С. Г. Михлин

НЕКОТОРЫЕ ВОПРОСЫ ТЕОРИИ ПОГРЕШНОСТЕЙ

ЛЕНИНГРАДСКИЙ ОРДЕНА ЛЕНИНА
И ОРДЕНА ТРУДОВОГО КРАСНОГО ЗНАМЕНИ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ имени А. А. ЖДАНОВА

С. Г. Михлин

НЕКОТОРЫЕ ВОПРОСЫ ТЕОРИИ ПОГРЕШНОСТЕЙ

ЛЕНИНГРАД
ИЗДАТЕЛЬСТВО ЛЕНИНГРАДСКОГО УНИВЕРСИТЕТА
1988

Рецензенты: проф. *А И Кошелев* (Ленингр. электротехн. ин-т им. В. И. Ульянова (Ленина)), проф. *В П. Ильин* (Ленингр. отд-ние Мат. ин-та им. В. А. Стеклова АН СССР)

Печатается по постановлению
Редакционно-издательского совета
Ленинградского университета

УДК 519.9

Михлин С. Г.

Некоторые вопросы теории погрешностей. — Л.: Издательство Ленинградского университета. 1988. — с.

ISBN 5—288—0050—6

В основе монографии лежит разработанная автором классификация погрешностей, возникающих при численном решении задач вычислительной математики. Эта классификация позволила получить априорные оценки погрешностей для ряда линейных и нелинейных задач и методов. Исследованы прогрешности методов Ритца, Бубнова — Галеркина, конечных элементов, Ньютона — Канторовича и др. Рассмотрены погрешности решения линейных алгебраических систем, интегральных уравнений, некоторых вариационных неравенств. Получены новые результаты по апостериорным оценкам погрешностей.

Книга предназначена для научных и инженерно-технических работников, занимающихся вычислительной математикой. Библиогр. 202 назв. Ил. 6. Табл. 17.

М $\frac{1702070000—097}{076(02)—88}$ 53—88

ISBN 5—288—0050—6

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Проблема погрешностей с давних пор привлекала внимание математиков. Значение ее еще более возросло с широким распространением ЭВМ. К классической проблеме погрешностей приближенных методов добавилась существовавшая ранее в зачаточном состоянии проблема погрешностей арифметических машинных действий.

Решение названных проблем вызвало необходимость классификации погрешностей, нашедшей отражение и в учебной литературе (см., например, [6, 7]). По этой классификации с каждой задачей вычислительной математики связывают три погрешности: 1) неустранимую погрешность, вызванную недостаточной точностью данных; 2) погрешность метода, возникающую при использовании для решения задачи приближенного метода; 3) вычислительную погрешность, называемую также погрешностью округления: она является следствием того, что как данные числа, так и результаты действий над ними невозможно, как правило, выразить точно и их приходится округлять.

Приведенная классификация в определенном смысле способствовала развитию вычислительной математики, ее понятий, методов и результатов. Однако эта классификация кажется автору недостаточно полной. Предлагаемая классификация предполагает разбиение погрешности метода и вычислительной погрешности на четыре: на погрешности аппроксимации, искажения, алгоритма и округления. (Подробнее об этом см. в главе 1.)

Погрешность аппроксимации проистекает оттого, что данная «точная» задача заменяется (аппроксимируется) другой, более простой «приближенной» задачей. Погрешность искажения появляется вследствие того, что данные приближенной задачи, которой заменяется исходная точная задача, вычисляются неточно. Погрешность алгоритма возникает в случае, когда алгоритм, применяемый для решения приближенной (в данном случае искаженной приближенной) задачи, дает ее точное решение лишь в результате некоторого предельного пе-

рехода, если его не удается выполнить эффективно. Погрешность округления является следствием того, что предписания упомянутого алгоритма выполняются неточно.

Автор не включил в предложенную им классификацию неустранимую погрешность, поскольку ее оценка, как правило, выходит за пределы возможностей математического анализа. Он придерживается классической точки зрения, по которой задача, к решению которой применяются математические методы, должна рассматриваться как поставленная точно. Конечно, существуют и другие мнения.

Расчленение погрешности метода и вычислительной погрешности на четыре более элементарные и поэтому легче поддающиеся анализу позволило получить оценки погрешностей для ряда линейных и нелинейных задач вычислительной математики. Так, исследованы погрешности методов Ритца, конечных элементов, Бубнова — Галеркина, коллокации, наискорейшего спуска, минимальных невязок, Ньютона — Канторовича; получены оценки погрешностей решений интегральных уравнений — фредгольмовых и, отчасти, сингулярных. Частично исследованы погрешности методов Ритца и Бубнова — Галеркина для нелинейных задач, погрешности решений одного класса односторонних вариационных задач, даны приложения к теории пластичности Сен-Венана-Мизеса.

В книге проведено также исследование погрешностей решений линейных алгебраических систем, основанное на описанной выше четырехчленной классификации погрешностей. На этом пути получены как некоторые известные (относящиеся, например, к методу Гаусса), так и некоторые новые результаты (например, о методе сопряженных направлений и др.).

Все сказанное относится к *априорным* оценкам погрешности, которые можно строить независимо от того, вычислено или не вычислено само приближенное решение. Немалый интерес представляют и *апостериорные* оценки погрешности, т. е. оценки, выводимые для уже построенного приближенного решения. О таких оценках довольно много сказано в монографии автора [84]; в данной же работе этим вопросам посвящены два параграфа главы 1 и глава 16, содержание которой составляют применение метода Ритца к трехмерным задачам упругопластического упрочнения и получение апостериорных оценок погрешности указанного метода. Глава написана по результатам диссертации, выполненной под руководством автора безвременно скончавшейся И. В. Алексиевой (Болгария).

Дополнение (§ 1—5) содержит несколько примеров, назначение которых показать, как вычислять оценки погрешностей исходя из результатов предшествующих разделов. Для простоты, а также из нежелания увеличивать чрезмерно объем книги, автор ограничился рассмотрением линейных задач, имеющих иллюстративный характер.

Раздел I

ВЫЧИСЛИТЕЛЬНЫЕ ПРОЦЕССЫ И ИХ ПОГРЕШНОСТИ

Глава I

ПОГРЕШНОСТИ МАШИННЫХ ДЕЙСТВИЙ И ВЫЧИСЛИТЕЛЬНЫХ ПРОЦЕССОВ

§ 1. АРИФМЕТИЧЕСКИЕ ДЕЙСТВИЯ

1°. Пусть a — какое-нибудь число. Наряду с его точным значением часто приходится рассматривать неточное значение, получаемое как результат некоторых вычислительных машинных операций. Обозначим это неточное значение через $(a)_м$ и через $\delta(a)$ погрешность $\delta(a)=(a)_м - a$.

Как известно (см., например, [39]), для каждой ЭВМ, в которой принято двоичное представление чисел и которая работает в режиме с плавающей запятой, можно указать характеризующие эту машину малые числа ε_1 и ε_2: наименьшее машинное число, большее единицы, равно $1+\varepsilon_1$, а наибольшее по модулю машинное число равно $1/\varepsilon_2$. Для ЭВМ БЭСМ-6 значения этих чисел таковы: $\varepsilon_1 = 2^{-39} \approx 0{,}182\cdot 10^{-11}$, $\varepsilon_2/2 = 2^{-64}(1-2^{-40})^{-1} \approx 0{,}543\cdot 10^{-19}$.

Числа ε_1 и ε_2 входят в оценки погрешностей машинных арифметических действий. Они малы по сравнению с единицей и по крайней мере для БЭСМ-6 ε_2 мало по сравнению с ε_1, и ε_2 появляется в оценках погрешности только тогда, когда результат выполнения машинной операции есть машинный нуль. Разумеется, мы при этом допускаем, что множители при ε_1 и ε_2 не слишком велики. Имея это в виду, будем отбрасывать в оценках погрешности члены порядка ε_2, а также члены, содержащие степени ε_1 с показателями, большими единицы. В результате из формул для оценки погрешностей машинных действий, приведенных в [39], вытекает общая формула

$$|\delta(a*b)| \leqslant \varepsilon_1 |a*b|, \qquad (1.1.1)$$

в которой звездочка обозначает любое из четырех арифметических действий. Формула, близкая к (1.1.1), дана в книге [142].

2°. Пользуясь формулой (1.1.1), можно вывести оценки погрешностей результатов многократного выполнения машинных

арифметических действий. Эти оценки имеют вид

$$\left|\delta\left(\sum_{i=1}^{m} a_i\right)\right| \leqslant \varepsilon_1 (m-1) \max_{2 \leqslant k \leqslant m} \left|\sum_{i=1}^{k} a_i\right|; \qquad (1.1.2)$$

$$\left|\delta\left(\prod_{i=1}^{m} a_i\right)\right| \leqslant \varepsilon_1 (m-1) \prod_{i=1}^{m} |a_i|. \qquad (1.1.3)$$

Они верны с точностью до членов высшего порядка малости.

Докажем неравенство (1.1.2). Машинное сложение нескольких чисел производится по рекуррентному соотношению

$$\left(\sum_{i=1}^{k} a_i\right)_{\text{м}} = \left(\left(\sum_{i=1}^{k-1} a_i\right)_{\text{м}} + a_k\right)_{\text{м}}. \qquad (1.1.4)$$

Обозначим $b_k = (a_1 + a_2 + \ldots + a_k)_{\text{м}}$ и $|b_k - a_1 - a_2 - \ldots - a_k| = |\delta(a_1 + a_2 + \ldots + a_k)| = \varepsilon_1 \gamma_k$. По формулам (1.1.1), (1.1.4)

$$\varepsilon_1 \gamma_{k+1} = \left|\left(\sum_{i=1}^{k+1} a_i\right)_{\text{м}} - \sum_{i=1}^{k+1} a_i\right| = \left|\left[(b_k + a_{k+1})_{\text{м}} - b_k - a_{k+1}\right] + \left[b_k - \sum_{i=1}^{k} a_i\right]\right| \leqslant \varepsilon_1 [|b_k + a_{k+1}| + \gamma_k]. \qquad (1.1.5)$$

Далее, $b_k + a_{k+1} = \sum_{i=1}^{k+1} a_i + \theta \varepsilon_1 \gamma_k$, $\theta = \pm 1$. Отсюда

$$|b_k + a_{k+1}| \leqslant \left|\sum_{i=1}^{k+1} a_i\right| + \varepsilon_1 \gamma_k.$$

Подставим это в (1.1.5) и отбросим члены с ε_1^2:

$$\gamma_{k+1} \leqslant \left|\sum_{i=1}^{k+1} a_i\right| + \gamma_k.$$

В этом неравенстве положим $k = 2, 3, \ldots, m-1$ и сложим полученные соотношения. Формула (1.1.1) показывает, что $\gamma_2 \leqslant |a_1 + a_2|$, и мы приходим к неравенству

$$\gamma_m \leqslant \left|\sum_{k=2}^{m} a_k\right| \leqslant (m-1) \max_{2 \leqslant k \leqslant m} \left|\sum_{i=1}^{k} a_i\right|,$$

равносильному неравенству (1.1.2). Заметим, что из (1.1.2) вытекает более простая, хотя и более грубая оценка

$$\left|\delta\left(\sum_{i=1}^{m} a_i\right)\right| \leqslant \varepsilon_1 (m-1) \sum_{i=1}^{m} |a_i|. \qquad (1.1.6)$$

Обратимся к неравенству (1.1.3). Оно очевидно, если $m = 2$. Допустим, что оно верно для $m-1$ множителей. Машинное умножение нескольких множителей совершается по рекуррентной формуле

$$\left(\prod_{i=1}^{k} a_i\right)_{\text{м}} = \left(\left(\prod_{i=1}^{k-1} a_i\right)_{\text{м}} a_k\right)_{\text{м}}. \qquad (1.1.7)$$

По индукционному предположению

$$\left(\prod_{i=1}^{m-1} a_i\right)_{\text{м}} = (1 + \varepsilon_1 (m-2)\theta) \prod_{i=1}^{m-1} a_i, \quad |\theta| \leqslant 1.$$

Отсюда следует, что с точностью до малых высших порядков

$$\left|\left(\delta\left(\prod_{i=1}^{m} a_i\right)\right)\right| = \left|\left(\prod_{i=1}^{m} a_i\right)_{\text{м}} - \prod_{i=1}^{m} a_i\right| \leqslant \left|\left(\left(\prod_{i=1}^{m-1} a_i\right)_{\text{м}} a_m\right)_{\text{м}} - \right.$$
$$\left. - \left[\prod_{i=1}^{m-1} a_i\right)_{\text{м}} a_m\right| + \left|\left(\prod_{i=1}^{m-1} a_i\right)_{\text{м}} a_m - \prod_{i=1}^{m} a_m\right| \leqslant$$
$$\leqslant \varepsilon_1 \left[\left|\left(\prod_{i=1}^{m-1} a_i\right)_{\text{м}} a_m\right| + (m-2) \prod_{i=1}^{m} |a_i|\right] \leqslant$$
$$\leqslant \varepsilon_1 \left[(1 + (m-2)\varepsilon_1) \prod_{i=1}^{m} |a_i| + (m-2) \prod_{i=1}^{m} |a_i|\right].$$

Отбросив справа слагаемые порядка ε_1^2, получим соотношение (1.1.3).

3°. Можно существенно уменьшить погрешности многократных машинных арифметических действий, а также улучшить и упростить оценки таких погрешностей, если выполнять промежуточные действия с повышенной точностью. Пусть величина a получается как результат нескольких арифметических действий, выполняемых с повышенной точностью. Тогда погрешность величины a будет оцениваться выражением вида $P\varepsilon_1^{1+\nu}$, $\nu > 0$, а результат вычислений имеет вид $a + \theta P\varepsilon_1^{\nu}$, $|\theta| \leqslant 1$. В этом результате отбросим последние значащие цифры так, чтобы относительная погрешность нового результата, который мы обозначим через $(a)_{\text{м}}$, не превосходила ε_1. Тогда

$$(a)_{\text{м}} = a + P\varepsilon_1^{1+\nu} + \varepsilon_1 \theta' (a + P\varepsilon_1^{1+\nu}), \quad |\theta'| \leqslant 1. \tag{1.1.8}$$

Допустим теперь, что число P не очень велико, так что $P\varepsilon_1^{\nu}$ мало по сравнению с единицей. Отбросив в (1.1.8) члены высшего порядка малости, найдем, что $(a)_{\text{м}} = a + \theta'\varepsilon_1 a$, $|\theta'| \leqslant 1$, и, следовательно,

$$|\delta(a)| \leqslant \varepsilon_1 |a|. \tag{1.1.9}$$

Оценка (1.1.9) остается в силе и тогда, когда вычисления с повышенной точностью заменяются вычислениями с использованием регистра младших разрядов [39].

4°. Рассмотрим погрешность извлечения квадратного корня. Обычно это действие выполняется по рекуррентной формуле

$$\forall x_0 > 0, \quad x_{n+1} = 2^{-1}(a/x_n + x_n). \tag{1.1.10}$$

Как хорошо известно, существует $\lim\limits_{n\to\infty} x_n = \sqrt{a}$. На практике вычисления заканчивают, когда достигается неравенство $|x_{n+1} - x_n| \leqslant \varepsilon$, где ε — заданное малое число. Покажем, что при этом с точностью до малых высших порядков выполняется неравенство

$$|x_n - \sqrt{a}| \leqslant \varepsilon. \tag{1.1.11}$$

Имеем $x_{n+1} = x_n + \theta\varepsilon$, $|\theta| \leqslant 1$. Подставив это в (1.1.10), получим квадратное уравнение $x_n^2 + 2\varepsilon\theta x_n - a = 0$ с положительным корнем $x_n = -\varepsilon\theta + \sqrt{a + \varepsilon^2\theta^2}$. С точностью до величин порядка ε^2 находим

$$|x_n - \sqrt{a}| = |\sqrt{a} + \varepsilon\theta - \sqrt{a + \varepsilon^2\theta^2}| \leqslant \varepsilon |\theta| \leqslant \varepsilon.$$

Для упрощения последующих выкладок положим $\varepsilon = \varepsilon_1 \sqrt{a}$. Тогда

$$|x_n - \sqrt{a}| \leqslant \varepsilon_1 \sqrt{a}. \tag{1.1.12}$$

К сказанному добавим некоторые, большей частью известные замечания. Пусть $\bar{x} > 0$ — некоторое приближение к $\sqrt{a}$, а $\bar{\bar{x}}$ — следующее за $\bar{x}$ приближение, построенное по формуле (1.1.10):

$$\bar{\bar{x}} = 2^{-1}(a/\bar{x} + \bar{x}).$$

Тогда

$$\bar{\bar{x}} - \sqrt{a} = (\sqrt{a} - \bar{x})^2/2\bar{x}. \tag{1.1.13}$$

Отсюда следует, что при любом выборе начального приближения $x_0 > 0$ все последующие приближения $x_1, x_2, \ldots, x_n, \ldots$ суть приближения с избытком к $\sqrt{a}$, за исключением тривиального случая, когда $x_0 = \sqrt{a}$. Примем, что $x_0 \neq \sqrt{a}$. Из (1.1.10) следует, что при $n \geqslant 2$

$$x_{n+1} - x_n = 2^{-1}(x_n - x_{n-1})\left(1 - \frac{a}{x_n x_{n-1}}\right). \tag{1.1.14}$$

Очевидно, что вторая скобка справа положительна, и из (1.1.14) вытекает, что все разности $x_{n+1} - x_n$, $n \geqslant 2$, имеют один и тот же знак. Более того, все эти разности отрицательны — в противном случае последовательность $(x_1, x_2, \ldots)$ была бы возрастающей, и так как все ее члены $x_n > \sqrt{a}$, то было бы $\lim\limits_{n\to\infty} x_n > \sqrt{a}$.

Выберем начальное приближение так, чтобы $\sqrt{a} < x_0 < 2\sqrt{a}$. По формуле (1.1.13)

$$x_n - \sqrt{a} = \frac{(x_{n-1} - \sqrt{a})^2}{2x_{n-1}} < \frac{(x_{n-1} - \sqrt{a})^2}{2\sqrt{a}} < \frac{(x_{n-2} - \sqrt{a})^4}{(2\sqrt{a})^3} < \ldots$$

$$\ldots < \frac{(x_0 - \sqrt{a})^{2^n}}{(2\sqrt{a})^{2^n - 1}} < 2^{-2^n+1}\sqrt{a}.$$

Неравенство (1.1.12) будет выполнено, если $2^{-2^n+1} \leqslant \varepsilon_1$; например, на ЭВМ БЭСМ-6, где $\varepsilon_1 = 2^{-39}$, достаточно выполнить по формуле (1.1.10) шесть итераций.

§ 2. СЛОЖЕНИЕ И УМНОЖЕНИЕ МАТРИЦ

1°. Пусть A и B — квадратные матрицы порядка m, составленные соответственно из элементов a_{ij} и b_{ij}; $i, j = 1, 2, \ldots, m$. Будем записывать это в виде $A = [a_{ij}]_{i,j=1}^{m}$, $B = [b_{ij}]_{i,j=1}^{m}$. Будем говорить, что матрица B мажорирует матрицу A, если $|a_{ij}| \leqslant b_{ij}$ при всех $i, j = 1, 2, \ldots, m$; соотношение мажорирования будем записывать в виде $A \omega B$. Введем еще обозначение: если $A = [a_{ij}]_{i,j=1}^{m}$, то $\hat{A} = [|a_{ij}|]_{i,j=1}^{m}$. Очевидно, что $A \omega \hat{A}$. Отметим еще, что если $A_1 \omega B_1$ и $A_2 \omega B_2$, то $A_1 + A_2 \omega B_1 + B_2$ и $A_1 A_2 \omega B_1 B_2$.

Любая матрица порядка m порождает линейный оператор, действующий в евклидовом пространстве R_m. Норму матрицы как оператора в R_m будем обозначать символом $\|\cdot\|$. Кроме того, мы иногда будем рассматривать евклидову норму матрицы, определяемую формулой

$$\|A\|_E^2 = \sum_{i,j=1}^{m} |a_{ij}|^2. \tag{1.2.1}$$

Известны соотношения [39]

$$\|A\| \leqslant \|A\|_E \leqslant \sqrt{m}\,\|A\| \tag{1.2.2}$$

Ясно, что если $A \omega B$, то $\|A\| \leqslant \|B\|$ и $\|A\|_E \leqslant \|B\|_E$. В частности, $\|A\| \leqslant \|\hat{A}\|$ и $\|A\|_E = \|\hat{A}\|_E$. Отсюда и из (1.2.2) следует, что

$$\|\hat{A}\| \leqslant \|\hat{A}\|_E = \|A\|_E \leqslant \sqrt{m}\,\|A\|. \tag{1.2.3}$$

2°. Рассмотрим произведение s матриц $A = A_1 A_2 \ldots A_s$; $A_l = [a_{ij}^{(l)}]_{i,j=1}^{m}$, $A = [a_{ij}]_{i,j=1}^{m}$. Имеем

$$a_{ij} = \sum_{k_1, k_2, \ldots, k_{s-1}=1}^{m} a_{ik_1}^{(1)} a_{k_1 k_2}^{(2)} \ldots a_{k_{s-1} j}^{(s)}.$$

Если вычисления производятся с повышенной точностью, то по формуле (1.1.9) $|\delta(a_{ij})| \leqslant \varepsilon_1 |a_{ij}|$. Это означает, что $\delta(A) \omega \varepsilon_1 \hat{A}$ и, следовательно,

$$\|\delta(A)\| \leqslant \varepsilon_1 \|\hat{A}\| \leqslant \varepsilon_1 \sqrt{m}\,\|A\| \leqslant \varepsilon_1 \sqrt{m} \prod_{j=1}^{s} \|A_j\|. \tag{1.2.4}$$

3°. Рассмотрим погрешность умножения матрицы на вектор. Допустим, что вычисления производятся с повышенной точностью. Если

$$A = [a_{ij}]_{i,j=1}^{m}, \quad x = (x_1, x_2, \ldots, x_m),$$

то

$$(Ax)_i = \sum_{j=1}^{m} a_{ij} x_j. \tag{1.2.5}$$

По формуле (1.1.9) $|\delta((Ax)_i)| \leqslant \varepsilon_1 |(Ax)_i|$ и далее

$$|\delta(Ax)_i|^2 \leqslant \varepsilon_1^2 \left| \sum_{j=1}^m a_{ij} x_j \right|^2 \leqslant \varepsilon_1^2 \sum_{j=1}^m |a_{ij}|^2 \cdot \|x\|^2. \quad (1.2.6)$$

Суммируя и извлекая корень, получаем

$$\|\delta(Ax)\| \leqslant \varepsilon_1 \|A\|_E \|x\| \leqslant \varepsilon_1 \sqrt{m} \|A\| \cdot \|x\|. \quad (1.2.7)$$

4°. Оценим погрешность сложения матриц, предполагая попрежнему, что вычисления производятся с повышенной точностью. Пусть $A_k = [a_{ij}^{(k)}]_{i,j=1}^m$ и $A = [a_{ij}]_{i,j=1}^m = A_1 + A_2 + \ldots + A_s$. По формуле (1.1.9)

$$|\delta(a_{ij})| \leqslant \varepsilon_1 \left| \sum_{k=1}^s a_{ij}^{(k)} \right|, \quad \delta(A) \text{ ю } \varepsilon_1 \sum_{k=1}^s \hat{A}_k.$$

Отсюда

$$\|\delta(A)\| \leqslant \varepsilon_1 \sum_{k=1}^s \|\hat{A}_k\| \leqslant \varepsilon_1 \sqrt{m} \sum_{k=1}^s \|A_k\|. \quad (1.2.8)$$

5°. Нетрудно оценить погрешность суммы произведений матриц

$$\delta\left(\sum_{k=1}^s B_k\right), \quad B_k = A_{k1} A_{k2} \ldots A_{kl_k}.$$

Сначала вычисляются произведения; в результате получаются матрицы $B_k + \delta(B_k)$; при этом

$$\delta\left(\sum_{k=1}^s B_k\right) = \left(\sum_{k=1}^s (B_k + \delta(B_k))\right)_{\text{м}} - \sum_{k=1}^s B_k =$$

$$= \left(\sum_{k=1}^s (B_k + \delta(B_k))\right)_{\text{м}} - \sum_{k=1}^s (B_k + \delta(B_k)) + \sum_{k=1}^s \delta(B_k),$$

и по неравенству (1.2.8)

$$\left\|\delta\left(\sum_{k=1}^s B_k\right)\right\| \leqslant \varepsilon_1 \sqrt{m} \sum_{k=1}^s \|B_k + \delta(B_k)\| + \sum_{k=1}^s \|\delta(B_k)\|.$$

Если теперь воспользоваться неравенством (1.2.4), то получится после отбрасывания членов более высокого порядка малости искомая оценка:

$$\left\|\delta\left(\sum_{k=1}^s B_k\right)\right\| \leqslant 2\varepsilon_1 \sqrt{m} \sum_{k=1}^s \|B_k\| \leqslant 2\varepsilon_1 \sqrt{m} \sum_{k=1}^s \prod_{j=1}^{l_k} \|A_{kj}\|. \quad (1.2.9)$$

§ 3. ВЫЧИСЛИТЕЛЬНЫЕ ПРОЦЕССЫ

1°. Многие задачи вычислительной математики подходят под следующую схему. Дан некоторый объект f; требуется найти другой объект x по двум условиям: 1) x принадлежит некоторому данному классу объектов X; 2) x находится в некотором заданном соотношении r с данным элементом f. Нашу задачу запишем так:

$$\{x \in X, \ xrf\}. \quad (1.3.1)$$

Объект x, удовлетворяющий условиям (1.3.1), есть решение задачи.

Пример 1. Пусть требуется вычислить интеграл

$$x = \int_0^1 f(t)\,dt, \tag{1.3.2}$$

где $f(t)$ — заданная функция. В этом случае $f = f(t)$, $X = R_1$, а соотношение r означает, что x и f связаны уравнением (1.3.2).

В несколько усложненном примере, когда требуется вычислить интеграл

$$x(z) = \int_D f(z, t)\,dt, \tag{1.3.3}$$

где D — множество в R_m, а z — параметр, который может принять любое значение из некоторого множества Ω, f есть функция двух точек z и t, X есть множество функций, определенных на Ω, а соотношение r определяется уравнением (1.3.3).

Пример 2. Пусть требуется решить систему линейных алгебраических уравнений

$$Ax = b, \tag{1.3.4}$$

где x и b — k-мерные векторы; A — квадратная матрица порядка k. Пусть для простоты все числа в нашей задаче вещественные. Тогда $f = b$, $X = R_k$, соотношение r определяется уравнением (1.3.4).

Пример 3. Рассмотрим смешанную задачу для волнового уравнения

$$\frac{\partial^2 x}{\partial t^2} = \Delta x + f(z, t);\ t \geqslant 0,\ z \in \Omega,$$
$$x(z, 0) = \varphi(z),\ x_t(z, 0) = \psi(z), \tag{1.3.5}$$
$$x(z,t)\,|_{z \in \partial\Omega} = \omega(z, t),$$

Δ — оператор Лапласа. Будем, например, искать решение этой задачи, принадлежащее соболевскому классу $W_p^{(2)}([0, \infty); \Omega)$. Тогда $X = W_p^{(2)}([0, \infty); \Omega)$, f есть совокупность функций $f(z, t)$, $\varphi(z)$, $\psi(z)$, $\omega(z, t)$, а соотношение r определяется совокупностью уравнений (1.3.5).

Пример 4. Пусть $f(x)$ — функционал, определенный на некотором множестве X, и требуется найти элемент $x \in X$, на котором функционал $f(x)$ достигает минимума. В данном случае $f = f(x)$, а соотношение r можно описать так: если x_* есть решение данной вариационной задачи, то $x_* rt$ означает, что $f(x_*) \leqslant f(x)$, $\forall x \in X$.

2°. Как правило, задача (1.3.1) решается приближенно. В довольно широком классе случаев это означает следующее. Каждому натуральному n приводятся в соответствие: объект $f^{(n)}$; класс объектов X_n; оператор p_n, отображающий X_n в X;

соотношение r_n. Задача (1.3.1) заменяется любой задачей из последовательности

$$\{v^{(n)} \in X_n,\ v^{(n)} r_n f^{(n)}\}; \tag{1.3.6}$$

мы предполагаем, что решение каждой из задач (1.3.6) не зависит от решений последующих задач, но может зависеть от предыдущих. Если $v_*^{(n)}$ есть решение n-й задачи (1.3.6), то элемент $x_*^{(n)} = p_n v_*^{(n)}$ рассматривается как приближенное решение задачи (1.3.1).

Будем говорить, что задача (1.3.6) аппроксимирует задачу (1.3.1), и допустим, что аппроксимирующая задача (1.3.6) в том или ином смысле проще, чем аппроксимируемая задача (1.3.1).

Процесс решения последовательности задач (1.3.6) назовем вычислительным процессом. Этот процесс назовем свободным, если при любом n решение $v_*^{(n)}$ определяется независимо от элементов $v_*^{(k)}$, $k < n$, и рекуррентным, если $v_*^{(n)}$ определяется через $v_*^{(k)}$, $k < n$.

Пример 5. Будем вычислять интеграл (1.3.2) примера 1, например, по формуле средних прямоугольников с n узлами. Тогда $f^{(n)}$ есть совокупность n ординат $f(t_k)$ в точках $t_k = (2k-1)/2n$, $k = 1, 2, \ldots, n$; $X_n = R_1$, p_n есть тождественный оператор в R_1, а соотношение r_n определяется выбранной квадратурной формулой

$$v^{(n)} r_n f^{(n)} := v^{(n)} = \frac{1}{n} \sum_{k=1}^{n} f\left(\frac{2k-1}{2n}\right). \tag{1.3.7}$$

Пример 6. В примере 2 положим $A = I - B$, где I — единичная матрица в R_k, и запишем уравнение (1.3.4) в виде $x = Bx + b$. Применим метод простой итерации: зададим какой-нибудь вектор $v^{(0)}$ и положим для любого натурального n

$$v^{(n)} = Bv^{(n-1)} + b. \tag{1.3.8}$$

В данном случае $f^{(n)}$ есть совокупность векторов b и $v^{(n-1)}$; $X_n = X = R_k$, $p_n = I$, соотношение r_n определяется формулой простых итераций

$$v^{(n)} r_n f^{(n)} := v^{(n)} = Bv^{(n-1)} + b. \tag{1.3.9}$$

Вычислительный процесс (1.3.9) — рекуррентный.

Приведем еще некоторые примеры. К свободным вычислительным процессам приводят методы Ритца, Бубнова — Галеркина, конечных элементов, разностные методы, метод прямых. К рекуррентным процессам приводят различные итерационные методы: метод простых итераций, релаксационные методы, градиентные методы, методы наискорейшего спуска и минимальных невязок, метод Ньютона — Канторовича и др.

3°. Предположим, что аппроксимирующую задачу (1.3.6) мы «умеем решать» при любом n. Под этим мы понимаем следующее. Существует набор «элементарных» операций, которые мы умеем выполнять, и мы располагаем алгоритмом, который при любом n после выполнения конечного числа элементарных операций приводит (если эти операции выполнены точно, без погрешности) либо к точному значению элемента $v_*^{(n)}$, решающего задачу (1.3.6), либо к некоторому элементу $\tilde{v}^{(n)}$, который в том или ином смысле близок к элементу $v_*^{(n)}$, причем степень этой близости можно оценить. Разумеется, мы предполагаем при этом, что для рассматриваемых n аппроксимирующая задача разрешима.

Поясним сказанное на примере. Пусть задача (1.3.1) имеет вид

$$Ax = f, \tag{1.3.10}$$

где A — положительно определенный оператор в некотором гильбертовом пространстве H; x и f — элементы этого пространства, искомый и данный. Если к этой задаче применить какой-нибудь вариант метода Ритца (например, метод конечных элементов), то мы придем к аппроксимирующей линейной алгебраической системе вида

$$A_n v^{(n)} = f^{(n)}, \tag{1.3.11}$$

где A_n — числовая квадратная матрица порядка $N \times N$; $v^{(n)}$ и $f^{(n)}$ суть N-мерные числовые векторы; N — некоторая целочисленная функция от n. Как известно, система (1.3.11) однозначно разрешима при любом n. Если для ее решения воспользоваться каким-нибудь точным методом, например методом Гаусса, то через конечное число шагов мы придем к точному решению; если же воспользоваться, например, методом простой итерации (для этого надо предварительно подвергнуть систему (1.3.11) некоторому элементарному преобразованию), то в результате конечного числа шагов мы получим вектор $\tilde{v}^{(n)}$, близкий к $v_*^{(n)}$. Набор элементарных операций, о котором говорилось выше, содержит арифметические действия; в случае метода Гаусса в этот набор входит еще определение наибольшего по модулю числа из заданного конечного множества чисел.

