурными формулами с постоянным шагом интегрирования h. До сих пор мы задавались некоторым шагом и затем оценивали точность численного интегрирования. Часто точность нужного интеграла известна заранее и требуется подобрать наибольший шаг, который обеспечивает заданную точность ε при использовании выбранной квадратурной формулы. Если значения подынтегральной функции приобретаются из измерений, что наиболее типично для геодезических и геофизических приложений, то выбор оптимального шага сетки оказывается очень важным, поскольку определяет собой, по существу, затраты на выполнение полевых работ.

Погрешность интегрирования по избранной квадратурной формуле складывается в основном из погрешности за счет дискретности подынтегральной функции и за счет ошибок в ее исходных значениях в узлах сетки. Обе эти погрешности должны быть согласованы между собой.

Рассмотрим два способа решения поставленной задачи.

Выбор шага по оценке остаточного члена. Пусть квадратурная формула для вычисления нужного интеграла с точностью ε выбрана. Используя формулу соответствующего остаточного члена r, подбираем h таким, чтобы выполнялось неравенство $|r| \leqslant \varepsilon/2$. Определив шаг, дальнейшие вычисления планируем проделывать с таким количеством значащих цифр, которое обеспечивает погрешность округления не более $\varepsilon/2$.

Неменьшее количество значащих цифр обязаны иметь и исходные значения функции, что определяет собой требование к точности их получения.

Указанный принцип равных влияний не обязателен. Его изменение зависит от того, что в данной задаче обходится, вообще говоря, дороже: увеличение количества исходных значений функции (измельчение шага) или увеличение их точности.

Пример 7.8. Пусть требуется вычислить $I=\int_0^1 e^x dx$ по формуле Симпсона с точностью не хуже $\varepsilon=0{,}001$. Здесь значения интегрируемой функции легко получать с высокой точностью, в частности с точностью до 4-го верного знака после запятой. Поэтому в качестве шага можно взять решение $|r| \leqslant \varepsilon$, где $|r|$ определяется правой частью неравенства (7.23): $h^4 \cdot 2{,}71/180 \leqslant \leqslant 0{,}001$, поскольку очевидно, что $M_4=e$. Решая неравенство, получаем $h \leqslant \leqslant 0{,}51$. При использовании формулы Симпсона шаг должен укладываться на отрезке интегрирования $[0,1]$ четное число раз. Поэтому окончательно выбираем значение шага $h=0{,}50$.

Упражнение 7.6. Известно, что $f''(x) \leqslant 6$ при $x \in [0,1]$, а значения $f(x)$ можно измерить не точнее чем до 0,01. С каким шагом надо выполнять измерения значений функции $f(x)$, чтобы вычислить $I=\int_0^1 f(x)dx$ по формуле прямоугольников с ошибкой, не превышающей $\varepsilon=0{,}02$?

Ответ: $h=0{,}2$.

Основной недостаток указанного способа состоит в том, что максимальный модуль производной нужного порядка, участвующий в выражении для предельной ошибки интегрирования, обычно трудно найти.

Более практичным часто оказывается другой способ, основанный на методе Рунге.

Двойной пересчет. Нужный интеграл вычисляют по выбранной квадратурной формуле дважды: с некоторым шагом h и затем с шагом $h/2$, т. е. удваивают номер последнего узла n. Обозначив результаты соответственно $I(h)$ и $I(h/2)$, оценивают главную часть погрешности последнего результата по формуле (7.29)

$$|r(h/2)| \approx |\tilde{r}(h/2)| = \frac{1}{2^m - 1} \times |I(h) - I(h/2)|, \qquad (7.32)$$

где $m=2$ для формул прямоугольников и трапеций и $m=4$ для формулы Симпсона. Если $\tilde{r}(h/2) < \varepsilon$, где ε — допустимая погрешность, то полагают $I \approx I(h/2)$ и вычисления заканчиваются. Если $\tilde{r}(h/2) \geqslant \varepsilon$, то шаг уменьшается еще в 2 раза и вычисляется $I(h/4)$. Главная часть $\tilde{r}(h/4)$ погрешности оценки $I(h/4)$ находится из сопоставления $I(h/4)$ с $I(h/2)$ и снова сравнивается с ε. Такой прием особенно удобен при вычислении на ЭВМ интегралов с аналитически заданной подынтегральной функцией, поскольку он позволяет автоматически подобрать шаг, обеспечивающий заданную точность с одновременным контролем вычислений. Если имеются результаты счета при трех шагах, то полезно для контроля вычислить по формуле (7.30) величину $\tilde{m}$ и сравнить ее с теоретическим значением m. Большое расхождение свидетельствует либо о недостаточной для используемой квадратурной формулы гладкости интегрируемой функции, либо об ошибке в программе.

Пример 7.9. Решим задачу из примера 7.8 применительно к формуле трапеций при $\varepsilon=0{,}01$. Начнем вычисления с максимально большого шага $h=1$. По формуле (7.15) легко получить $I(1)=1{,}8591$. Далее считаем тот же интеграл с шагом 0,5. Имеем (см. пример 7.2) $I(0{,}5)=1{,}7539$. Согласно формуле (7.32), $\tilde{r}(0{,}5)=0{,}1052/3=0{,}035>\varepsilon=0{,}01$. Поэтому проделаем новые вычисления с шагом 0,25 (см. упражнение 7.1). Точность результата $I(0{,}25)=1{,}7272$ снова оценим по формуле (7.32): $\tilde{r}(0{,}25)=0{,}0267/3=0{,}009<\varepsilon=0{,}01$. Требуемая точность обеспечена. Для контроля вычислим величину $\tilde{m}$ по формуле (7.30). Полученное значение $\tilde{m}=1{,}98$ (см. упражнение 7.5) почти совпадает с теоретическим значением $m=2$.

В качестве окончательного значения интеграла полезно брать уточненное в соответствии с формулой (7.29) число $\tilde{\tilde{I}}$ (см. упражнение 7.4).

В заключение этого параграфа отметим следующую общую рекомендацию. Если есть возможность подробно анализировать интегрируемую функцию, то шаг интегрирования полезно выбрать меньше расстояния между соседними нулями функции и

ее производной. Полезно разбивать отрезок интегрирования на части, где $f(x)$ и $f'(x)$ знакопостоянны, и интегрировать по каждой такой части отдельно со своим шагом.

Простейшая, но грубая прикидка требуемого шага может быть выполнена по формуле $h \approx \sqrt[m]{\varepsilon}$.

§ 7.7. КВАДРАТУРНЫЕ ФОРМУЛЫ НАИВЫСШЕЙ АЛГЕБРАИЧЕСКОЙ ТОЧНОСТИ

Случай, когда положение узлов можно выбирать. При составлении конкретных квадратурных формул вида (7.5) в §§ 7.2—7.4 мы всегда заранее задавали узлы, не придавая особого значения их расположению на отрезке интегрирования, и затем находили соответствующие веса. Такой порядок автоматически обеспечивал алгебраический порядок точности квадратуры не меньше n (см. замечание 7.2). Однако иногда он оказывался равным n, а иногда — больше n. Вспомним, например, формулу средних прямоугольников (7.8), где $n=0$. В § 7.2 подчеркивалось (см. замечание 7.3), что точность (7.8) неожиданно оказалась выше ожидаемой за счет удачного выбора единственного узла интегрирования. В общем случае, когда на отрезке интегрирования имеется $k=n+1$ узлов, но располагать их допускается как угодно, точность квадратурной формулы можно существенно повысить за счет разумного выбора сетки. Формула (7.5) содержит $2(n+1)=2k$ параметров (это веса и узлы). Столько же коэффициентов имеет всякий полином степени $2k-1$. Следовательно, можно так подобрать параметры, чтобы квадратурная формула (7.5) была точна для любого полинома степени не выше $2k-1$. Если этот предел достигнут, то соответствующую квадратурную формулу можно назвать *формулой наивысшей алгебраической точности.*

Чтобы уяснить себе метод, позволяющий достигнуть поставленную цель, вернемся к § 7.2 и проанализируем причину неожиданного повышения точности. Итак, пусть $n=0$, $k=1$ и $x_0=\bar{x}=(a+b)/2$. Предположим, что известно значение интегрируемой функции еще в одном узле. Обозначим этот узел x_1. Его положение на отрезке $[a, b]$ произвольно, лишь бы $x_1 \neq x_0$. Два узла позволяют нам построить интерполяционную квадратурную формулу, порядок точности которой равен единице. Ее весовые коэффициенты определяются формулой (7.4):

$$c_0^* = \int_a^b \Lambda_0^*(x)\,dx,$$

$$c_1 = \int_a^b \Lambda_1(x)\,dx.$$

Здесь* указывает, что речь идет о полиноме влияния, измененного за счет появившегося узла x_1. Однако если x_0 находится

в середине отрезка $[a, b]$, то вычисления по квадратурной формуле с двумя узлами x_0, x_1 и по квадратурной формуле (7.8) с одним узлом дают то же самое. В самом деле

$$\Lambda_0^*(x)=\Lambda_0(x)\frac{x-x_1}{x_0-x_1}=\Lambda_0(x)\left(1+\frac{x-x_0}{x_0-x_1}\right), \qquad (7.33)$$
$$\Lambda_1(x)=(x-x_0)/(x_1-x_0),$$

и поэтому

$$c_0^*=c_0, \text{ а } c_1=0. \qquad (7.34)$$

Это означает, что порядок точности формулы с одним узлом $x_0=\bar{x}$ равен единице, т. е. тот же, что и для формулы с двумя узлами!