4°. Для приложений особенно интересны случаи, когда X и X_n суть банаховы пространства. Предположим еще, что f и $f^{(n)}$ также суть элементы некоторых банаховых пространств Y и Y_n соответственно. Эти предположения мы сохраним на протяжении всей книги.

§ 4. ИСТОЧНИКИ ПОГРЕШНОСТЕЙ

1°. Если задача (1.3.1) решается каким-либо методом, подходящим под схему § 3, то, как кажется автору, речь может идти о четырех возможных источниках погрешности. При этом

мы считаем, что задача (1.3.1) поставлена точно. Тем самым мы пренебрегаем так называемыми неустранимыми погрешностями, связанными с постановкой упомянутой задачи как задачи естествознания или социальнонаучных дисциплин (погрешности измерений, недостаточная точность основных гипотез и т. п.). По мнению автора, в общем случае устранение или хотя бы уменьшение подобных погрешностей лежат вне пределов действия математических методов, и самое большее, на что здесь можно рассчитывать, это при наличии достаточной информации о степени точности упомянутых факторов воспользоваться методами теории возмущений и дать оценку погрешности, обусловленной неточностью этих факторов.

2°. Переходим к описанию источников погрешностей.

1. Погрешность аппроксимации. Переход от точной задачи (1.3.1) к аппроксимирующей задаче (1.3.6) с выбранным номером n приводит к тому, что элемент x_* пространства X заменяется элементом $x_*^{(n)} = p_n v_*^{(n)}$ того же пространства. Возникающую при этом погрешность $x_* - p_n v_*^{(n)}$ естественно оценить ее нормой

$$\rho_n = \| x_* - p_n v_*^{(n)} \| = \| x_* - x_*^{(n)} \|. \tag{1.4.1}$$

Величину ρ_n назовем *погрешностью аппроксимации задачи* (1.3.1).

Вопрос о погрешности аппроксимации и ее оценке — один из важнейших в вычислительной математике, и ему посвящена обширная литература.

2. Погрешность искажения. Как правило, переход от точной задачи (1.3.1) к аппроксимирующей задаче (1.3.6) совершается с погрешностью. Элемент $f^{(n)}$ не задан заранее, а вычисляется по данным точной задачи. Вычислительные операции почти всегда выполняются с той или иной погрешностью, и вместо элемента $f^{(n)}$ мы на самом деле получим некоторый «искаженный» элемент $\tilde{f}^{(n)}$. Может случиться, что и соотношение r_n заменится «искаженным» соотношением $\tilde{r}_n$. Элемент $f^{(n)}$ и соотношение r_n на самом деле остаются неизвестными; известными являются $\tilde{f}^{(n)}$ и $\tilde{r}_n$. В результате вместо последовательности аппроксимирующих задач (1.3.6) получается последовательность «искаженных» аппроксимирующих задач

$$\{ z^{(n)} \in X_n,\ z^{(n)} \tilde{r}_n \tilde{f}^{(n)} \}. \tag{1.4.2}$$

Допустим, что эти задачи разрешимы, и пусть $z_*^{(n)}$ — их решения. Величину

$$\tilde{\xi}_n = \| z_*^{(n)} - v_*^{(n)} \|_{X_n} \tag{1.4.3}$$

назовем *X_n-погрешностью искажения.* Больший интерес представляет величина

$$\xi_n = \| p_n z_*^{(n)} - p_n v_*^{(n)} \|_X, \tag{1.4.4}$$

которую мы назовем *X-погрешностью искажения* (или просто *погрешностью искажения*).

3. Искаженную задачу (1.4.2) предстоит решать. Будем считать, как об этом сказано в § 3, что мы располагаем некоторым алгоритмом, который после выполнения конечного числа действий приводит, если эти действия выполнены без погрешностей, либо к точному значению элемента $z_*^{(n)}$, либо к такому элементу $w^{(n)} \in X_n$, что погрешность

$$\eta_n = \| p_n w^{(n)} - p_n z_*^{(n)} \|_X \tag{1.4.5}$$

можно оценить и можно добиться того, чтобы величина, оценивающая эту погрешность, была сколь угодно малой. Первый случай — частный по отношению ко второму, и мы будем рассматривать именно второй случай, допуская, что величина (1.4.5) может иногда равняться нулю. Эту величину назовем *X-погрешностью* алгоритма (или просто *погрешностью алгоритма*); можно ввести и *X_n-погрешность алгоритма*:

$$\tilde{\eta}_n = \| w^{(n)} - z_*^{(n)} \|_{X_n}. \tag{1.4.6}$$

4. Практически любой алгоритм содержит ряд операций, выполнение которых сопряжено с погрешностями. Таковы арифметические действия, операции вычисления значений функции и т. п. Названные погрешности приведут к тому, что вместо элемента $w^{(n)} \in X_n$ мы получим другой элемент $\tilde{w}^{(n)} \in x_n$. Величину

$$\tilde{\xi}_n = \| \tilde{w}^{(n)} - w^{(n)} \|_{x_n} \tag{1.4.7}$$

назовем *X_n-погрешностью округления*, а величину

$$\zeta_n = \| p_n \tilde{w}^{(n)} - p_n w^{(n)} \|_{X_n} \tag{1.4.8}$$

X-погрешностью округления (или просто *погрешностью округления*).

3°. Не претендуя на полноту, приведем некоторые литературные указания. Вопросы оценки погрешности аппроксимации возникли почти одновременно с возникновением анализа. Из многочисленных работ последних десятилетий, в которых изучается погрешность аппроксимации, отметим монографии В. Вазова и Дж. Форсайта [9], Г. М. Вайникко [13], Р. С. Варга [15], Б. Г. Габдулхаева [30], М. К. Гавурина [32], Л. В. Канторовича и Г. П. Акилова [64], Л. В. Канторовича и В. И. Крылова [65], М. А. Красносельского с соавторами [73], Г. И. Марчука [80], автора этой книги [87, 96, 182, 108], Ж.-П. Обэна [125], Л. А. Оганесяна и Л. А. Руховца [126], С. Пресдорфа и Б. Зильбермана [192], С. Л. Соболева [132], Дж. Стрэнга и Дж. Фикса [199], Ф. Сьярле [134]. Понятие погрешности искажения было первоначально введено в работах автора [85, 86, 89]: в первых двух исследована устойчивость метода Ритца

по отношению к погрешностям искажения, в третьей получены некоторые общие теоремы об устойчивости вычислительных процессов, которые выше были названы свободными. Самый термин «искажение» ввела Л. Н. Довбыш в работе [54], посвященной погрешностям искажения собственных чисел и собственных подпространств самосопряженных операторов. В работе Г. Н. Ясковой и М. Н. Яковлева [153] исследована устойчивость метода Бубнова — Галеркина относительно погрешностей искажения. М. А. Велиев [16, 17] исследовал устойчивость метода Бубнова — Галеркина для ряда нестационарных задач, линейных и нелинейных. Перечисленные здесь работы, относящиеся к устойчивости относительно искажения и опубликованные не позднее 1965 г., суммированы в книге автора [90], в которой, кроме того, содержатся некоторые результаты, относящиеся к нелинейным стационарным задачам. Дальнейшие результаты по погрешностям искажения свободных вычислительных процессов содержатся в работах автора [94, 96, 122, 181, 107, 108], Ю. К. Демьяновича [45], Т. С. Таккера [201], Б. Дж. Омоден [189], С. Г. Суворова [133]. Тот же вопрос для рекуррентных вычислительных процессов рассмотрен в работах автора [105, 106]. Велиев продолжал исследование устойчивости для нестационарных задач в работах [18—24].

Вопрос об оценке погрешности алгоритма во многих случаях достаточно прост и, насколько известно автору, специально не исследовался (если не считать учебной литературы). Впрочем, имеются случаи, когда погрешность алгоритма пока не удается оценить. Таковы, например, некоторые односторонние вариационные задачи.

Много работ посвящено погрешностям округления; эти работы относятся главным образом к различным методам решения линейных алгебраических систем. Отметим здесь работы Г. М. Голуба [172], Дж. Х. Уилкинсона [138, 202]. Некоторые другие случаи рассмотрены автором [105, 106].

В заключение отметим, что некоторая классификация погрешностей приведена в учебниках [6, 7]. В этой классификации различаются три типа погрешностей: 1) неустранимая погрешность, вытекающая из неточностей постановки задачи (1.3.1); 2) погрешность метода — она в существенном совпадает с определенной выше погрешностью аппроксимации; 3) вычислительная погрешность (в [7] она названа погрешностью округления), включающая в себя введенные выше погрешности искажения, алгоритма и округления.

Соображения автора относительно неустранимой погрешности изложены выше, в п. 1° настоящего параграфа. Что касается вычислительной погрешности, то нам хотелось бы привести некоторые соображения в пользу нашей более подробной классификации. Мы уже отмечали в Предисловии, что введенные нами погрешности более элементарны по своей структуре

и потому легче поддаются анализу. В частности, погрешность искажения зависит лишь от того, насколько точно вычислены объекты, которые являются данными в приближенной задаче (1.4.2); она не зависит от того, какой алгоритм применяется для решения этой последней задачи и с какой точностью выполняются предписания этого алгоритма. От его выбора зависит погрешность алгоритма. Наконец, погрешность округления (в том смысле, какой мы придаем этому термину) при выбранном алгоритме полностью определяется той точностью, с которой выполняются предписания этого алгоритма. Таким образом, вычислительная погрешность распадается на три слагаемых, каждое из которых определяется независимо.

Наша классификация впервые дана в статье [105].

§ 5. ПОГРЕШНОСТИ КВАДРАТУРНЫХ ФОРМУЛ

Цель настоящего параграфа — проиллюстрировать на сравнительно простом примере общие соображения § 4.

1°. Погрешность аппроксимации для квадратурных формул изучалась с давних времен. Мы приведем здесь оценки этой погрешности для наиболее распространенных квадратурных формул. Через (a, b) обозначен промежуток интегрирования, через n — число промежутков разбиения, через f — подынтегральная функция.

1) Формула прямоугольников

$$\rho_n \leqslant \frac{(b-a)^2}{2n} \|f'\|_{C[a, b]};$$

2) формула средних прямоугольников

$$\rho_n \leqslant \frac{(b-a)^3}{24n^2} \|f''\|_{C[a, b]};$$

3) формула трапеций

$$\rho_n \leqslant \frac{(b-a)^3}{12n^2} \|f''\|_{C[a, b]};$$

4) формула Симпсона

$$\rho_n \leqslant \frac{(b-a)^5}{180n^4} \|f^{(4)}\|_{C[a, b]};$$

5) формула Гаусса $(a, b) = (-1, 1)$,

$$\rho_n \leqslant \frac{2^{2n+1}(n!)^4}{(2n+1)[(2n)!]^3} \|f^{(n)}\|_{C[-1, 1]};$$

6) формула Эйлера-Маклорена

$$\rho_n \leqslant \frac{nh^{2m+1}B_{2m}}{(2m)!} \|f^{(2m)}\|_{C[a, b]}.$$

В последней формуле m — произвольно выбранное натуральное число, от которого зависит точность формулы Эйлера — Маклорена, $h = (b-a)/n$, B_{2m} — числа Бернулли.

Отметим еще построенные Соболевым [132] кубатурные формулы, которые для подынтегральных функций класса $W_2^{(m)}(\Omega)$ (Ω — область интегрирования) приводят к наименьшей или близкой к наименьшей погрешности аппроксимации.

2°. Исследование остальных погрешностей мы проведем для класса квадратурных формул вида

$$\int_a^b f(t)\,dt \approx \sum_{k=1}^{n} c_k f(t_k), \tag{1.5.1}$$

обладающих следующими свойствами: $c_k \geqslant 0$; формула (1.5.1) точна, если $f(t) \equiv \text{const}$, так что

$$\sum_{k=1}^{n} c_k = b - a. \tag{1.5.2}$$

На узлы t_k мы не накладываем никаких ограничений. Подынтегральную функцию предполагаем ограниченной: $|f(t)| \leqslant M = \text{const}$. Пусть коэффициенты c_k и ординаты $y_k = f(t_k)$ вычислены неточно, так что на самом деле нам известны их искаженные значения $\tilde{c}_k = c_k + \gamma_k$ и $\tilde{y}_k = y_k + \alpha_k$. Мы предполагаем при этом, что

$$|\alpha_k| \leqslant \alpha, \quad \sum_{k=1}^{n} |\gamma_k| \leqslant \gamma, \tag{1.5.3}$$

где α и γ — малые числа, и что искаженные коэффициенты $\tilde{c}_k$ удовлетворяют условиям, сформулированным выше:

$$\tilde{c}_k \geqslant 0, \quad \sum_{k=1}^{n} \tilde{c}_k = b - a.$$

Погрешность искажения формулы (1.5.1) равна

$$\xi_n = \left| \sum_{k=1}^{n} [(c_k + \gamma_k)(y_k + \alpha_k) - c_k y_k] \right| =$$

$$= \left| \sum_{k=1}^{n} (\gamma_k y_k + \tilde{c}_k \alpha_k) \right| \leqslant M\gamma + \alpha(b-a). \tag{1.5.4}$$

Таким образом, малые искажения коэффициентов и ординат приводят к малому искажению интеграла, вычисленного по квадратурной формуле.

После того как ординаты определены, алгоритм вычисления правой части формулы (1.5.1) содержит только конечное число действий сложения и умножения, поэтому для квадратурной формулы погрешность алгоритма равна нулю.

3°. Оценим, наконец, погрешность округления. Мы вычисляем интеграл по искаженной квадратурной формуле

$$\int_a^b f(t)\,dt \approx \sum_{k=1}^{n} c_k \tilde{y}_k.$$

Допуская, что вычисления производятся с повышенной точностью, находим по формуле (1.1.9)

$$\left|\delta\left(\sum_{k=1}^{n}\tilde{c}_k\tilde{y}_k\right)\right|\leqslant\varepsilon_1\left|\sum_{i=1}^{n}\tilde{c}_k\tilde{y}_k\right|\leqslant\varepsilon_1\sum_{i=1}^{n}|\tilde{c}_k|\,|\tilde{y}_k|$$

и с точностью до малых высшего порядка

$$\left|\delta\left(\sum_{k=1}^{n}\tilde{c}_k\tilde{y}_k\right)\right|\leqslant\varepsilon_1\int_a^b|f(t)|\,dt. \qquad (1.5.5)$$

Формула (1.5.5), точнее, ее правая часть, и дает оценку погрешности округления квадратурной формулы. Общая оценка погрешности квадратурной формулы имеет вид

$$\rho_n+M\gamma+\alpha+\varepsilon_1\int_a^b|f(t)|\,dt. \qquad (1.5.6)$$

4°. **Пример.** Допустим, что мы вычисляем на ЭВМ БЭСМ-6 по формуле Симпсона интеграл

$$\int_1^2 t^{-1}\,dt=\ln 2.$$

Пусть число промежутков деления $2n=2^k$, где k невелико. Погрешность аппроксимации имеет оценку

$$\frac{24\cdot 2^{-4k}}{180}=\frac{2}{15}\cdot 2^{-4k}.$$

Коэффициенты квадратурной формулы в данном случае записываются точно, так что $\gamma_k=0$. Далее, $|\alpha_k|\leqslant\varepsilon_1|y_k|\leqslant\varepsilon_1$. По формуле (1.5.4) погрешность искажения $\xi_n\leqslant\varepsilon_1$. Наконец, формула (1.5.5) дает оценку погрешности округления ε_1. Общая погрешность оценивается так:

$$\delta_n\leqslant\frac{2}{15}\cdot 2^{-4k}+2\varepsilon_1.$$

Если мы возьмем $k=8$, то получим оценку $\delta_n\leqslant 3^{-1}\cdot 10^{-10}+$ $+0{,}364\cdot 10^{-11}$, т. е. величина $\ln 2$ будет вычислена с десятью верными знаками. Дальнейшее увеличение k, по-видимому, нецелесообразно из-за наличия слагаемого $2\varepsilon_1$ в оценке погрешности.

5°. Нетрудно проанализировать также погрешности формул типа формулы Эйлера — Маклорена, которая содержит некоторое число отрицательных коэффициентов, причем последние умножаются не только на значения подынтегральной функции, но и на значения некоторых ее производных. Не вносит новых трудностей и исследование погрешностей кубатурных формул при любом числе координат.

§ 6. О ПОГРЕШНОСТЯХ РЕШЕНИЙ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ СИСТЕМ

1°. Вопрос об оценке погрешностей, возникающих при решении задач линейной алгебры, будет подробно освещен в разделе II. Здесь мы ставим более ограниченную задачу: дать еще на одном примере иллюстрацию общих понятий § 4. С этой целью мы рассмотрим задачу о решении линейной алгебраической системы с квадратной неособенной матрицей и выясним, с какими погрешностями приходится иметь дело при численном решении упомянутой задачи. Разумеется, ответ зависит от применяемого метода.

2°. Обычные методы решения линейных алгебраических систем таковы, что они применяются либо непосредственно к данной системе, либо к системе, полученной из данной тем или иным эквивалентным преобразованием. Отсюда следует, что в задаче о численном решении линейной алгебраической системы погрешность аппроксимации равна нулю.

3°. Как известно [139], методы решения линейных алгебраических систем распадаются на две группы: точные (их называют также прямыми) и итеративные методы. Из точных особенно важны методы исключения, к которым относится метод Гаусса, его частный случай — метод прогонки, метод матриц вращения, метод матриц отражения и др. Отличительная черта этих методов состоит в том, то данная система с матрицей общего вида преобразуется путем последовательного исключения неизвестных к системе с треугольной матрицей; эта операция называется *прямым ходом* данного метода. *Обратный ход* состоит в решении полученной треугольной системы. Таким образом, прямой ход сводит данную задачу — решение системы с произвольной матрицей — к эквивалентной, но более простой задаче — к решению системы с треугольной матрицей. Сопоставляя сказанное с определениями § 4, заключаем, что прямой ход в методах исключения связан с погрешностью искажения. Обратный ход, как легко видеть, связан с погрешностью округления. Погрешность алгоритма в точных методах равна нулю.

По методу Холецкого (методу квадратных корней), идейно близкому к методам исключения, решение системы с симметричной положительно определенной матрицей сводится к решению двух сопряженных систем с треугольными матрицами. Можно считать, что в этом методе прямой ход состоит в построении упомянутых треугольных систем, а обратный — в решении этих систем. Как и в методах исключения, в методе Холецкого с прямым ходом связана погрешность искажения, с обратным — погрешность округления. Метод квадратных корней — точный, и погрешность аппроксимации для него равна нулю.

4°. Рассмотрим теперь итерационные методы. Если система такова, что к ней можно непосредственно применить метод последовательных приближений, то в этом случае мы будем иметь дело с погрешностями алгоритма и округления. Чаще бывает так, что систему надо предварительно преобразовать к такому виду, чтобы применение метода последовательных приближений стало возможным. В этом случае к названным выше погрешностям добавляется погрешность искажения — она связана с упомянутым преобразованием системы.

Для итерационных методов решения линейных алгебраических систем, так же, как и для точных методов, погрешность аппроксимации равна нулю.

§ 7. АПОСТЕРИОРНЫЕ ОЦЕНКИ ПОГРЕШНОСТЕЙ

1°. Оценки погрешностей, о которых шла речь в предшествующих параграфах, могут быть построены до того, как само приближенное решение построено; поэтому упомянутые оценки принято называть априорными. Априорность оценок является их большим преимуществом: оценив суммарную погрешность до того, как задача решена, можно судить о том, приемлемо или нет приближенное решение с такой погрешностью; если нет, то надо предварительно перестроить точность вычислений так, чтобы новая суммарная оценка погрешности оказалась приемлемой.

2°. В ряде случаев, однако, вычисление априорных оценок погрешности оказывается весьма трудоемким — иногда существенно более трудоемким, чем вычисление самого приближенного решения. В этом случае может оказаться целесообразным построение так называемых апостериорных оценок погрешности, которые определяются по уже построенным приближенным решениям. Ниже мы кратко опишем два случая, когда построение апостериорных оценок оказывается возможным; подробнее об этом сказано в монографии автора [84].

3°. Пусть A — линейный оператор, действующий из банахова пространства B_1 в другое банахово пространство B_2, и пусть требуется решить уравнение

$$Ax = f. \tag{1.7.1}$$

Допустим, что тем или иным способом нам удалось построить элемент $\tilde{x} \in D(A) \subset B_1$, который мы рассматриваем как приближенное решение уравнения (1.7.1). Если существует обратный оператор A^{-1} и известна оценка сверху его нормы, то можно оценить норму разности $x_* - \tilde{x}$, где x_* — точное решение уравнения (1.7.1). Именно

$$x_* - \tilde{x} = A^{-1}A(x_* - \tilde{x}) = A^{-1}(Ax_* - A\tilde{x}) = A^{-1}(f - A\tilde{x});$$

отсюда и вытекает апостериорная оценка погрешности

$$\|x_* - \bar{x}\|_{B_1} \leqslant \|A^{-1}\|_{B_2 \to B_1} \cdot \|f - A\bar{x}\|_{B_2}. \qquad (1.7.2)$$

Оценка (1.7.2) не учитывает погрешности вычисления элемента $A\bar{x}$, поэтому, используя эту оценку, желательно позаботиться о том, чтобы этот элемент был вычислен достаточно точно.

Сходную апостериорную оценку можно получить и для некоторого класса нелинейных уравнений. Пусть в уравнении (1.7.1) A — нелинейный оператор, действующий из банахова пространства B в сопряженное пространство B^*, и пусть этот оператор — сильно монотонный. Это значит, что

$$\forall x,\ y \in B,\ (Ax - Ay,\ x - y) \geqslant \mu(\|x - y\|) \cdot \|x - y\|, \qquad (1.7.3)$$

где $\mu(t)$ — непрерывная строго монотонная функция на R_1^+, причем $\mu(0) = 0$ [70]. Пусть, как и выше, x_* — точное решение уравнения (1.7.1), а $\bar{x} \in D(A)$ — элемент пространства B, рассматриваемый как приближенное решение того же уравнения. По неравенству (1.7.3)

$$\mu(\|x_* - \bar{x}\|)\|x_* - \bar{x}\| \leqslant (f - A\bar{x},\ x_* - \bar{x}) \leqslant \|f - A\bar{x}\| \cdot \|x_* - \bar{x}\|.$$

Отсюда следует искомая оценка

$$\|x_* - \bar{x}\| \leqslant \mu^{-1}(\|f - A\bar{x}\|). \qquad (1.7.4)$$

4°. Необходимо указать, что оценки (1.7.2), (1.7.4) могут оказаться грубыми: может случиться, что $\bar{x}$ близко к x_*, тогда как $A\bar{x}$ не близко к f. Так, если оператор A неограничен, то может оказаться, что $\bar{x} \notin D(A)$ (так бывает в методе Ритца и, в частности, в методе конечных элементов), тогда названные оценки просто теряют смысл. Если же $\bar{x} \in D(A)$, то в общем случае норма $\|f - A\bar{x}\|$ может оказаться сколь угодно большой при сколь угодно малой норме $\|x_* - \bar{x}\|$. Отмеченная здесь опасность не возникает, если уравнение (1.7.1) линейное, а приближенное решение построено по методу наименьших квадратов. Подробнее об этом сказано в [84].

5°. Формулы (1.7.2), (1.7.4) могут оказаться недостоверными, если нельзя пренебречь погрешностью элемента $\bar{f} = A\bar{x}$. Пусть $\delta(\bar{f})$ — погрешность элемента $\bar{f}$, и $\|\delta(\bar{f})\| \leqslant \varepsilon$. Обозначим вычисленное значение $\bar{f}$ через $\hat{f}$, так что $\hat{f} = \bar{f} + \delta(\bar{f})$. Тогда $\|f - \bar{f}\| \leqslant |f - \hat{f}\| + \varepsilon$ и достоверная апостериорная оценка принимает вид

$$\|x_* - \bar{x}\| \leqslant \|A^{-1}\|\,[\|f - \hat{f}\| + \varepsilon], \qquad (1.7.5)$$

если оператор A — линейный, и

$$\|x_* - \bar{x}\| \leqslant \mu(\|f - \hat{f}\| + \varepsilon), \qquad (1.7.6)$$

если этот оператор нелинейный и сильно монотонный.

6°. Изложим еще один способ построения апостериорных оценок. Этот способ имеет более ограниченную область применения, чем способ п. 3°, но он свободен от недостатка, отмеченного в п. 4°. Пусть A — линейный положительно определенный оператор, действующий из рефлективного банахова пространства B в сопряженное пространство B^*. Тогда задача (1.7.1) равносильна задаче о минимуме функционала

$$F(x)=(Ax,\ x)-2(f,\ x),\quad x\in B_0, \tag{1.7.7}$$

где B_0 — энергетическое пространство оператора A. По поводу используемых в этом пункте понятий см. [8], а также [90].

Обозначая по-прежнему точное решение задачи (1.7.1) через x_*, имеем $F(x)=|x-x_*|^2-|x_*|^2$; здесь $|\cdot|$ — норма в B_0. Пусть $\bar{x}\in B_0$ — приближенное решение задачи (1.7.1); таким приближенным решением может служить, например, любой член последовательности, минимизирующей функционал (1.7.7). Тогда

$F(\bar{x})=|\bar{x}-x_*|^2-|x_*|^2$ и

$$|\bar{x}-x_*|=\sqrt{F(\bar{x})+|x_*|^2}. \tag{1.7.8}$$

Однако это точное равенство не позволяет судить о погрешности приближенного решения, так как норма $|x_*|$ неизвестна.

Допустим, что нам удалось построить в некотором (может быть, новом) пространстве E функционал $\Phi(y)$, такой, что

$$\inf\Phi(y)=|x_*|^2. \tag{1.7.9}$$

Для любого $\bar{y}\in E$ будет $\Phi(\bar{y})\geqslant|x_*|^2$: если при этом в качестве $\bar{y}$ взять достаточно далекий член последовательности, минимизирующей функционал Φ, то

$$|\bar{x}-x_*|\leqslant\sqrt{F(\bar{x})+\Phi(\bar{y})}, \tag{1.7.10}$$

и правая часть неравенства (1.7.10) сколь угодно мало превосходит левую.

Функционалы F и Φ мы называем *встречными*. В [84], в случае, когда B — гильбертово пространство, приведены два функционала, встречных для функционала (1.7.7). Один из них связан с методом ортогональных проекций, впервые предложенным Зарембой в 1909 г.; второй функционал связан с методом Трефца (1926 г.) и с принципом Кастильяно в теории упругости.

Метод встречных функционалов может быть использован в известных условиях и для нелинейных задач. Одной такой задаче посвящена глава 16 данной монографии.

§ 8. АПОСТЕРИОРНАЯ ОЦЕНКА ПОГРЕШНОСТИ РЕШЕНИЯ ЛИНЕЙНОЙ АЛГЕБРАИЧЕСКОЙ СИСТЕМЫ

1°. Пусть в евклидовом пространстве R_m дано уравнение

$$Ax = f, \tag{1.8.1}$$

где A — неособенная квадратная матрица, $A = [a_{ij}]_{i,j=1}^{m}$; $f = (f_1, f_2, \ldots, f_m)'$, и пусть каким-нибудь способом найден вектор $\bar{x} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m)'$, который мы рассматриваем как приближенное решение уравнения (1.8.1). Оценим норму $\|\bar{x} - x_*\|$, где x_* — точное решение уравнения (1.8.1). Прежде всего вычислим вектор $\bar{f} = A\bar{x}$. Имеем

$$\bar{f}_j = \sum_{k=1}^{m} a_{jk}\bar{x}_k. \tag{1.8.2}$$

Машинное вычисление по формуле (1.8.2) приводит к некоторой погрешности. Оценим ее. Допустим сначала, что промежуточные вычисления производятся с обычной точностью. Тогда по формулам (1.1.6) находим

$$(a_{jk}\bar{x}_k)_{\text{м}} = a_{jk}\bar{x}_k + \theta_{jk}\varepsilon_1 |a_{jk}\bar{x}_k|, \quad |\theta_{jk}| \leqslant 1, \tag{1.8.3}$$

и

$$\left|\left(\sum_{k=1}^{m}(a_{jk}\bar{x}_k)_{\text{м}}\right)_{\text{м}} - \sum_{k=1}^{m}(a_{jk}\bar{x}_k)_{\text{м}}\right| \leqslant \varepsilon_1 (m-1)\sum_{k=1}^{m}|(a_{jk}\bar{x}_k)_{\text{м}}|.$$

В последней формуле заменим $(a_{jk}\bar{x}_k)_{\text{м}}$ по формуле (1.8.3) и отбросим члены порядка $O(\varepsilon_1^2)$. В результате получим

$$\left|\left(\sum_{k=1}^{m}(a_{jk}\bar{x}_k)_{\text{м}}\right)_{\text{м}} - \sum_{k=1}^{m}a_{jk}\bar{x}_k\right| \leqslant \varepsilon_1 m \sum_{k=1}^{m}|a_{jk}\bar{x}_k|. \tag{1.8.4}$$

Величина $\left(\sum_{k=1}^{m}(a_{jk}\bar{x}_k)_{\text{м}}\right)_{\text{м}} = (\bar{f}_j)_{\text{м}}$ есть машинное значение числа $\bar{f}_j$. Отсюда $|\delta(\bar{f}_j)| \leqslant \varepsilon_1 m \sum_{k=1}^{m}|a_{jk}\bar{x}_k|$, а погрешность машинного значения нормы $\|\delta(\bar{f})\|$ не превосходит величины

$$\varepsilon_1 m\left[\sum_{j=1}^{m}\left(\sum_{k=1}^{m}|a_{jk}\bar{x}_k|\right)^2\right]^{1/2} \leqslant \varepsilon_1 m \|A\|_E \cdot \|\bar{x}\|.$$

Далее,

$$\|A\bar{x} - f\| = \|\bar{f} - f\| \leqslant \|(\bar{f})_{\text{м}} - f\| + \|\delta(\bar{f})\| \leqslant$$
$$\leqslant \|(\bar{f})_{\text{м}} - f\| + \varepsilon_1 m \|A\|_E \cdot \|\bar{x}\|. \tag{1.8.5}$$

Обозначая через x_*, как и выше, точное решение уравнения (1.8.1), получаем

$$\|\bar{x} - x_*\| \leqslant \|A^{-1}\| \cdot \|A\bar{x} - f\| \leqslant$$
$$\leqslant \|A^{-1}\| \cdot \|(\bar{f})_{\text{м}} - f\| + \varepsilon_1 m \|A^{-1}\| \cdot \|A\|_E \cdot \|\bar{x}\|. \tag{1.8.6}$$

2°. Если вычисления производятся с повышенной точностью, то с точностью до членов высшего порядка малости находим по формуле (1.1.9) $|\delta(\bar{f}_j)| \leqslant \varepsilon_1 |\bar{f}_j|$ и $\|\delta(\bar{f})\| \leqslant \varepsilon_1 \|\bar{f}\|$. Вектор $\bar{f}$ на самом деле неизвестен; заменим его на $(\bar{f})_м$, что приведет к изменению правых частей последних неравенств на величины порядка $O(\varepsilon_1^2)$. Поэтому можно принять, что $\|\delta(\bar{f})_м\| \leqslant \varepsilon_1 \|(\bar{f})_м\|$, а тогда

$$\|A\bar{x} - f\| \leqslant \|(\bar{f})_м - f\| + \varepsilon_1 \|(\bar{f})_м\|$$

и окончательно

$$\|\bar{x} - x_*\| \leqslant \|A^{-1}\| \cdot \|(\bar{f})_м - f\| + \varepsilon_1 \|A^{-1}\| \cdot \|(\bar{f})_м\|. \qquad (1.8.7)$$

3°. Вторые члены справа в формулах (1.8.6), (1.8.7) малы; если малы и первые члены, то приближенное решение $\bar{x}$ оказывается удовлетворительным. Нетрудно видеть, что это приближенное решение неудовлетворительно, если норма $\|(\bar{f}) - f\|$ не мала. Действительно, $\|A\bar{x} - f\| \leqslant \|A\| \cdot \|\bar{x} - x_*\|$. Отсюда $\|\bar{x} - x_*\| \geqslant \|A\|^{-1} \cdot \|\bar{f} - f\|$. Справа заменим f на $(\bar{f})_м$; тогда получим неравенство $\|\bar{x} - x_*\| \geqslant \|A\|^{-1} \cdot \|(\bar{f})_м - f\|$, верное с точностью до малых высших порядков, и разность $\bar{x} - x_*$ оказывается не малой.

Глава 2

ПОГРЕШНОСТЬ АППРОКСИМАЦИИ

Ниже на протяжении всей книги мы будем пользоваться следующими обозначениями. Норму в $W_2^{(s)}(\Omega)$ будем обозначать через $\|\cdot\|_{(s),\Omega}$ или, если это не может вызвать недоразумений, через $\|\cdot\|_{(s)}$. Обозначения $\|\cdot\|_s$ или $\|\cdot\|_{s,\Omega}$ мы сохраним для нормы в $L_s(\Omega)$.

Степенью одночлена $t_1^{\alpha_1} t_2^{\alpha_2} \ldots t_m^{\alpha_m}$ в R_m назовем мультииндекс $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$; степенью многочлена в R_m будем называть верхнюю грань степеней составляющих его одночленов. Вектор или мультииндекс, все составляющие которого равны одному и тому же числу a, будем обозначать через $\underline{a}$. Таким образом, если $P(t)$ — многочлен в R_m и по каждой из координат его степень не превосходит некоторого числа n, то степень $P(t)$ в R_m не превосходит $\underline{n}$. Будем записывать это так: $\mathrm{gr}\, P_n \leqslant \underline{n}$.

Буквой C без индексов будем обозначать постоянные, вообще говоря, различные, точные значения которых для нас несущественны.

§ 1. МЕТОД РИТЦА. ОЦЕНКИ В ЭНЕРГЕТИЧЕСКОЙ МЕТРИКЕ

1°. Напомним основы метода Ритца для линейных задач; подробно этот метод изложен в [84]. Пусть симметричный билинейный функционал $A(x, y)$ и соответствующий ему однородный квадратичный функционал $A(x)=A(x, x)$ определены на некотором линейном множестве M. Допустим, что функционал $A(x)$ положителен; это значит, что $A(x)>0$, если x отличен от нулевого элемента множества M. Превратим M в предгильбертово пространство, введя скалярное произведение $[x, y]=A(x, y)$ и норму $|x|=\sqrt{[x, x]}=\sqrt{A(x)}$. Замкнув M в этой норме, получим гильбертово пространство, которое обозначим через H_A. Для упрощения последующих записей примем, что пространство H_A — вещественное. В приложениях чаще встречается случай, когда M есть линейное плотное множество некоторого гильбертова (реже банахова) пространства H, и $A(x, y)=(Ax, y)$, где A — положительный (и симметричный — эта оговорка необходима, если H — вещественное пространство) оператор в H. В этом случае мы назовем H_A энергетическим пространством оператора A, а скалярное произведение и норму в H_A — соответственно энергетическим произведением и энергетической нормой.

Пусть $f(x)$ — линейный функционал, ограниченный в H_A и определенный на всем этом пространстве. Поставим задачу о минимуме функционала

$$F(x)=A(x)-2(f, x). \tag{2.1.1}$$

Эта задача теоретически решается элементарно. По известной теореме Ф. Риса, в H_A существует один и только один элемент x_*, такой, что $f(x)=[x, x_*]$, $\forall x \in H_A$. Теперь $F(x)=|x|^2-2[x, x_*]=|x-x_*|^2-|x_*|^2$, и очевидно, что функционал (2.1.1) достигает минимума при $x=x_*$.

2°. Задачу (2.1.1) можно приближенно решить, например, по классическому методу Ритца. Выбираем в H_A какое-нибудь конечномерное подпространство $H_A^{(n)}$ и ищем минимум функционала $F(x)$, или, что равносильно, нормы $|x-x_*|$, на этом подпространстве. Элемент $x_*^{(n)} \in H_A^{(n)}$, на котором последний минимум достигается, существует и единствен. Очевидно, $x_*^{(n)}$ есть ортогональная (в метрике H_A) проекция элемента x_* — точного решения задачи (2.1.1) — на подпространство $H_A^{(n)}$. Сравнивая это с общей схемой § 3 главы 1, видим, что в данном случае $X=H_A$, $X_n=H_A^{(n)}$, p_n есть тождественный оператор, соотношения r и r_n определяются формулами

$$xrf:=|x|^2-2f(x)=\min,\ x \in H_A;$$

$$x^{(n)} r_n f^{(n)}:=|x^{(n)}|-2f^{(n)}(x^{(n)})=\min,\ x^{(n)} \in H_A^{(n)}$$

и, наконец, $f^{(n)}$ есть сужение функционала f на подпространство $H_A^{(n)}$. Отсюда видно, что погрешность аппроксимации приближенного решения по Ритцу для задачи (2.1.1), равная $\rho_n = |x_* - x_*^{(n)}| = \min\limits_{x^{(n)} \in H_A^{(n)}} |x_* - x^{(n)}|$, совпадает с наилучшим (в энергетической норме) приближением точного решения x_* элементами подпространства $H_A^{(n)}$.

В ближайших параграфах настоящей главы мы применим высказанные здесь общие соображения к краевым задачам для дифференциальных уравнений.

Разумеется, применять метод Ритца целесообразно только тогда, когда есть уверенность, что $|x_*^{(n)} - x_*| \to 0$. Очевидно, что последнее соотношение будет выполнено, если последовательность подпространств $\{H_A^{(n)}\}$ полна в H_A, иначе говоря, если по любому $\varepsilon > 0$ можно найти такое n_0, что при любом $n \geqslant n_0$ существует элемент $x^{(n)} \in H_A^{(n)}$, удовлетворяющий неравенству $|x_* - x^{(n)}| < \varepsilon$.

В основополагающей работе Ритца [194] подпространства $H_A^{(n)}$ строятся так, что при возрастании n они расширяются. Более определенно Ритц выбирает последовательность элементов $\{\psi_n\} \subset H_A$, полную в H_A; предполагается еще, что эти элементы, взятые в любом конечном числе, линейно независимы. За $H_A^{(n)}$ Ритц принимал подпространство с базисом $(\varphi_1, \varphi_2, \ldots, \varphi_n)$. Позднее эта система была названа координатной. Отметим, что Ритц не вводил энергетического пространства и требовал сходимости приближенных решений к точному в $C^{(s)}$ при подходящем выборе s. Это ограничивало применение его метода и осложняло рассуждения. Энергетическая метрика была введена Фридрихсом; она была применена автором этой книги для исследования методов Ритца и Бубнова — Галеркина. (Подробно см. [84].)

Р. Курант [161] показал, что необязательно брать расширяющиеся подпространства $H_A^{(n)}$, и предложил такое видоизменение метода Ритца: пространства $H_A^{(n)}$ суть произвольные конечномерные подпространства энергетического пространства, подчиненные единственному требованию полноты их последовательности в H_A. Если выбрать в $H_A^{(n)}$ базис

$$(\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nN}), \tag{2.1.2}$$

где $N = N(n)$ есть размерность $H_A^{(n)}$, то элементы φ_{nk} необязательно принадлежат подпространству $H_A^{(n')}$, $n' > n$; таким образом, может случиться, что $H_A^{(n)}$ и $H_A^{(n')}$ имеют общим только нулевой элемент.

В случае расширяющихся подпространств будем говорить о классическом методе Ритца, в общем случае — об обобщен-

ном методе Ритца. Классический метод Ритца имеет то преимущество, что погрешность $|x_* - x_*^{(n)}|$ убывает с ростом n [84]; следует, однако, напомнить, что обобщенный метод Ритца является идейной основой метода конечных элементов — одного из самых распространенных в настоящее время методов решения краевых задач.

В заключение напомним, что если в $H_A^{(n)}$ выбран базис (2.1.2), то приближенное решение имеет вид

$$x_*^{(n)} = \sum_{k=1}^{n} a_k^{(n)} \varphi_{nk}, \tag{2.1.3}$$

причем коэффициенты $a_k^{(n)}$ определяются из алгебраической системы

$$\sum_{k=1}^{n} [\varphi_{nk}, \varphi_{nj}] a_k^{(n)} = f(\varphi_{nj}), \quad j = 1, 2, \ldots, N. \tag{2.1.4}$$

Последовательность базисов

$$(\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nN}), \quad n = 1, 2, \ldots, \tag{2.1.5}$$

будем называть координатной системой обобщенного метода Ритца, а элементы φ_{nk} — координатными элементами этого метода.

§ 2. ОБОБЩЕНИЯ ТЕОРЕМЫ ДЖЕКСОНА НА ФУНКЦИИ МНОГИХ ПЕРЕМЕННЫХ

1°. Наиболее важное, как кажется автору, обобщение теоремы Джексона на функции многих переменных получила И. Ю. Харрик [143—145]. Ниже мы сформулируем ее основной результат.

Пусть Ω — конечная область в R_m, r и $s < r$ — натуральные числа. Далее, пусть $\varphi(t)$ — функция класса $C^{(\nu)}$ в некоторой области $\Omega_1 \supset \bar{\Omega}$, причем ν достаточно велико, Ω_1 гомеоморфна шару, $\varphi|_{\partial\Omega} = 0$, $\operatorname{grad}\varphi|_{\partial\Omega} \neq 0$ и $\varphi(t) \neq 0$, $t \in \Omega$.

Теорема 2.2.1. *Пусть* $x \in C^{(r)}(\bar{\Omega})$ и $D^\alpha x|_{\partial\Omega} = 0$, $0 \leqslant |\alpha| \leqslant s - 1$. *Тогда при любом натуральном* n *существует полином* $P_n(t)$, $\operatorname{gr} P_n \leqslant n$, *удовлетворяющий неравенству*

$$\|x - \varphi^{(s)} P_n\|_{C^{(\bar{r})}(\bar{\Omega})} \leqslant C\omega^{(r)}(x, 1/n)\, n^{-(r-\bar{r})}, \quad 0 \leqslant \bar{r} \leqslant r. \tag{2.2.1}$$

Здесь $\omega^{(r)}$ — *наибольший из модулей непрерывности производных порядка* r *от функции* $x(t)$ *в метрике* $C(\bar{\Omega})$, *а величина* C *не зависит ни от* x, *ни от* n.

Доказательство теоремы 2.2.1 основано на построении операторов, аналогичных операторам Джексона. Постоянную C можно оценить.

2°. Т. О. Шапошникова [150] обобщила теорему 2.2.1 на пространства $W_2^{(r)}$. Приводим это обобщение.

Теорема 2.2.2. *Пусть область* Ω *и функция* $\varphi(t)$ *удовлетворяют условиям* п. 1°. *Пусть, далее,* $x \in W_2^{(r)}(\Omega)$ *и* $D^\alpha x|_{\partial\Omega} = 0$, $0 \leqslant |\alpha| \leqslant s-1$. *При любом натуральном* n *существует полином* $P_n(t)$, $\operatorname{gr} P_n \leqslant \underline{n}$, *удовлетворяющий неравенству*

$$\|x - \varphi^s\|_{(\bar{r}),\,\Omega} \leqslant C\omega_{L_2}^{(r)}(x,\, 1/n)\, n^{-(r-\bar{r})},\quad 0 \leqslant \bar{r} \leqslant r, \qquad (2.2.2)$$

где C *не зависит от* x *и* n, *а* $\omega_{L_2}^{(r)}$ *есть наибольший из* L_2*-модулей непрерывности производных* $D^\alpha x$, $|\alpha| = r$.

Доказательство теоремы 2.2.2 построено на том, что основные свойства операторов, введенных Харрик, сохраняются при замене $C^{(k)}(\Omega)$ на $W_2^{(k)}(\Omega)$. Постоянную C в неравенстве (2.2.2) также можно оценить.

3°. **Следствие 2.2.2.** *Пусть граница* $\partial\Omega$ *представляет собой* $(m-1)$*-мерную липшицеву поверхность, и* $x \in W_2^{(r)}(\Omega)$. *Тогда существует полином* $Q_n(t)$, $\operatorname{gr} P_n \leqslant \underline{n}$, *удовлетворяющий неравенству*

$$\|x - Q_n\|_{(\bar{r}),\,\Omega} \leqslant C_1 \max_{|\alpha|=r} \|D^\alpha x\|_{2,\,\Omega} n^{-(r-\bar{r})},\quad 0 \leqslant \bar{r} \leqslant r. \qquad (2.2.3)$$

Если поверхность $\partial\Omega$ достаточно гладкая, так что существует функция $\varphi(t)$ со свойствами, описанными в п. 1°, то для доказательства следствия 2.2.1 достаточно положить в теореме 2.2.2 $s = 0$, заметив при этом, что $\omega_{L_2}^{(r)}(x, \delta) \leqslant 2 \max_{|\alpha|=r} \|x^{(\alpha)}\|_{2,\,\Omega}$.

В общем случае продолжим функцию $x(t)$ с сохранением класса на все пространство R_m так, чтобы она обращалась в нуль вне некоторого шара $\Omega_0 = \{t: |t| \leqslant R\}$ — такое продолжение возможно (см. [160], а также [180]). Пусть $x^*(t)$ — продолженная функция; тем же символом $x^*(t)$ обозначим ее сужение на Ω_0. Граница $\partial\Omega_0$ — сфера — есть бесконечно гладкая поверхность, и, по сказанному выше, для функции $x^*(t)$ неравенство (2.2.3) справедливо: существует такой полином $Q_n(t)$, $\operatorname{gr} P_n \leqslant \underline{n}$, что

$$\|x^* - Q_n\|_{(\bar{r}),\,\Omega} \leqslant C_1 \max_{|\alpha|=r} \|D^\alpha x^*\|_{2,\,\Omega_0} n^{-(r-\bar{r})}. \qquad (2.2.4)$$

Функция $x^*(t)$ продолжает функцию $x(t)$ с сохранением класса, поэтому существует такая постоянная продолжения $\varkappa$, что если $\|\alpha\| \leqslant r$, то

$$\|D^\alpha x^*\|_{2,\,\Omega_0} \leqslant \varkappa \|D^\alpha x\|_{2,\,\Omega};$$

кроме того, очевидно

$$\|x - Q_n\|_{(\bar{r}),\,\Omega} \leqslant \|x^* - Q_n\|_{(\bar{r}),\,\Omega_0}.$$

Последние два соотношения позволяют заменить неравенство (2.2.4) на (2.2.3) с постоянной $\varkappa C_1$.

4°. Вернемся к теореме 2.2.2. Пусть $m=1$ и область Ω есть промежуток $(0,1)$. В этом случае можно положить $\varphi(t)=t(1-t)$, и неравенство (2.2.2) дает следующее: если $x\in W_2^{(r)}(0,1)\cap \overset{\circ}{W}{}_2^{(s)}(0,1)$, то при любом натуральном n существует такой полином $P_n(t)$, $\operatorname{gr} P_n\leqslant n$, что

$$\|x-P_{n,s}\|_{(\bar r),(0,1)}\leqslant C\omega_{L_2}^{(r)}(x,1/n)\,n^{-(r-\bar r)}, \tag{2.2.5}$$

$$0\leqslant\bar r\leqslant r,\quad P_{n,s}(t)=t^s(1-t)^s P_n(t).$$

Точно так же из следствия 2.2.1 вытекает, что если $x\in W_2^{(r)}(0,1)$, то при любом натуральном n существует такой полином $Q_n(t)$, $\operatorname{gr} P_n\leqslant n$, что справедливо неравенство

$$\|x-Q_n\|_{(\bar r),(0,1)}\leqslant C\|D^r x\|_{2,(0,1)}\,n^{-(r-\bar r)},\quad 0\leqslant\bar r\leqslant r. \tag{2.2.6}$$

§ 3. ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

1°. Рассмотрим краевую задачу для уравнения второго порядка

$$Ax=-\frac{d}{dt}\left(p(t)\frac{dx}{dt}\right)+q(t)x=f(t),\quad 0<t<1;\qquad x(0)=x(1)=0. \tag{2.3.1}$$

Примем следующие допущения:

$$f\in L_2(0,1),\quad p\in C^{(1)}[0,1],\quad p_0\leqslant p(t)\leqslant p_1,\quad 0<p_0\leqslant p_1<\infty;$$
$$q\in L_\infty(0,1),\quad q(t)>0.$$

При этих допущениях оператор A — положительно определенный в $L_2(0,1)$, и слабое решение задачи (2.3.1) есть решение вариационной задачи

$$|x|^2-2\int_0^1 f(t)x(t)\,dt=\min,\quad x(0)=x(1)=0, \tag{2.3.2}$$

где

$$|x|^2=\int_0^1\left[p(t)\left(\frac{dx}{dt}\right)^2+q(t)x^2(t)\right]dt. \tag{2.3.3}$$

Пространство H_A состоит из функций $x(t)$, для которых интеграл (2.3.3) конечен и которые обращаются в нуль при $x=0$ и $x=1$. Полезна следующая оценка энергетической нормы: если $q(t)\leqslant q_1$, то

$$|x|^2\leqslant(p_1+q_1/\pi^2)\|x'\|^2:=c_0\|x'\|^2; \tag{2.3.4}$$

здесь $\|\cdot\|$ означает норму в $L_2(0,1)$. Доказательство следует из того, что наименьшее собственное число оператора $-d^2/dt^2$ при краевых условиях $x(0)=x(1)=0$ равно π^2.

2°. За $H_A^{(n)}$ примем n-мерное подпространство с базисом $\{\sin k\pi t\}$, $1 \leqslant k \leqslant n$. Если $x^{(n)}$ — произвольный элемент из $H_A^{(n)}$, то

$$\rho_n^2 = |x_* - x_*^{(n)}|^2 \leqslant |x_* - x^{(n)}|^2 \leqslant c_0^2 \| x' - x^{(n)\prime} \|^2. \qquad (2.3.5)$$

Допустим, что $x_* \in W_2^{(r)}(0, 1)$, $r \geqslant 3$. Для этого достаточно, чтобы $p \in W_2^{(r-1)}(0, 1)$, а q, $f \in W_2^{(r-2)}(0, 1)$. Функция $x_*(t)$ допускает разложение в ряд Фурье

$$x_*(t) = \sum_{k=1}^{\infty} a_k \sin kt.$$

За $x^{(n)}$ примем отрезок этого ряда:

$$x^{(n)} = \sum_{k=1}^{n} a_k \sin k\pi t.$$

Оценим коэффициенты

$$a_k = 2 \int_0^1 x_*(t) \sin k\pi t \, dt.$$

Интегрируя по частям и учитывая краевые условия (2.3.1), получаем

$$a_k = -\frac{2}{k^3\pi^3} \int_0^1 x_*'''(t) \cos k\pi t \, dt + \frac{2}{k^3\pi^3} x_*''(t) \cos k\pi t \Big|_0^1.$$

Отсюда

$$|a_k| \leqslant \frac{C}{k^3} \left[\| x_*''' \|_{L_2} + \| x_*'' \|_{L_\infty} \right].$$

Дальнейшее интегрирование по частям (возможное, если $r > 3$) не улучшает последней оценки. Теперь

$$\| x_*' - x^{(n)\prime} \|^2 = \frac{1}{2} \sum_{k=n+1}^{\infty} k^2\pi^2 a_k^2 \leqslant C \left[\| x_*''' \|_{L_2}^2 + \| x_*'' \|_{L_\infty}^2 \right] \times$$

$$\times \sum_{k=n+1}^{\infty} \frac{1}{k^4} \leqslant C \left[\| x_*''' \|_{L_2}^2 + \| x_*'' \|_{L_\infty}^2 \right] \int_n^{\infty} \frac{dk}{k^4} =$$

$$= C \left[\| x_*''' \|_{L_2}^2 + \| x_*'' \|_{L_\infty}^2 \right] \frac{1}{n^3}$$

и формула (2.3.5) приводит к оценке погрешности аппроксимации:

$$\rho_n \leqslant C \left[\| x_*''' \|_{L_2} + \| x_*'' \|_{L_\infty} \right] n^{-3/2}. \qquad (2.3.6)$$

3°. По-прежнему пусть $x_* \in W_2^{(r)}(0, 1)$. При $r > 3$ можно получить лучшую оценку аппроксимации, если за $H_A^{(n)}$ принять подпространство полиномов вида

$$Q_n(t) = t(1 - t) \sum_{l=0}^{n} a_l t^l.$$

В силу неравенства (2.2.5) существует такой полином $Q_n(t)$, что

$$\|x_* - Q_n\|_{(\bar{r})} \leqslant C\omega_L^{(r)}(x_*,\ 1/n)\, n^{-(r-\bar{r})}, \qquad 0 \leqslant \bar{r} \leqslant r,$$

поэтому, если $x_*^{(n)}$ — приближенное решение из пространства $H_A^{(n)}$, то

$$\rho_n = |x_* - x_*^{(n)}| \leqslant C\omega_{L_2}^{(r)}(x_*,\ 1/n)\, n^{-r+1}. \tag{2.3.7}$$

§ 4. УРАВНЕНИЯ ЭЛЛИПТИЧЕСКОГО ТИПА. ОЦЕНКИ В ЭНЕРГЕТИЧЕСКОЙ НОРМЕ

1°. Рассмотрим первую краевую задачу для уравнения эллиптического типа (или эллиптической системы уравнений) вида

$$\sum_{|\alpha|,\,|\beta|=0}^{s} (-1)^{\alpha} D^{\alpha}\left(A_{\alpha\beta} D^{\beta} x\right) = f(t), \qquad t \in \Omega; \tag{2.4.1}$$

$$D^{\nu} x\big|_{\partial\Omega} = 0, \qquad 0 \leqslant |\nu| \leqslant s-1. \tag{2.4.2}$$

Пусть $\partial\Omega$ удовлетворяет условиям п. 1° § 2: поверхность $\partial\Omega$ можно задать уравнением $\varphi(t) = 0$, причем $\varphi \in C^{(\nu)}(\Omega_1), \Omega_1 \supset \Omega$ и Ω_1 гомеоморфна шару; $\varphi(t) \neq 0$, $t \notin \partial\Omega$ и $\operatorname{grad}\varphi|_{\partial\Omega} \neq 0$. Допустим, что оператор задачи (2.4.1), (2.4.2) — симметричный и положительно определенный в $L_2(\Omega)$; его энергетическая норма определяется формулой

$$|x|^2 = \int_{\Omega} \sum_{|\alpha|,\,|\beta|=0}^{s} A_{\alpha\beta} D^{\alpha} x D^{\beta} x\, dt. \tag{2.4.3}$$

Будем считать, что коэффициенты $A_{\alpha\beta}$ суть функции от t, ограниченные и измеримые в Ω. Тогда, очевидно,

$$|x| \leqslant C_0 \|x\|_{(s),\,\Omega}. \tag{2.4.4}$$

Задачу (2.4.1), (2.4.2) будем решать по методу Ритца, приняв за $H_A^{(n)}$ подпространство функций $\varphi^s(t) P_n(t)$, $\operatorname{gr} P_n \leqslant \underline{n}$. В общем случае x_* — решение нашей задачи — принадлежит пространству $W_2^{(s)}(\Omega)$. Допустим, что оно на самом деле более гладкое: пусть $x_* \in W_2^{(r)}(\Omega)$, где r — достаточно большое число. Тогда существует такое произведение $\varphi^s(t)\,\tilde{P}_n(t) \in H_A^{(n)}$, что

$$\|x_* - \varphi^s \tilde{P}_n\|_{(s)} \leqslant C\omega_{L_2}^{(r)}(x_*,\ 1/n)\, n^{-(r-s)}.$$

Если $x_*^{(n)}$ — приближенное решение задачи (2.4.1), (2.4.2) по Ритцу, то, как это вытекает из сказанного в § 1, $|x_* - x_*^{(n)}| \leqslant |x_* - \varphi^s \tilde{P}_n| \leqslant C_0 \|x_* - \varphi^s \tilde{P}_n\|_{(s)}$ и, следовательно,

$$\rho_n = |x_* - x_*^{(n)}| \leqslant C\omega_{L_2}^{(r)}(x_*,\ 1/n)\, n^{-(r-s)}. \tag{2.4.5}$$

2°. Пусть теперь дифференциальное уравнение (2.4.1) решается при естественных краевых условиях. По-прежнему будем считать, что оператор рассматриваемой задачи — симметричный и положительно определенный в $L_2(\Omega)$, энергетическая норма удовлетворяет неравенству (2.4.4) и решение, которое мы по-прежнему обозначим через x_*, принадлежит к классу $W_2^{(r)}(\Omega)$. За $H_A^{(n)}$ примем подпространство полиномов степени не выше n в R_m. Используя следствие 2.2.1 и повторяя рассуждения п. 1° настоящего параграфа, мы придем к оценке

$$\left|x_* - x_*^{(n)}\right| \leqslant C \|x_*\|_{(r)} n^{-(r-s)}. \tag{2.4.6}$$

Ту же оценку (2.4.6) можно получить, если краевые условия имеют вид (2.4.2) при $0 \leqslant |\gamma| \leqslant k-1$, $k < s$, а остальные условия естественные. В этом случае за $H_A^{(n)}$ следует принять подпространство функций $\varphi^k(t)P_n(t)$, где $P_n(t)$ — полином степени не выше n в R_m.

3°. Условия, при которых точное решение $x_* \in W_2^{(r)}(\Omega)$, см. в [1, 154].

§ 5. ОЦЕНКИ ПОГРЕШНОСТЕЙ ПРОИЗВОДНЫХ

1°. Будем рассматривать задачу (2.4.1), (2.4.2). В § 4 мы вывели оценку погрешности аппроксимации для приближенного решения этой задачи в энергетической норме. По существу, это оценка для производных, порядок которых равен половине порядка уравнения (2.4.1). Представляет значительный интерес получение аналогичных оценок для производных других порядков. Этому посвящен ряд работ, в которых рассматриваются координатные функции, описанные в п. 2° § 4. Отметим некоторые из этих работ.

2°. В. П. Ильин [57, 58] рассмотрел основные краевые задачи для невырождающегося эллиптического уравнения второго порядка; для этих задач получены оценки нормы $\left\|x_* - x_*^{(n)}\right\|_{C(\bar{\Omega})}$ при условии, что известен порядок убывания некоторых величин A_n, определенным образом зависящих от приближенного решения $x_*^{(n)}$. В аналогичных предположениях получены оценки погрешности аппроксимации в $C(\bar{\Omega})$ и $C^{(1)}(\bar{\Omega})$ в случае первой краевой задачи для бигармонического уравнения. В [58] улучшены оценки для эллиптического уравнения второго порядка. Для того же уравнения И. Ю. Харрик [143] получила оценку величин A_n; аналогичные оценки для бигармонического уравнения даны в [144]. В [60] рассмотрены самосопряженные задачи для эллиптического уравнения любого четного порядка $2s$ и получены оценки погрешности аппроксим-

мации в $W_p^{(\tilde{s})}(\Omega)$, $1 < p < \infty$, $\tilde{s}$ — не очень большое неотрицательное целое число.

3°. В статье Л. В. Канторовича [63] (см. также [64]) предложена схема исследования погрешности проекционных методов в банаховых пространствах. В статье автора [93] предложено видоизменение этой схемы, которое позволило оценить погрешность аппроксимации метода Ритца в $W_2^{(\tilde{s})}$, где $\tilde{s} > 2s$ — не очень большое натуральное число. Следует сказать, что оценки этой статьи хуже оценок статьи [60], соответствующих значению $p = 2$. Другое видоизменение схемы Канторовича было использовано Т. О. Шапошниковой [149], которая таким способом получила оценки погрешности для производных порядка $\tilde{s} > 2s$. Ее оценки хуже оценок статьи [60] при $p = 2$ для $\tilde{s} < s$. В [151] получены оценки в некоторых банаховых пространствах; в частности, в этой статье показано, что при $\tilde{s} \leqslant s$ оценки работы [60] могут быть улучшены и при $p \neq 2$.

В статье автора [104] методы работ [93, 149] перенесены на несамосопряженные задачи, допускающие решение по методу Бубнова — Галеркина. Благодаря использованию уточненных марковских неравенств удалось улучшить результаты работ [93] и [149]. Новые оценки совпадают с оценками Ильина (при $p = 2$) для $\tilde{s} > s$ и с оценками Шапошниковой для $\tilde{s} < s$.

Ниже в данном параграфе в основном излагается содержание статьи [104] применительно к методу Ритца. В пп. 4—7° даны некоторые марковские неравенства, на которые опирается дальнейшее изложение. В пп. 8—11° выводятся оценки для производных порядка $\tilde{s} \geqslant 2s$; аналогичные оценки для производных порядка $\tilde{s} < 2s$ выводятся в пп. 12—15°.

4°. Пусть A — линейный неограниченный оператор, действующий из множества, плотного в некотором банаховом пространстве X, в банахово пространство Y. Пусть, далее, $\{X_n\}$, $n = 1, 2, \ldots,$ — последовательность конечномерных подпространств пространства X, полная в этом пространстве, причем элементы любого из подпространств X_n принадлежат области определения оператора A. Сужение A_n оператора A на подпространство X_n, очевидно, ограничено; пусть σ_n — какая-нибудь верхняя граница нормы $\|A_n\|$, так что

$$\forall x \in X_n, \quad \|A_n x\|_Y \leqslant \sigma_n \|x\|_X. \qquad (*)$$

Неравенства вида (*) и суть марковские неравенства; первое такое неравенство было получено А. А. Марковым в 1889 г. В современной записи можно представить упомянутое неравенство Маркова в таком виде: если $X = Y = C[a, b]$, $-\infty < a < b + \infty$, и X_n — пространство полиномов степени не выше n, а A — оператор дифференцирования, то $\|A_n\| = 2n^2/(b-a)$.

Марковские неравенства играют большую роль во многих вопросах анализа, и им посвящено большое количество работ.

5°. Пусть $P_n(t)$ — полином степени не выше n от вещественной переменной t на промежутке $a \leqslant t \leqslant b$. Обозначим через $q_k(t)$, $k = 0, 1, 2, \ldots$, полиномы степени k, ортонормированные в $L_2(a, b)$: они лишь множителем $[2(b-a)]^{1/2}$ отличаются от нормированных полиномов Лежандра, преобразованных к промежутку (a, b). Справедливо тождество

$$P_n(t) = \sum_{k=0}^{n} a_k q_k(t), \qquad a_k = \text{const}.$$

Отсюда

$$P_n^2(t) \leqslant \sum_{k=0}^{n} a_k^2 \sum_{k=0}^{n} q_k^2(t) = \|P_n\|^2_{L_2(a,\,b)} \sum_{k=0}^{n} q_k^2(t).$$

Далее, $q_k(t) = [2(b-a)]^{1/2} p_k(\tau)$, где $p_k(\tau)$ — полиномы Лежандра, ортонормированные на промежутке $(-1, 1)$, и $\tau = (b-a)^{-1}(2t - a - b)$. Так как $\max_{|\tau|=1} p_k(\tau) = \sqrt{(2k+1)/2}$, то

$$\max_{t \in [a,\,b]} \sum_{k=1}^{n} q_k^2(t) \leqslant \frac{2}{b-a} \sum_{k=0}^{n} \frac{2k+1}{2}$$

и, следовательно,

$$\|P_n\|_{C[a,\,b]} \leqslant \frac{n-1}{\sqrt{b-a}} \|P_n\|_{L(a,\,b)}. \tag{2.5.1}$$

Это марковское неравенство хорошо известно.

6°. Хорошо известно также обобщение классического марковского неравенства на случай пространства $L_p(a, b)$ [174]. Оно имеет вид

$$\|P_n'\|_{L_p(a,\,b)} \leqslant Cn^2 \|P_n\|_{L_p(a,\,b)}, \tag{2.5.2}$$

где C зависит только от p и от $b - a$. Выведем аналогичное неравенство для полиномов от многих переменных; для простоты ограничимся случаем $p = 2$.