Причина состоит в том, что $x_0=\bar{x}$, а

$$\int_a^b (x-\bar{x})dx=0. \qquad (7.35)$$

Равенство (7.35) легко проверить непосредственно. Оно является прямым следствием симметричного расположения узла $x_0=\bar{x}$, но мы проинтерпретируем это равенство в более сложной форме: $\bar{x}$ является корнем полинома 1-й степени $\Pi_1(x)=x-\bar{x}$, ортогонального любому полиному 0-й степени (т. е. константе). Ортогональность понимается в смысле скалярного произведения $(\psi_1, \psi_2)=\int_a^b \psi_1(x)\psi_2(x)dx$ (см. §§6.1, 6.3). Преимущество такой формулировки условия (7.35) состоит в том, что ею можно руководствоваться в принципе и для оптимального распределения любого числа $k=n+1$ узлов интегрирования.

Переходим к общему случаю, когда на отрезке интегрирования можно выбрать количество узлов $k=n+1>1$. Как надо расположить k узлов, чтобы соответствующая им квадратурная формула (7.5) с весовыми коэффициентами (7.4) имела наивысший алгебраический порядок точности $2k-1$? Ответом может служить обобщение вышесформулированного: в качестве k узлов $x_1, x_2, \ldots, x_k$ надо взять k корней такого полинома $\Pi_k(x)$, который ортогонален любому другому полиному степени не выше $k-1$, т. е.

$$\int_a^b \Pi_k(x)x^\varkappa dx=0; \qquad \varkappa=0, 1, \ldots, k-1. \qquad (7.36)$$

Доказательство можно выполнить по следующей схеме. Обозначим через $\Lambda_1(x), \ldots, \Lambda_k(x)$ полиномы влияния (5.14) оптимально располагаемых узлов, и пусть в нашем распоряжении имеется еще k дополнительных узлов $x_{k+1}, \ldots, x_{2k}$, расположенных произвольно на $[a, b]$. Полиномы влияния, соответствующие всем $2k$ узлам, обозначим $\Lambda_1^*(x)$, ..., $\Lambda_k^*(x)$, $\Lambda_{k+1}(x)$, ...,

$\Lambda_{2k}(x)$. Можно проверить, что квадратурная формула с $2k$ узлами, имеющая порядок точности $2k-1$, дает точно такие же результаты интегрирования, как и квадратурная формула с k узлами, расположенными в соответствии с правилом (7.36), поскольку

$$c_i^* = \int_a^b \Lambda_i^*(x)dx = \int_a^b \Lambda_i(x)dx = c_i, \quad i=1,\ldots,k, \tag{7.37}$$

$$c_i = \int_a^b \Lambda_i(x)dx = 0 \text{ при } i=k+1,\ldots,2k. \tag{7.38}$$

(Равенства (7.34) являются частным случаем.)

В самом деле, согласно формуле (5.14),

$$\Lambda_i(x) = \Pi_k(x)\lambda_{k-1}(x),$$

где $\lambda_{k-1}(x)$ — некоторый полином степени $k-1$, а $\Pi_k(x) = (x-x_1)(x-x_2), \ldots, (x-x_k)$ — полином k-й степени, имеющий своими корнями k первых узлов интегрирования. Поэтому равенство (7.38) сразу следует из условия (7.36).

Для доказательства равенства (7.37) заметим, что

$$\Lambda_i^*(x) = \Lambda_i(x)\frac{x-x_{k+1}}{x_i-x_{k+1}}\frac{x-x_{k+2}}{x_i-x_{k+2}}\ldots\frac{x-x_{2k}}{x_i-x_{2k}}, \tag{7.39}$$

но

$$\frac{x-x_{k+\varkappa}}{x_i-x_{k+\varkappa}} = 1+\frac{x-x_i}{x_i-x_{k+\varkappa}}, \quad \varkappa=1,\ldots,k.$$

Поэтому (7.39) можно записать в виде

$$\Lambda_i^*(x) = \Lambda_i(x)+\Pi_k(x)\lambda_{k-1}(x), \tag{7.40}$$

откуда и следует, учитывая (7.36), равенство (7.37). Обратите внимание, что выражение (7.33), является частным видом общего выражения (7.40).

Полиномы $\Pi_k(x)$, удовлетворяющие условиям (7.36), хорошо известны и называются *полиномами Лежандра*. Имеются точные таблицы корней полиномов Лежандра для разных k. Взяв эти корни в качестве узлов интегрирования, весовые коэффициенты вычисляют по обычной формуле (7.4) (см. замечание 7.1). Впервые формулы наивысшей алгебраической точности построил Гаусс. Результаты вычислений узлов и весов квадратурных формул имеются в различных справочниках, например в [9]. Вычисления выполнены для стандартного отрезка интегрирования $[a,b]=[-1,1]$, и потому результаты зависят только от количества узлов k. Узлы расположены на интервале $(-1,1)$ всегда симметрично относительно нуля. Все весовые коэффици-

енты положительны. Сумма их равна двум. При вычислении интеграла $I=\int_a^b f(x)dx$ следует сделать замену переменной

$$x=\frac{b+a}{2}+\frac{b-a}{2}t. \tag{7.41}$$

Тогда квадратурная формула Гаусса имеет вид

$$I\approx\tilde{I}=\frac{b-a}{2}\sum_{i=1}^{k}c_i f(x_i),\ x_i=\frac{b+a}{2}+\frac{b-a}{2}t_i. \tag{7.42}$$

Погрешность интегрирования характеризуется остаточным членом

$$r^*_k=\left(\frac{b-a}{2}\right)^{2k+1}r_k. \tag{7.43}$$

Ниже приводятся узлы t_i, веса c_i и выражения r_k для некоторых значений k:
при $k=1$: $t_1=0$, $c_1=2$, $r_1=f''(\xi)/3$, $-1<\xi<1$ (совпадает с формулой среднего прямоугольника);
при $k=2$: $-t_1=t_2=0{,}577350269$, $c_1=c_2=1$, $r_2=f^{(4)}(\xi)/135$;
при $k=3$: $-t_1=t_3=0{,}774596669$, $t_2=0$, $c_1=c_3=5/9$, $c_2=8/9$, $r_3=f^{(6)}(\xi)/15750$;
при $k=4$: $-t_1=t_4=0{,}861136312$, $-t_2=t_3=0{,}339981044$, $c_1=c_4=0{,}347854845$, $c_2=c_3=0{,}652145155$, $r_4=f^{(8)}(\xi)/3472875$;
при $k=5$: $-t_1=t_5=0{,}906179846$, $-t_2=t_4=0{,}538469310$, $c_1=c_5=0{,}236926885$, $c_2=c_4=0{,}478628670$, $t_3=0$, $c_3=0{,}568888889$, $r_5=f^{(10)}(\xi)8\cdot10^{-10}$.

Пример 7.10. По формуле Гаусса вычислить интеграл $I=\int_0^1 dx/(1+x^2)$, взяв три узла.

Решение. Согласно формуле (7.42),

$$I\approx\tilde{I}=\frac{1}{2}\left[\frac{5}{9}f(x_1)+\frac{8}{9}f(x_2)+\frac{5}{9}f(x_3)\right].$$

Здесь $x_1=0{,}5-0{,}774597/2=0{,}112702$; $x_2=0{,}5$; $x_3=0{,}5+0{,}774597/2=0{,}887298$. В результате $\tilde{I}=0{,}7853$. Точное значение $I=0{,}7854$. Относительная погрешность меньше 0,02%.

Квадратурная формула Гаусса дает очень высокую точность даже при небольшом количестве узлов, но лишь в том случае, когда интегрируемая функция имеет достаточно высокие производные. Если, например, интегрируемая функция имеет ограниченной только 4-ю производную, то формула Гаусса с $k=4$ и $b-a=2$ характеризуется предельной погрешностью порядка $0{,}0003\,M_4$, что в 2 раза меньше погрешности формулы Симпсона с $h=0{,}5$. С уменьшением порядка ограниченной производной на единицу коэффициент пропорциональности предельной погреш-

ности увеличивается примерно на порядок. Для функций малой гладкости формула Гаусса может дать меньшую точность, чем формула трапеций.

Упражнение 7.7. Вычислите $I=\int\limits_{-1}^{1}|x|dx$ по формуле трапеций с $h=1$ и по формуле Гаусса с тремя узлами. Сравните результаты с точным значением $I=1$.

Ответ: формула трапеций дает точный результат, а формула Гаусса — ошибку, равную 0,14.

При применении формулы Гаусса трудно оценить точность, что является ее недостатком. Приведенные выше выражения для остаточного члена r применяются редко, так как необходимо знать максимальные значения производных высокого порядка. Методом Рунге пользоваться также затруднительно, поскольку при удвоении числа узлов ранее вычисленные значения функции не нужны — узлы гауссовых квадратур с k и $2k$ узлами не совпадают между собой.

Для того чтобы воспользоваться методом Рунге без существенного увеличения объема вычислений, приходится ответить на следующий вопрос. Пусть некоторый интеграл вычислен по формуле Гаусса с k узлами. Как построить новую, возможно более точную квадратурную формулу с $2k+1$ узлами так, чтобы использовать уже имеющиеся k узлов, а также вычисленные значения функции в этих узлах и добавить лишь $k+1$ новых? Решением этого вопроса является уточняющая квадратурная формула, предложенная А. С. Кронродом. Значения соответствующих узлов и весовых коэффициентов можно найти в его монографии «Узлы и веса квадратурных формул» (М., Наука, 1964).

Обобщение формул Гаусса на интегралы вида $I=\int\limits_{a}^{b}p(x)f(x)dx$ выполнил Кристоффель. Здесь $p(x)$ — известная функция, называемая весовой; предполагается, что $p(x)>0$ и непрерывна при $x\in(a,b)$. На концах этого отрезка $p(x)$ может обращаться в нуль или в бесконечность, но $\int\limits_{a}^{b}p(x)dx<\infty$. Квадратурная формула Гаусса — Кристоффеля имеет вид (7.42). Узлы t_i расположены на стандартном интервале $(-1,1)$ и представляют собой корни полиномов $\Pi_{p,k}$ степени k, ортогональных на $[a,b]$ с весом $p(x)$ в том смысле, что $\int\limits_{-1}^{1}\Pi_{p,k}(x)\times\Pi_{p\,m}(x)p(x)dx=0$ при $k\neq m$. Весовые коэффициенты вычисляются по формуле $c_i=\int\limits_{-1}^{1}p(x)\Lambda_i(x)dx$. Можно доказать, что $c_i>0$ и $\sum\limits_{i=1}^{k}c_i=\int\limits_{-1}^{-1}p(x)dx$. Формулы Гаусса являются частным случаем, соответствующим $p(x)=1$.