Неравенство (2.5.2) легко обобщается на полиномы степени не выше n в R_m, если отрезок (a, b) заменяется параллелепипедом. Действительно, можно считать, что ребра параллелепипеда параллельны координатным осям. Пусть $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ — мультииндекс, у которого $\alpha_k = \delta_{jk}$ и индекс j фиксирован, и пусть j-е ребро параллелепипеда определено неравенствами $a_j \leqslant t_j \leqslant b_j$. По неравенству (2.5.2)

$$\int_{a_j}^{b_j} |D^\alpha P_n(t)|^2 \, dt_j \leqslant Cn^4 \int_{a_j}^{b_j} |P_n(t)|^2 \, dt_j.$$

Интегрируя по остальным координатам, получим (Π — параллелепипед)

$$|\alpha| = 1, \qquad \|D^\alpha P_n(t)\|^2_{L_2(\Pi)} \leqslant Cn^4 \|P_n\|^2_{L(\Pi)}. \tag{2.5.3}$$

Отсюда сразу вытекает для произвольного мультииндекса соответствующее марковское неравенство:

$$\forall\alpha, \quad \|D^\alpha P_n\|_{L_\infty(\Pi)} \leqslant C_\alpha n^{2|\alpha|} \|P_n\|_{L_2(\Pi)}. \tag{2.5.4}$$

Рассмотрим теперь конечную область $\Omega \in R_m$ с границей, удовлетворяющей условиям п. 1° § 2. Представим $\overline{\Omega}$ в виде $\overline{\Omega} = \Omega' \cup \Omega_\delta$, $\Omega' \cap \Omega_\delta = \varnothing$. Здесь Ω_δ — пограничная полоса шириной δ и δ — достаточно малое фиксированное число. Каждую точку замкнутой области $\overline{\Omega}'$ сделаем центром куба, ребра которого параллельны осям координат и который целиком вместе со своей границей лежит в Ω. По лемме Бореля выделим конечное число таких кубов, образующих покрытие для Ω'; пусть это будут кубы $Q_1, Q_2, \ldots, Q_k$. В силу неравенства (2.5.4)

$$\forall\alpha, \quad \|D^\alpha P_n\|_{2, Q_j} \leqslant C n^{2|\alpha|} \|P_n\|_{2, Q_j}.$$

Просуммируем это по j. Слева заменим полученную сумму меньшей величиной $\|D^\alpha P_n\|_{2, \Omega'}$, а справа — большей величиной $k\|P_n\|_{2, \Omega}$. В результате получим

$$\forall\alpha, \quad \|D^\alpha P_n\|_{2, \Omega'} \leqslant C n^{2|\alpha|} \|P_n\|_{2, \Omega'}. \tag{2.5.5}$$

Теперь рассмотрим пограничную полосу Ω_δ. Ширину δ выберем следующим образом. Пусть $t_0 \in \partial\Omega$. Построим сферу с центром в t_0, обладающую тем свойством, что любая прямая, параллельная нормали к $\partial\Omega$ в точке t_0, пересекает часть Γ_0 поверхности $\partial\Omega$, заключенную внутри упомянутой сферы, не более одного раза. Известно [92], что радиус d этой сферы можно выбрать независящим от t_0. Участок Γ_0 границы $\partial\Omega$ можно задать уравнением вида $\xi_m = g(\xi')$, где $\xi' = (\xi_1, \xi_2, \ldots, \xi_{m-1})$, ось ξ_m направлена по нормали к $\partial\Omega$ в точке t_0 и начало новой системы координат совпадает с t_0. Так как функция $\varphi(t)$ (см. п. 1° § 2) достаточно гладкая, а $\operatorname{grad}\varphi|_{\partial\Omega} \neq 0$, то функция $g(\xi')$ также достаточно гладкая, и можно найти такое $\delta > 0$, чтобы лежащая в Ω_δ замкнутая область

$$E_{t_0, \delta} = \{t: -\delta \leqslant \xi_i \leqslant \delta, \ 1 \leqslant i \leqslant m-1; \ g(\xi') - \delta \leqslant \xi_m \leqslant g(\xi_i)\}$$

находилась внутри сферы радиусом d с центром в t_0.

Введем координаты $\eta_i = \xi_i$, $1 \leqslant i \leqslant m-1$, $\eta_m = g(\xi') - \xi_m$. В этих координатах область $E_{t_0\delta}$ переходит в параллелепипед

$$F_\delta = \{\eta: -\delta \leqslant \eta_i \leqslant \delta, \ 1 \leqslant i \leqslant m-1; \ 0 \leqslant \eta_m \leqslant \delta\}.$$

Очевидно, что якобиан преобразования от координат ξ_i в координаты η по абсолютной величине равен единице, а производные первого порядка по t_i оцениваются через производные того же порядка по η_i, и наоборот.

Каждой точке $t_0 \in \partial\Omega$ можно привести в соответствие область $E_{t_0,\delta}$, и эти области образуют покрытие для Ω_δ. По лемме Бореля можно выделить конечное число областей $E_{t_0,\delta}$, также образующих покрытие для Ω_δ. Для каждого из параллелепипедов F_δ справедливо неравенство (2.5.4), из которого, в свою очередь, следует новое неравенство

$$\forall\alpha, \quad \|D^\alpha P_n\|_{2,\Omega_\delta} \leqslant C\|P_n\|_{2,\Omega} n^{2|\alpha|}.$$

Сложив последнее неравенство с (2.5.5), получим искомое марковское неравенство

$$\forall\alpha, \quad \|D^\alpha P_n\|_{2,\Omega} \leqslant Cn^{2|\alpha|}\|P_n\|_{2,\Omega}. \tag{2.5.6}$$

Нетрудно получить аналогичное неравенство и для $p \neq 2$.

7°. Докажем, что справедливо неравенство

$$\|\varphi^{s-1}P_n\|_{2,\Omega} \leqslant Cn^2\|\varphi^s P_n\|_{2,\Omega}; \tag{2.5.7}$$

свойства функции $\varphi(t)$ описаны в п. 1° § 4 данной главы. Начнем со случая одного измерения. В этом случае можно свести дело к сравнению двух интегралов:

$$\int_0^\delta P_n^2(t)\,dt, \qquad \int_0^\delta t^2 P_n^2(t)\,dt,$$

где δ — фиксированное число. Интегрируя по частям, получаем

$$\int_0^\delta P_n^2(t)\,dt = \delta P_n^2(\delta) - 2\int_0^\delta tP_n(t)\,P_n'(t)\,dt.$$

Оценим правую часть. Прежде всего по формуле (2.5.1)

$$\delta P_n^2(\delta) \leqslant \delta^{-1}\|tP_n\|^2_{C[0,\delta]} \leqslant Cn^2\|tP_n\|^2_{2,(0,1)}.$$

Далее,

$$\left|2\int_0^\delta P_n(t)\,P_n'(t)\,dt\right| \leqslant \varepsilon^{-1}\int_0^\delta t^2P_n^2(t)\,dt +$$

$$+\varepsilon\int_0^\delta P_n'^2(t)\,dt \leqslant \frac{1}{\varepsilon}\int_0^\delta t^2P_n^2(t)\,dt + C\varepsilon n^4\int_0^\delta P_n^2(t)\,dt.$$

Положив $\varepsilon = (2Cn^4)^{-1}$, найдем искомую оценку

$$\|P_n\|_{2,(0,\delta)} \leqslant Cn^2\|tP_n\|_{2,(0,\delta)}. \tag{2.5.8}$$

Обратимся к общему случаю. Внутри Ω оценка тривиальна: если Ω' — внутренняя подобласть и $|\varphi|(t)| \geqslant \beta$, $\forall t \in \Omega'$, то

$$\|\varphi^{k-1}P_n\|_{2,\Omega'} \leqslant \beta^{-1}\|\varphi^k P_n\|_{2,\Omega'}. \tag{2.5.9}$$

Поэтому достаточно рассмотреть область вида $E_{t_0,\delta}$ и сравнить интегралы

$$\int_{E_{t_0,\delta}} \varphi^{(2k-2)}(t)\,P_n^2(t)\,dt, \qquad \int_{E_{t_0,\delta}} \varphi^{2k}(t)\,P_n^2(t)\,dt.$$

От переменных t_i перейдем к переменным η_i (п. 6°). Область $E_{t_0,\delta}$ перейдет в параллелепипед F_δ; при этом интеграл вида

$$\int_{E_{t_0,\delta}} \varphi^{2s}(t) P_n^2(t)\, dt \tag{2.5.10}$$

оценится сверху и снизу через интеграл

$$\int_{F_\delta} \eta_m^{2s} P_n^2(t)\, dt \tag{2.5.11}$$

с постоянными, не зависящими от n. Выражение $\eta_m^{k-1} P_n(t)$ есть полином степени не выше $n+k-1$ от η_m, и по неравенству (2.5.8)

$$\int_0^\delta \eta_m^{2k-2} P_n^2(t)\, d\eta_m \leqslant C n^4 \int_0^\delta \eta_m^{2k} P_n^2(t)\, d\eta_m.$$

Интегрируя последнее неравенство по η_i, $1 \leqslant i \leqslant m-1$, в пределах от $-\delta$ до δ и используя эквивалентность интегралов (2.5.10) и (2.5.11), получаем

$$\int_{E_{t_0,\delta}} \varphi^{2k-2}(t) P_n^2(t)\, dt \leqslant C n^4 \int_{E_{t_0,\delta}} \varphi^{2k}(t) P_n^2(t)\, dt.$$

Отсюда и из неравенства (2.5.9) вытекает оценка (2.5.7).

8°. Выше было упомянуто, что Канторович предложил некоторую схему построения и исследования проекционных методов. Мы изложим здесь эту схему в условиях не самых общих, но достаточных для того, чтобы ее можно было применить к методу Ритца.

Рассмотрим уравнение

$$Ax = f, \tag{2.5.12}$$

где A — замкнутый линейный оператор, взаимно-однозначно отображающий банахово пространство X на банахово пространство Y; тогда оба оператора A и A^{-1} ограничены. Построим полную в X последовательность конечномерных подпространств $\{X_n\}$; напомним, что это означает следующее:

$$\forall x \in X, \quad \mathscr{E}_n(x) = \inf_{x^{(n)} \in X_n} \|x - x^{(n)}\| \xrightarrow[n\to\infty]{} 0. \tag{2.5.13}$$

Положим $Y_n = AX_n$; для каждого n выберем оператор Q_n, проектирующий Y на Y_n. Приближенное решение уравнения (2.5.12) построим как элемент $x^{(n)} \in X_n$, удовлетворяющий уравнению

$$Q_n A x^{(n)} = Q_n f. \tag{2.5.14}$$

Если x_* и $x_*^{(n)}$ — точные решения задач (2.5.12) и (2.5.14) соответственно, то, как доказано в цитированных выше работах Канторовича,

$$\|x_* - x_*^{(n)}\|_X \leqslant (1 + \|A^{-1} Q_n A\|_X)\, \mathscr{E}_n(x_*). \tag{2.5.15}$$

Отсюда следует оценка

$$\|x_* - x_*^{(n)}\| \leqslant C \|Q_n\|_Y \mathscr{E}_n(x_*). \qquad (2.5.16)$$

9°. В работе [93] рассмотрено некоторое видоизменение схемы Канторовича. Пусть $X \subset Y \subset Z$; здесь Y и Z — гильбертовы пространства, X — банахово пространство; знак $\subset$ означает плотное вложение. В уравнении (2.5.12) будем рассматривать A как оператор в Z и будем считать его положительно определенным, так что

$$\forall x \in D(A), \quad (Ax, x)_Z \geqslant \gamma^2 \|x\|_Z^2, \quad \gamma^2 = \text{const} > 0.$$

В то же время в соответствии со сказанным в п. 8° будем считать, что A взаимно-однозначно отображает X на Y.

Выберем полную в X последовательность конечномерных подпространств $\{X_n\}$, и пусть $(\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nN})$ — базис в X_n. Положим $Y_n = AX_n$ и определим оператор Q_n, проектирующий Y на Y_n, соотношением

$$\forall f \in Y, \quad (f - Q_n f, \varphi_{nj})_Z = 0, \quad j = 1, 2, \ldots, N. \qquad (2.5.17)$$

Докажем, что это соотношение действительно определяет проектор из Y на Y_n. По определению Q_n: $Y \to Y_n$, поэтому, если $f \in Y$, то $Q_n f \in Y_n$ и, следовательно,

$$Q_n f = \sum_{k=1}^N a_k^{(n)} A\varphi_{nk}, \quad a_k^{(n)} = \text{const}.$$

Далее, по тому же определению

$$\sum_{k=1}^N a_k^{(n)} (A\varphi_{nk}, \varphi_{nj})_Z = (f, \varphi_{nj})_Z, \quad 1 \leqslant j \leqslant N. \qquad (2.5.18)$$

Оператор A — положительно определенный в Z, поэтому матрица системы (2.5.18) положительно определенная; эта система разрешима единственным образом, и оператор Q_n определен для всех натуральных n. Остается доказать, что $Q_n^2 = Q_n$. Если $f \in Y$, то $Q_n f \in Y_n$ и $Q_n^2 f \in Y_n$, или $Q_n^2 f = \sum_{k=1}^N b_k^{(n)} A\varphi_{nk}$. Далее, $(Q_n^2 f - Q_n f, \varphi_{nj})_Z = 0$, поэтому $(Q_n^2 f, \varphi_{nj})_Z = (Q_n f, \varphi_{nj})_Z = (f, \varphi_{nj})_Z$, или $\sum_{k=1}^N b_k^{(n)} (A\varphi_{nk}, \varphi_{nj})_Z = (f, \varphi_{nj})_Z$. Система (2.5.18) имеет единственное решение, поэтому $b_k^{(n)} = a_k^{(n)}$ и $Q_n^2 = Q_n$.

Заменив f на $Ax^{(n)}$ под знаком проектора Q_n в (2.5.17), получим систему уравнений Бубнова — Галеркина для уравнения (2.5.12); так как A — оператор, положительно определенный в Z, то (2.5.17), или равносильная система (2.5.18), есть также система Ритца для того же уравнения. Из приведенного выше представления $Q_n f$ следует

$$\|Q_n f\|_Y^2 = \Big\| \sum_{k=1}^N a_k^{(n)} A\varphi_{nk} \Big\|_Y^2 \leqslant \Lambda_n^{(N)} \sum_{k=1}^N |a_{nk}|^2, \qquad (2.5.19)$$

где $\Lambda_n^{(N)}$ — наибольшее собственное число матрицы скалярных произведений $(A\varphi_{nk}, A\varphi_{nj})_Y$, $j, k = 1, 2, \ldots, N$. Заметим, что $a_k^{(n)}$ суть коэффициенты Ритца приближенного решения $x_*^{(n)}$

$$x_*^{(n)} = \sum_{k=1}^{N} a_k^{(n)} \varphi_{nk}.$$

Обозначая, как обычно, через $[\ ,\ ]$ и $|\ |$ скалярное произведение и норму в энергетическом пространстве Z_A оператора A, получаем из последнего тождества

$$|x_*^{(n)}|^2 = \sum_{j,\,k=1}^{N} a_j^{(n)} \bar{a}_k^{(n)} [\varphi_{nj}, \varphi_{nk}] =$$
$$= \sum_{j,\,k=1}^{N} a_j^{(n)} \bar{a}_k^{(n)} (A\varphi_{nj}, \varphi_{nk})_Z \geqslant \lambda_n^{(1)} \sum_{k=1}^{N} |a_k^{(n)}|^2;$$

здесь $\lambda_n^{(1)}$ — наименьшее собственное число матрицы скалярных произведений $(A\varphi_{nj}, \varphi_{nk})_Z$, $j, k = 1, 2, \ldots, N$. Теперь [106]

$$\sum_{k=1}^{N} |a_k^{(n)}|^2 \leqslant \frac{|x_*^{(n)}|^2}{\lambda_n^{(1)}} \leqslant \frac{|x_*|^2}{\lambda_n^{(1)}} \leqslant \frac{\|f\|_Z^2}{\gamma^2 \lambda_n^{(1)}}.$$

Из вложения $Y \subset Z$ вытекает существование такой постоянной C, что $\|f\|_Z \leqslant C\|f\|_Y$, и из (2.5.19) следует оценка нормы оператора Q_n

$$\|Q_n\|_{Y \to Y_n} \leqslant C \sqrt{\Lambda_n^{(N)} / \lambda_n^{(1)}}. \tag{2.5.20}$$

Если базис подпространства X_n ортонормирован в Z_A, то последняя оценка упрощается:

$$\|Q_n\|_{Y \to Y_n} \leqslant C \sqrt{\Lambda_n^{(N)}}. \tag{2.5.21}$$

Из (2.5.16), (2.5.21) получается оценка погрешности аппроксимации в X-норме

$$\|x_* - x_*^{(n)}\| \leqslant C \sqrt{\Lambda_n^{(N)}}\, \mathscr{E}_n(x_*). \tag{2.5.22}$$

10°. Вернемся к задаче (2.4.1), (2.4.2). Допустим, что система (2.4.1) равномерно эллиптична в Ω, а оператор рассматриваемой задачи положительно определен в $L_2(\Omega)$. Решение $x_*(t)$ будет сколь угодно гладким, если достаточно гладкими будут функции $\varphi(t), f(t), A_{\alpha\beta}(t)$. Будем считать, что $x_* \in W_2^{(r)}(\Omega)$, где r — достаточно большое число. Рассмотрим сначала более простой случай, когда $\varphi(t)$ есть полином. Задачу (2.4.1), (2.4.2) будем решать методом Ритца, взяв в качестве координатных функций полиномы $\varphi^s(t) P_n^{(j)}(t)$, $\operatorname{gr} P_n^{(j)} \leqslant \underline{n}$. Тогда приближенное решение примет вид $x_*^{(n)}(t) = \varphi^s(t) P_n(t)$, где P_n — также полином степени не выше $\underline{n}$ в R_m.

Введем в рассмотрение пространства $X = \mathring{W}_2^{(s)}(\Omega)$, $Y = W_2^{(l)}(\Omega)$, $Z = L_2(\Omega)$; $l \geqslant 0$, $s = 2s + l \leqslant r$. Нолик сверху здесь означает,

что речь идет о подпространстве пространства $W_2^{(s)}(\Omega)$, определенном условиями (2.4.2).

Координатные функции будем считать ортонормированными в $W_2^{(s)}(\Omega)$, тогда, как нетрудно убедиться (подробнее см. [90]), числа $\lambda_n^{(1)}$ положительно ограничены снизу и верна оценка (2.5.22).

Пусть N — число функций φ_{nj} при фиксированном n. Положим $\tau=(\tau_1, \tau_2, \ldots, \tau_N)$ и $\|\tau\|^2=\tau_1^2+\tau_2^2+\ldots+\tau_n^2$. Тогда $\Lambda_n^{(N)}=\max\limits_{\|\tau\|=1}\|A\varphi_n(\cdot, \tau)\|_Y^2$, где $\phi_n(t, \tau)=\sum_{j=1}^N \tau_j\varphi_{nj}(t)$. Для любой функции $u\in\overset{\circ}{W}{}_2^{(s)}(\Omega)$ справедливо неравенство $\|Au\|_Y\leqslant C\|u\|_X$; в частности, $\|A\phi_n(\cdot, \tau)\|_Y\leqslant C\|\phi_n(\cdot, \tau)\|_X=C\|\phi_n(\cdot, \tau)\|_{(s),\Omega}$. Но $\phi_n(t, \tau)$ есть полином степени не выше $n+\text{const}$ в R_m, и по неравенству (2.5.3)

$$\|\phi_n\|_{(s),\Omega}\leqslant Cn^{2(\tilde{s}-s)}\|\phi_n\|_{(s)}.$$

Теперь

$$\Lambda_n^{(N)}\leqslant Cn^{4(\tilde{s}-s)}. \tag{2.5.23}$$

Остается оценить наилучшее приближение

$$\mathscr{E}_n(x_*)=\inf_{x^{(n)}\in X_n}\|x_*-x^{(n)}\|_X=\inf_{x^{(n)}\in X_n}\|x_*-x^{(n)}\|_{(\tilde{s}),\Omega}.$$

Функция $x^{(n)}$ есть полином степени не выше $n+\text{const}$ в R_m, а $x_*\in W_2^{(r)}(\Omega)$; по теореме 2.2.2 существует такой полином $\bar{x}^{(n)}\in X_n$, что

$$\|x_*-\bar{x}^{(n)}\|_{(\tilde{s}),\Omega}\leqslant Cn^{-(r-\tilde{s})}\omega_{L_2}^{(r)}(x_*, 1/n).$$

Это значит, что $\mathscr{E}_n(x_*)\leqslant Cn^{-(r-s)}\omega_{L_2}^{(r)}(x_*, 1/n)$. Теперь из формул (2.5.11) и (2.5.23) следует

$$\rho_n=\|x_*-x_*^{(n)}\|_{(\tilde{s}),\Omega}=Cn^{-r+3\tilde{s}-2s}\omega_{L_2}^{(r)}(x_*, 1/n) \tag{2.5.24}$$

это и есть оценка погрешности аппроксимации для производных порядка $\tilde{s}>2s$.

11°. Вернемся к общему случаю и примем, что функция $\varphi(t)$ достаточно гладкая, но необязательно полином. Заново произведем оценку величины $\Lambda_n^{(N)}=\max\limits_{\|\tau\|=1}\|A\phi(\cdot, \tau)\|_Y$. Для любой функции $u\in\overset{\circ}{W}{}_2^{(\tilde{s})}(\Omega)$ верно неравенство $\|Au\|_{(l)}\leqslant C\|u\|_{(s)}$; в частности, $\|A\phi_n\|_{(l)}\leqslant C\|\phi_n\|_{(s)}$. Норму в $W_2^{(\tilde{s})}(\Omega)$ зададим формулой

$$\|u\|_{(\tilde{s})}=\sum_{\sigma,\zeta}\|D^{\sigma+\zeta}u\|_2;$$

суммирование производится по всем мультииндексам σ и ζ, таким, что $|\sigma| \leqslant s + l$ и $|\zeta| \leqslant s$. Докажем неравенство

$$\| D^{\sigma+\zeta} \phi_n \|_2 \leqslant C n^{2|\sigma|} \| \phi_n \|_{(s)}. \tag{2.5.25}$$

Имеем $\phi_n(t, \tau) = \varphi^s(t) P_n(t, \tau)$, где $P_n(t, \tau)$ — полином относительно t, $\operatorname{gr} P_n \leqslant \underline{n}$. Построим полином $q_n(t)$, $\operatorname{gr} q_n \leqslant \underline{n}$, такой, что

$$\forall \nu, \quad |\nu| \leqslant |\sigma + \zeta|, \quad |D^\nu [\varphi^s(t) - q_n(t)]| = O(n^{-v+\nu});$$

это возможно в силу теоремы 2.2.2. Будем считать, что v достаточно велико. Далее,

$$\| D^{\sigma+\zeta} \phi_n(\cdot, \tau) \|_2 \leqslant \| D^{\sigma+\zeta} (q_n P_n(\cdot, \tau)) \|_2 + \| D^{\sigma+\zeta} ((\varphi^s - q_n) P_n(\cdot, \tau)) \|_2. \tag{2.5.26}$$

Выражение $D^\zeta (q_n(t) P_n(t, \tau))$ есть полином степени не выше $2\underline{n} - \zeta \leqslant 2\underline{n}$ в R_m, и по неравенству (2.5.4)

$$\| D^{\sigma+\zeta} (q_n P_n(\cdot, \tau)) \|_2 \leqslant C n^{2|\sigma|} \| D^\zeta (q_n P_n(\cdot, \tau)) \|_2. \tag{2.5.27}$$

Но

$$\| D^\zeta (q_n P_n(\cdot, \tau)) \|_2 \leqslant \| D^\zeta (\varphi^s P_n(\cdot, \tau)) \|_2 + \| D^\zeta [(\varphi^s - q_n) P_n(\cdot, \tau)] \|_2. \tag{2.5.28}$$

Оценим второе слагаемое:

$$\| D^\zeta [(\varphi^s - q_n) P_n(\cdot, \tau)] \|_2 = \left\| \sum\nolimits_{\alpha \leqslant \zeta} D^\alpha (\varphi^s - q_n) D^{\zeta-\alpha} P_n(\cdot, \tau) \right\|_2 \leqslant$$

$$\leqslant \frac{C}{n^{v-\zeta}} \sum\nolimits_{\alpha \leqslant \zeta} \| D^{\zeta-\alpha} P_n(\cdot, \tau) \|_2 \leqslant C n^{-v+3|\zeta|} \| P_n(\cdot, \tau) \|_2 \leqslant$$

$$\leqslant C n^{-v+3|\zeta|+2s} \| \varphi^s P_n(\cdot, \tau) \|_2 \leqslant C \| \varphi^s P_n(\cdot, \tau) \|;$$

здесь использованы неравенства (2.5.6) и (2.5.11), а также тот факт, что v достаточно велико. Соболевские теоремы вложения дают неравенство

$$\| \varphi^s P_n(\cdot, \tau) \|_2 \leqslant C \| \varphi^s P_n(\cdot, \tau) \|_{(s)},$$

а следовательно, и неравенство

$$\| D^\zeta [(\varphi^s - q_n) P_n(\cdot, \tau)] \|_2 \leqslant C \| \varphi_s P_n(\cdot, \tau) \|_{(s)}. \tag{2.5.29}$$

Точно так же находим

$$\| D^{\zeta+\sigma} [(\varphi^s - q_n) P_n(\cdot, \tau)] \|_2 \leqslant C \| \varphi^s P_n(\cdot, \tau) \|_{(s)}. \tag{2.5.30}$$

Из неравенств (2.5.26)—(2.5.30), очевидно, следует соотношение (2.5.25), из которого, в свою очередь, следует, что

$$\| \phi_n \|^2_{(\tilde{s})} \leqslant C n^{4\tilde{s}-2s} \| \phi_n \|_{(s)} \leqslant C n^{4\tilde{s}-2s}.$$

В силу теоремы 2.2.1 $\mathscr{E}_n(x_*) \leqslant Cn^{-r+s}\omega_L^{(r)}(x_*, 1/n)$. Последние два неравенства вместе с формулой (2.5.21) приводят к оценке (2.5.24).

Если уравнение (2.4.1) решается при естественных краевых условиях, то можно получить оценку, близкую к (2.5.24). За координатные функции можно взять полиномы, при этом все предшествующие рассуждения в существенном сохраняются, но для наилучшего приближения получается оценка (2.2.3), и окончательно

$$\rho_n = \|x_* - x_*^{(n)}\|_{(\bar{s})} \leqslant Cn^{-r+3\bar{s}-2s}. \qquad (2.5.31)$$

12°. Как было отмечено в п. 3° данного параграфа, оценка погрешности аппроксимации младших производных получила, в частности, Шапошникова, используя еще одно видоизменение схемы Канторовича. По-прежнему к пространствам X и Y добавляется еще одно пространство Z и по-прежнему пространства Y и Z гильбертовы, но названные пространства в данном случае связаны соотношениями $X \subset Z \subset Y$. Допустим, что существует такое сужение $\tilde{A}$ оператора задачи (2.5.12), что $D(\tilde{A}) \subset Z$, $R(\tilde{A}) \subset Z$ и $\tilde{A}$ — положительно определенный оператор в Z. Энергетическое пространство оператора $\tilde{A}$ обозначим через $\hat{Z}$. Возможны два случая расположения пространств:

$$X \subset \hat{Z} \subset Z \subset Y; \qquad (2.5.32)$$

$$\hat{Z} \subset X \subset Z \subset Y. \qquad (2.5.33)$$

Введем в рассмотрение пространство Y^*, сопряженное с Y относительно скалярного умножения в Z; примем, что $Y^{**} = Y$, где Y^{**} — пространство, сопряженное с Y^* относительно того же скалярного умножения в Z.

Выберем последовательность конечномерных подпространств $\{X_n\}$, полную в $X \cap Y^*$, и пусть, как и выше, $(\varphi_{n1}, \varphi_{n2}, \ldots \varphi_{nN})$ — базис в X_n. Будем считать, что элементы этого базиса ортонормированы в $\hat{Z}$. Обозначим через $\tilde{A}$ сужение оператора A на $X \cap Y^*$ и допустим, что $\tilde{A}$ — положительно определенный оператор из Y^* в Y [8, 90]. Допустим еще, что соответствующее энергетическое пространство совпадает с $\hat{Z}$.

Положим $Y_n = AX_n$. Проекторы Q_n определим соотношениями

$$\forall f \in Y, \quad Q_n f \in Y_n, \quad (f - Q_n f, \varphi_{nj}) = 0, \quad j = 1, 2, \ldots, N, \qquad (2.5.34)$$

левая часть уравнения (2.5.34) рассматривается как значение функционала $(f - Q_n f) \in Y$ на элементе $\varphi_{nj} \in Y^*$. Как и выше, доказывается, что Q_n есть проектор из Y на Y_n.

Предполагая, что выполнено соотношение (2.5.32), оценим норму $\|Q_n\|_Y$. Приближенное решение имеет вид

$$x_*^{(n)} = \sum_{k=1}^{N} a_k^{(n)} \varphi_{nk},$$

при этом $Ax_*^{(n)} = Q_n f$ и, следовательно,

$$\|Q_n f\|_Y^2 = \sum_{j,k=1}^{n} (A\varphi_{nk}, A_{nj})_Y a_k^{(n)} \bar{a}_{jn}^{(n)} \leqslant \Lambda_n^{(N)} \sum_{k=1}^{n} |a_k^{(n)}|^2; \quad (2.5.35)$$

здесь, как и выше, $\Lambda_n^{(N)}$ есть наибольшее собственное число матрицы чисел $(A\varphi_{nk}, A\varphi_{nj})_Y$, $j, k = 1, 2, \ldots, N$. Элементы φ_{nk} ортонормированы в Z, поэтому $\sum_{k=1}^{N} |a_k^{(n)}|^2 = |x_*^{(n)}|^2 := \|x_*^{(n)}\|_{\hat{Z}}^2$

Нетрудно убедиться, что x_* есть решение задачи $\hat{A}x = f$, а $x_*^{(n)}$ — приближенное решение той же задачи по Ритцу. В [84] показано, что $|x_*^{(n)}| \leqslant |x_*|$; отсюда

$$\sum_{k=1}^{N} |a_k^{(n)}|^2 \leqslant |x_*|^2 \leqslant C \|x_*\|_X^2 = C \|A^{-1} f\|_X \leqslant C \|f\|_Y.$$

Теперь из (2.5.35) следует, что $\|Q_n\|_Y \leqslant C\sqrt{\Lambda_n^{(N)}}$ и

$$\|x_* - x_*^{(n)}\|_X \leqslant C\sqrt{\Lambda_n^{(N)}}\, \mathscr{E}_n(x_*). \quad (2.5.36)$$

13°. Рассмотрим случай, когда выполняется соотношение (2.5.33). Неравенство (2.5.16) остается справедливым. Для произвольных элементов $h \in Y$ и $g \in Y^*$ имеет место неравенство $|(h, g)| \leqslant \|h\|_Y \cdot \|g\|_{Y^*}$. Используя уравнения (2.5.34) и ограниченность оператора $\tilde{A}^{-1}$, обратного положительно определенному оператору $\tilde{A}$, можно утверждать, что

$$\sum_{k=1}^{n} |a_k^{(n)}|^2 = |x_*^{(n)}|^2 \leqslant |x_*|^2 = (\hat{A}x_*, x_*) = (f, x_*) \leqslant$$
$$\leqslant \|f\|_Y \cdot \|x_*\|_{Y^*} = \|f\|_Y \cdot \|A^{-1} f\|_{Y^*} \leqslant C \|f\|_Y^2.$$

Теперь из (2.5.35) вытекает, что $\|Q_n f\|_Y \leqslant C\sqrt{\Lambda_n^{(N)}}\, \|f\|_Y$, и, следовательно,

$$\|x_* - x_*^{(n)}\|_X \leqslant C\sqrt{\Lambda_n^{(N)}}\, \mathscr{E}_n(x_*). \quad (2.5.37)$$

14°. Вернемся к задаче (2.4.1), (2.4.2); сохраним все сделанные в предшествующих пунктах настоящего параграфа допущения о данных и решении этой задачи. Возьмем целое число l, $0 < l \leqslant 2s$, и обозначим $\tilde{s} = 2s - l$. Введем пространства $X = \overset{\circ}{W}{}_2^{(\tilde{s})}(\Omega)$, $Y = W_2^{(-l)}(\Omega)$, $Z = L_2(\Omega)$. Тогда $X \subset Z \subset Y$; здесь нолик сверху означает, что функции из X удовлетворяют условиям (2.4.2) при $|\alpha| \leqslant s - 1$. В данном случае $\hat{Z} \approx \overset{\circ}{W}{}_2^{(s)}(\Omega)$, поэтому расположения пространств (2.5.32) и (2.5.33) соответствуют значениям $0 < l < s$ и $s < l < 2s$. Чтобы доказать, что наша задача соответствует схеме п. 12° настоящего параграфа, достаточно установить справедливость следующих утверждений: а) операторы A и A^{-1} ограничены; б) оператор $\tilde{A}$ положительно определенный; в) энергетическое пространство оператора $\tilde{A}$ совпадает с $\hat{Z}$. Ограниченность оператора A очевидна.