Пример 7.11. $I=\int_{-1}^{1}(f(x)/\sqrt{1-x^2})dx$, $p(x)=1/\sqrt{1-x^2}$. Роль Π_k играют полиномы Чебышева первого рода. Узлы интегрирования $t_i=\cos[\pi(i-1/2)/k]$, где $i=1, \ldots, k$. Все весовые коэффициенты одинаковы, $c_i=\pi/k$. Погрешность характеризуется неравенством $|r|\leqslant\pi\, M_{2k}[2^{k-1}(2k)!]$. Соответствующую квадратурную формулу часто называют *формулой Эрмита*. Она позволяет вычислять несобственные интегралы. Пусть, например, $f(x)=1/\sqrt{1+x^2}$ и $k=6$. Тогда

$$\tilde{I}=\frac{\pi}{6}\left[\frac{2}{\sqrt{1+\cos^2 15^\circ}}+\frac{2}{\sqrt{1+\cos^2 45^\circ}}+\frac{2}{\sqrt{1+\cos^2 75^\circ}}\right]=$$
$$=2{,}22133.$$

Точное значение $I=2{,}22144$.

Пример 7.12. В теории фигуры Земли часто приходится иметь дело с интегралами вида $I=\int_0^R \rho^m F(\rho)d\rho$, которые возникают при вычислении двойных интегралов $\int\int_\sigma f(\rho, \varphi)\rho d\rho d\varphi$ в полярных координатах ($m=1$) и тройных интегралов $\int\int\int_\sigma f(\rho, \varphi, \theta)\rho^2\cdot\sin\theta d\rho d\theta d\varphi$ в шаровых координатах ($m=2$). При помощи подстановки $\rho=R(1-x)/2$ интеграл I приводится к виду

$$I=\left(\frac{R}{2}\right)^{m+1}\int_{-1}^{1}(1-x)^m f(x)\,dx,\ f(x)=F(\rho),$$

удобному для использования формул Гаусса — Кристоффеля с весовой функцией $p(x)=(1-x)^m$.

Узлы и весовые коэффициенты квадратурных формул Гаусса — Кристоффеля для наиболее распространенных на практике весовых функций $p(x)$ можно найти в учебном пособии [9].

Случай, когда положение узлов фиксировано. Предположим теперь, что k узлов $x_0, x_1, \ldots, x_n$, где $n=k-1$, произвольным образом фиксированы на отрезке интегрирования и их положение невозможно изменить. Как выбирать весовые коэффициенты, чтобы квадратурная формула (7.5) имела максимально возможную алгебраическую точность? Один путь решения этой задачи уже известен: надо пользоваться формулами (7.4), (5.14). Так мы и поступали при выводе формул (7.8), (7.15), (7.20), соответствующих значениям $n=0$, 1, 2. Однако при больших n интегрирование (7.4) становится громоздким и удобнее поступать следующим образом.

При фиксированных узлах в формуле (7.5) остается $n+1$ свободных параметров — весовых коэффициентов c_i. Столько же коэффициентов имеет любой полином степени n. Обеспечение n-й степени точности квадратурной формулы (7.5) равносильно требованию о том, чтобы эта формула была точна на всех степенных функциях x^j, степень j которых не превосхо-

дит n. Следовательно, искомые весовые коэффициенты должны быть корнями C следующей системы $n+1$ линейных уравнений:

$$\underset{(n+1)\cdot(n+1)}{A}\ \underset{(n+1)\cdot 1}{C} = \underset{(n+1)\cdot 1}{b}, \tag{7.44}$$

где

$$b_j = \int_a^b x^j dx = \frac{b^{j+1} - a^{j+1}}{j+1},$$

A — квадратная матрица с элементами $a_{ij} = x^i{}_j$.

Определитель этой матрицы есть определитель Вандермонда и отличен от нуля [см. формулу (5.14)]. Решение системы дает нужные весовые коэффициенты c_i формулы (7.5).

Упражнение 7.8. Вычислите $I = \int_0^1 f(x)dx$, где $f(x)$ — таблично заданная функция из примера 5.1. Указание: весовые коэффициенты c_0, c_1, c_2, c_3 определите из решения системы четырех уравнений вида (7.44).

Заметим, что описанный способ составления квадратурных формул применим и для решения задач § 7.7. Однако система (7.44) оказывается уже системой нелинейной с $2(n+1)$ неизвестными узлами и весами, и решить ее затруднительно.

§ 7.8. КРАТКИЙ ОБЗОР ДРУГИХ ФОРМУЛ ЧИСЛЕННОГО ИНТЕГРИРОВАНИЯ В СОБСТВЕННОМ СМЫСЛЕ

Аппроксимируя подынтегральную функцию интерполяционным полиномом Лагранжа степени $n=0$, 1, 2, мы построили в §§ 7.2, 7.3, 7.4 квадратурные формулы соответственно прямоугольников, трапеций и парабол. Аналогично можно действовать и с более высокой степенью интерполяции $n>2$. Соответствующие квадратурные формулы с постоянным шагом называются формулами Котеса. Применяются они редко.

В геодезической практике значения интегрируемой функции часто подвержены случайным ошибкам, поскольку получены из измерений. Результат интегрирования $\tilde{I}$ по формуле (7.5) имеет тогда наименьшую дисперсию при равных весовых коэффициентах (если их сумма фиксирована). Соответствующие квадратурные формулы разработаны П. Л. Чебышевым. Узлы в этих формулах подобраны так, что обеспечивается максимальная при равных весовых коэффициентах алгебраическая точность (на единицу больше количества узлов k, если k нечетно, и на две единицы — если k четно; $k=1, 2, \ldots, 7, 9$).

Формулы наивысшей алгебраической точности Гаусса являются формулами открытого типа. При необходимости включить в число узлов один или оба конца отрезка интегрирования можно пользоваться формулами А. А. Маркова. Порядок их точно-

сти на 2 или 3 меньше порядка точности соответствующей формулы Гаусса. Для интегрирования быстроизменяющихся функций с большими значениями производных рекомендуется специальный метод Филона [9].

Для случая, когда в узлах интегрирования известны не только значения функции, но и значения ее первой производной, применяются квадратурные формулы Эйлера. Обобщение их можно получить, заменяя производные разностными выражениями (5.4). Соответствующие формулы называются формулами Грегори. Формула Эйлера является частным примером решения задачи приближенного интегрирования в условиях, когда о функции известны не только ее значения на сетке, но и другие функционалы на ней. В общем случае приближенное вычисление определенного интеграла представляет собой одну из локальных задач коллокации: на функции $f(x)$ заданы несколько линейных функционалов, надо оценить новый функционал на этой функции — интеграл (см. § 5.6, в частности пример 5.5). Одно из возможных решений задачи численного интегрирования с помощью сплайна минимальной нормы рассмотрено в упражнении 5.5.

Упражнение 7.9. Вычислите $I=\int_{-\pi}^{\pi} e^x dx$ с шагом $h=\pi/2$ по формуле, полученной в упражнении 5.5, полагая $\nu=1$ и $l_i=e^{x_i}i$, $x_i=-\pi+i(\pi/2)$; $i=0, 1, \ldots, 4$.

Ответ: $I=23{,}097467$; точное значение $I=23{,}097476$.

Хорошие результаты получаются при использовании дифференциальных сплайнов. В этом смысле все сказанное в предыдущих двух главах относительно коллокации имеет непосредственное отношение к задачам приближенного интегрирования.

§ 7.9. ПРИБЛИЖЕННОЕ ВЫЧИСЛЕНИЕ НЕСОБСТВЕННЫХ ИНТЕГРАЛОВ

Несобственные интегралы 1-го рода (с бесконечными пределами) обычно сводят к интегралам либо с помощью подстановки, либо искусственным усечением луча интегрирования. Что касается методов численного интегрирования, то они остаются, по существу, без изменений по сравнению с уже разобранными в этой главе.

Поэтому мы будем рассматривать приближенное вычисление несобственных интегралов только 2-го рода (от неограниченных функций) со степенной особенностью. Именно такие интегралы представляют наибольший интерес с точки зрения геодезических приложений.

Аддитивное выделение особенности. Наиболее общий прием вычисления несобственных интегралов 2-го рода предложен Л. В. Канторовичем и называется *аддитивным выделением осо-*

бенности. Он состоит в следующем. Из подынтегральной функции $f(x)$ выделяют в отдельное слагаемое некоторую функцию $\varphi(x)$, имеющую ту же особенность, но элементарно интегрируемую. При этом разность $f(x)-\varphi(x)$ должна оказаться уже достаточно гладкой, чтобы воспользоваться какой-либо стандартной квадратурной формулой. Таким образом, исходный интеграл разбивается на два интеграла (I_1+I_2)

$$\int_a^b f(x)\,dx=\int_a^b \varphi(x)\,dx+\int_a^b [f(x)-\varphi(x)]\,dx, \tag{7.45}$$

из которых I_1 вычисляется аналитически, а I_2 — численно, но уже известными нам способами (§§ 7.1—7.8).

Пример 7.13. $I=\int_0^{0,5} dx/\sqrt{x(1-x)}$. Здесь $f(x)=1/\sqrt{x}\sqrt{1-x}=(1/\sqrt{x})+(1/\sqrt{x}\sqrt{1-x}-1/\sqrt{x})=(1/\sqrt{x})+(1-\sqrt{1-x}/\sqrt{x}\sqrt{1-x})$. Потому $I=I_1+I_2$, где $I_1=\int_0^{0,5} dx/\sqrt{x}=\sqrt{2}$, а $I_2=\int_0^{0,5} 1-\sqrt{1-x}/\sqrt{x}\sqrt{1-x}\,dx$ можно вычислить, например, по формуле (7.8): $I_2=0{,}1547$. Окончательно $I=I_1+I_2=1{,}5689$. Точное значение $I=\pi/2=1{,}5708$.

Рассмотрим систематический прием построения функции $\varphi(x)$ в разложении (7.45) для случая, когда

$$f(x)=\frac{g(x)}{(x-c)^{\beta}}, \tag{7.46}$$

$$\text{где } 0<\beta<1, \quad a\leqslant c\leqslant b,$$

а $g(x)$ — многократно дифференцируемая функция на $[a,b]$.

Пусть при вычислении I_2 предполагается использовать квадратурную формулу, требующую, чтобы функция $f(x)-\varphi(x)$ имела непрерывную производную по крайней мере k-го порядка. Разложим $g(x)$ в конечный ряд Тейлора в окрестности особой точки c, ограничиваясь членом $(k+1)$-го порядка, и обозначим его сумму $g_{k+1}(x)$, т. е.

$$g_{k+1}(x)=g(c)+g'(c)(x-c)+\ldots+\frac{g^{(k+1)}(x)}{(k+1)!}(x-c)^{k+1}.$$

Нужную функцию $\varphi(x)$ возьмем в виде

$$\varphi(x)=g_{k+1}(x)/(x-c)^{\beta}. \tag{7.47}$$

При этом интеграл I_1 легко вычислить аналитически как сумму интегралов от степенных функций. Функция $g(x)-g_{k+1}(x)$ имеет порядок малости относительно $(x-c)$ при $x\to c$ более высокий, чем $k+1$. Следовательно, функция

$$f(x)-\varphi(x)=\frac{g(x)-g_{k+1}(x)}{(x-c)^{\beta}}=o[(x-c)^k] \tag{7.48}$$

и поэтому непрерывно дифференцируема не менее k раз.

Пример 7.14. Вычислим по указанной методике интеграл из примера 7.13. Для численного интегрирования I_2 будем использовать формулу трапеции. Поэтому зададим $k=2$. Имеем:

$$f(x)=\frac{(1-x)^{-1/2}}{x^{1/2}},\quad g(x)=(1-x)^{-1/2},\quad \beta=1/2,\quad c=0;$$

$$g_3(x)=1+x/2+3x^2/8+5x^3/16;$$
$$\varphi(x)=x^{-1/2}+x^{1/2}/2+3x^{3/2}/8+5x^{5/2}/16.$$

$I_1=\int_0^{0,5}\varphi(x)dx$ вычисляем аналитически по обычной формуле Ньютона — Лейбница: $I_1=1{,}5665$. $I_2=\int_0^{0,5}[f(x)-\varphi(x)]dx$ считаем по формуле трапеции (7.15): $I_2=0{,}25(0+0{,}0444)=0{,}0111$. Окончательно $I=I_1+I_2=1{,}5776$.

Мультипликативное выделение особенности. Пусть подынтегральная функция $g(x)$ с особой точкой $c=a$ представлена в виде произведения $g(x)=p(x)f(x)$, где $p(x)$ имеет ту же особенность, что и $g(x)$, причем $p(x)>0$ при $x\in[a,b]$ и $\int_a^b p(x)dx<\infty$, а $f(x)$ — достаточно гладкая функция. Тогда можно воспользоваться квадратурными формулами наивысшей алгебраической точности Гаусса — Кристоффеля, поскольку их остаточный член не зависит от $p(x)$, т. е. от разрывного сомножителя функции $f(x)$. Так же можно поступать и в том случае, когда $c=b$ или a и b являются особыми точками одновременно.

Пример 7.15. $I=\int_{-1}^{1}dx/\sqrt{1-x^4}$ имеет две особые точки $+1$ и -1. Но $g(x)=1/\sqrt{1-x^4}=(1/\sqrt{1-x^2})(1/\sqrt{1+x^2})=p(x)f(x)$. Поэтому в качестве весовой функции выбираем первый сомножитель и используем соответствующую формулу Гаусса — Кристоффеля. Практически дело сводится к формуле Эрмита с $f(x)=1/\sqrt{1+x^2}$ (см. пример 7.11).

Упражнение 7.10. Вычислите интеграл из примеров 7.13, 7.14 по формуле Гаусса — Кристоффеля (7.42) с четырьмя узлами. Указание: $p(x)=x^{-1/2}$, $f(x)=(1-x)^{-1/2}$; узлы и весовые коэффициенты взять из таблиц*.

Заметим, что при вычислении интегралов в собственном смысле от недостаточно гладких функций также бывает полезно выделить особенность, но только теперь не самой подынтегральной функции, а ее разрывной производной.

Пример 7.16. Вычислим $I=\int_0^1\sqrt{x}dx=2/3$ по формуле Гаусса (7.42) с двумя узлами: $I=(1/2)(\sqrt{0{,}5+0{,}5\cdot 0{,}57735}+\sqrt{0{,}5-0{,}5\cdot 0{,}57735})=0{,}6739$. Точность невысока, поскольку уже 1-я производная подынтегральной функции терпит разрыв. Запишем I в виде $\int_0^1 p(x)f(x)dx$, полагая $p(x)=\sqrt{x}$, $f(x)=1$,

---

* *Крылов В. И., Шульгина Л. Т.* Справочная книга по численному интегрированию. М., Наука, 1966.

и воспользуемся соответствующей формулой Гаусса — Кристоффеля с двумя узлами. Значения весовых коэффициентов выбираем из таблиц:
$\tilde{I}=0{,}277556\cdot 1+0{,}389111\cdot 1=0{,}666667$.

Аналогичный прием бывает удобно применять и для вычисления несобственных интегралов 1-го рода (с бесконечными пределами).

§ 7.10. ПРИБЛИЖЕННОЕ ВЫЧИСЛЕНИЕ ДВОЙНЫХ ИНТЕГРАЛОВ. ПОНЯТИЕ О КУБАТУРНОЙ ФОРМУЛЕ

Теперь рассмотрим методы приближенного вычисления двойных интегралов по ограниченной плоской области D. Так же как и в предыдущих параграфах этой главы, предположим, что подынтегральная функция $g(P)=g(x, y)$ будет интегрируемой, но такой, что вычислить интеграл аналитически не удается. Это означает, что $g(P)$ либо достаточно сложна, либо ее значения известны только в отдельных точках $P_1, \ldots, P_n$ — узлах сетки области D. Приближенное значение $\tilde{I}$ нужного интеграла вычисляется по формулам вида

$$\iint_D g(P)dxdy=I\approx\tilde{I}=\sum_{i=1}^{n}C_i g(P_i), \tag{7.49}$$

называемым ***кубатурными формулами.*** Задача состоит в разумном выборе узлов интегрирования $P_1, \ldots, P_n$ (если узлами можно распоряжаться) и определении весовых коэффициентов $C_1, \ldots, C_n$. Оценка остаточного члена кубатурной формулы характеризует точность приближенного интегрирования.

В геодезической практике подынтегральная функция $g(x, y)$ часто представляет собой произведение двух функций $p(x, y)\times$ $\times f(x, y)$. При этом $p(x, y)$ — аналитически заданная в области D функция, называемая *весовой.* Она не обязательно будет непрерывной, но предполагается, что существуют ее моменты — интегралы

$$p_{kl}=\iint_D p(x, y)x^k y^l dxdy, \tag{7.50}$$

где $k, l=0, 1, 2\ldots$. Значения функции $f(x, y)$, называемой *обкладкой*, обычно приобретаются из измерений и поэтому известны лишь в узлах некоторой сетки (не обязательно регулярной).

Пример 7.17. В точках $P_1, \ldots, P_n$ области D, показанной на рис. 27, известны значения аномалии силы тяжести $f(x, y)$ (в геодезии эту функцию принято обозначать Δg). В точке Q требуется вычислить в плоской аппроксимации: 1) высоту квазигеоида ζ по формуле Стокса

$$\zeta(Q)=\frac{1}{2\pi\gamma}\iint_D\frac{f(P)}{r(Q, P)}dxdy; \tag{7.51}$$

2) компоненты ξ, η уклонения отвесной линии по формуле Венинг-Мейнеса

$$\begin{Bmatrix}\xi(Q)\\ \eta(Q)\end{Bmatrix}=-\frac{1}{2\pi\gamma}\iint_D\frac{f(P)}{r^2(Q, P)}\begin{Bmatrix}\cos\alpha\\ \sin\alpha\end{Bmatrix}dxdy. \tag{7.52}$$

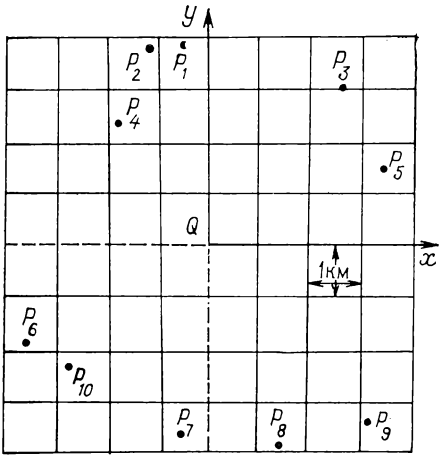


Рис. 27. Расположение узлов с исходными данными для интегрирования по формуле Стокса

Здесь γ — известная константа (нормальная сила тяжести); $r(P, Q)$ — расстояние от текущей точки интегрирования P до фиксированной точки Q; α — угол между осью абсцисс и вектором $\overrightarrow{QP}$. Так как точка Q фиксирована, то r и α представляют собой известные функции двух переменных x, y.