Нетрудно также убедиться, что A — симметричный и положительный оператор из $X = Y^*$ в Y и что его энергетическое пространство совпадает с Z. Если будет доказано, что оператор A^{-1} ограничен, то положительный оператор $\tilde{A}$ окажется положительно определенным. Окончательно, достаточно доказать ограниченность оператора A^{-1}.

По определению A есть расширение оператора $\tilde{A}$, выполненное по соотношению

$$(Ax,\ y)_Z = (x,\ A^*y)_Z = (x,\ \tilde{A}y)_Z; \qquad x \in \overset{\circ}{W}{}_2^{(s)}(\Omega), \quad y \in W_2^{(2s)}(\Omega). \tag{2.5.38}$$

В пространстве $Z = L_2(\Omega)$ оператор $\tilde{A}$ положительно определенный, поэтому он имеет только тривиальный нуль. Докажем, что у расширенного оператора A других нулей нет.

Пусть $Ax_0 = 0$. Тогда $(x_0, \tilde{A}y)_Z = 0$. Это значит, что x_0, рассматриваемый как функционал в $Z = L_2(\Omega)$, аннулируется на множестве всех значений оператора $\tilde{A}$, т. е. на всем пространстве $L_2(\Omega)$. Но тогда $x_0 = 0$, оператор A не имеет нетривиальных нулей и существует обратный оператор A^{-1}. Далее, оператор $\tilde{A}$, очевидно, нетеров, и его индекс равен нулю. Как доказано в [91] (см. также [95]), оператор A также нетеров, и его индекс равен индексу оператора $\tilde{A}$, т. е. равен нулю. Только что было доказано, что подпространство нулей оператора A имеет нулевую размерность, а тогда нулевую размерность имеет и подпространство нулей сопряженного оператора A^*. Отсюда следует, что оператор A^{-1} определен на всем пространстве $W_2^{(-l)} = Y$. В то же время он замкнут как обратный нетерову (и, следовательно, замкнутому) оператору A. Из сказанного следует, что оператор A^{-1} ограничен.

15°. Как и выше, $x_*^{(n)}(t) = \varphi^s(t) P_n(t)$, $\operatorname{gr} P_n \leqslant \underline{n}$, и

$$\Lambda_n^{(N)} = \max_{\|r\|=1} \left\| \sum_{k=1}^n \tau_k A\varphi_{nk} \right\|_{(-l)}^2 = \max_{\|\tau\|=1} \|A\phi_n(\cdot, \tau)\|_{(-l)}^2.$$

Оператор A ограничен, поэтому

$$\|A\phi_n(\cdot, \tau)\|_{(-l)} \leqslant C \|\phi_n(\cdot, \tau\|_{(\tilde{s})}.$$

Рассмотрим сначала случай $\tilde{s} > s$, или $l < s$, что соответствует расположению (2.5.32). По неравенству (2.5.6)

$$\|\phi_n(\cdot, \tau)\|_{(s)} \leqslant Cn^{2(\tilde{s}-s)} \|\phi_n(\cdot, \tau)\|_{(s)} \leqslant Cn^{2(\tilde{s}-s)}. \tag{2.5.39}$$

По теореме 2.2.2

$$\mathscr{E}_n(x_*) \leqslant Cn^{-r+\tilde{s}} \omega_L^{(r)}(x_*,\ 1/n),$$

и из формулы (2.5.36) следует, что

$$s < \tilde{s} < 2s, \qquad \rho_n = \|x_* - x_*^{(n)}\|_{(\tilde{s})} \leqslant Cn^{-r+3\tilde{s}-2s} \omega_{L_2}^{(r)}(x_*,\ 1/n). \tag{2.5.40}$$

Теперь рассмотрим случай $0 \leqslant \tilde{s} < s$. Неравенство (2.5.39) заменяется более простым:

$$\|\phi_n(\cdot, \tau)\|_{(\tilde{s})} \leqslant C\|\phi_n(\cdot, \tau)\|_{(s)} = C.$$

Отсюда следует неравенство

$$0 \leqslant \tilde{s} < s, \quad \rho_n = \|x_* - x_*^{(n)}\|_{(\tilde{s})} \leqslant Cn^{-r+s}\omega_L^{(r)}(x_*, 1/n). \quad (2.5.41)$$

Из оценок (2.5.40) и (2.4.41) первая точнее, а вторая совпадает с соответствующей оценкой статьи [149].

Некоторые оценки погрешности метода Ритца в банаховых пространствах даны в статье [152].

§ 6. МЕТОД БУБНОВА — ГАЛЕРКИНА

1°. Основы метода Бубнова — Галеркина изложены в монографиях [82, 84]; будем считать, что эти основы известны читателю. Оценка погрешности аппроксимации для метода Бубнова — Галеркина впервые была дана М. А. Красносельским [72] в весьма общей ситуации — для уравнений, вообще говоря, нелинейных. С подробными доказательствами эти оценки изложены в монографии [73], поэтому здесь мы ограничимся тем, что приведем их для простейшего случая.

Пусть A_0 — самосопряженный положительно определенный оператор в некотором гильбертовом пространстве H и пусть $H_0 = H_{A_0}$ — его энергетическое пространство. Пусть, далее, K — линейный оператор в H, такой, что $D(K) \supset D(A_0)$ и произведение $T = A_0^{-1}K$ вполне непрерывно в H_0. Положим $A = A_0 + K$ и допустим, что уравнение $Ax = 0$ имеет только тривиальное решение $x = 0$. Рассмотрим неоднородное уравнение

$$Ax = f, \quad f \in H; \quad (2.6.1)$$

при перечисленных выше условиях оно имеет одно и только одно обобщенное решение, которое является обычным решением уравнения

$$x + Tx = A_0^{-1}f. \quad (2.6.2)$$

Обозначим это решение через x_*. Как показано в [82, 84], при достаточно больших n существует одно и только одно приближенное решение по Бубнову — Галеркину, которое мы обозначим через $x_*^{(n)}$, и $|x_* - x_*^{(n)}| \xrightarrow[n\to\infty]{} 0$, где $|\cdot|$ — норма в H_0. Оценка погрешности, данная Красносельским, имеет вид

$$\mathscr{E}_n(x_*) \leqslant |x_* - x_*^{(n)}| \leqslant C\mathscr{E}_n(x_*); \quad (2.6.3)$$

здесь $\mathscr{E}_n(x_*)$ есть наилучшее приближение (в метрике H_0) элемента x_* линейными комбинациями координатных элементов, использованных при построении приближенного решения.

2°. А. В. Джишкариани и Г. М. Вайникко посвятили ряд работ исследованию погрешности аппроксимации метода Бубнова — Галеркина в различных более частных предположениях, а также в задаче о собственных значениях. В настоящем пункте мы изложим результаты работы Джишкариани [48].

Уравнение (2.6.1) рассмотрим при дополнительном предположении, что оператор $T_1 = KA_0^{-1}$ также вполне непрерывен в H_0. Допустим, что спектр оператора A_0 дискретен: это значит, что A_0 имеет счетное множество собственных чисел с единственной точкой сгущения на бесконечности и что последовательность собственных элементов оператора A_0 полна как в H, так и в H_0.

Пусть B — оператор, сходный с A_0. Это значит, что B — самосопряженный оператор, положительно определенный в H, и что $D(B) = D(A_0)$. Как известно, спектр оператора B также дискретен. Допустим, что этот спектр нам известен, и пусть ω_k и v_k — собственные числа и соответствующие им собственные элементы оператора B. Так как это оператор положительно определенный, то существует обратный оператор B^{-1}, и числа $\omega_k > 0$. Последовательность $\{v_k\}$ примем за координатную.

Пусть
$$x_*^{(n)} = \sum_{k=1}^{n} a_k v_k \tag{2.6.4}$$
— приближенное по Бубнову — Галеркину решение задачи (2.6.1). Заметим прежде всего, что операторы $A^{-1}B$, $B^{-1}A$, а также сопряженные с ними операторы BA^{-1}, AB^{-1} ограничены в H. Далее, $x_*^{(n)} \in D(B) = D(A)$; отсюда следует, что невязка приближенного решения $\delta_n = Ax_*^{(n)} - f$ есть элемент пространства H. Опираясь на результаты работ [83] и [12], можно доказать, что $\|\delta_n\|_H \xrightarrow[n\to\infty]{} 0$.

Имеем $x_*^{(n)} + Tx_*^{(n)} - A_0^{-1}f = A_0^{-1}\delta_n$ и $x_* + Tx_* - A_0^{-1}f = 0$. Вычитая второе равенство из первого, получим $(I+T)(x_* - x_*^{(n)}) = A_0^{-1}\delta_n$. Отсюда
$$\|x_* - x_*^{(n)}\| \leqslant \|(I+T)^{-1}\| \cdot \|A_0^{-1}\delta_n\|. \tag{2.6.5}$$
Далее,
$$A_0^{-1}\delta_n = \sum_{ik=1}^{\infty} c_{ik}(\delta_n, v_k) v_i, \qquad c_{ik} = (A_0^{-1}v_k, v_l). \tag{2.6.6}$$

Уравнения Бубнова — Галеркина можно представить в виде $(\delta_n, v_k) = 0$, $k = 1, 2, \ldots, n$, и соотношение (2.6.6) сводится к такому:
$$A_0^{-1}\delta_n = \sum_{i=1}^{\infty}\sum_{k=n+1}^{\infty} c_{ik}(\delta_n, v_k) v_i.$$
Отсюда с помощью несложных преобразований получаем
$$\|A_0^{-1}\delta_n\| \leqslant \frac{\|A_0^{-1}B\| \cdot \|\delta_n\|}{\omega_{n+1}}.$$

Подставив это в (2.6.5), найдем искомую оценку:

$$\|x_* - x_*^{(n)}\| \leqslant \frac{\|(I+T)^{-1}\| \cdot \|A_0^{-1}B\| \|\delta^n\|}{\omega_{n+1}} = O\left(\frac{1}{\omega_{n+1}}\right). \quad (2.6.7)$$

Нетрудно получить оценку и в норме H_0:

$$|x_* - x_*^{(n)}| \leqslant \frac{\|(I+T)^{-1}\| \cdot \|(I+T_1)^{-1}\| \cdot \|A_0^{-1}B\|^{1/2} \cdot \|\delta_n\|}{\omega_{n+1}^{1/2}}. \quad (2.6.8)$$

Если $\delta_n \in D(B^r)$, $r > 0$, то последние оценки можно улучшить по порядку: справа в (2.6.7) и (2.6.8) можно заменить $\|\delta_n\|$ на $\|B^r\delta_n\|$, а ω_{n+1} и $\omega_{n+1}^{1/2}$ — на ω_{n+1}^{r+1} и $\omega_{n+1}^{r+1/2}$ соответственно.

§ 7. РАЗНОСТНЫЕ МЕТОДЫ

Погрешность аппроксимации разностных методов хорошо освещена в монографической литературе, и мы ограничимся здесь самыми общими указаниями. Достаточно подробное изложение можно найти, например, в монографиях [9, 4, 80, 40].

1°. Рассмотрим задачу Дирихле для уравнения Лапласа в двумерной области Ω с достаточно гладкой границей:

$$\Delta x = \frac{\partial^2 x}{\partial t_1^2} + \frac{\partial^2 x}{\partial t_2^2} = 0, \quad t = (t_1, t_2) \in \Omega, \quad x|_{\partial\Omega} = \varphi. \quad (2.7.1)$$

Построим квадратную сетку с шагом h. Обозначим через Ω_1 замкнутое подмножество Ω, обладающее тем свойством, что если $t \in \Omega_1$, то при любых численных значениях ξ, $|\xi| \leqslant 1$, будет $(t_1 + \xi h, t_2) \in \bar{\Omega}$ и $(t_1, t_2 + \xi h) \in \bar{\Omega}$. В точках множества Ω_1 уравнение Лапласа аппроксимируется разностным уравнением

$$X(t_1, t_2) = 4^{-1}[X(t_1 + h, t_2) + X(t_1 - h, t_2) + X(t_1, t_2 + h) + X(t_1, t_2 - h)]; \quad (2.7.2)$$

в точках множества $\Omega \setminus \Omega_1$ для функции X строится некоторая линейная интерполяционная формула, позволяющая имитировать краевое условие.

Узлы сетки, лежащие в Ω_1, называются внутренними, а узлы, лежащие в $\Omega \setminus \Omega_1$, — граничными. Во внутренних узлах, как было сказано, запишем уравнение (2.7.2), а в граничных — упомянутую интерполяционную формулу. В результате получим некоторую линейную алгебраическую систему (сеточную систему), о которой известно, что она имеет единственное решение X_h.

Доказывается [9], что если точное решение задачи (2.7.1) $x_* \in C^{(4)}(\bar{\Omega})$ и x_h — функция, определенная на узлах сетки, ле-

жащих в $\bar{\Omega}$, и совпадающая там с x_*, то

$$\|x_h - X_h\|_{L_\infty} \leqslant (12^{-1} M_4 \sup_{\Omega} Q + M_2) h^2. \qquad (2.7.3)$$

Здесь Q — некоторая неотрицательная функция, непрерывная в $\bar{\Omega}$, M_k — верхняя граница модулей частных производных от x_* порядка k.

Близкие результаты содержатся в [9] для общего эллиптического уравнения 2-го порядка в области любой размерности и для более общих краевых условий.

В [4] даны оценки погрешности аппроксимации разностного метода для бигармонического уравнения на плоскости.

2°. В. С. Рябенький и А. Ф. Филиппов доказали весьма интересную общую теорему об оценке погрешности аппроксимации разностных методов. Эта теорема подробно доказывается в [40, 80]. Рассмотрим вычислительную задачу

$$Lx = f, \quad lx = \varphi. \qquad (2.7.4)$$

Здесь x и f — функции, искомая и заданная, определенные в области $\Omega \subset R_m$, φ — функция, определенная на $\partial\Omega$, L и l — линейные операторы. Пусть задача (2.7.4) заменена разностной задачей

$$L_h x^h = f^h, \quad l_h x^h = \varphi^h, \qquad (2.7.5)$$

в которой x^h, f^h, φ^h — сеточные функции. В общем случае эти функции могут принадлежать различным сеточным пространствам, так что $x^h \in X_h$, $f^h \in F_h$, $\varphi^h \in \phi_h$.

Пусть $(\cdot)_h$ — оператор, преобразующий функцию, заданную на Ω или на $\partial\Omega$, в сеточную функцию, заданную на соответствующем множестве узлов. Теорема Рябенького — Филиппова утверждает следующее.

Пусть справедливы аппроксимационные неравенства

$$\begin{gathered} \|(Lx)_h - L_h x^h\|_{X_h} \leqslant Ch^k, \quad \|(lx)_h - l_h x^h\|_{\phi_h} \leqslant Ch^k, \\ \|(f)_h - f^h\|_{F_h} \leqslant Ch^k, \quad \|(\varphi)_h - \varphi^h\|_{\phi_h} \leqslant Ch^k \end{gathered} \qquad (2.7.6)$$

и пусть обратный оператор задачи (2.7.5) ограничен независимо от h, так что

$$\|x_*^h\|_{X_h} \leqslant C\left[\|f^h\|_{F_h} + \|\varphi^h\|_{\phi_h}\right]. \qquad (2.7.7)$$

Тогда справедлива оценка погрешности аппроксимации

$$\|(x_*)_h - x_*^h\|_{X_h} \leqslant Ch^k. \qquad (2.7.8)$$

§ 8. МЕТОД КОЛЛОКАЦИИ

1°. Метод коллокации впервые был предложен Канторовичем в [61]. Сущность метода такова. Пусть требуется приближенно решить линейное уравнение (требование линейности несущественно)

$$Ax = f, \quad x \in X, \quad f \in Y, \tag{2.8.1}$$

в котором X и Y — банаховы пространства; A — оператор, действующий из X в Y, причем элементы пространства Y образуют подмножество множества функций, непрерывных на некотором компакте $K \in R_m$. Выберем последовательность конечномерных подпространств $\{X_n\}$, полную в X, и положим $\dim X_n = N(n) = N$, $Y_n = AX_n$.

Пусть существует оператор A^{-1}, определенный и непрерывный на всем пространстве Y, тогда $\dim Y_n = N$ и последовательность $\{Y_n\}$ полна в Y. Далее, если $\{\varphi_{nk}\}$, $k = 1, 2, \ldots, N$, — базис в X_n и $\psi_{nk}(t) = A\varphi_{nk}$, $t \in K$, то $\{\psi_{nk}\}$ есть базис в Y_n.

Выберем в K точки $t_j = t_j^{(n)}$, $j = 1, 2, \ldots, N$, называемые узлами коллокации. Приближенное решение ищем как элемент подпространства X_n:

$$x^{(n)} = \sum_{k=1}^{N} a_k^{(n)} \varphi_{nk} \tag{2.8.2}$$

и определяем коэффициенты $a_k^{(n)}$ из условия, чтобы уравнение (2.8.1) выполнялось в узлах коллокации:

$$\sum_{k=1}^{N} a_k^{(n)} \psi_{nk}(t_j) = f(t_j), \qquad j = 1, 2, \ldots, n. \tag{2.8.3}$$

2°. Погрешность аппроксимации метода коллокации исследовалась в ряде работ. Приведем важнейшие результаты некоторых из этих работ. Сразу же отметим, что некоторые оценки упомянутой здесь погрешности, а также дополнительная библиография по этому вопросу приведены в [64] и в [72].

В работах [66—68] изучена погрешность аппроксимации коллокационного метода для обыкновенного дифференциального уравнения порядка $2s$ при условиях первой краевой задачи. Пусть уравнение задано на промежутке $a \leqslant t \leqslant b$ и пусть точное решение $x_* \in C^{(r,\alpha)}[a, b]$, $0 < \alpha \leqslant 1$. Приближенное решение $x_*^{(n)}$ строится в виде полинома степени $n + 2s$, удовлетворяющего краевым условиям. Допустим, что коэффициенты и свободный член дифференциального уравнения принадлежат к классу $C^{(r,\alpha)}[a, b]$, и пусть задача имеет не более одного решения. Если в качестве узлов коллокации выбраны узлы Гаусса или Чебышева, то соответственно

$$\| x_* - x_*^{(n)} \|_{C^{(s)}} \leqslant \frac{C}{n^{r+\alpha-1/2}} \tag{2.8.4}$$

и

$$\|x_* - x_*^{(n)}\|_{C^{(s)}} \leqslant \frac{C \ln n}{n^{r+\alpha}}. \qquad (2.8.5)$$

В работах [66—68] рассмотрена также первая краевая задача для уравнения

$$\Delta x + \lambda p(t)\, x = f(t) \qquad (2.8.6)$$

в квадрате R: $0 \leqslant t_1,\ t_2 \leqslant \pi$. За координатные функции взяты собственные функции той же задачи для уравнения Лапласа:

$$\varphi_{jk}(t) = \sin jt_1 \sin kt_2, \quad 0 \leqslant j \leqslant m, \quad 0 \leqslant k \leqslant n;$$

за узлы коллокации — точки $\left(\frac{2j-1}{2m+1}\pi, \frac{2k-1}{2n+1}\pi\right)$. Если λ отлично от собственных чисел рассматриваемой задачи, а $p,\ f \in \in \mathrm{Lip}_\alpha(R)$, то ($x_*^{(m,\,n)}$ — приближенное решение)

$$\|\Delta(x_* - x_*^{(m,\,n)})\|_{C(R)} \leqslant C \ln^2 m \ln^2 n\,(1/m^2 + 1/n^2). \qquad (2.8.7)$$

В статьях [67—68] рассмотрены и некоторые другие задачи.

3°. В статье [28] рассматривается интегральное уравнение Фредгольма

$$(Ax)(t) := \mu x(t) - (1/2\pi)\int_0^{2\pi} K(t,\ s)\, x(s)\, ds = f(t) \qquad (2.8.8)$$

в пространстве 2π-периодических функций с C-нормой. Приближенное решение строится в виде кусочно-линейной функции с угловыми точками $t_j = t_j^{(n)} = 2\pi j/n$; эти же точки суть узлы коллокации. Доказывается следующее утверждение. Пусть μ отлично от собственных чисел уравнения (2.8.8), а n таково, что

$$r := \omega_t(K,\ 2\pi/n)\,\|A^{-1}\| < 1. \qquad (2.8.9)$$

Тогда алгебраическая система метода коллокации имеет единственное решение. При этом, если $x_*(t)$ — точное, а $x_*^{(n)}(t)$ — приближенное решение уравнения (2.8.8), то

$$\|x_* - x_*^{(n)}\|_C \leqslant (1 + P_A/(1-r))\,\omega(x_*,\ 2\pi/n). \qquad (2.8.10)$$

Здесь $P_A = \|A\| \cdot \|A^{-1}\|$ — число обусловленности оператора A.

В той же статье [28] содержатся и некоторые результаты, относящиеся к одномерным сингулярным интегральным уравнениям.

4°. В книге [192] (см. также [183]) метод коллокации применен к одномерному сингулярному интегральному уравнению

$$a(t)\, x(t) + \frac{b(t)}{\pi i}\int_\Gamma \frac{x(s)}{s-t}\, ds + Tx = f(t). \qquad (2.8.11)$$

Здесь Γ — окружность $|t| = 1$, $a(t)$ и $b(t)$ непрерывны и символ не вырождается, так что $a^2(t) - b^2(t) \neq 0$, $t \in \Gamma$; T — опе-

ратор, вполне непрерывный в $L_p(\Gamma)$, $1 < p < \infty$. Координатные функции — тригонометрические полиномы порядка n относительно $\theta = \arg t$, узлы коллокации совпадают с точками $t_j^{(n)} = \exp(2\pi ij/n)$. Пусть $\hat{a}(t) = a(t) + b(t)$, $\hat{b}(t) = a(t) - b(t)$; $\operatorname{ind} \hat{a}(t) = \operatorname{ind} \hat{b}(t)$; $a(t)$, $b(t) \in C^{(r,\mu)}(\Gamma)$, $T : L_p(\Gamma) \to C^{(r,\nu)}(\Gamma)$. Тогда

$$\|x_* - x_*^{(n)}\|_p = O(n^{-r-\lambda}), \quad \lambda = \min(\mu, \nu). \tag{2.8.12}$$

Некоторые результаты получены и для уравнений вида (2.8.11) с вырождающимся символом.

В статье [193] рассмотрен метод коллокации для уравнения (2.8.11) при $T = 0$ с узлами $t_j^{(n)} = \exp(2\pi ij/n)$ и с кусочно-линейными координатными функциями, угловые точки которых совпадают с узлами коллокации. Основной результат статьи [193] состоит в следующем. Пусть коэффициенты $a(t)$ и $b(t)$ непрерывны. Для того чтобы описанный здесь метод коллокации сходился в $L_2(\Gamma)$, необходимо и достаточно, чтобы

$$a(t) + \tilde{\lambda} b(t) \neq 0, \quad \forall t \in \Gamma, \quad \forall \tilde{\lambda} \in [-1, 1]. \tag{2.8.13}$$

Если при этом a, $b \in \operatorname{Lip}_\mu(\Gamma)$, $f \in \operatorname{Lip}_\lambda(\Gamma)$, то

$$\|x_* - x_*^{(n)}\|_{\operatorname{Lip}_\nu(\Gamma)} = O(n^{-\nu} \ln n), \quad \nu = \min(\lambda, \mu). \tag{2.8.14}$$

В [193] получены некоторые результаты и для разрывных символов.

5°. Ряд результатов по оценке погрешности аппроксимации для уравнений Фредгольма и одномерных сингулярных интегральных уравнений содержится в монографии [30].

Глава 3

ПОГРЕШНОСТЬ ИСКАЖЕНИЯ

§ 1. ПОГРЕШНОСТЬ ИСКАЖЕНИЯ СВОБОДНОГО ВЫЧИСЛИТЕЛЬНОГО ПРОЦЕССА

1°. Проблема, указанная в названии главы, исследована в ряде работ автора и его учеников; первая из работ автора в этом направлении опубликована в 1960 г. Работы периода 1960—1965 гг. суммированы в монографии [90]; эти работы тесно связаны с понятием, тогда же введенным автором, об устойчивости вычислительных процессов относительно искажения. В [90] содержатся и не публиковавшиеся ранее результаты, в основном касающиеся нелинейных вычислительных процессов; эти последние будут рассмотрены в разделе IV. Ряд результатов был получен позднее для погрешности искажения

метода конечных элементов; эти результаты будут изложены в главе 9.

В § 1, 2 данной главы будут изложены результаты, относящиеся к искажению линейных вычислительных процессов и полученные частично в упомянутых выше более ранних, частично в более поздних работах автора [105, 181, 107], а также в статье С. Г. Суворова [133].

2°. Рассмотрим линейный свободный вычислительный процесс, состоящий в решении последовательности независимых уравнений

$$A_n x^{(n)} = f^{(n)}, \quad x^{(n)} \in X_n, \quad f^{(n)} \in Y_n, \quad n = 0, 1, 2, \ldots \qquad (3.1.1)$$

Здесь X_n и Y_n — банаховы пространства, A_n — замкнутый линейный оператор, отображающий X_n на Y_n. Мы принимаем, что операторы A_n ограниченно обратимы, так что каждый из операторов A_n^{-1} существует, ограничен и определен на всем пространстве Y_n.

Искаженный вычислительный процесс, который также естественно считать линейным, имеет вид

$$(A_n + \Gamma_n) z^{(n)} = f^{(n)} + \delta^{(n)}, \quad n = 0, 1, 2, \ldots \qquad (3.1.2)$$

Можно считать, что оператор Γ_n — искажение оператора A_n — в том или ином смысле мал по сравнению с оператором A_n. Примем, что величина $\| A_n^{-1}\Gamma_n \|$ может быть сделана сколь угодно малой; ниже мы будем пользоваться и другими, хотя и близкими ограничениями. Норму элемента $\delta^{(n)}$ также можно считать сколь угодно малой, однако в пределах линейной теории это допущение не играет существенной роли для последующих рассуждений.

Оценим погрешность искажения процесса (3.1.1). Зададим число β, $0 \leqslant \beta < 1$, и потребуем, чтобы $\| A_n^{-1}\Gamma_n \| \leqslant \beta$. При таком условии оператор $A_n + \Gamma_n$ ограниченно обратим. Действительно, пусть I_n — тождественный оператор в X_n. Тогда $A_n + \Gamma_n = A_n [I_n + A_n^{-1}\Gamma_n]$. Первый множитель ограниченно обратим по предположению, второй — по известной теореме Банаха, и

$$(A_n + \Gamma_n)^{-1} = (I_n + A_n^{-1}\Gamma_n)^{-1} A_n^{-1}. \qquad (3.1.3)$$

Решение уравнения (3.1.2) существует и единственно:

$$z_*^{(n)} = (I_n + A_n^{-1}\Gamma_n)^{-1} A_n^{-1} (f^{(n)} + \delta^{(n)}).$$

Решение уравнения (3.1.1) есть $x_* = A_n^{-1} f^{(n)}$, поэтому

$$z_*^{(n)} - x_*^{(n)} = [(I_n + A_n^{-1}\Gamma_n)^{-1} - I_n] x_*^{(n)} + (I_n + A_n^{-1}\Gamma_n)^{-1} A_n^{-1}\delta^{(n)}. \qquad (3.1.4)$$

Если I — тождественный, а $I + \gamma$ — ограниченно обратимый оператор в некотором банаховом пространстве, то $(I + \gamma)[(I +$

$+\gamma)^{-1}-I]=-\gamma$; следовательно, $(I+\gamma)^{-1}-I=-(I+\gamma)^{-1}\gamma$. В частности,

$$(I_n+A_n^{-1}\Gamma_n)^{-1}-I_n=-(I_n+A_n^{-1}\Gamma_n)^{-1}A_n^{-1}\Gamma_n; \qquad (3.1.5)$$

замечая еще, что $\|(I_n+A_n^{-1}\Gamma_n)^{-1}\|\leqslant(1-\beta)^{-1}$, мы приходим к искомой оценке погрешности искажения:

$$\hat{\xi}_n=\|z_*^{(n)}-x_*^{(n)}\|\leqslant(1-\beta)^{-1}[\|A_n^{-1}\Gamma_n x_*^{(n)}\|+\|A_n^{-1}\delta^{(n)}\|]. \qquad (3.1.6)$$

Отсюда вытекает несколько более простая и более удобная для применения оценка:

$$\hat{\xi}_n\leqslant(1-\beta)^{-1}[\|A_n^{-1}\Gamma_n\|\cdot\|x_*^{(n)}\|+\|A_n^{-1}\delta^{(n)}\|]. \qquad (3.1.7)$$

§ 2. УСТОЙЧИВОСТЬ СВОБОДНОГО ПРОЦЕССА ОТНОСИТЕЛЬНО ИСКАЖЕНИЯ

1°. Представляется интересным установить условия, при которых погрешность искажения остается малой независимо от n, если в том или ином смысле малы искажения Γ_n и $\delta^{(n)}$. Такие условия связаны с понятием устойчивости вычислительного процесса относительно искажения.

Целесообразными оказываются два несколько различных определения устойчивости. Согласно с первым из них мы называем процесс (3.1.1) устойчивым относительно погрешности искажения в последовательности пар пространств (X_n, Y_n), если существуют такие положительные постоянные p, q, r, что из неравенства $\|A_n^{-1}\Gamma_n\|\leqslant r$ следует однозначная разрешимость уравнения (3.1.2) при любом n и неравенство

$$\|z_*^{(n)}-x_*^{(n)}\|\leqslant p\|A_n^{-1}\Gamma_n\|+q\|\delta^{(n)}\|. \qquad (3.2.1)$$

Второе определение имеет смысл, если искажения Γ_n суть ограниченные операторы. В этом случае можно назвать процесс (3.1.1) устойчивым относительно искажения в последовательности пар пространств (X_n, Y_n), если существуют такие положительные постоянные p, q, r, что из неравенства $\|\Gamma_n\|\leqslant r$ вытекает однозначная разрешимость уравнения (3.1.2) и неравенство

$$\|z_*^{(n)}-x_*^{(n)}\|\leqslant p\|\Gamma_n\|+q\|\delta^{(n)}\|. \qquad (3.2.2)$$

Ниже мы для краткости будем говорить, что вычислительный процесс устойчив (или неустойчив), если он устойчив (или неустойчив) относительно погрешностей искажения. Там, где это не может вызвать недоразумений, мы будем опускать слова «в последовательности пар пространств (X_n, Y_n)».

2°. **Теорема 3.2.1.** *Для того чтобы вычислительный процесс* (3.1.1) *был устойчив в смысле первого определения, необхо-*

димо и достаточно, чтобы

$$\| A_n^{-1} \| \leqslant C_1, \quad \| x_*^{(n)} \| \leqslant C_2; \quad C_1,\ C_2 = \text{const}. \tag{3.2.3}$$

Достаточность условий (3.2.3) сразу вытекает из оценки (3.1.7): если эти условия выполнены, то можно положить $r = \beta$, $0 < \beta < 1$, и из (3.1.7) следует неравенство (3.2.1) с постоянными

$$p = C_2/(1-\beta), \quad q = C_1/(1-\beta).$$

Необходимость. Пусть процесс (3.1.1) устойчив в смысле первого определения. Допустим, что $\Gamma_n = 0$, так что искажаются только свободные члены уравнений (3.1.1). Неравенство (3.2.1) принимает вид $\|y^{(n)}\| \leqslant q\|\delta^{(n)}\|$, где $y^{(n)} = z_*^{(n)} - x_*^{(n)}$. В данном случае $A_n z_*^{(n)} = f^{(n)} + \delta^{(n)}$, поэтому $A_n y^{(n)} = \delta^{(n)}$ и, следовательно, $\|A_n y^{(n)}\| \geqslant q^{-1}\|y^{(n)}\|$. Это означает, что нормы $\| A_n^{-1} \|$ ограничены независимо от n; можно положить $C_1 = q$.

Чтобы доказать необходимость второго условия (3.2.3), допустим, что $\delta^{(n)} = 0$. Тогда, если $\| A_n^{-1}\Gamma_n \| \leqslant r$, то

$$\| z_*^{(n)} - x_*^{(n)} \| \leqslant p \| A_n^{-1}\Gamma_n \|. \tag{3.2.4}$$

Положим $\Gamma_n = rA_n$. Уравнение (3.1.2) принимает вид $(1+r) \times A_n z^{(n)} = f^{(n)}$. Отсюда $z_*^{(n)} = (1+r)^{-1} x_*^{(n)}$, и из (3.2.4) следует, что $\| x_*^{(n)} \| \leqslant p(1+r)$.