Если $p(x, y) = 1$, то понятия подынтегральной функции и обкладки совпадают: $g(x, y) = f(x, y)$.

Повторное применение квадратурных формул. Пусть область D определена неравенствами

$$y_1(x) \leqslant y \leqslant y_2(x), \qquad a \leqslant x \leqslant b, \qquad (7.53)$$

а весовая функция для простоты положена равной единице. Тогда кубатурную формулу легко построить путем повторного применения уже известных нам квадратурных формул для приближенного вычисления однократных интегралов. В самом деле,

$$I = \int_a^b dx \int_{y_1(x)}^{y_2(x)} f(x, y)\, dy = \int_a^b F(x)\, dx, \qquad (7.54)$$

где

$$F(x) = \int_{y_1(x)}^{y_2(x)} f(x, y)\, dy. \qquad (7.55)$$

Обозначим через $x_0, x_1, \ldots, x_{n_x}$ узлы той квадратурной формулы, которую мы выбрали для вычисления однократного интеграла (7.54), т. е.

$$I = \sum_{j=0}^{n_x} c_{jx} F(x_j). \qquad (7.56)$$

Здесь c_{jx} — известные весовые коэффициенты выбранной квадратурной формулы, а в качестве $F(x_j)$ возьмем приближенные значения соответствующего интеграла (7.55), полученные по той же или другой квадратурной формуле, т. е.

$$F(x_j) \approx \bar{F}(x_j) = \sum_{i=0}^{n_y} c_{iy}^{(j)} f(x_j, y_i^j). \qquad (7.57)$$

Подставляя (7.57) в (7.56), получим кубатурную формулу

$$I = \sum_{j=0}^{n_x} \sum_{i=0}^{n_{yj}} C_{ij} f(x_j,\ y_i^{(j)}), \tag{7.58}$$

где $C_{ij} = c_{jx} c_{iy}$. Если, в частности, D — прямоугольник $[a, b;\ e, d]$, т. е. $y_1(x) \equiv e$ и $y_2(x) \equiv d$ в (7.55), то формулы (7.56), (7.57), (7.58) несколько упрощаются:

$$F(x) = \int_e^d f(x,\ y)dy \approx \bar{F}(x_j) = \sum_{i=0}^{y} c_{iy} f(x_j, y_i), \tag{7.59}$$

$$I = \sum_{j=0}^{n_x} \sum_{i=0}^{n_y} C_{ij} f(x_j,\ y_i), \tag{7.60}$$

где $C_{ij} = c_{jx} c_{iy}$. Геометрически этот метод эквивалентен вычислению объема с помощью поперечных сечений на сетке (см. рис. 14).

Формулы (7.58), (7.60) удобны с точки зрения программирования для ЭВМ, так как достаточно ограничиться стандартными подпрограммами только для одномерных интервалов.

Рассмотрим основные частные случаи кубатуры (7.60).

Предположим сначала, что в прямоугольнике D имеется всего один узел $P_0(x_0,\ y_0)$, расположенный в центре $\bar{P}$: $x_0 = \bar{x} = (b-a)/2$, $y_0 = \bar{y} = (d-e)/2$. Тогда формулу (7.60) легко получить как произведение двух квадратурных формул среднего прямоугольника (7.8):

$$I = sf(\bar{x},\ \bar{y}), \tag{7.61}$$

где $s = (b-a)(d-e)$ — площадь D. Для повышения точности строят составную формулу: область D делится на N прямоугольных ячеек, для каждой такой ячейки применяется формула (7.61) и результаты складываются:

$$I = \sum_{i=1}^{N} s_i f(\bar{x}_i,\ \bar{y}_i). \tag{7.62}$$

Здесь s_i — площадь i-й ячейки; $\bar{x}_i$, $\bar{y}_i$ — координаты ее центра; $N = n_x n_y$, $n_x = (b-a)/h_x$, $n_y = (d-e)/h_y$, где h_x, h_y — шаг интегрирования по x и по y соответственно. Погрешность составной формулы (7.62) имеет вид

$$r \approx \tilde{r} = \frac{1}{24}\left[h_x^2 \iint_D f''_{xx}\, dxdy + h_y^2 \iint_D f''_{yy}\, dxdy\right]. \tag{7.63}$$

Это выражение редко удается использовать на практике, но зато оно показывает, что формула (7.62) имеет второй порядок малости относительно шагов интегрирования. Поэтому для выделения главной части погрешности интегрирования рекоменду-

ется двойной пересчет по схеме Рунге (см. § 7.5). При этом шаги h_x и h_y должны изменяться в одно и то же число раз.

Таким же порядком малости характеризуется и составная кубатурная формула трапеций. Сама формула имеет вид (7.49) с весовыми коэффициентами, равными h_xh_y, $h_xh_y/2$, $h_xh_y/4$ соответственно для внутренних, граничных и угловых узлов сетки. Алгебраический порядок точности кубатурных формул средних и трапеций равен единице.

Кубатурная формула Симпсона имеет вид (7.49). Ее весовые коэффициенты равны $16h_xh_y/9$, $4h_xh_y/9$, $h_xh_y/9$ соответственно для внутреннего, четырех граничных и четырех угловых узлов в случае несоставной формулы ($n=9$). Составную кубатурную формулу Симпсона удобно записывать в виде (7.60). Ее весовые коэффициенты c_{ij} являются соответствующими элементами двумерного массива

$$c=\frac{1}{9}h_xh_y\begin{vmatrix} 1 & 4 & 2 & 4 & 2 & \dots & 4 & 2 & 4 & 1 \\ 4 & 16 & 8 & 16 & 8 & \dots & 16 & 8 & 16 & 4 \\ 2 & 8 & 4 & 8 & 4 & \dots & 8 & 4 & 8 & 2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 2 & 8 & 4 & 8 & 4 & \dots & 8 & 4 & 8 & 2 \\ 4 & 16 & 8 & 16 & 8 & \dots & 16 & 8 & 16 & 4 \\ 1 & 4 & 2 & 4 & 2 & \dots & 4 & 2 & 4 & 1 \end{vmatrix},$$

имеющего n_y+1 строк и n_x+1 столбцов (нумерация с нуля). Предполагается, что узлы интегрирования $P_{ij}(x_j, y_i)$ составляют регулярную прямоугольную сетку с шагом h_x по оси абсцисс и шагом h_y по оси ординат: $n_x=(b-a)/h_x$, $n_y=(e-d)/h_y$ — числа обязательно четные: $x_j=x_0+jh_x$, где $x_0=a$, $j=0, 1, \dots, n_x$; $y_i=y_0+ih_y$, где $y_0=e$, $i=0, 1, \dots, n_y$; количества узлов интегрирования (n_x+1) и (n_y+1) вдоль каждой оси координат представляют собой числа обязательно нечетные.

Формула Симпсона точна на всех полиномах двух переменных степени не выше трех и имеет 4-й порядок малости относительно шага интегрирования.

Пример 7.18. Вычислим $I=\int_0^{\pi/2}\int_0^{\pi/2}\sin(x+y)dxdy$ по кубатурной формуле Симпсона с $h_x=h_y=\pi/8$. Имеем: $n_x=[(\pi/2)-0]/(\pi/8)=4$, $n_y=[(\pi/4)-0]/(\pi/8)=2$, $I=\sum_{i=0}^{4}\sum_{j=0}^{4}c_{ij}\sin[\pi/8)(i+j)]=1{,}0003$. Здесь весовые коэффициенты представляют собой умноженные на $\pi^2/576$ числа, которые на рис. 28 выписаны рядом с узлами интегрирования. Точное значение интеграла $I=1$.

Нетрудно подобрать положение линий сетки и весовые коэффициенты так, чтобы каждая одномерная квадратурная формула имела наивысшую алгебраическую точность, т. е. была бы формулой Гаусса (7.42). На этом пути можно заметно снизить

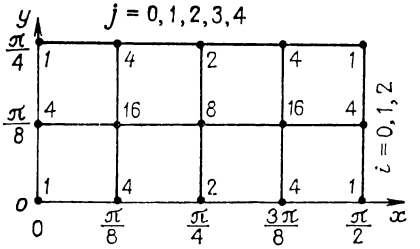


Рис. 28. Весовые коэффициенты кубатурной формулы Симпсона

погрешность интегрирования, но лишь в том случае, когда узлами можно распорядиться, а интегрируемая функция характеризуется достаточно высокой гладкостью.

Интерполяционные кубатурные формулы. Пусть требуется вычислить интеграл вида

$$I = \iint\limits_D p(x, y) f(x, y)\, dx dy, \tag{7.64}$$

где значения функции $f(x, y)$ известны только в отдельных точках $P_1, \ldots, P_n$, образующих в области D некоторую сетку, в общем случае нерегулярную. Для построения кубатурной формулы

$$I = \sum_{i=1}^{n} c_i f(P_i) \tag{7.65}$$

достаточно заменить в (7.64) обкладку f интерполяционным полиномом Лагранжа

$$\varphi(x, y) = \sum_{i=1}^{n} f(P_i) \Lambda_i(P),$$

где $\Lambda_i(P)$ — полином влияния i-го узла. Тогда, очевидно, весовые коэффициенты

$$c_i = \iint\limits_D p(x, y) \Lambda_i(x, y)\, dx dy \tag{7.66}$$

зависят только от весовой функции и положения узлов интегрирования, но не зависят от обкладки. Кубатурная формула (7.65) с весовыми коэффициентами (7.66) называется *интерполяционной*. Ее алгебраический порядок точности a связан с количеством узлов n соотношением

$$(a+1)(a+2) = 2n. \tag{7.67}$$

Алгебраический порядок точности a кубатурной формулы (7.65) означает, что эта формула точна для всех тех случаев, когда $f(x, y)$ представляет собой любой полином двух переменных, степень которого не превышает a.

Для практического определения весовых коэффициентов c_i рекомендуется метод, описанный нами применительно к одномерным интегралам в § 7.7.