3°. **Теорема 3.2.2.** *Для того чтобы процесс* (3.1.1) *был устойчив в смысле второго определения в последовательности пар пространств* (X_n, Y_n), *необходимо и достаточно, чтобы были выполнены следующие условия*: 1) $\| A_n^{-1} \| \leqslant C_1 = \text{const}$; 2) *существует такая постоянная* C_3, *что каков бы ни был линейный оператор* B_n *с единичной нормой, действующий из* X_n *в* Y_n, *справедливо неравенство*

$$\| A_n^{-1} B_n x_*^{(n)} \| \leqslant C_3. \tag{3.2.5}$$

Необходимость условия 1) доказывается так же, как в теореме 3.2.1. Докажем необходимость условия (3.2.5). Пусть $\delta^{(n)} = 0$. Положим $\Gamma_n = \varepsilon B_n$, $\varepsilon = \text{const} < r$. Тогда $(A_n + \Gamma_n) z_*^{(n)} = f^{(n)}$; отсюда $z_*^{(n)} + A_n^{-1}\Gamma_n z_*^{(n)} = x_*^{(n)}$ и

$$z_*^{(n)} - x_*^{(n)} + A_n^{-1}\Gamma_n (z_*^{(n)} - x_*^{(n)}) = -A_n^{-1}\Gamma_n x_*^{(n)}. \tag{3.2.6}$$

Далее,

$$\| A_n^{-1}\Gamma_n x_*^{(n)} \| = \varepsilon \| A_n^{-1} B_n x_*^{(n)} \| = \| \Gamma_n \| \cdot \| A_n^{-1} B_n x_*^{(n)} \|;$$

$$\| A_n^{-1}\Gamma_n (z_*^{(n)} - x_*^{(n)}) \| = \varepsilon \| A_n^{-1} B_n (z_*^{(n)} - x_*^{(n)}) \| \leqslant$$

$$\leqslant \varepsilon \| A_n^{-1} \| \cdot \| z_*^{(n)} - x_*^{(n)} \| \leqslant \varepsilon C_1 \| z_*^{(n)} - x_*^{(n)} \|,$$

и из (3.2.6) следует

$$\| z_*^{(n)} - x_*^{(n)} \| \geqslant \frac{\| \Gamma_n \| \cdot \| A_n^{-1} B_n x_*^{(n)} \|}{1 + \varepsilon C_1}.$$

Сравнив это с неравенством (3.2.2), которое в данном случае имеет вид $\| z_*^{(n)} - x_*^{(n)} \| \leqslant p \| \Gamma_n \|$, получим

$$\| A_n^{-1} B_n x_*^{(n)} \| \leqslant p (1 + \varepsilon C_1).$$

Полагая $\varepsilon \to 0$, находим, что условие (3.2.5) выполнено со значением $C_3 = p$.

Докажем достаточность условий теоремы. Положим $r = \beta / C_1$, где $\beta = \text{const}$, $0 < \beta < 1$, и потребуем, чтобы $\| \Gamma_n \| \leqslant r$. Если $\Gamma_n = 0$, то $z_*^{(n)} = x_*^{(n)} + A_a^{-1} \delta^{(n)}$, $\| z_*^n - x_*^{(n)} \| \leqslant C_1 \| \delta^{(n)} \|$, и теорема доказана. Если же $\Gamma_n \neq 0$, то положим $B_n = \Gamma_n / \| \Gamma_n \|$ и $z_*^{(n)} = u^{(n)} + v^{(n)}$, где

$$(A_n + \Gamma_n) u^{(n)} = f^{(n)}, \quad (A_n + \Gamma_n) v^{(n)} = \delta^{(n)}. \tag{3.2.7}$$

Имеем

$$\| v^{(n)} \| \leqslant \| (I_n + A_n^{-1} \Gamma_n)^{-1} \| \cdot \| A_n^{-1} \| \cdot \| \delta^{(n)} \| \leqslant C_1 (1 - \beta)^{-1} \| \delta^{(n)} \|. \tag{3.2.8}$$

По формуле (3.1.5) находим из первого уравнения (3.2.7)

$$\| u^{(n)} - x_*^{(n)} \| = \| (I_n + A_n^{-1} \Gamma_n)^{-1} A_n^{-1} \Gamma_n x_*^{(n)} \| \leqslant$$

$$\leqslant \frac{\| A_n^{-1} B_n x_*^{(n)} \|}{1 - \beta} \| \Gamma_n \| \leqslant \frac{C_3}{1 - \beta} \| \Gamma_n \|,$$

неравенство устойчивости выполняется при значениях постоянных $p = (1 - \beta)^{-1} C_3$, $q = (1 - \beta)^{-1} C_1$.

4°. **Следствие 3.2.1.** *Если нормы* $\| x_*^{(n)} \|$ *ограничены в совокупности, то для устойчивости процесса* (3.1.1) *в смысле второго определения необходимо и достаточно, чтобы* $\| A_n^{-1} \| \leqslant C_1 = \text{const}$.

Отметим частный случай, обычно встречающийся при реализации приближенных методов. Пусть X_n суть подпространства некоторого пространства X и пусть в норме этого пространства $x_*^{(n)} \to_{n \to \infty} x_* \in X$. В этом случае назовем процесс (3.1.1) сходящимся. Для такого процесса величины $\| x_*^{(n)} \|$ ограничены, и для его устойчивости необходимо и достаточно, чтобы $\| A_n^{-1} \| \leqslant C_1 = \text{const}$.

Следствие 3.2.2 [139]. *Для устойчивости процесса* (3.1.1) *в смысле второго определения необходимо и достаточно, чтобы*

$$\| A_n^{-1} \| \leqslant C_1, \quad \| A_n^{-1} \| \cdot \| x_*^{(n)} \| \leqslant C_3; \quad C_1, C_3 = \text{const}. \tag{3.2.9}$$

Достаточность условий (3.2.9) и необходимость первого из них очевидны; докажем необходимость второго условия.

Пусть процесс устойчив. В силу теоремы 3.2.2 $\| A_n^{-1} B_n x_*^{(n)} \| \leqslant C_2$; $\forall B_n$, $\| B_n \| = 1$. Допустим, что произведение $\| A_n^{-1} \| \cdot \| x_*^{(n)} \|$

неограничено. Перейдя к подпоследовательности, можно считать, что $\|A_n^{-1}\| \cdot \|x_*^{(n)}\| \to_{n\to\infty} \infty$. Для каждого n найдется такой элемент $y^{(n)}$, что $\|y^{(n)}\| = 1$ и $\|A_n^{-1}y^{(n)}\| \geqslant 2^{-1}\|A_n^{-1}\|$. Построим линейный функционал l_n, определенный на X_n, со свойствами $\|l_n\| = 1$ и $|l_n x_*^{(n)}| = \|x_*^{(n)}\|$. Положим $B_n x^{(n)} = l_n x^{(n)} \cdot y^{(n)}$; ясно, что $\|B_n\| = 1$. Имеем $A_n^{-1}B_n x_*^{(n)} = \|x_*^{(n)}\| \cdot A_n^{-1}y^{(n)}$ и $\|A_n^{-1}B_n x_*^{(n)}\| \geqslant 2^{-1}\|A_n^{-1}\| \cdot \|x_*^{(n)}\| \xrightarrow[n\to\infty]{} \infty$, что противоречит предположению.

Из следствия 3.2.2 сразу вытекает такое утверждение.

Следствие 3.2.3. *Если нормы* $\|A_n^{-1}\|$ *положительно ограничены снизу, то для устойчивости процесса* (3.1.1) *в смысле второго определения необходимо и достаточно, чтобы* $\|A_n^{-1}\| \leqslant C_1$, $\|x_*^{(n)}\| \leqslant C_2$; $C_1, C_2 = \text{const}$. *Если еще процесс* (3.1.1) *сходящийся, то для устойчивости необходимо и достаточно, чтобы* $\|A_n^{-1}\| \leqslant C_1$.

Следствие 3.2.3 можно доказать независимо от следствия 3.2.2 [90].

5°. Если $\sup_n \|A_n^{-1}\| = \infty$, то процесс (3.1.1) неустойчив в последовательности (X_n, Y_n). Рассмотрим общий случай, когда $\sup_n \|A_n^{-1}\|_{Y_n \to X_n} \leqslant \infty$. Введем новые пространства $\bar{X}_n$ и $\bar{Y}_n$, совпадающие соответственно с X_n и Y_n как множества элементов, но нормы в них определяются формулами

$$\|\cdot\|_{\bar{X}_n} = \frac{1}{\sqrt{\|A_n^{-1}\|}}\|\cdot\|_{X_n}, \quad \|\cdot\|_{\bar{Y}_n} = \sqrt{\|A_n^{-1}\|} \cdot \|\cdot\|_{Y_n}. \tag{3.2.10}$$

Теорема 3.2.3. *Для того чтобы процесс* (3.1.1) *был устойчив в смысле второго определения в последовательности* $(\bar{X}_n, \bar{Y}_n)$, *необходимо и достаточно, чтобы*

$$\|x_*^{(n)}\|_{\bar{X}_n} \leqslant C = \text{const}. \tag{3.2.11}$$

Доказательство сразу вытекает из соотношений (3.2.9), написанных для пространств $\bar{X}_n$ и $\bar{Y}_n$, и из очевидного тождества $\|A_n^{-1}\|_{\bar{Y}_n \to \bar{X}_n} = 1$.

Замечание 3.2.1. Теорема 3.2.3 остается в силе, если нормы в $\bar{X}_n$ и $\bar{Y}_n$ определить формулами

$$\|\cdot\|_{\bar{X}_n} = \gamma(n)\|\cdot\|_{\bar{X}_n}, \quad \|\cdot\|_{\bar{Y}_n} = \frac{1}{\gamma(n)}\|\cdot\|_{\bar{Y}_n}, \tag{3.2.12}$$

где $\gamma(n)$ — любая неотрицательная функция от n, удовлетворяющая неравенству

$$\left[\|A_n^{-1}\|_{Y_n \to X_n}\right]^{-1} \geqslant \tilde{C}\gamma^2(n), \quad \tilde{C} = \text{const}. \tag{3.2.13}$$

§ 3. УСТОЙЧИВОСТЬ ПРОЦЕССА РИТЦА

1°. Метод Ритца описан выше, в § 1 главы 2. Обозначим соответственно через $a^{(n)}$ и $f^{(n)}$ векторы с составляющими $a_k^{(n)}$ и (f, φ_{nk}), $k = 1, 2, \ldots, N$, а через M_n — матрицу Ритца, т. е. матрицу энергетических произведений $[\varphi_{nk}, \varphi_{nj}]$, $k, j = 1, 2, \ldots, N$, и запишем систему (2.1.4) в виде

$$M_n a^{(n)} = f^{(n)}, \quad n = 1, 2, \ldots \tag{3.3.1}$$

Будем рассматривать $a^{(n)}$ и $f^{(n)}$ как элементы евклидова пространства R_N, а M^n — как оператор в R_N. Если применить в данном случае общую схему § 3 главы 1, то $X = H_A$, $X_N = R_N$, $x^{(n)} = a^{(n)}$; соответственно r_n задается системой (2.4.1), или, что то же, — уравнением (3.3.1). Наконец, оператор p_n, отображающий R_N в H_A, определяется формулой (2.1.3).

2°. С методом Ритца связаны два вычислительных процесса: процесс (3.3.1), приводящий к вычислению $a^{(n)}$, и процесс вычисления приближенного решения $x_*^{(n)}$. Если обозначить через $a_*^{(n)}$ решение уравнения (3.3.1), то

$$x_*^{(n)} = p_n a_*^{(n)}; \tag{3.3.2}$$

имея в виду применить общие теоремы § 2 настоящей главы, запишем последний процесс иначе. Будем рассматривать p_n как оператор из R_N в $X_n = H_A^{(n)}$; тогда существует обратный оператор $S_n = p_n^{-1}$; $S_n x^{(n)} = a^{(n)}$. Подставив это в (3.3.1), получим процесс, определяющий приближенное решение

$$M_n S_n x^{(n)} = f^{(n)}. \tag{3.3.3}$$

Ниже в данном параграфе исследуется устойчивость обоих названных процессов в смысле второго определения.

3°. Пусть в некотором гильбертовом пространстве H задана последовательность вида (2.1.5). Пусть G_n — матрица Грама элементов $\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nN}$ и $\lambda_n^{(1)}$ — наименьшее собственное число этой матрицы. Последовательность (2.1.5) называется *сильно минимальной в H*, если числа $\lambda_n^{(1)}$ ограничены снизу положительным числом, не зависящим от n.

Теорема 3.3.1. *Для того чтобы процесс* (3.3.1) *вычисления коэффициентов Ритца был устойчив в последовательности пар пространств* (R_N, R_N), *необходимо и достаточно, чтобы координатная система была сильно минимальна в энергетическом пространстве рассматриваемой задачи.*

Матрица Ритца совпадает с матрицей Грама (в энергетической метрике) координатных элементов $\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nN}$. Эта матрица — симметричная и положительно определенная, и ее собственные числа положительны; пусть $\lambda_n^{(1)}$ — наименьшее

из них. Обозначим через A_n оператор, порожденный матрицей M_n и действующий в R_N. В данном случае элемент $x_*^{(n)}$ (§ 4 главы 1) следует отождествить с $a_*^{(n)}$ — решением системы (3.3.1). Очевидно, что $\|A_n^{-1}\| = 1/\lambda_n^{(1)}$. По теореме 3.2.1 для устойчивости необходимо, чтобы $\|A_n^{-1}\| \leqslant C_1 = \text{const}$, или $\lambda_n^{(1)} \geqslant 1/C_1$. Чтобы доказать достаточность условия теоремы, остается проверить (следствие 3.2.1), что нормы $\|a_*^{(n)}\|$ ограничены. Процесс Ритца — сходящийся в H_A, поэтому $|x_*^{(n)}| \leqslant C = \text{const}$. Далее,

$$|x_*^{(n)}|^2 = [\textstyle\sum_{k=1}^N a_k^{(n)}\varphi_{nk},\ \sum_{j=1}^N a_j^{(n)}\varphi_{nj}] = (M_n a^{(n)},\ a^{(n)}) \geqslant \lambda_n^{(1)}\|a^{(n)}\|^2.$$

Отсюда $\|a^{(n)}\| \leqslant \sqrt{C_1}C$. Теорема доказана.

4°. **Теорема 3.3.2.** *Для того чтобы процесс* (3.3.3) *вычисления приближенных решений по Ритцу был устойчив в последовательности пар пространств* $(H_A^{(n)}, R_N)$, *необходимо и достаточно, чтобы координатная система была сильно минимальной в* H_A.

Процесс (3.3.3) — сходящийся в H_A. По следствию 3.2.1 для доказательства теоремы необходимо и достаточно установить, что нормы $\|A_n^{-1}\|_{R_N \to H_A}$ ограничены. В данном случае следует отождествить оператор A_n с $M_n S_n = M_n p_n^{-1}$, так что $A_n^{-1} = p_n M_n^{-1}$. Имеем $x_*^{(n)} = p_n a_*^{(n)}$. Как мы видели выше, $|x_*^{(n)}|^2 = (M_n a_*^{(n)},\ a_*^{(n)})$, поэтому

$$|p_n a_*^{(n)}|^2 = (M_n a_*^{(n)},\ a_*^{(n)}) = \|M_n^{1/2} a_*^{(n)}\|_{R_N}^2.$$

Но $a_*^{(n)} = M_n^{-1} f^{(n)}$, следовательно,

$$|p_n M_n^{-1} f^{(n)}| = \|M_n^{-1/2} f^{(n)}\|_{R_N},$$

или

$$\|p_n M_n^{-1}\|_{R_N \to H_A^{(n)}} = \|M_n^{-1/2}\|_{R_N} = [\lambda_n^{(1)}]^{-1/2}.$$

Последняя величина ограничена тогда и только тогда, когда координатная система сильно минимальна в энергетической норме. Теорема доказана.

Устойчивость метода Ритца в задачах о спектре самосопряженного оператора (спектр предполагается дискретным) исследована в статье [54]; результаты этой статьи изложены с некоторыми изменениями в [90].

5°. Исследуем устойчивость процессов Ритца в общем случае, когда $\inf \lambda_n^{(1)} \geqslant 0$. В последовательностях пар пространств (R_N, R_N) и $(H_A^{(n)}, R_N)$ эти процессы могут быть неустойчивыми (это будет, если $\inf \lambda_n^{(1)} = 0$), и мы введем вместо R_N новые пространства.

Пусть $\gamma(n)$ — такая положительная функция от n, что $\lambda_n^{(1)} \geqslant c_0^2\gamma^2(n)$, $c_0 = \text{const}$; в частности, можно принять $\gamma^2(n) = \lambda_n^{(1)}$. Введем N-мерные гильбертовы пространства $\bar{X}_n$ и $\bar{Y}_n$, элементы которых суть N-мерные векторы, а нормы заданы формулами

$$\forall b \in R_N, \quad \|b\|_{\bar{X}_n} = \gamma(n)\|b\|_{R_N}, \quad \|b\|_{\bar{Y}_n} = \frac{1}{\gamma(n)}\|b\|_{R_N}. \tag{3.3.4}$$

На этот раз обозначим через A_n оператор, порожденный матрицей M_n и действующий из $\bar{X}_n$ в $\bar{Y}_n$, и будем считать, что $a^{(n)} \in \bar{X}_n$, $f^{(n)} \in \bar{Y}_n$.

Теорема 3.3.3. *Процесс* (3.3.1) *устойчив в последовательности пар пространств* $(\bar{X}_n, \bar{Y}_n)$.

Достаточно проверить, что выполнены неравенства

$$\|A_n^{-1}\|_{\bar{Y}_n \to \bar{X}_n} \leqslant C_1, \quad \|a_*^{(n)}\|_{\bar{X}_n} \leqslant C_2. \tag{3.3.5}$$

Оценим величину $\|A_n^{-1}\|_{\bar{Y}_n \to \bar{X}_n}$:

$$\|A_n^{-1}\|_{\bar{Y}_n \to \bar{X}_n} = \sup_{b \in R_N} \frac{\|A_n^{-1}b\|_{\bar{X}_n}}{\|b\|_{\bar{Y}_n}} = \gamma^2(n) \sup_{b \in R_N} \frac{\|M_n^{-1}b\|_{R_N}}{\|b\|_{R_N}} =$$

$$= \frac{\gamma^2(n)}{\lambda_n^{(1)}} \leqslant \frac{1}{c_0^2},$$

и первое из неравенств (3.3.5) установлено. Заметим, что если принять $\gamma(n) = \sqrt{\lambda_n^{(1)}}$, то получится $\|A_n^{-1}\|_{\bar{Y}_n \to \bar{X}_n} = 1$.

Пусть $v^{(n)}$ — произвольный элемент из $H_A^{(n)}$;

$$v^{(n)} = \sum_{k=1}^{N} b_k \varphi_{nk}.$$

Полагая $b = (b_1, b_2, \ldots, b_N)$, имеем

$$|v^{(n)}|^2 = (M_n b, b)_{R_N} \geqslant \lambda_n^{(1)}\|b\|_{R_N}^2 \geqslant c_0^2\|b\|_{\bar{X}_n}^2. \tag{3.3.6}$$

Метод Ритца сходится в H_A, поэтому $|x_*^{(n)}| \leqslant C = \text{const}$. По неравенству (3.3.6) $\|a_*^{(n)}\|_{\bar{X}_n} \leqslant C/c_0$, и второе неравенство (3.3.5) также установлено. Теорема доказана.

Замечание 3.3.1. Пусть $\mu_n^{(1)}$ — наименьшее собственное число матрицы $\hat{M}_n$ скалярных произведений $(\varphi_{nk}, \varphi_{nj})_H$, $j, k = 1, 2, \ldots, N$. Достаточно подчинить $\gamma(n)$ неравенству $\mu_n^{(1)} \geqslant$

$\geqslant c^2\gamma^2(n)$, $c = \text{const} > 0$. Для доказательства достаточно установить, что $\mu_n^{(1)} \leqslant c'\lambda_n^{(1)}$, $c' = \text{const}$. Действительно,

$$\mu_n^{(1)} = \inf_{b \in R_N} \frac{(\widehat{M}_n b,\, b)_{R_N}}{\|b\|_{R_N}^2} = \inf_{b \in R_N} \frac{\left\|\sum_{k=1}^N b_k \Phi_{nk}\right\|_H^2}{\|b\|_{R_N}^2};$$

$$\lambda_n^{(1)} = \inf_{b \in R_N} \frac{(M_n b,\, b)_{R_N}}{\|b\|_{R_N}^2} = \inf_{b \in R_N} \frac{\left|\sum_{k=1}^N b_k \Phi_{nk}\right|^2}{\|b\|_{R_N}^2}.$$

Пусть λ_0 — нижняя грань оператора A и пусть второй инфимум достигается при $b = \bar{b} = (\bar{b}_1, \bar{b}_2, \ldots, \bar{b}_N)$. Тогда

$$\mu_n^{(1)} \leqslant \frac{\left\|\sum_{k=1}^N \bar{b}_k \Phi_{nk}\right\|^2}{\|\bar{b}\|_{R_N}} \leqslant \frac{1}{\lambda_0} \frac{\left|\sum_{k=1}^N \bar{b}_k \Phi_{nk}\right|^2}{\|\bar{b}\|_{R_N}} = \frac{\lambda_n^{(1)}}{\lambda_0}$$

и можно положить $c' = 1/\lambda_0$.

6°. **Теорема 3.3.4.** *Процесс* (3.3.3) *устойчив в последовательности пар пространств* $(H_A^{(n)}, \overline{Y}_n)$.

Уравнение (3.3.1) запишем в виде

$$A_n a^{(n)} = f^{(n)}, \tag{3.3.7}$$

где мы считаем, что $a^{(n)} \in \overline{X}_n$, $f^{(n)} \in \overline{Y}_n$, и обозначаем через A_n оператор, действующий из $\overline{X}_n$ в $\overline{Y}_n$ и порожденный матрицей M_n. Наряду с уравнением (3.3.7) рассмотрим искаженное уравнение

$$(A_n + \Gamma_n) c^{(n)} = f^{(n)} + \delta^{(n)}. \tag{3.3.8}$$

Искаженное приближенное решение по Ритцу имеет вид

$$z_*^{(n)} = \sum_{k=1}^N c_k^{(n)} \varphi_{nk} = p_n c^{(n)}.$$

Отсюда

$$\left|z_*^{(n)} - x_*^{(n)}\right|^2 = \left(M_n\left(c^{(n)} - a^{(n)}\right),\ c^{(n)} - a^{(n)}\right) \leqslant$$
$$\leqslant \|A_n(c^{(n)} - a^{(n)})\|_{\overline{Y}_n} \cdot \|c^{(n)} - a^{(n)}\|_{\overline{X}_n}. \tag{3.3.9}$$

По теореме 3.3.3 существуют такие числа p, q, $r > 0$, что если $\|\Gamma_n\|_{\overline{X}_n \to \overline{Y}_n} \leqslant r$, то

$$\|c^{(n)} - a^{(n)}\|_{\overline{X}_n} \leqslant p\|\Gamma_n\|_{\overline{X}_n \to \overline{Y}_n} + q\|\delta^{(n)}\|_{\overline{Y}_n}. \tag{3.3.10}$$

Из (3.3.7) и (3.3.8) следует

$$(A_n + \Gamma_n)\left(c^{(n)} - a^{(n)}\right) = \left(I_n + \Gamma_n A_n^{-1}\right) A_n\left(c^{(n)} - a^{(n)}\right) = \delta^{(n)} - \Gamma_n a^{(n)};$$

I_n — тождественный оператор в $\overline{Y}_n$. Как было доказано в п. 5 $\|A_n^{-1}\|_{\overline{Y}_n \to \overline{X}_n} \leqslant c_0^{-2}$. Потребуем, чтобы $\|\Gamma_n\|_{\overline{X}_n \to \overline{Y}_n} \leqslant \beta c_0^{-2}$, где β — какое-нибудь число из интервала (0, 1). Тогда $\left\|\left(I_n + \Gamma_n A_n^{-1}\right)^{-1}\right\| \leqslant$

$\leqslant (1-\beta)^{-1}$, и так как нормы $\|a^{(n)}\|_{\overline{X}_n}$ ограничены, то

$$\|A_n(c^{(n)}-a^{(n)})\|_{\overline{Y}_n} \leqslant (1-\beta)^{-1}\left[\|\delta^{(n)}\|_{\overline{Y}_n} + c'\|\Gamma_n\|_{\overline{X}_n\to\overline{Y}_n}\right];\ c'=\text{const}. \tag{3.3.11}$$

Из формул (3.3.9) — (3.3.11) очевидно следует, что

$$|z_*^{(n)} - x_*^{(n)}| \leqslant p_1\|\Gamma_n\|_{\overline{X}_n\to\overline{Y}_n} + q_1\|\delta^{(n)}\|_{\overline{Y}_n}, \tag{3.3.12}$$

и за r_1 можно принять число $r_1 = \min(r,\ \beta c_0^{-2})$.

7°. Частными случаями теорем 3.3.3 и 3.3.4 являются доказанные ниже, в главе 9, теоремы об устойчивости процессов, связанных с методом конечных элементов.

8°. Пусть $\inf \lambda_n^{(1)} = 0$. Тогда процесс вычисления коэффициентов Ритца неустойчив в (R_N, R_N), а процесс вычисления приближенных решений неустойчив в $(H_A^{(n)}, R_N)$. Это значит, что погрешности $\|c^{(n)} - a^{(n)}\|_{R_N}$ и $|z_*^{(n)} - x_*^{(n)}|$ могут быть велики, даже если малы погрешности $\|\Gamma_n\|_{R_N\to R_N}$ и $\|\delta^{(n)}\|_{R_N}$. Нетрудно, однако, убедиться, что оценка для $|z_*^{(n)} - x_*^{(n)}|$ в этом случае существенно меньше, чем для величины $\|c^{(n)} - a^{(n)}\|_{R_N}$. Действительно, из (3.3.10) и (3.3.12) следует

$$\|c^{(n)} - a^{(n)}\|_{R_N} \leqslant \frac{1}{\gamma(n)}\left[\frac{p}{\gamma^2(n)}\|\Gamma_n\|_{R_N\to R_N} + \frac{q}{\gamma(n)}\|\delta^{(n)}\|_{R_N}\right] \leqslant$$

$$\leqslant \frac{\bar{p}}{\gamma(n)}\left[\|\Gamma_n\|_{R_N\to R_N}\gamma^{-2}(n) + \|\delta^{(n)}\|_{R_N}\gamma^{-1}(n)\right];$$

$$\bar{p} = \max(p,\ q);$$

$$|z_*^{(n)} - x_*^{(n)}| \leqslant \left[\frac{p'}{\gamma^2(n)}\|\Gamma_n\|_{R_N\to R_N} + \frac{q'}{\gamma(n)}\|\delta^{(n)}\|_{R_N}\right] \leqslant$$

$$\leqslant \bar{p}'\left[\|\Gamma_n\|_{R_N\to R_N}\gamma^{-2}(n) + \|\delta^{(n)}\|_{R_N}\gamma^{-1}(n)\right];\quad \bar{p}' = \max(p',\ q');$$

правая часть первого неравенства содержит малый делитель $\gamma(n)$, откуда и следует наше утверждение.

§ 4. ПРИМЕРЫ

1°. Рассмотрим уравнение

$$(t+1)x(t) = 1,\quad 0 < t < 1. \tag{3.4.1}$$

Оператор умножения на $t+1$ — положительно определенный в $L_2(0, 1)$, и уравнение (3.4.1) равносильно задаче о минимуме функционала

$$\int_0^1 [(t+1)x^2(t) - 2x(t)]\,dt. \tag{3.4.2}$$

Эту последнюю задачу можно решать по методу Ритца. В качестве координатных функций возьмем функции t^n, $n = 0,\ 1,$

2, ... По известной теореме Г. М. Мюнца (см., например, [123]) эта система полна в $L_2(0, 1)$, так как ряд обратных показателей, $\sum_{n=1}^{\infty} n^{-1}$, расходится. Из той же теоремы легко вытекает [90], что эта система не сильно минимальна, и процессы Ритца с такой координатной системой неустойчивы в обычных пространствах. Посмотрим, как эта неустойчивость проявится в счете.

2°. При выбранной нами координатной системе приближенное решение имеет вид

$$x^{(n)}(t) = \sum_{k=1}^{n} a_k^{(n)} t^{k-1}, \qquad (3.4.3)$$

что приводит к неискаженной системе Ритца

$$\sum_{k=1}^{n} \frac{2k+2j+3}{(k+j+1)(k+j+2)} a_k^{(n)} = \frac{1}{j+1}, \quad 0 \leqslant j \leqslant n-1. \qquad (3.4.4)$$

Напишем расширенную систему Ритца 5-го порядка

3/2	5/6	7/12	9/20	11/30	1
5/6	7/12	9/20	11/30	13/42	1/2
7/12	9/20	11/30	13/42	15/56	1/3
9/20	11/30	13/42	15/56	17/72	1/4
11/30	13/42	15/56	17/72	19/90	1/5

Как обычно, расширенные матрицы Ритца низших порядков получаются усечением.

Системы Ритца порядков от 3 до 5 решались на ЭВМ БЭСМ-6. При этом оказалось, что для системы 5-го порядка наступает переполнение и получить решение не удается. Тем более не имело смысла решать системы более высокого порядка. Точные значения коэффициентов Ритца $a_k^{(n)}$, $n = 1, 2, 3, 4$, приведены в табл. 1. В табл. 2 приведены десятичные приближения тех же коэффициентов с шестью верными знаками, а также десятичные приближения коэффициентов с $n = 5$, полученные следующим образом. Так как точные значения этих коэффициентов получить не удалось, то для $n = 5$ элементы расширенной матрицы Ритца были заменены их десятичными приближениями с 12 верными знаками; после этого были получены десятичные приближения коэффициентов, в которых мы сохранили шесть знаков.

Таблица 1

n \ k	1	2	3	4
1	2/3			
2	12/13	−6/13		
3	62/63	−50/63	20/63	
4	320/321	−100/107	70/107	−70/321

Таблица 2

n \ k	1	2	3	4	5
1	0,666667				
2	0,923077	−0,461538			
3	0,981270	−0,793651	0,317460		
4	0,996885	−0,934579	0,654026	−0,218069	
5	0,992859	−0,864970	0,367892	0,200555	−0,199993

Представляется интересным выяснить, насколько неискаженные приближения близки к точному решению $x_*(t) = (1 + t)^{-1}$. В данном примере энергетическая норма определяется формулой

$$|x|^2 = \int_0^1 (1 + t)\, x^2(t)\, dt \qquad (3.4.5)$$

и эквивалентна L_2-норме. Были вычислены значения $|x_* - x_*^{(n)}|$ для n от 1 до 5. Результаты приведены в табл. 3.

Таблица 3

n	1	2	3	4	5
ξ	0,1627	0,0290	0,0051	0,0008	0,0016

Из табл. 3 видно, что приближенные решения, вычисленные без погрешности, довольно хорошо сходятся в энергетической метрике к точному решению. В классическом методе Ритца погрешность аппроксимации в энергетической метрике убывает с ростом n [84]. Последнее число в табл. 3 показывает, что при $n = 5$ коэффициенты Ритца оказались вычисленными недостаточно точно.

В заключение настоящего пункта приведем таблицу значений искомой функции, точных и неискаженных приближенных (табл. 4); последние соответствуют значениям $n = 3, 4, 5$. Все значения даны с шестью верными десятичными знаками.

Как мы видим, приближенные значения решений в фиксированых точках приближаются к точным значениям довольно медленно и не монотонно. Это связано с тем, что метод Ритца сходится, вообще говоря, только в энергетической метрике, которая в данном случае эквивалентна метрике L_2.

3°. Исказим расширенную матрицу Ритца, заменив все ее элементы их десятичными приближениями с четырьмя верными знаками после запятой. Приведем таблицу искаженных коэффициентов Ритца, вычисленных с шестью верными знаками после запятой (табл. 5).