Суть состоит в следующем. Вычислим аналитически интегралы (7.50), совпадающие с (7.64) при $f(x, y)=x^k y^l$. Здесь k и l равны 0, 1, 2, ..., a так, чтобы $k+l\leqslant a$, где a — натуральное число, устанавливаемое на основании соотношения (7.67). Знание одного числа p_{kl} дает одно уравнение с n неизвестными c_i:

$$\sum_{i=1}^{n} c_i x_i^k y_i^l = p_{kl}. \tag{7.68}$$

Если a удовлетворяет соотношению (7.67), то получаем систему n таких уравнений. Решив эту систему с n неизвестными, мы и получим нужные весовые коэффициенты. Заметим, что правые части p_{kl} уравнений не зависят от заданных узлов, а зависят только от области интегрирования. Поэтому их можно вычислить заранее для стандартных областей.

Пример 7.19. Пусть в задаче 2 из примера 7.17 область интегрирования D — круг с центром в точке Q и радиусом R. Вычислим соответствующие числа p_{kl}. Опуская постоянный коэффициент $-1/2\pi\gamma$, имеем

$$p_{kl}=\iint\limits_D \frac{x^k y^l \cos\alpha}{r^2}\,dxdy=\int\limits_0^{2\pi}\int\limits_0^R \frac{r^k \cos\alpha^{k+1} r^l (\sin\alpha)^l}{r^2}\,rdrd\alpha=$$

$$=\frac{R^{k+l}}{k+l}\int\limits_0^{2\pi}(\cos\alpha)^{k+1}(\sin\alpha)^l\,d\alpha,$$

где $k+l>0$. Если $k=0, 2, 4, \ldots$, то $p_{kl}=0$ для всякого натурального l. Если $k=1, 3, 5, \ldots$, то $p_{kl}=0$ для всякого нечетного l. Если $k=l=0$, то $p_{kl}=0$ (в смысле главного значения по Коши). Остальные интегралы вычисляются непосредственно. Так, например, если $k+l\leqslant a=4$, то $p_{10}=\pi R$, $p_{12}=\pi R^3/12$, $p_{30}=\pi R^3/4$, все остальные интегралы равны нулю.

Упражнение 7.11. Решите ту же задачу для случая, когда D — квадрат $[-s, s; -s, s]$ $k+l\leqslant a=4$.

Ответ: $p_{kl}=s^{k+l}b_{kl}$, $b_{10}=4\ln(1+\sqrt{2})$, $b_{12}=4[2\ln(1+\sqrt{2})-\sqrt{2}]$, $b_{30}=2[\sqrt{2}-\ln(1+\sqrt{2})]$, все остальные интегралы равны нулю.

Пример 7.20. Вычислить в точке Q по формуле (7.52) составляющую в меридиане уклонения отвеса ξ по значениям аномалии силы тяжести $f(P)$, заданным в 10 точках $P_1, \ldots, P_{10}$, $f_1=f(P_1)=5{,}0$ мгал, $f_2=15{,}1$, $f_3=-23{,}9$, $f_4=12{,}5$, $f_5=-34{,}4$, $f_6=11{,}7$, $f_7=-17{,}4$, $f_8=-32{,}0$, $f_9=-42{,}8$, $f_{10}=-2{,}4$.

Область интегрирования и расположение в ней точек Q, $P_1, \ldots, P_{10}$ указаны на рис. 27.

Решение можно выполнять в следующей последовательности.

1) полагая $n=10$, из соотношения (7.67) определяем $a=3$;

2) вне зависимости от расположения узлов $P_1, \ldots, P_{10}$ вычисляем интегралы (7.50) с $p(x, y)=-(\cos\alpha)2/\pi\gamma r^2=-x/2\pi\gamma(x^2+y^2)^{3/2}$ для всех k и $l=0, 1, 2, 3$, таких, что $k+l\leqslant 3$;

3) составляем 10 уравнений вида (7.68);

4) решив эту систему, получаем весовые коэффициенты: $c_1=-0{,}033$; $c_2=0{,}041$; $c_3=-0{,}268$; $c_4=0{,}131$; $c_5=0{,}374$; $c_6=-0{,}853$; $c_7=-2{,}371$; $c_8=1{,}524$; $c_9=-0{,}488$; $c_{10}=2{,}251$;

5) по формуле (7.65) получаем $\tilde{\xi}=-6'',2$.

Достоинством описанного метода является возможность построения кубатурной формулы наивысшей алгебраической точности в условиях, распространенных на практике, когда узлы интегрирования составляют хаотичную сетку, но изменять их положение нельзя. Однако при $n>10$ этот метод становится слишком громоздким. Кроме того, сказываются негативные стороны полиномиальной интерполяции высокой степени, о которых говорилось в § 5.4. Поэтому при большом количестве хаотично расположенных узлов рекомендуется пользоваться методами коллокации, описанными в § 5.6. При этом исходные значения функции $f(x, y)$ в узлах P_i трактуются как значения дельта-функционалов $L_i f=\delta_{P_i} f=f(P_i)=l_i$ на обкладке, а нужный интеграл (7.64) представляет собой искомый функционал F на $f(P)$ (см. пример 7.17). Таким образом, задача численного интегрирования представляет собой одну из локальных задач коллокации и решается согласно § 5.6.

В геодезической практике часто приходится решать такие задачи интегрирования, в которых исходными числами являются не значения подынтегральной функции (или обкладки) в узлах, а значения других функционалов $L_1, \ldots, L_n$ на этой функции (см. пример 5.6). Кроме того, исходные числа возмущены случайными ошибками. Решение подобных задач описано в §§ 6.4—6.6 и составляет основу современной теории математической обработки геодезических измерений (ТМОГИ).

Оптимальные кубатурные формулы. Если узлами можно распоряжаться, то естественно подбирать такие узлы и весовые коэффициенты, при которых соответствующая кубатурная формула имеет наивысшую алгебраическую точность. Однако практически такие оптимальные кубатурные формулы удается построить только для областей интегрирования наиболее простой формы (прямоугольник, круг, сфера). Таблицы соответствующих кубатурных формул имеются в монографии [15].

Пример 7.21. Пусть D в формуле Стокса (7.51) — круг радиусом R с центром в точке Q. Чтобы выполнить интегрирование с алгебраическим порядком точности $a=3$, надо знать значения обкладки $f(P)$ по крайней мере в четырех точках. При $R=1$ оптимальными узлами являются точки $(\pm\sqrt{3}/3;\ 0)$, $(0;\ \pm\sqrt{3}/3)$. Все весовые коэффициенты равны $(\pi/4)\,(1/2\pi\gamma)=1/8\gamma$ и соответствуют весовой функции $p(x)=(2\pi\gamma\sqrt{x^2+y^2})^{-1}$.

Пример 7.22. Пусть D в формуле (7.49) — круг радиусом R с центром в начале координат. Чтобы выполнить интегрирование с $a=5$, надо знать значения подынтегральной функции $g(P)$ по крайней мере в 7 точках. При $R=1$ оптимальная кубатурная формула имеет вид

$$I=\pi\left[(1/4)\,g(0,0)+(1/8)\sum_{i=1}^{6} g(P_i)\right],$$

где

$$P_i\left(\sqrt{2/3}\cos(\pi i/3),\ \sqrt{2/3}\sin(\pi i/3)\right).$$

Заметим, что для обеспечения того же порядка точности двукратное применение квадратных формул Гаусса в примерах 7.21, 7,22 потребовало бы соответственно четырех и девяти узлов.

§ 7.11. ДВОЙНЫЕ НЕСОБСТВЕННЫЕ ИНТЕГРАЛЫ СО СТЕПЕННОЙ ОСОБЕННОСТЬЮ

В задачах теории фигуры Земли часто приходится иметь дело с интегралами вида

$$I = \iint\limits_D \frac{\psi(\alpha)}{r^\beta} f(x, y)\, dxdy. \tag{7.69}$$

Здесь $r = |\overrightarrow{OP}|$ — расстояние от начала координат O до текущей точки интегрирования $P(x, y)$; α — угол между осью абсцисс и вектором $\overrightarrow{OP}$; $\psi(\alpha)$ — заданная ограниченная функция от α, называемая *характеристикой*; β — заданное положительное число; $f(P)$ — обкладка, значения которой обычно получают из измерений; D — круг радиусом R с центром в точке O.

Поскольку весовая функция $p(x, y) = \psi(\alpha)/r^\beta$ терпит в начале координат бесконечный разрыв, то интеграл (7.69) несобственный. Можно доказать, что при $\beta < 2$ этот интеграл сходится, т. е. имеет слабую особенность. Если же $\beta = 2$, то интеграл сингулярен, т. е. расходится в классическом смысле и лишь при определенных условиях, указанных ниже теоремой 7.2, имеет главное значение в смысле Коши.

Предположим сначала, что $\beta < 2$ и перейдем к полярным координатам r, α с полюсом в точке O. При этом $dxdy = rdrd\alpha$. Новая область интегрирования G — прямоугольник $[O, R; 0, 2\pi]$, и проще всего действовать повторной квадратурой § 7.10: сначала интегрировать по r, а затем по α. Имеем

$$I = \iint\limits_G \frac{\psi(\alpha)}{r^{\beta-1}} f_1(r, \alpha)\, drd\alpha = \int\limits_0^{2\pi} \psi(\alpha) F(\alpha)\, d\alpha, \tag{7.70}$$

где

$$F(\alpha) = \int\limits_0^R \frac{f_1(r, \alpha)}{r^{\beta-1}}\, dr. \tag{7.71}$$

Если $\beta \leqslant 1$, то особенность устраняется: выражения (7.70), (7.71) являются интегралами в собственном смысле этого слова и вычисляются методами §§ 7.1—7.8. Если подынтегральная функция в (7.70) периодична, то при ее вычислении удобно пользоваться формулой прямоугольников (см. теорему 7.1 и § 7.2).

Для $1<\beta<2$ при вычислении (7.71) рекомендуется выделить особенность средствами § 7.9:

$$F(\alpha)=\int_0^R \frac{f_{1k}(r,\alpha)}{r^{\beta-1}}\,dr+\int_0^R \frac{f_1(r,\alpha)-f_{1k}(r,\alpha)}{r^{\beta-1}}\,dr, \qquad (7.72)$$

где

$$f_{1k}(r,\alpha)=f_1(0,\alpha)+(\partial f_1/\partial r)_{0,\alpha}r+\ldots+\frac{1}{k!}(\partial^k f_1/\partial r^k)_{0,\alpha}r^k. \qquad (7.73)$$

Здесь $(\partial^k f/\partial r^k)_{0,\alpha}$ — значение k-й производной f_1 по r при $r=0$. Таким образом, все коэффициенты в правой части (7.73) — функции только от α и потому 1-й интеграл в (7.72) легко берется аналитически, а 2-й — численно.

Упражнение 7.12. Составьте схему вычисления высоты квазигеоида по формуле Стокса (7.51), полагая в (7.69), что $\psi(\alpha)\equiv 1$, $\beta=1$, $R=1$, и имея в виду, что интегрирование (7.71) будет выполняться с помощью квадратуры Гаусса § 7.7 с тремя узлами, а интегрирование (7.70) — с помощью составной формулы прямоугольников § 7.2 с шагом $\pi/3$.

Указание. Воспользуйтесь результатами примера 7.10.

Ответ: $\zeta=(1/6\gamma)\sum_{i=1}^{6}F(\alpha_i)$, где $\alpha_i=(\pi/6)+(i-1)\pi/3$; $F(\alpha_i)=$ $=(1/2)[(5/9)f_1(r_1,\alpha_i)+(8/9)f_1(r_2,\alpha_i)+(5/9)f_1(r_3,\alpha_i)]$; $r_1=0{,}112702$; $r_2=0{,}5$; $r_3=0{,}887298$.

Рассмотрим теперь случай, когда $\beta=2$. Важную роль при этом играет следующая теорема.

Теорема 7.2. [13]. Пусть в интеграле (7.69) характеристика $\psi(\alpha)$ ограничена, а обкладка $f(P)$ удовлетворяет условию Липшица (т. е. существуют такие положительные константы μ и $\lambda<1$, что $|f(P)-f(Q)|\leqslant\mu r^{\lambda}$, где $r=\overrightarrow{|PQ|}$). Тогда для существования главного значения этого интеграла в смысле Коши необходимо и достаточно, чтобы

$$\int_0^{2\pi}\psi(\alpha)d\alpha=0. \qquad (7.74)$$

Следствие. В условиях теоремы 7.2

$$\iint_D \frac{\psi(\alpha)}{r^2}f(x,y)\,dxdy=\iint_D \frac{\psi(\alpha)}{r^2}[f(x,y)-C]\,dxdy, \qquad (7.75)$$

где C — произвольная константа.

В самом деле,

$$C\iint_D \frac{\psi(\alpha)}{r_2}\,dxdy=C\lim_{\varepsilon\to 0}\left[\int_\varepsilon^R \frac{dr}{r}\int_0^{2\pi}\psi(\alpha)\,d\alpha\right]=C\lim_{\varepsilon\to 0}\left[\ln\frac{R}{\varepsilon}\cdot 0\right]=0,$$

так как сначала надо умножить на 0, а потом переходить к пределу.

В частности,

$$\iint\limits_D \frac{\psi(\alpha)}{r^2} f(x, y)\, dxdy = \iint\limits_D \frac{\psi(\alpha)}{r^2} [f(x, y) - f(0, 0)]\, dxdy =$$

$$= \iint\limits_G \frac{\psi(\alpha)}{r} [f_1(r, \alpha) - f_1(0, \alpha)]\, drd\alpha. \qquad (7.76)$$

Здесь $f_1(0, 0) = f_1(0, \alpha)$ для всякого α, так как при $|\vec{r}| = 0$ направление не определено.

Вычисляем (7.76), интегрируя сначала по r, а потом по α. Имеем

$$I = \int\limits_0^{2\pi} \psi(\alpha) F(\alpha) d\alpha, \qquad (7.77)$$

где

$$F(\alpha) = \int\limits_0^R \frac{f_1(r, \alpha) - f_1(0, \alpha)}{r}\, dr.$$

Так как по условию теоремы функция $f_1(r, \alpha)$ удоветворяет условию Липшица, то $F(\alpha)$ в (7.77) — несобственный, но регулярный интеграл (со слабой особенностью) и вычисляется уже рассмотренными нами средствами § 7.9, т. е.

$$F(\alpha) = F_1(\alpha) + F_2(\alpha),$$

где

$$F_1(\alpha) = \int\limits_0^R \frac{f_{1k}(r, \alpha)}{r}\, dr, \quad F_2(\alpha) = \int\limits_0^R \frac{f_1(r, \alpha) - f_1(0, \alpha) - f_{1k}(r, \alpha)}{r}\, dr,$$

$$f_{1k}(r, \alpha) = (\partial f_1/\partial r)_{0,\alpha} r + \frac{1}{2!} (\partial^2 f_1/\partial r^2)_{0,\alpha} r^2 + \ldots + \frac{1}{k!} (\partial^k f_1/\partial r^k)_{0,\alpha} r^k.$$

Пример 7.23. Воспользуемся теоремой 7.2 и ее следствием для вычисления интеграла (7.52) по кругу малого радиуса R с центром в точке Q, совпадающей с началом координат. Условие (7.74) выполняется, поскольку $\int\limits_0^{2\pi} \cos\alpha d\alpha = \int\limits_0^{2\pi} \sin\alpha d\alpha = 0$. Ограничиваясь $k = 1$, имеем

$$\xi = -\frac{1}{2\pi\gamma} \int\limits_0^{2\pi} \cos\alpha F(\alpha)\, d\alpha, \quad f_{11}(r, \alpha) = (\partial f_1/\partial r)_{0,\alpha} r,$$

$$F_1(\alpha) = R(\partial f_1/\partial r)_{0,\alpha},$$

$$F_2(\alpha) = \int\limits_0^R \frac{f_1(r, \alpha) - f_1(0, \alpha) - (\partial f_1/\partial r)_{0,\alpha} r}{r}\, dr,$$

$$F(\alpha) = F_1(\alpha) + F_2(\alpha).$$

В геодезической практике при малом значении радиуса R обычно интегралом $F_2(\alpha)$ пренебрегают и тогда

$$\xi \approx -\frac{R}{2\pi\gamma}\int_0^{2\pi} (\partial f_1/\partial r)_{0,\alpha}\cos\alpha d\alpha = -\frac{R}{2\pi\gamma}\int_0^{2\pi}(\partial f_1/\partial r)\frac{\partial r}{\partial x}d\alpha =$$

$$= -\frac{R}{2\pi\gamma}\int_0^{2\pi}(\partial f_1/\partial x)_0\, d\alpha = -\frac{R}{\gamma}(\partial f_1/\partial x)_0.$$

Аналогично $\eta \approx (-R/\gamma)(\partial f_1/\partial y)_0$. Необходимые производные оценивают графически по карте изоаномал. Иногда поступают еще проще, полагая

$$\left(\frac{\partial f_1}{\partial r}\right)_{0,\alpha} \approx [f_1(R,\alpha) - f_1(0,0)]/R.$$

Тогда

$$\xi \approx -\frac{1}{2\pi\gamma}\int_0^{2\pi}[f_1(R,\alpha)-f_1(0,0)]\cos\alpha d\alpha = -\frac{1}{2\pi\gamma}\int_0^{2\pi} f_1(R,\alpha)\cos\alpha d\alpha$$

и последний интеграл считается, например, по составной формуле прямоугольников § 7.2.

Замечание 7.4. Если обкладка $f(P)$ не только удовлетворяет условию Липшица, как это требуется в теореме 7.2, но и является дифференцируемой функцией по r в окрестности начала координат, то

$$\lim_{r\to 0}\frac{f_1(r,\alpha)-f_1(0,\alpha)}{r} = (\partial f_1/\partial r)_{0,\alpha} < \infty$$

и поэтому $F(\alpha)$ в (7.77) является интегралом в собственном смысле этого слова (подынтегральная функция хотя и терпит разрыв, но этот разрыв является устранимым). Поэтому при вычислении обоих интегралов (7.77) достаточно пользоваться обычными квадратурными формулами без предварительного выделения особенности (см. §§ 7.1—7.8).

До сих пор предполагалось, что исходными данными для вычисления интегралов являются значения подынтегральной функции в отдельных точках области интегрирования. В практике геодезических, геофизических и других работ область интегрирования представляет собой конкретный район земной территории, а информацию о подынтегральной функции составляет каталог, содержащий средние интегральные значения $\bar{f}_i$ обкладки по отдельным ячейкам d_i области интегрирования D. Поскольку при любом выборе полюса O используются одни и те же ячейки (обычно прямоугольной формы), то интегрировать целесообразно не в полярной, а в прямоугольной системе координат. Имеем:

$$I = \iint_D p(x,y)f(x,y)dxdy \approx \sum_{i=1}^{n} c_i\bar{f}_i, \qquad (7.78)$$

где $\bar{f}_i = \frac{1}{|d_i|} \iint\limits_{d_i} f(x, y)\, dxdy$, $c_i = \iint\limits_{d_i} p(x,y)\, dxdy$,

$|d_i|$ — площадь ячейки d_i. Весовая функция $p(x, y)$ известна аналитически, и поэтому весовые коэффициенты c_i можно вычислить заранее применительно к ячейкам d_i того каталога значений $\bar{f}_i$, который предполагается использовать.