Таблица 4

t	Точные значения	Приближенные значения		
		$n=3$	$n=4$	$n=5$
0,0	1,000000	0,984310	0,995807	0,994871
0,1	0,909091	0,908003	0,909744	0,910432
0,2	0,833333	0,838072	0,835098	0,833902
0,3	0,769231	0,774517	0,773706	0,776198
0,4	0,714286	0,717339	0,721801	0,708751
0,5	0,666667	0,666538	0,678548	0,663192
0,6	0,625000	0,622114	0,641773	0,630532
0,7	0,588235	0,584066	0,608823	0,609512
0,8	0,555556	0,552395	0,576560	0,593628
0,9	0,526316	0,527100	0,541364	0,566283
1,0	0,500000	0,508182	0,499133	0,493396

Таблица 5

n \ k	1	2	3	4	5
3	0,98310	−0,794960	0,317048		
4	0,995807	−0,923202	0,627642	−0,201115	
5	0,994871	−0,882381	0,377633	0,258677	−0,255133

Сравнивая данные табл. 5 и 2, видим, что искажение коэффициентов сказывается уже, как правило, во втором или в третьем знаке после запятой, тогда как в расширенной матрице Ритца искажение коснулось только пятых знаков. Это связано с неустойчивостью процесса Ритца в данном примере в последовательности пар пространств (R_N, R_N).

4°. Рассмотрим теперь процесс решения задачи (3.4.1) по Ритцу, когда за координатные функции взяты полиномы Лежандра, умноженные на множители, по порядку роста совпадающие с нормирующими множителями

$$\varphi_n(t) = [\sqrt{n}]\, p_{n-1}(2t-1).$$

Последовательность $\{\varphi_n(t)\}$ сильно минимальна в $L_2(0,1)$ [90], а следовательно, и в энергетическом пространстве задачи (3.4.1). При $n \leqslant 6$ получены точные (в виде простых дробей) значения коэффициентов; в табл. 6 приведены десятичные приближения этих коэффициентов с шестью верными знаками после запятой. По данным таблицы можно догадаться, что коэффициенты Ритца $a_k^{(n)}$ стремятся при возрастании n к некоторым пределам, а при возрастании k довольно быстро убывают. В [90] показано, что если координатная система классического метода Ритца сильно минимальна в энергетической

Таблица 6

n \ k	1	2	3	4	5	6	7
3	0,693146	—0,238318	0,054517	—0,005452			
4	0,693147	—0,238324	0,054565	—0,005615	0,001070		
5	0,693147	—0,238325	0,054567	—0,005620	0,001101	—0,000204	
6	0,693147	—0,238325	0,054567	—0,005620	0,001102	—0,000210	0,000038

норме, то существуют пределы $a_k = \lim\limits_{n\to\infty} a_k^{(n)}$, и последовательность $\{a_k\} \in l_2$.

Расширенные матрицы Ритца были подвергнуты искажению, такому же, как в п. 3°: их элементы были заменены десятичными приближениями с четырьмя верными знаками, и полученные системы были решены с возможно большей точностью. Значения искаженных коэффициентов Ритца с шестью верными знаками приведены в табл. 7. Сравнение данных табл. 6 и 7 показывает, что искажения коэффициентов Ритца проявляются только в пятых и шестых знаках и, следовательно, являются величинами того же порядка, что и искажения элементов матрицы Ритца.

Таблица 7

n \ k	1	2	3	4	5	6	7
3	0,693146	—0,238314	0,054516	—0,005452			
4	0,693146	—0,238320	0,054564	—0,005615	0,001070		
5	0,693146	—0,238321	0,054566	—0,005620	0,001101	—0,000204	
6	0,693146	—0,238321	0,054566	—0,005620	0,001102	—0,000210	0,000038

5°. В пп. 5 и 6° настоящего параграфа мы вкратце изложим результатв вычислений по двум примерам, подробно рассмотренным в [90]. Будем решать по методу Ритца задачу, положительно определенную в $L_2(0, 1)$:

$$0 < t < 1, \quad -\frac{d^2x}{dt^2} = \frac{1}{1+t}; \quad x(0) = x'(1) = 0; \qquad (3.4.6)$$

в качестве координатной системы возьмем последовательность $\varphi_n(t) = t^n$, $n = 1, 2, \ldots$; опираясь на уже упомянутую теорему Мюнца, легко доказать, что эта система полна, но не сильно минимальна в энергетической метрике задачи (3.4.6).

Первые восемь приближений можно получить, решая систему Ритца восьмого порядка, а также все системы, полученные из нее усечением. Коэффициенты Ритца — решения упомянутых систем — обозначим через $a_k^{(n)}$. Решим эти системы точно, без по-

грешностей, а затем заменим полученные таким образом неискаженные значения коэффициентов Ритца их десятичными приближениями с четырьмя верными знаками (табл. 8).

Таблица 8

n \ k	1	2	3	4	5	6	7	8
1	0,3069							
2	0,6480	—0,3411						
3	0,6869	—0,4579	0,0788					
4	0,6922	—0,4899	0,1312	—0,0267				
5	0,6930	—0,4977	0,1547	—0,0541	0,0110			
6	0,6931	—0,4995	0,1631	—0,0708	0,0260	—0,0050		
7	0,6931	—0,4999	0,1657	—0,0786	0,0378	—0,0136	0,0025	
8	0,6931	—0,5000	0,1664	—0,0817	0,0446	—0,0218	0,0075	—0,0013

Исказим теперь расширенную матрицу, заменив в ней свободные члены их десятичными приближениями с четырьмя верными знаками после запятой. Искаженные коэффициенты Ритца, которые мы получим, решая систему с искаженной расширенной матрицей и соответствующие усеченные системы, обозначим через $b_k^{(n)}$. Новые системы решим точно, а затем заменим коэффициенты $b_k^{(n)}$ их десятичными приближениями с четырьмя верными знаками (табл. 9).

Таблица 9

n \ k	1	2	3	4	5	6	7	8
1	0,3069							
2	0,6483	—0,3414						
3	0,6883	—0,4614	0,080					
4	0,6974	—0,5160	0,171	—0,0455				
5	0,7127	—0,6690	0,630	—0,5810	0,2142			
6	0,7633	—1,428	4,172	—7,665	6,590	—2,125		
7	0,9141	—4,595	25,28	—71,00	101,6	—71,79	19,91	
8	1,627	—24,54	204,8	—819,1	1747	—2047	1243	—305,7

Сравнение табл. 8 и 9 показывает, что погрешность искажения коэффициентов Ритца в данном случае сильно возрастает при возрастании n.

6°. Рассмотрим еще задачу

$$-1<t<1, \quad -\frac{d}{dt}\left[(2+t)\frac{dx}{dt}\right]=1, \quad x(-1)=x(1)=0, \quad (3.4.7)$$

положительно определенную в $L_2(-1, 1)$. В качестве координатной системы выберем последовательность интегралов от

нормированных полиномов Лежандра

$$\varphi_n(t) = \sqrt{(2n+1)/2} \int_{-1}^{t} P_n(\tau)\, d\tau. \qquad (3.4.8)$$

Эта система сильно минимальна в энергетическом пространстве задачи (3.4.7) (см. [90]). Как и в предыдущем примере, были составлены и точно решены системы Ритца от 1-го до 8-го порядка. В табл. 10 приведены десятичные приближения этих

Таблица 10

n \ k	1	2	3	4	5	6	7	8
1	—0,4082							
2	—0,4374	0,1129						
3	—0,4395	0,1213	—0,0308					
4	—0,4397	0,1219	—0,0330	0,0083				
5	—0,4397	0,1219	—0,0332	0,0089	—0,0022			
6	—0,4397	0,1220	—0,0332	0,0090	—0,0024	0,0006		
7	—0,4397	0,1220	—0,0332	0,0090	—0,0024	0,0006	—0,0002	
8	—0,4397	0,1220	—0,0332	0,0090	—0,0024	0,0006	—0,0002	0,0001

решений с четырьмя верными знаками после запятой. Далее системы Ритца были искажены: все элементы расширенных матриц были заменены их десятичными приближениями с четырьмя верными знаками после запятой. В этом случае получение точных решений оказалось затруднительным; поэтому, чтобы по возможности избежать ошибок округления, вычисления производились с сравнительно большим числом знаков — с семью знаками для систем до 4-го порядка включительно и с 12 знаками для систем порядка от 5-го до 8-го. В табл. 11

Таблица 11

n \ k	1	2	3	4	5	6	7	8
1	—0,4083							
2	—0,4374	0,1129						
3	—0,4396	0,1213	—0,0308					
4	—0,4397	0,1219	—0,0330	0,0083				
5	—0,4398	0,1220	—0,0332	0,0089	—0,0022			
6	—0,4398	0,1220	—0,0332	0,0089	—0,0024	0,0007		
7	—0,4398	0,1220	—0,0332	0,0089	—0,0024	0,0007	—0,0002	
8	—0,4398	0,1220	—0,0332	0,0089	—0,0024	0,0007	—0,0002	0,0001

приведены десятичные значения искаженных коэффициентов Ритца с четырьмя верными знаками после запятой. Сравнение

двух последних таблиц показывает, что в данном случае искажения коэффициентов Ритца суть величины того же порядка, что и искажения элементов расширенной матрицы Ритца.

§ 5. ОБ УСТОЙЧИВОСТИ ПРОЦЕССА БУБНОВА — ГАЛЕРКИНА

1°. В сепарабельном гильбертовом пространстве, которое для упрощения обозначений считаем вещественным, рассмотрим уравнение

$$Ax := A_0x + Kx = f, \tag{3.5.1}$$

где A_0 — положительно определенный самосопряженный оператор и произведение $T = A_0^{-1}K$ вполне непрерывно в энергетическом пространстве H_0 оператора A_0. В работе автора от 1948 г. (подробнее см. [84]) установлены следующие теоремы.

а) Уравнение (3.5.1) эквивалентно уравнению

$$x + Tx = A_0^{-1}f, \tag{3.5.2}$$

рассматриваемому в H_0, в следующем смысле: любое решение уравнения (3.5.1) удовлетворяет уравнению (3.5.2), а любое решение уравнения (3.5.2) является обобщенным решением уравнения (3.5.1).

б) При заданной координатной системе метод Бубнова — Галеркина, примененный к уравнениям (3.5.1) и (3.5.2), приводит к одной и той же алгебраической системе.

в) Если координатная система полна в H_0, то метод Бубнова — Галеркина для уравнений (3.5.1), (3.5.2) сходится в H_0.

2°. Устойчивость метода Бубнова — Галеркина достаточно исследовать для уравнения (3.5.2). Рассмотрим оба вычислительных процесса: для вектора коэффициентов и для приближенного решения. В [153] для более общего метода Галеркина — Петрова доказано, что устойчивость названных процессов имеет место, если координатная система сильно минимальна в H_0. В [10] доказано, что условие сильной минимальности необходимо для устойчивости процесса вычисления приближенного решения; там же упрощено доказательство достаточности. В данном параграфе мы изложим содержание статьи [10] для метода Бубнова — Галеркина и дадим сравнительно простое доказательство теоремы об устойчивости процесса вычисления коэффициентов Бубнова — Галеркина.

3°. Пусть для построения приближенного решения уравнения (3.5.2) по Бубнову — Галеркину использована координатная система $\{\varphi_n\} \subset H_0$, подчиненная обычным условиям; в частности, при любом n элементы φ_k, $k = 1, 2, \ldots, n$, предполагаются линейно независимыми. Пусть $\{\omega_n\}$ — система, полученная ортогонализацией системы $\{\varphi_n\}$ в H_0. Обе системы свя-

заны соотношениями вида

$$\omega_k = \sum_{j=1}^{k} c_{kj}\varphi_j, \quad \varphi_k = \sum_{j=1}^{k} \gamma_{kj}\omega_j. \tag{3.5.3}$$

Обозначим через C_n матрицу n первых соотношений (3.5.3).

Лемма 3.5.1. *Если последовательность* $\{\varphi_n\}$ *сильно минимальна в* H_0, *то нормы* $\|C_n\|_{R_n \to R_n}$ *ограничены; в противном случае они монотонно стремятся к бесконечности вместе с* n.

Имеем $[\omega_k, \omega_j] = \delta_{kj}$; используя представление (3.5.3), получаем

$$C_n M_n C_n^* = I_n, \tag{3.5.4}$$

где M_n имеет тот же смысл, что и в § 3; звездочка обозначает сопряженную матрицу, а I_n — единичная матрица в R_n. Из (3.5.4) следует, что $M_n^{-1} = C_n^* C_n$ и, далее, $\|C_n\| = \sqrt{\|M_n^{-1}\|} = [\lambda_n^{(1)}]^{-1/2}$. Дальнейшее очевидно.

Теорема 3.5.1. *Для устойчивости процесса вычисления приближенного по Бубнову — Галеркину решения уравнения* (3.5.2) *в последовательности пар пространств* $(H_0^{(n)}, R_n)$ *необходимо и достаточно, чтобы координатная система была сильно минимальна в* H_0.

Через $H_0^{(n)}$ здесь обозначено подпространство пространства H_0 с базисом $(\varphi_1, \varphi_2, \ldots, \varphi_n)$.

Обозначим соответственно неискаженный и искаженный векторы коэффициентов Бубнова — Галеркина через $a_*^{(n)}$ и $b_*^{(n)}$, неискаженное и искаженное приближенные решения — через $x_*^{(n)}$ и $z_*^{(n)}$. Далее, положим

$$f^{(n)} = ((f, \varphi_1), (f, \varphi_2), \ldots, (f, \varphi_n)), \quad \varphi^{(n)} = (\varphi_1, \varphi_2, \ldots, \varphi_n);$$

$$\alpha^{(n)} = (C_n^*)^{-1} a_*^{(n)}, \quad \beta^{(n)} = (C_n^*)^{-1} b_*^{(n)}.$$

Наконец, через B_n обозначим матрицу элементов $[\varphi_k + T\varphi_k, \varphi_j]$, $j, k = 1, 2, \ldots, n$. В этих обозначениях $x_*^{(n)} = a_*^{(n)}\varphi^{(n)}$. Далее, $a_*^{(n)} = C_n^* \alpha_n$, отсюда

$$x_*^{(n)} = \sum_{k=1}^{n} \sum_{j=k}^{n} c_{jk}\alpha_j^{(n)}\varphi_k = \sum_{j=1}^{n} \alpha_j^{(n)} \sum_{k=1}^{j} c_{jk}\varphi_k = \sum_{j=1}^{n} \alpha_j^{(n)}\omega_j.$$

Аналогично

$$z_*^{(n)} = \sum_{j=1}^{n} \beta_j^{(n)}\omega_j$$

и, следовательно,

$$z_*^{(n)} - x_*^{(n)} = \sum_{j=1}^{n} (\beta_j^{(n)} - \alpha_j^{(n)})\omega_j. \tag{3.5.5}$$

Системы уравнений Бубнова — Галеркина, искаженная и неискаженная, соответственно имеют вид

$$(B_n + \Gamma_n) b^{(n)} = f^{(n)} + \delta^{(n)}; \quad B_n a^{(n)} = f^{(n)}$$

или $$(C_nB_nC_n^* + C_n\Gamma_nC_n^*)\,\beta^{(n)} = C_n(f^{(n)} + \delta^{(n)}),$$

$$C_nB_nC_n^*\alpha^{(n)} = C_nf^{(n)}.$$

Отсюда

$$(C_nB_nC_n^* + C_n\Gamma_nC_n^*)(\beta^{(n)} - \alpha^{(n)}) = C_n\delta^{(n)} - C_n\Gamma_nC_n^*\alpha^{(n)}. \quad (3.5.6)$$

Обозначим через Π_n ортопроектор из H_0 на $H_0^{(n)}$. Докажем, что существует постоянная $\tau > 0$, удовлетворяющая неравенству

$$\forall n,\ \forall v^{(n)} \in (I+T)H_0^{(n)},\ |\Pi_n v^{(n)}| \geqslant \tau |v^{(n)}|. \quad (3.5.7)$$

Допустим противное. Тогда можно построить такую подпоследовательность $v^{(n)} \in (I+T)H_0^{(n)}$, $n = 1, 2, \ldots$, что $|v^{(n)}| = 1$ и $|\Pi_n v^{(n)}| \xrightarrow[n\to\infty]{} 0$. Пусть $v^{(n)} = (I+T)w^{(n)}$; тогда $w^{(n)} \in H_0^{(n)}$. При этом $|w^{(n)}| \leqslant |(I+T)^{-1}|$ и $|\Pi_n(I+T)w^{(n)}| = |w^{(n)} + \Pi_nTw^{(n)}| \xrightarrow[n\to\infty]{} 0$. Последовательность $\{w^{(n)}\}$ ограничена и потому слабо компактна в H_0. Выделим из нее подпоследовательность, которую опять обозначим через $\{w^{(n)}\}$ и которая слабо сходится в H_0 к некоторому элементу $w^{(0)}$. Тогда $Tw^{(n)} \to Tw^{(0)}$ сильно в H_0. Далее,

$$\Pi_nTw^{(n)} = \Pi_nTw^{(0)} + \Pi_nT(w^{(n)} - w^{(0)}).$$

Первое слагаемое справа стремится к $Tw^{(0)}$ в норме H_0, потому что последовательность подпространств $\{H_0^{(n)}\}$ полна в H_0; второе слагаемое стремится к нулю в той же норме. Таким образом, $\Pi_nTw^{(n)} \to Tw^{(0)}$. Теперь из соотношения $|w^{(n)} + \Pi_nTw^{(n)}| \to 0$ следует, что $(I+T)w^{(0)} = 0$. Отсюда вытекает, что

$$\lim v^{(n)} = \lim (I+T)w^{(n)} = (I+T)w^{(0)} = 0,$$

а это противоречит условию $|v^{(n)}| = 1$.

Докажем теперь, что при достаточно больших n справедливо неравенство

$$\forall c^{(n)} \in R_n,\ \frac{1}{\mu}\|c^{(n)}\| \leqslant \|C_nB_nC_n^*c^{(n)}\| \leqslant$$

$$\leqslant |I+T|\,\|c^{(n)}\|,\quad \mu = \tau^{-1}|(I+T)^{-1}|^{-1}. \quad (3.5.8)$$

Как легко видеть, $C_nB_nC_n^*$ есть матрица скалярных произведений $[\omega_k + T\omega_k,\ \omega_j]$, $j,\ k = 1, 2, \ldots, n$. Отсюда

$$\|C_nB_nC_n^*c^{(n)}\|^2 = \sum_{j=1}^n \left|\sum_{k=1}^n [\omega_k + T\omega_k,\ \omega_j]\, c_k^{(n)}\right|^2 =$$

$$= \sum_{j=1}^n \left|\left[(I+T)\sum_{k=1}^n c_k^{(n)}\omega_k,\ \omega_j\right]\right|^2 \leqslant \left|\Pi_n(I+T)\sum_{k=1}^n c_k^{(n)}\omega_k\right|^2 \leqslant$$

$$\leqslant |I+T|^2 \cdot \left|\sum_{k=1}^n c_k^{(n)}\omega_k\right|^2 = |I+T|^2 \cdot \|c^{(n)}\|^2.$$

В то же время неравенство (3.5.7) дает

$$\left|\Pi_n(I+T)\sum_{k=1}^{n}c_k^{(n)}\omega_k\right|^2 \geqslant \tau^2\left|(I+T)\sum_{k=1}^{n}c_k^{(n)}\omega_k\right|^2 \geqslant$$
$$\geqslant \frac{1}{\mu^2}\left|\sum_{k=1}^{n}c_k^{(n)}\omega_k\right|^2 = \frac{1}{\mu^2}\|c^{(n)}\|^2.$$

4°. Переходим к доказательству утверждений теоремы 3.5.1. Пусть координатная система сильно минимальна в H_0. По лемме 3.5.1 $\|C_n\| = \|C_n^*\| \leqslant c_0 = \text{const}$. Положим $r = (2\mu c_0^2)^{-1}$ и пусть $\|\Gamma_n\| \leqslant r$. Тогда $\|C_n\Gamma_n C_n^*\| \leqslant 2^{-1}\mu$. В то же время, как показывает левое неравенство (3.5.8), $\|C_nB_nC_n^*\| > 1/\mu$ и матрица в левой части уравнения (3.5.6) обратима. При этом $\|[C_n(B_n+\Gamma_n)\times \times C_n^*]^{-1}\| \leqslant 2\mu$; из соотношений (3.5.6) и из факта сходимости метода Бубнова — Галеркина следует

$$|z_*^{(n)} - x_*^{(n)}| = \|\beta^{(n)} - \alpha^{(n)}\| \leqslant p\|\Gamma_n\| + q\|\delta^{(n)}\|, \qquad (3.5.9)$$

где p и q — некоторые постоянные. Достаточность условий теоремы доказана.

Пусть теперь координатная система не сильно минимальна в H_0. По лемме 3.5.1 $\|C_n\| \to \infty$. Выберем $\delta^{(n)}$ так, чтобы $\|C_n\delta^{(n)}\| \geqslant 2^{-1}\|C_n\| \cdot \|\delta^{(n)}\|$. Правое неравенство (3.5.8) показывает, что при $\Gamma_n = 0$

$$\|\beta^{(n)} - \alpha^{(n)}\| \geqslant \frac{\|C_n\| \cdot \|\delta^{(n)}\|}{2|I+T|},$$

и так как коэффициент при $\|\delta^{(n)}\|$ бесконечно возрастает, то рассматриваемый процесс неустойчив. Этим доказана необходимость условия теоремы.

5°. **Теорема 3.5.2.** *Если координатная система сильно минимальна в H_0, то процесс вычисления коэффициентов Бубнова — Галеркина устойчив в последовательности пар пространств (R_n, R_n).*

Доказательство очень просто: так как $b^{(n)} - a^{(n)} = C_n^*(\beta^{(n)} - \alpha^{(n)})$, а $\|C_n^*\| \leqslant c_0$, то, по неравенству (3.5.9),

$$\|b^{(n)} - a^{(n)}\| \leqslant c_0 p\|\Gamma_n\| + c_0 q\|\delta^{(n)}\|.$$

Замечание. В работах [16—24] исследована устойчивость метода Бубнова — Галеркина (точнее, метода Фаэдо — Галеркина) для довольно широкого класса нестационарных операторных уравнений. Часть упомянутых результатов изложена (и частично усилена) в [90]; некоторые из этих результатов усилены в [135]. В [78, 79] исследована устойчивость метода Фаэдо — Галеркина для задач, в которых квазилинейное параболическое уравнение решается при краевом условии (вообще говоря, нелинейном), содержащем производную по вре-

мени от искомой функции. В работах [3] и [56] изучается устойчивость метода Бубнова — Галеркина в банаховых пространствах.

§ 6. ПОГРЕШНОСТЬ ИСКАЖЕНИЯ РЕКУРРЕНТНОГО ВЫЧИСЛИТЕЛЬНОГО ПРОЦЕССА

1°. Рассмотрим бесконечный рекуррентный вычислительный процесс

$$\sum_{k=0}^{n} A_{nk}x^{(k)} = f^{(n)}, \quad n = 0, 1, 2, \ldots \tag{3.6.1}$$

Примем, что $x^{(n)} \in X_n$, $f^{(n)} \in Y_n$, где X_n и Y_n — банаховы пространства, и что A_{nk} — линейные операторы, действующие из X_k в Y_n. Будем считать, что каждый из операторов A_{nn} ограниченно обратим, а произведения $A_{nn}^{-1}A_{nk}$, $0 \leqslant k \leqslant n-1$, ограничены. При этих предположениях существует последовательность элементов $x^{(n)} \in X_n$, удовлетворяющих бесконечной системе (3.6.1).

Пусть каждый из операторов A_{nk} вычислен с погрешностью Γ_{nk}, а каждый из элементов $f^{(n)}$ — с погрешностью $\delta^{(n)}$, так что вместо системы (3.6.1) на самом деле решается система

$$\sum_{k=0}^{n} (A_{nk} + \Gamma_{nk}) z^{(k)} = f^{(n)} + \delta^{(n)}, \quad n = 0, 1, 2, \ldots \tag{3.6.2}$$

Допустим, что погрешности Γ_{nk} можно сделать сколь угодно малыми в следующем смысле: каково бы ни было число $\varepsilon > 0$, можно добиться того, чтобы

$$\forall n \geqslant 0, \quad \sum_{k=0}^{n} \| A_{nn}^{-1}\Gamma_{nk} \| \leqslant \varepsilon. \tag{3.6.3}$$

2°. Оценим погрешность искажения $\hat{\xi}_n = \| z_*^{(n)} - x_*^{(n)} \|$. Уравнения (3.6.1) и (3.6.2) преобразуем к виду

$$A_{nn}x^{(n)} = g^{(n)} := f^{(n)} - \sum_{k=0}^{n-1} A_{nk}x^{(k)}; \tag{3.6.4}$$

$$(A_{nn} + \Gamma_{nn}) z^{(n)} = g^{(n)} + \varkappa^{(n)},$$
$$\varkappa^{(n)} := \delta^{(n)} - \sum_{k=0}^{n-1} A_{nk}(z^{(k)} - x^{(k)}) - \sum_{k=0}^{n-1} \Gamma_{nk}z^{(k)}. \tag{3.6.5}$$

Переход к этим уравнениям преобразует рекуррентный процесс в эквивалентный свободный, и для оценки погрешности искажения можно воспользоваться неравенством (3.1.7), которое в данном случае означает следующее: если $\| A_{nn}^{-1}\Gamma_{nn} \| \leqslant \beta$, $0 \leqslant \beta < 1$, то

$$\hat{\xi}_n \leqslant (1-\beta)^{-1} [\, \| A_{nn}^{-1}\Gamma_{nn}x_*^{(n)} \| + \| A_{nn}^{-1}\varkappa^{(n)} \|\,] \leqslant$$
$$\leqslant (1-\beta)^{-1} \{ \| A_{nn}^{-1}\Gamma_{nn} \| \cdot \| x_*^{(n)} \| + \sum_{k=0}^{n-1} \| A_{nn}^{-1}\Gamma_{nk} \| \cdot \| x_*^{(k)} \| +$$
$$+ \sum_{k=0}^{n-1} [\, \| A_{nn}^{-1}A_{nk} \| + \| A_{nn}^{-1}\Gamma_{nk} \|\,] \hat{\xi}_k + \| A_{nn}^{-1} \| \cdot \| \delta^{(n)} \| \} =$$
$$= (1-\beta)^{-1} \{ \sum_{k=0}^{n} \| A_{nn}^{-1}\Gamma_{nk} \| \cdot \| x_*^{(k)} \| + \| A_{nn}^{-1} \| \cdot \| \delta^{(n)} \| +$$
$$+ \sum_{k=0}^{n-1} [\, \| A_{nn}^{-1}A_{nk} \| + \| A_{nn}^{-1}\Gamma_{nk} \|\,] \hat{\xi}_k \}. \tag{3.6.6}$$

Формула (3.6.6) позволяет последовательно вычислять оценки для ξ_0, ξ_1, ξ_2, ..., если известны оценки соответствующих норм $\|x_*^{(n)}\|$, $n=0,1,2, \ldots,10$,

§ 7. УСТОЙЧИВОСТЬ РЕКУРРЕНТНОГО ВЫЧИСЛИТЕЛЬНОГО ПРОЦЕССА

1°. Процесс (3.6.1) назовем *устойчивым относительно искажения в последовательности пар пространств* (X_n, Y_n), если существуют такие постоянные p_1, p_2, $p_3 > 0$, что из неравенства

$$\forall n \geqslant 0, \quad \sum_{k=0}^{n} \| A_{an}^{-1} \Gamma_{nk} \| \leqslant p_3 \tag{3.7.1}$$

вытекает следующая оценка погрешности искажения:

$$\hat{\xi}_n \leqslant p_1 \max_{j \leqslant n} \sum_{k=0}^{j} \| A_{jj}^{-1} \Gamma_{jk} \| + p_2 \max_{j \leqslant n} \| A_{jj}^{-1} \delta^{(j)} \|. \tag{3.7.2}$$

Это определение аналогично первому из определений устойчивости свободного вычислительного процесса.

Замечание. В предшествующем параграфе было отмечено, что любой рекуррентный вычислительный процесс легко преобразуется в эквивалентный свободный процесс. Можно было бы определить устойчивость рекуррентного процесса как устойчивость соответствующего свободного процесса, и тогда были бы применимы утверждения § 2 данной главы. Однако условия этих утверждений оказываются в данном случае довольно трудно проверяемыми и громоздкими, почему мы и предпочли дать для рекуррентного процесса независимое определение устойчивости.

Теорема 3.7.1. *Пусть* $h^{(n)} \in Y_n$, $n=0, 1, 2, \ldots$, *и пусть* $t^{(n)}$, $n=0, 1, 2, \ldots$, *есть последовательность решений бесконечной системы*

$$\forall n \geqslant 0, \quad \sum_{k=0}^{n} A_{nk} t^k = h^{(n)}. \tag{3.7.3}$$

Если выполнены неравенства

$$\| t^{(n)} \| \leqslant C_0 \max_{j \leqslant n} \| A_{jj}^{-1} h^{(j)} \|, \quad C_0 = \text{const}, \tag{3.7.4}$$

и

$$\| x_*^{(n)} \| \leqslant C_1 = \text{const}, \tag{3.7.5}$$

где $\{x_*^{(n)}\}$ *— последовательность решений системы* (3.6.1), *то вычислительный процесс* (3.6.1) *устойчив.*

Уравнения (3.6.2) преобразуем к виду

$$\sum_{k=0}^{n} A_{nk} (z^{(k)} - x^{(k)}) = \delta^{(n)} - \sum_{k=0}^{n} \Gamma_{nk} (z^{(k)} - x^{(k)}) - \sum_{k=0}^{n} \Gamma_{nk} x^{(k)}.$$

По неравенствам (3.7.4), (3.7.5)

$$\hat{\xi}_n \leqslant C_0 \Big\{ \max_{j \leqslant n} \| A_{jj}^{-1} \delta^{(j)} \| + \max_{j \leqslant n} \sum_{k=0}^{j} \| A_{jj}^{-1} \Gamma_{jk} \| \times$$

$$\times \max_{j \leqslant n} \hat{\xi}_j + C_1 \max_{j \leqslant n} \sum_{k=0}^{j} \| A_{jj}^{-1} \Gamma_{jk} \| \Big\}. \qquad (3.7.6)$$

Обозначим

$$\sigma_n = \max_{j \leqslant n} \| A_{jj}^{-1} \delta^{(j)} \| + C_1 \max_{j \leqslant n} \sum_{k=0}^{j} \| A_{jj}^{-1} \Gamma_{jk} \|. \qquad (3.7.7)$$

Очевидно, что с возрастанием n величина σ_n не убывает. Возьмем p_3 столь малым, чтобы $\beta = C_0 p_3 < 1$. Тогда

$$\hat{\xi}_n \leqslant \beta \max_{j \leqslant n} \hat{\xi}_j + \sigma_n. \qquad (3.7.8)$$

Докажем, что

$$\forall n \geqslant 0, \ \hat{\xi}_n \leqslant \sigma_n/(1-\beta). \qquad (3.7.9)$$

Неравенство (3.7.9) верно при $n = 0$ — в этом случае соотношение (3.7.8) принимает вид $\xi_0 \leqslant \beta \xi_0 + \sigma_0$. Пусть неравенство (3.7.9) верно для $n \leqslant N-1$; докажем, что оно верно и для $n = N$. Может случиться, что

$$\max_{j \leqslant N} \hat{\xi}_j = \xi_N; \qquad (3.7.10)$$

тогда по (3.7.8) $\xi_N \leqslant \beta \xi_N + \sigma_N$ и неравенство (3.7.9) доказано. Если же неравенство (3.7.10) неверно, то $\max\limits_{j \leqslant N} \hat{\xi}_j = \hat{\xi}_k$, где k — одно из чисел 0,1, ...,$N-1$, а тогда

$$\hat{\xi}_N < \hat{\xi}_k \leqslant \sigma_k/(1-\beta) \leqslant \sigma_N/(1-\beta),$$

что и требовалось доказать. Ясно, что неравенство (3.7.2) справедливо при значениях постоянных

$$p_1 = C/(1-\beta), \ p_2 = 1/(1-\beta), \ p_3 = \beta/C_0, \ \forall \beta \in (0, 1). \qquad (3.7.11)$$

2°. Нетрудно указать простое условие, при котором неравенство (3.7.4) справедливо: достаточно, чтобы

$$\sum_{k=1}^{n-1} \| A_{nn}^{-1} A_{nk} \| \leqslant q = \text{const} < 1. \qquad (3.7.12)$$

Действительно, из уравнения (3.7.3) следует, что

$$\| t^{(n)} \| \leqslant q \max_{k \leqslant n-1} \| t^{(k)} \| + \| A_{nn}^{-1} h^{(n)} \|. \qquad (3.7.13)$$

Пусть $\max\limits_{k \leqslant n} \| t^{(k)} \| = \| t^{(l)} \|$, $l \leqslant n$. Если $l = n$, то из (3.7,13) находим $\| t^{(n)} \| \leqslant q \| t^{(n)} \| + \| A_{nn}^{-1} h^{(n)} \|$, или $\| t^{(n)} \| \leqslant (1-q) \| A_{nn}^{-1} h^{(l)} \|$. Если

же $l < n$, то, заменив в (3.7.13) n на l, получим $\|t^{(l)}\| \leqslant q\|t^{(l)}\| + \|A_{ll}^{-1}h^{(l)}\|$ и, следовательно,

$$\|t^{(n)}\| < \|t^{(l)}\| \leqslant (1-q)^{-1}\|A_{ll}^{-1}h^{(l)}\| \leqslant (1-q)^{-1}\max_{j\leqslant n}\|A_{jj}^{-1}h^{(j)}\|.$$

3°. **Теорема 3.7.2.** *Рекуррентный процесс* (3.6.1) *устойчив, если выполнены условия* (3.7.5) *и* (3.7.12).