Пример 7.24. Пусть область интегрирования D в интервалах Стокса (7.51) и Венинг-Мейнеса (7.52) — прямоугольник, поделенный на n конгруэнтных квадратных ячеек d_i. Тогда

$$\zeta \approx \sum_{i=1}^{n} c_i^{\xi} \bar{f}_i, \quad \xi \approx \sum_{i=1}^{n} c_i^{\xi} \bar{f}_i, \quad \eta \approx \sum_{i=1}^{n} c_i^{\eta} \bar{f}_i,$$

где

$$c_i^{\zeta} = \frac{1}{2\pi\gamma} \left[x_1 \ln \left| \frac{\operatorname{tg}\left(\frac{A_1}{2} + \frac{\pi}{4}\right)}{\operatorname{tg}\left(\frac{A_4}{2} + \frac{\pi}{4}\right)} \right| + x_2 \ln \left| \frac{\operatorname{tg}\left(\frac{A_3}{2} + \frac{\pi}{4}\right)}{\operatorname{tg}\left(\frac{A_2}{2} + \frac{\pi}{4}\right)} \right| + \right.$$

$$\left. + y_1 \ln \left| \frac{\operatorname{tg}\frac{A_2}{2}}{\operatorname{tg}\frac{A_1}{2}} \right| + y_2 \ln \left| \frac{\operatorname{tg}\frac{A_4}{2}}{\operatorname{tg}\frac{A_3}{2}} \right| \right],$$

$$c_i^{\xi} = \frac{1}{2\pi\gamma} \ln \left| \frac{\operatorname{tg}\left(\frac{A_1}{2} + \frac{\pi}{4}\right) \operatorname{tg}\left(\frac{A_3}{2} + \frac{\pi}{4}\right)}{\operatorname{tg}\left(\frac{A_2}{2} + \frac{\pi}{4}\right) \operatorname{tg}\left(\frac{A_4}{2} + \frac{\pi}{4}\right)} \right|,$$

$$c_i^{\eta} = \frac{1}{2\pi\gamma} \ln \left| \frac{\operatorname{tg}\frac{A_2}{2} \operatorname{tg}\frac{A_4}{2}}{\operatorname{tg}\frac{A_1}{2} \operatorname{tg}\frac{A_3}{2}} \right|,$$

x, y — координаты четырех вершин i-й ячейки (рис. 29); A — соответствующий азимут, $\operatorname{tg} A_j = y_j/x_j$, $j = 1, 2, 3, 4$.

Полюс O обычно выбирают внутри какой-нибудь ячейки и соответствующий этой ячейке весовой коэффициент c_i полагают равным нулю. Такое вырезание особой точки из области интегрирования представляет собой своеобразную регуляризацию несобственных интегралов.

Точность приближенного равенства (7.78) можно оценить на основании известной в интегральном исчислении теоремы о среднем. Предположим, что $p(x, y)$ — знакопостоянная функция в ячейке d_i. Тогда в этой ячейке существует такая точка M_i, что

$$\iint\limits_{d_i} p(x, y) f(x, y) dxdy = f(M_i) \iint\limits_{d_i} p(x, y) dxdy.$$

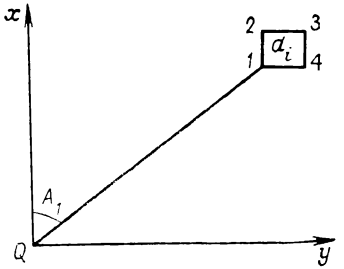


Рис. 29. Ячейка интегрирования в формулах Стокса и Венинга Мейнеса

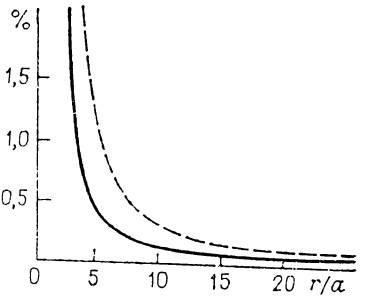


Рис. 30. График выраженного в процентах отличия $\tilde{c}_i$ от c_i для интегралов Стокса (сплошная кривая) и Венинг-Мейнеса (прерывистая кривая)

Поэтому

$$I = \sum_{i=1}^{n} \int_{d_i} \int p(x, y) f(x, y) dx dy = \sum_{i=1}^{n} c_i f(M_i) \approx \sum_{i=1}^{n} c_i \bar{f}_i$$

и погрешность вычислений по формуле (7.78) обусловлена разностью $\bar{f}_i - f(M_i)$. Точка M_i может занимать произвольное положение в ячейке M_i. Поэтому средний квадрат отклонения $\bar{f}_i$ от $f(M_i)$ представляет собой то, что в геодезической практике принято называть дисперсией представительства. Величина дисперсии представительства зависит от размеров ячеек, гладкости обкладки и подсчитывается известными в геодезии методами [16]. Значение дисперсии представительства позволяет легко подсчитать дисперсию вычисленного значения интеграла. В общем случае в ближайшей окрестности полюса целесообразно использовать мелкие ячейки. С удалением ячеек от полюса их размеры можно увеличивать без заметной потери точности.

Дальнейшие упрощения при вычислении интеграла (7.78) состоят в том, что вместо c_i берутся числа $\tilde{c}_i$, представляющие собой значения весовой функции $p(x, y)$ в центре ячейки d_i, умноженные на площадь этой ячейки $|d_i|$.

Соответствующая дополнительная погрешность для интегралов типа (7.69) зависит от удаленности ячеек от начала координат и их размеров. Выраженное в процентах отличие $\tilde{c}_i$ от c_i для интегралов Стокса (7.51) и Венинг-Мейнеса (7.52) показано на рис. 30.

По оси абсцисс отложены расстояния r от начала координат, выраженные в единицах стороны a ячеек. Сплошная кривая соответствует интегралу Стокса, прерывистая — интегралу Венинг-Мейнеса.

СПИСОК ЛИТЕРАТУРЫ

1. *Беклемишев Д. В.* Дополнительные главы линейной алгебры. М., Наука, 1983.

2. *Большаков В. Д., Бывшев В. А., Нейман Ю. М.* Об использовании регуляризации при уравнивании геодезических сетей. — Изв. вузов. Геодезия и аэрофотосъемка, 1985, № 1, с. 3—9.

3. *Василенко В. А.* Сплайн-функции: теория, алгоритмы, программы. Наука. Новосибирск, 1983.

4. *Вычислительная математика*/В. А. Вергасов, И. Г. Журкин, Ю. М. Нейман и др. М., Недра, 1976.

5. *Воеводин В. В.* Вычислительные основы линейной алгебры. М., Наука, 1977.

6. *Воеводин В. В.* Линейная алгебра. М., Наука, 1980.

7. *Жидков Н. П.* Линейные аппроксимации функционалов. М., Изд-во МГУ, 1977.

8. *Завьялов Ю. С., Квасов Б. И., Мирошниченко В. Л.* Методы сплайн-функций. М., Наука, 1980.

9. *Калиткин Н. Н.* Численные методы. М., Наука, 1978.

10. *Ланкастер П.* Теория матриц. М., Наука, 1982.

11. *Лоусон Ч., Хенсон Р.* Численное решение задач метода наименьших квадратов. М., Наука, 1982.

12. *Марчук Г. И.* Методы вычислительной математики. М., Наука, 1977.

13. *Михлин С. Г.* Линейные уравнения в частных производных. М., Высшая школа, 1977.

14. *Мориц Г.* Современная физическая геодезия/Пер. с англ.; Под ред. Ю. М. Неймана. М., Недра, 1983.

15. *Мысовских И. П.* Интерполяционные кубатурные формулы. М., Наука, 1981.

16. *Нейман Ю. М.* Вариационный метод физической геодезии. М., Недра, 1979.

17. *Нейман Ю. М., Лебедев С. В.* Приближенное решение задач коллокации методом конечных элементов. — Изв. вузов. Геодезия и аэрофотосъемка, 1986, вып. 1, 2, с. 14—28.

18. *Пеллинен Л. П., Нейман Ю. М.* Физическая геодезия. — Итоги науки и техники. ВИНИТИ. 1980, т. 18, с. 132.

19. *Розанов Ю. А.* Введение в теорию случайных процессов. М., Наука, 1982.

20. *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. М., Наука, 1979.

21. *Тихонов А. Н., Большаков В. Д., Нейман Ю. М.* Некорректные задачи геодезии. — Изв. вузов. Геодезия и аэрофотосъемка, 1980, вып. 1, с. 54—63.

22. *Треногин В. А.* Функциональный анализ. М., Наука, 1980.

23. *Тьюарсон Р.* Разреженные матрицы. М., Мир, 1977.

24. *Хейгеман Л., Янг Д.* Прикладные итерационные методы. М., Мир, 1986.

25. *Эстербю О., Златев З.* Прямые методы для разреженных матриц. М., Мир, 1987.

ОГЛАВЛЕНИЕ

УЧЕБНОЕ ИЗДАНИЕ

Журкин Игорь Георгиевич
Нейман Юрий Михайлович

МЕТОДЫ ВЫЧИСЛЕНИЙ В ГЕОДЕЗИИ

Заведующий редакцией *Л. Г. Иванова*
Редакторы издательства *С. А. Базаева, А. В. Куприянова*
Технический редактор *Л. Г. Лаврентьева*
Корректор *Е. М. Федорова*

ИБ № 7635

Сдано в набор 17.05.88. Подписано в печать 12.10.88. Т-19659. Формат 60×88$^1/_{16}$. Бумага книжно-журнальная. Гарнитура Литературная. Печать высокая. Усл. печ. л. 18,62. Усл. кр.-отт. 18,62. Уч.-изд. л. 19,26. Тираж 7250 экз. Заказ 299/1670—8. Цена 95 коп.

Ордена «Знак Почета» издательство «Недра», 125047, Москва, пл. Белорусского вокзала, д. 3.

Московская типография № 11 Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли. 113105, Москва, Нагатинская ул., д. 1.