Доказательство следует из неравенства (3.6.6). Допуская, что

$$\sum_{k=0}^{n}\|A_{nn}^{-1}\Gamma_{nk}\| \leqslant \beta,\quad 0 \leqslant \beta < 1,$$

получаем из (3.6.6)

$$\hat{\xi}_n \leqslant (1-\beta)^{-1}\Big[C_1\sum_{k=0}^{n}\|A_{nn}^{-1}\Gamma_{nk}\| + \|A_{nn}^{-1}\delta^{(n)}\| + (q+\beta)\max_{k\leqslant n}\hat{\xi}_k\Big].$$

Выберем β столь малым, чтобы $q_1 := (q+\beta)(1-\beta)^{-1} < 1$. Имеем

$$\hat{\xi}_n \leqslant q_1\max_{k\leqslant n}\hat{\xi}_k + (1-\beta)^{-1}\Big[C_1\sum_{k=0}^{n}\|A_{nn}^{-1}\Gamma_{nk}\| + \|A_{nn}^{-1}\delta^{(n)}\|. \tag{3.7.14}$$

Уравнения (3.7.14) и (3.7.8) совпадают с точностью до обозначений, и аналогично формуле (3.7.9) получаем

$$\hat{\xi}_n \leqslant p_1\sum_{k=0}^{n}\|A_{nn}^{-1}\Gamma_{nk}\| + p_2\|A_{nn}^{-1}\delta^{(n)}\|. \tag{3.7.15}$$

Замечание. Из замечания п. 1° ясно, что теорема 3.7.2 есть частный случай теоремы 3.7.1. Мы выделили этот частный случай, потому что оценка (3.7.15) несколько лучше по порядку, чем это требуется формулой (3.7.2), входящей в определение устойчивости.

§ 8. ПОГРЕШНОСТЬ ИСКАЖЕНИЯ И УСТОЙЧИВОСТЬ МЕТОДА НАИСКОРЕЙШЕГО СПУСКА

1°. Результаты § 6, 7 применим к методу наискорейшего спуска, предложенному в свое время Дж. Темплем [200] и независимо от него Л. В. Канторовичем [62] (см. также [64]), который нашел также оценку скорости сходимости метода (см. ниже, формула (3.8.3)); в статье Темпля установлен только факт сходимости. Коротко напомним основные черты рассматриваемого метода.

Пусть A — ограниченный самосопряженный положительно определенный оператор, действующий в некотором гильбертовом пространстве H, и пусть требуется решить уравнение

$$Ax = f. \tag{3.8.1}$$

Зададим произвольный элемент $x^{(0)} \in H$, который будем рассматривать как нулевое приближение к искомому решению x_*

уравнения (3.8.1). Пусть построено $(n-1)$-е приближение $x^{(n-1)}$; тогда следующее приближение $x^{(n)}$ строится по формулам

$$y^{(n-1)} = Ax^{(n-1)} - f, \quad \sigma_{n-1} = \frac{\| y^{(n-1)} \|^2}{(Ay^{(n-1)}, y^{(n-1)})}, \tag{3.8.2}$$

$$x^{(n)} = x^{(n-1)} - \sigma_{n-1} y^{(n-1)}.$$

Как показано в [62, 64], $x^{(n)} \xrightarrow[n\to\infty]{} x_*$ с оценкой

$$\| x_* - x^{(n)} \| \leqslant \frac{\| y_1 \|}{\mu} \left(\frac{M-\mu}{M+\mu} \right)^n, \tag{3.8.3}$$

где M и μ суть соответственно верхняя и нижняя грани оператора A:

$$M = \sup_{\| x \|=1} (Ax, x), \quad \mu = \inf_{\| x \|=1} (Ax, x).$$

Таким образом, метод наискорейшего спуска связан с рекуррентным процессом

$$x^{(n)} - B_n x^{(n-1)} = f^{(n)}, \tag{3.8.4.}$$

где

$$B_n = I - \sigma_{n-1} A, \quad f^{(n)} = \sigma_{n-1} f. \tag{3.8.5.}$$

2°. Исследуем погрешность искажения и устойчивость процесса (3.8.4). Очевидно, что $M^{-1} \leqslant \sigma_{n-1} \leqslant \mu^{-1}$; верхняя и нижняя грани оператора B_n суть, следовательно, $1 - \sigma_{n-1}\mu$ и $1 - \sigma_{n-1}M$. Далее,

$$1 - \sigma_{n-1} M \geqslant -\frac{M-\mu}{\mu}, \quad 1 - \sigma_{n-1}\mu \leqslant \frac{M-\mu}{M}.$$

Отсюда

$$\| B_n \| \leqslant \max \left(\frac{M-\mu}{\mu}, \frac{M-\mu}{M} \right) = \frac{M-\mu}{\mu}. \tag{3.8.6}$$

В обозначениях § 6 здесь $A_{nn} = I$, $A_{n,\,n-1} = -B$, $A_{nk} = 0$, $k \leqslant n-2$. Обозначим через ϑ_n погрешность вычисления величины σ_{n-1}, тогда $\Gamma_{n,\,n-1} = \vartheta_n A$ и $\Gamma_{nk} = 0$ при $k \neq n-1$; кроме того, $\delta^{(n)} = \vartheta_n f$.

Обратимся к формуле (3.6.6). Условие $\| A_{nn}^{-1} \Gamma_{nn} \| \leqslant \beta$, $0 \leqslant \beta < 1$, здесь, очевидно, выполняется, потому что $\Gamma_{nn} = 0$. Можно считать, что $\beta = 0$, и формула (3.6.6) принимает вид

$$\hat{\xi}_{n-1} \leqslant ((M-\mu)/\mu + |\, \upsilon_n \,| M) \hat{\xi}_{n-1} + |\, \upsilon_n \,| (C_0 M + \| f \|). \tag{3.8.7}$$

Здесь C_0 — какая-нибудь верхняя граница значений $x_*^{(n)}$; такая конечная граница существует, потому что процесс наискорейшего спуска сходится. Эту границу нетрудно вычислить, исходя из неравенства (3.8.3).

Проанализируем формулу (3.8.7). Пусть погрешность ϑ_n достаточно мала, $|\vartheta_n| < \varepsilon$. Если при этом $M < 2\mu$, то

$$(M-\mu)/\mu + |v_n| M \leqslant (M-\mu)/\mu + \varepsilon M := q,$$

и при достаточно малом ε будет $q < 1$. Теперь

$$\hat{\xi}_n \leqslant q\hat{\xi}_{n-1} + \varepsilon a, \quad a = C_0 M + \|f\|. \tag{3.8.8}$$

Заметим еще, что

$$\hat{\xi}_0 = 0, \tag{3.8.9}$$

потому что элемент $x^{(0)}$ не вычисляется, а задается, и естественно считать, что он задан точно. Очевидно, что $\hat{\xi}_n \leqslant \tau_n$, где τ_n — решение разностного уравнения

$$\tau_n = q\tau_{n-1} + \varepsilon a, \quad \tau_0 = 0. \tag{3.8.10}$$

Решение этого уравнения есть $\tau_n = \varepsilon a (1-q)^{-1}(1-q^n)$, и оценка погрешности имеет вид

$$\hat{\xi}_n \leqslant \frac{\varepsilon a}{1-q}(1-q^n) < \frac{\varepsilon a}{1-q}. \tag{3.8.11}$$

Таким образом, если: 1) $M < 2\mu$ и 2) ε — верхняя граница погрешностей величин σ_{n-1} — достаточно мала, то погрешность искажения метода наискорейшего спуска есть величина порядка $O(\varepsilon)$, и эта оценка — равномерная относительно n.

Если хотя бы одно из условий 1), 2) нарушено, то может случиться, что погрешности искажения будут неограниченно возрастать вместе с n.

3°. Близкие результаты дает и анализ устойчивости. Условие (3.7.5) очевидно выполнено, а условие (3.7.12) в данном случае означает, что $\|B_n\| \leqslant q < 1$; это условие выполнено, если $M < 2\mu$, — тогда можно положить $q = (M-\mu)/\mu$. По теореме 3.7.2 процесс наискорейшего спуска устойчив, если $M < 2\mu$. Тот же результат дает и теорема 3.7.1.

4°. Если $M \geqslant 2\mu$, то для оценки погрешности искажения можно воспользоваться непосредственно формулой (3.8.7). Допустим, что величина ϑ_n ограничена; тогда $(M-\mu)/(\mu + \vartheta_n M) \geqslant r = \text{const}$, и

$$\hat{\xi}_n \leqslant r\hat{\xi}_{n-1} + |v_n| a, \quad \hat{\xi}_0 = 0. \tag{3.8.12}$$

Решение этого неравенства есть

$$\hat{\xi}_n \leqslant a \sum_{k=1}^{n} |v_k| r^{n-k}. \tag{3.8.13}$$

Зафиксируем натуральное число N. Если желательно, чтобы было $\xi_N \leqslant \xi$, где ξ — заданное малое число, то достаточно вычислять σ_{k-1}, $1 \leqslant k \leqslant N$, с такой точностью, чтобы

$$|v_k| \leqslant \xi/(Nar^{N-k}). \tag{3.8.14}$$

§ 9. ОБ УСТОЙЧИВОСТИ МЕТОДА КОЛЛОКАЦИИ

1°. Неискаженная алгебраическая система метода коллокации имеет вид

$$M_n a^{(n)} = f^{(n)}, \tag{3.9.1}$$

где M_n — матрица чисел $\psi_{kn}(t_j^{(n)}) = A\varphi_{kn}(t_j^{(n)})$, $a^{(n)} = (a_1^{(n)}, a_2^{(n)}, \ldots, a_N^{(n)})$, $f^{(n)} = (f(t_1^{(n)}), f(t_2^{(n)}), \ldots, f(t_N^{(n)}))$.

Пусть K — компакт в m-мерном евклидовом пространстве R_m. Точки $t_j^{(n)}$ — узлы коллокации — выберем в вершинах некоторой параллелепипедальной сетки, лежащих в K. Ребра сетки, параллельные k-й координатной оси, пусть имеют длину $h_k^{(n)}$, $k = 1, 2, \ldots, m$; допустим, что $c_1 h_n \leqslant h_k^{(n)} \leqslant c_2 h_n$, где c_1 и c_2 — постоянные, и $h_n \xrightarrow[n\to\infty]{} 0$.

Будем рассматривать $a^{(n)}$ как элемент пространства R_N, а $f^{(n)}$ — как элемент пространства Y_{ns}, также N-мерного, в котором норма определяется так:

$$\|\cdot\|_{Y_{ns}} = h^{s/2}\|\cdot\|_{R_N} \tag{3.9.2}$$

s — вещественная постоянная. Оператор, порожденный матрицей M_n и действующий из R_N в Y_{ns}, обозначим через A_n; систему (3.9.1) можно записать в операторной форме:

$$A_n a^{(n)} = f^{(n)}; \tag{3.9.3}$$

соответствующая искаженная система имеет вид

$$(A_n + \Gamma_n) b^{(n)} = f^{(n)} + \delta^{(n)}. \tag{3.9.4}$$

Оператор, порожденный матрицей M_n и действующий в R_N, обозначим тем же символом M_n.

2°. Пусть $s_n^{(k)}$ — собственные числа неотрицательной матрицы $M_n^* M_n$. Мы будем называть числа $s_n^{(k)}$ сингулярными числами матрицы M_n, хотя такое словоупотребление и не является общепринятым; в частности, в [39] сингулярными числами матрицы называются величины $\sqrt{s_n^{(k)}}$. Расположим числа $s_n^{(k)}$ в порядке неубывания: $s_n^{(1)} \leqslant s_n^{(2)} \leqslant \ldots s_n^{(N)}$.

Теорема 3.9.1. *Если*

$$s_n^{(1)} \geqslant C_0 h_n^{-s}, \quad C_0 = \text{const} > 0, \tag{3.9.5}$$

и $\|a^{(n)}\| \leqslant C_1 = \text{const}$, *то вычислительный процесс метода коллокации* (3.9.3) *устойчив в смысле второго определения* (§ 2, *глава* 3) *в последовательности пар пространств* (R_N, Y_{Ns}). *Если же*

$$s_n^{(1)} \leqslant \gamma(h_n) h_n^{-s}, \quad \gamma(t) \xrightarrow[t\to 0]{} 0, \tag{3.9.6}$$

то в той же последовательности процесс (3.9.3) *неустойчив.*

По следствию 3.2.2 для устойчивости процесса (3.9.3) в упомянутом смысле необходимо и достаточно, чтобы: 1) $\|A_n^{-1}\| \leqslant C_1$, 2) $\|A_n^{-1}\| \cdot \|a_n\| \leqslant C_2$, где C_1 и C_2 — постоянные. Оценим величину

$$\|A_n^{-1}\| = \sup_{g \in Y_{Ns}} \frac{\|A_n^{-1} g\|_{R_N}}{\|g\|_{Y_{Ns}}} = h_n^{-s/2} \sup_{g \in R_N} \frac{\|M_n^{-1} g\|_{R_N}}{\|g\|_{R_N}} = h_n^{-s/2} \|M_n^{-1}\|.$$

Из соотношения $M_n^{-1}(M_n^*)^{-1} = (M_n^* M_n)^{-1}$ следует, что сингулярные числа матрицы M_n^{-1} суть $1/s_n^{(N)}$, $1/s_n^{(N-1)}$, ..., $1/s_n^{(1)}$. Отсюда

$$\|M_n^{-1}\| = \|(M_n^* M_n)^{-1}\|^{1/2} = 1/\sqrt{s_n^{(1)}}$$

и окончательно

$$\|A_n^{-1}\| = (h_n^s s_n^{(1)})^{-1/2}. \tag{3.9.7}$$

Если имеет место неравенство (3.9.6), то $\|A_n^{-1}\| \geqslant [\gamma(h_n)]^{-1/2} \xrightarrow[n \to \infty]{} \infty$, и процесс (3.9.3) неустойчив.

Пусть выполнено неравенство (3.9.5). Тогда $\|A_n^{-1}\| \leqslant C_0^{-1/2}$, и условие 1) выполнено. Условие 2) выполнено по предположению, и процесс коллокации устойчив.

Отметим частный случай. Прежде всего заметим следующее: так как $f \in C(K)$, то интеграл $\int_K |f^2(t)|\,dt$ является пределом своих римановых интегральных сумм, а тогда

$$h_n^m \sum_{j=1}^N |(t_j^{(n)})|^2 \leqslant C = \text{const},$$

или, что то же, $\|f^{(n)}\|_{Y_{Nm}} \leqslant C$. Пусть теперь неравенство (3.9.5) выполнено при $s = m$; тогда норма $\|A_n^{-1}\|$ ограничена постоянной $C^{-1/2}$. Как было только что отмечено, нормы $\|f^{(n)}\|_{Y_{Nm}}$ также ограничены, а тогда ограничены и нормы $\|a^{(n)}\|_{R_N} \leqslant \|A_n^{-1}\| \times \|f^{(n)}\|_{Y_{Nm}}$. По следствию 3.2.2 процесс (3.9.3) устойчив. Таким образом, при $s = m$ условие (3.9.5) необходимо и достаточно для устойчивости процесса коллокации (3.9.3).

3°. Выше (§ 3, глава 3) было дано определение сильной минимальности последовательности систем

$$(\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nN}); \quad N = N(n), \quad n = 1, 2, \ldots \tag{3.9.8}$$

Введем еще одно определение: последовательность систем (3.9.8) назовем *почти ортонормированной*, если собственные числа матриц Грама этих систем положительно ограничены сверху и снизу. Очевидно, что почти ортонормированная последовательность также и сильно минимальна.

Теорема 3.9.2. *Пусть X — сепарабельное гильбертово пространство, последовательность* (3.9.8) *почти ортонормирована в X, а процесс* (3.9.3) *устойчив в смысле второго определения* (§ 2 *главы* 3). *Тогда процесс вычисления приближенного решения по методу коллокации устойчив в смысле того же определения в последовательности пар пространств (X_N, Y_{Ns}). Если последовательность* (3.9.8) *сильно минимальна в X и процесс* (3.9.3) *неустойчив, то процесс вычисления приближенного решения также неустойчив.*

Если последовательность (3.9.8) почти ортонормирована в X, то

$$\|x_*^{(n)} - z_*^{(n)}\|_X = (M_n(a^{(n)} - b^{(n)}),\ a^{(n)} - b^{(n)}) \leqslant \Lambda_0 \|a^{(n)} - b^{(n)}\|^2_{R_N}.$$

Здесь M_n — матрица Грама элементов (3.9.8) при фиксированном n, Λ_0 — верхняя граница собственных чисел этих матриц при всех n, $x_*^{(n)}$ и $z_*^{(n)}$ соответственно неискаженное и искаженное приближенные решения, с которыми векторы $a^{(n)}$ и $b^{(n)}$ связаны соотношениями

$$x_*^{(n)} = \sum_{k=1}^{N} a_k^{(n)} \varphi_{nk},\quad z_*^{(n)} = \sum_{k=1}^{N} b_k^{(n)} \varphi_{nk}.$$

Процесс вычисления коэффициентов по предположению устойчив, поэтому существуют такие постоянные $p, q, r > 0$, что при $\|\Gamma_n\| \leqslant r$ будет

$$\|x_*^{(n)} - z_*^{(n)}\| \leqslant \sqrt{\Lambda_0}\left(p\|\Gamma_n\|_{R_N \to Y_{Ns}} + q\|\delta^{(n)}\|_{Y_{Ns}}\right)$$

последнее неравенство означает, что процесс вычисления приближенного решения устойчив.

Если последовательность (3.9.3) сильно минимальна в X, то, обозначая положительную нижнюю границу собственных чисел матрицы M_n через λ_0, имеем

$$\|x_*^{(n)} - z_*^{(n)}\| \geqslant \sqrt{\lambda_0}\,\|a^{(n)} - b^{(n)}\|_{R_N}.$$

По предположению процесс (3.9.3) неустойчив, поэтому при любых, фиксированных независимо от n нормах $\|\Gamma_n\|$ и $\|\delta^{(n)}\|$ нормы $\|a^{(n)} - b^{(n)}\|$ неограничены, а тогда неограничены и нормы $\|x_*^{(n)} - z_*^{(n)}\|$, и процесс вычисления приближенных решений неустойчив.

4°. Некоторые оценки наименьшего сингулярного числа $s_n^{(1)}$ можно получить, исследуя определитель $D_n := \det M_n$. Ограничимся случаем $m = 1$, $K = [0, 1]$, и пусть узлы $t_j^{(n)} = j/n$, $j = 1, 2, \ldots, n$. Допустим, что функции $\psi_{nk}(t) = A\varphi_{nk}(t)$ удовлетворяют условию Липшица с положительным показателем γ и с постоянной, которая оценивается некоторой степенью k:

$$|\psi_{nk}(t') - \psi_{nk}(t'')| \leqslant Ck^{\beta}|t' - t''|^{\gamma},\quad \beta = \text{const} \geqslant 0; \quad (3.9.9)$$

предположим еще, что

$$|\psi_{nk}(t)| \leqslant Ck^{\alpha}, \quad \alpha = \text{const} \geqslant 0. \tag{3.9.10}$$

Имеем (для упрощения обозначений пишем ψ_k вместо ψ_{nk} и t_j вместо $t_j^{(n)}$)

$$D_n = \begin{vmatrix} \psi_1(t_1) & \psi_2(t_1) & \dots & \psi_n(t_1) \\ \psi_1(t_2) & \psi_2(t_2) & \dots & \psi_n(t_2) \\ \dots & \dots & \dots & \dots \\ \psi_1(t_n) & \psi_2(t_n) & \dots & \psi_n(t_n) \end{vmatrix}. \tag{3.9.11}$$

Из первой строки определителя D_n вычтем вторую, из второй — третью и т. д.; из $(n-1)$-й строки вычтем n-ю. Элементы первых $n-1$ строк примут вид $\psi_k(t_j) - \psi_k(t_{j+1})$ и оценятся по неравенству (3.9.9): $|\psi_k(t_j) - \psi_k(t_{j+1})| \leqslant Cn^{-\gamma}k^{\beta}$; элементы последней строки получат оценку Ck^{α}. По известной теореме Адамара (см., например, [130]) получаем

$$D_n \leqslant \frac{C^n}{n^{\gamma(n-1)}} \left[\sum_{k=1}^{n} k^{2\alpha}\right]^{1/2} \left[\sum_{k=1}^{n} k^{2\beta}\right]^{(n-1)/2}.$$

Так как $\alpha \geqslant 0$, то $k^{2\alpha} \leqslant \int_k^{k+1} u^{2\alpha}\,du$ и

$$\sum_{k=1}^{n} k^{2\alpha} \leqslant \int_1^{n+1} u^{2\alpha}\,du < (2\alpha+1)^{-1}(n+1)^{2\alpha+1};$$

аналогично

$$\sum_{k=1}^{n} k^{2\beta} \leqslant (2\beta+1)^{-1}(n+1)^{2\beta+1}.$$

Теперь

$$|D_n| \leqslant \frac{C_1^n}{n^{\gamma(n-1)}}(n+1)^{\frac{2\alpha+1}{2}+\frac{(2\beta+1)(n-1)}{2}}, \quad C_1 = \text{const},$$

откуда легко получается и несколько более простая оценка

$$|D_n| \leqslant C_2^n n^{(\beta+1/2-\gamma)(n-1)+\alpha+1/2}, \quad C_2 = \text{const}. \tag{3.9.12}$$

Далее, $D_n^2 = \det(M_n^* M_n) = s_n^{(1)} s_n^{(2)} \dots s_n^{(n)} \geqslant [s_n^{(1)}]^n$, отсюда

$$s_n^{(1)} \leqslant D_n^{2/n} \leqslant C_2^2 n^{(2\beta+1-2\gamma)(1-1/n)+(2\alpha+1)/n} \leqslant C_3^2 n^{2(\beta-\gamma)+1}, \quad C_3 = \text{const}. \tag{3.9.13}$$

Теперь из теоремы 3.9.1 вытекает, что если $m = s = 1$ и $\beta > \gamma$, то процесс (3.9.3) неустойчив. В частности, если $\gamma = 1$, то процесс (3.9.3) неустойчив при $\beta > 1$.

5°. **Пример.** Пусть A — тождественный оператор. Тогда задача состоит в интерполяции функции $f(t)$ линейной комбинацией

$$\sum_{k=1}^{N} a_k^n \varphi_{nk}(t)$$

на заданном компакте K. Пусть $m=1$, $K=[0,1]$ и $\varphi_{nk}(t)=\varphi_k(t)=t^{k-1}$, $k=1, 2, \ldots$ Положим еще $t_j=t_j^{(n)}=j/n$, $j=0, 1, 2, \ldots, n-1$. Определитель коллокационной системы

$$D_n=\begin{vmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & t_1 & t_2 & \ldots & t_{n-1} \\ 1 & t_1^2 & t_2^2 & \ldots & t_{n-1}^2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & t_1^{n-1} & t_2^{n-1} & \ldots & t_{n-1}^{n-1} \end{vmatrix}=t_1t_2\ldots t_{n-1}\begin{vmatrix} 1 & 1 & \ldots & 1 \\ t_1 & t_2 & \ldots & t_{n-1} \\ \ldots & \ldots & \ldots & \ldots \\ t_1^{n-2} & t_2^{n-2} & \ldots & t_{n-1}^{n-2} \end{vmatrix}=$$

$$=\frac{(n-1)!}{n^{n-1}}\prod_{1\leqslant k<l\leqslant n}(t_l-t_k)<1.$$

В данном случае $s_1^{(n)}<1=hh^{-1}$, $h=1/n$, $\gamma(h)=h\to 0$ и коллокационный процесс неустойчив в последовательности пар пространств (R_n, Y_{n1}).

6°. В заключение параграфа коротко изложим содержание работы [11], в которой, по-видимому, впервые рассматривался вопрос об устойчивости метода коллокации. В этой работе рассмотрено обыкновенное дифференциальное уравнение

$$x^{(m)}(t)+\sum_{k=1}^{m-1}c_k(t)\,x^{(k)}(t)=f(t),\quad a<t<b, \tag{3.9.14}$$

при краевых условиях

$$\sum_{k=0}^{m-1}[\alpha_{jk}x^{(k)}(a)+\beta_{jk}x^{(k)}(b)]=0,\quad j=0, 1, \ldots, 2m-1;$$
$$\alpha_{jk},\ \beta_{jk}=\text{const}. \tag{3.9.15}$$

Предположим, что соответствующая однородная задача имеет только тривиальное решение. Тогда существует последовательность полиномов $P_k(t)$ степени $m+k$,

$$P_k(t)=\sum_{l=0}^{m+k}c_{kl}t^l,\quad k=0, 1, 2, \ldots,$$

удовлетворяющая условиям (3.9.15). Эту последовательность примем за координатную систему метода коллокации. За узлы коллокации при данном n примем корни $t_j^{(n)}$ полинома $\omega_n(t)$, где $\{\omega_n(t)\}$ — система полиномов, ортонормированная и полная в метрике $L_2(a, b; \rho(t))$, причем вес $\rho(t)$ есть функция, неотрицательная и суммируемая на (a, b), такая, что $\int_a^b[\rho(t)]^{-1}\,dt<\infty$.

Таким образом, приближенное решение задачи ищется как функция вида

$$x_n(t)=\sum_{k=1}^{n-1}a_k^{(n)}P_k(t); \tag{3.9.16}$$

коэффициенты $a_k^{(n)}$ определяются из системы уравнений

$$\sum_{k=0}^{n}\left[P_k^{(m)}(t_j^{(n)})+\sum_{i=0}^{m-1}c_i(t_j^{(n)})P_k^{(i)}(t_j^{(n)})\right]a_k^{(n)}=$$
$$=f(t_j^{(n)}),\ j=1,\ 2,\ \ldots,\ n. \qquad (3.9.17)$$

В предположении, что коэффициенты $c_k(t)$ и свободный член $f(t)$ непрерывны, доказывается, что система (3.9.17) однозначно разрешима при достаточно больших n; приближенные решения $x_n(t)$ сходятся к точному решению $x_*(t)$ в метрике $L_2(a, b; \rho(t))$. Погрешность аппроксимации имеет оценку

$$\|x_*^{(m)}-x_n^{(m)}\|_{L_{(a,\,b;\,\rho)}}\leqslant C\mathscr{E}_n(x_*^{(m)}), \qquad (3.9.18)$$

где $\mathscr{E}_n(u)$ есть наилучшее приближение функции $u(t)$ полиномами степени не выше n в метрике $L_2(a, b; \rho)$.

Обозначим через H_n подпространство пространства $L_2(a, b; \rho)$, образованное полиномами вида (3.9.16) с произвольными коэффициентами $a_k^{(n)}$. Доказывается, что процесс вычисления приближенных решений $x_n(t)$ устойчив в последовательности пар пространств (H_n, R_n) тогда и только тогда, когда последовательность $\left\{\frac{d^m P_k}{dt^m}\right\}$ сильно минимальна в $L_2(a, b; \rho)$.

§ 10. О ПОГРЕШНОСТЯХ СОСТАВЛЕНИЯ СИСТЕМ РИТЦА И БУБНОВА — ГАЛЕРКИНА

1°. В § 1 данной главы получена оценка искажения (формула (3.1.7)) для свободного вычислительного процесса; чтобы воспользоваться этой формулой, необходимо иметь оценки норм операторов Γ_n — погрешностей операторов A_n и элементов $\delta^{(n)}$ — погрешностей правых частей уравнений (3.1.1). В настоящем параграфе будут даны некоторые такого рода оценки для случая, когда вычислительный процесс представляет собой последовательность систем Ритца или Бубнова — Галеркина для линейного дифференциального уравнения, обыкновенного или в частных производных. Метод Ритца можно рассматривать как частный случай метода Бубнова — Галеркина, когда данная задача — положительно определенная, поэтому ниже в данном параграфе мы будем говорить о методе Бубнова — Галеркина.

2°. Рассмотрим линейное эллиптическое дифференциальное уравнение в области $\Omega \in R_m$, $m \geqslant 1$,

$$\sum_{|\alpha|=0}^{2s} A_\alpha(t)\,D^\alpha x(t)=f(t). \qquad (3.10.1)$$

Для простоты ограничимся первой краевой задачей:

$$\forall t \in \partial\Omega;\ \forall\alpha,\ 0\leqslant|\alpha|\leqslant s-1,\ D^\alpha x(t)=0. \qquad (3.10.2)$$

Примем, что оператор задачи (3.10.1), (3.10.2) можно представить в виде $A_0 + K$, где A_0 положительно определен в $L_2(\Omega)$, и $A_0^{-1}K$ вполне непрерывен в энергетическом пространстве H_0 оператора A_0. В таком случае эту задачу можно решать приближенно по методу Бубнова — Галеркина. Выберем в H_0 полную систему подпространств $\{H_n\}$; пусть $N = N(n)$ есть размерность H_n и пусть

$$(\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nN}) \tag{3.10.3}$$

— базис в H_n. Система Бубнова — Галеркина имеет вид

$$A_n a^{(n)} = f^{(n)}, \tag{3.10.4}$$

где $a^{(n)}$ и $f^{(n)}$ — векторы в R_N; A_n — матрица $\left[\alpha_{jk}^{(n)}\right]_{j,\,k=1}^{N}$, причем

$$\alpha_{jk}^{(n)} = \int_\Omega \varphi_{nj}(t) \sum_{|\alpha|=0}^{2s} A_\alpha(t)\, D^\alpha \varphi_{nk}(t)\, dt \tag{3.10.5}$$

и

$$f_j^{(n)} = (f, \varphi_{jn}) = \int_\Omega f(t)\, \varphi_{nj}(t)\, dt. \tag{3.10.6}$$

Интегрируя (3.10.5) по частям и используя краевые (3.10.2), можно преобразовать интеграл (3.10.5) так, чтобы его подынтегральная функция содержала производные от φ_{nj} и φ_{nk} порядка не выше s.

3°. Пусть A_α, f, φ_{nj} суть полиномы от t; относительно области Ω допустим, что интегралы $\int_\Omega t^\alpha\, dt$ вычисляются точно, каков бы ни был мультииндекс α. Если $m = 1$, то достаточно допустить, что Ω — конечный промежуток; если $m > 1$, то нашему требованию удовлетворяет, например, любой многогранник, а также эллипсоид и некоторые другие области.

Будем считать, что вычисления производятся с повышенной точностью. Тогда по формуле (1.1.9) получаются оценки

$$\left|\delta\left(\alpha_{jk}^{(n)}\right)\right| \leqslant \varepsilon_1 \left|\alpha_{jk}^{(n)}\right|, \quad \left|\delta f_j^{(n)}\right| \leqslant \varepsilon \left|f_j^{(n)}\right|. \tag{3.10.7}$$

Если обозначить через Γ_n и $\delta^{(n)}$ погрешности матрицы A_n и вектора $f^{(n)}$ соответственно, то из (3.10.7) следует

$$\|\delta^{(n)}\| \leqslant \varepsilon_1 \|f^{(n)}\|, \quad \|\Gamma_n\| \leqslant \varepsilon_1 \|A_n\|_E \leqslant \varepsilon_1 \sqrt{N}\, \|A_n\|. \tag{3.10.8}$$

Теперь для рассматриваемой здесь задачи можно придать оценке (3.1.7) более конкретный вид, именно, ее можно выразить через данные задачи. Обозначим через $P_n = \|A_n\| \cdot \|A_n^{-1}\|$ число обусловленности матрицы A_n и заменим через $\|f^{(n)}\| \cdot \|A_n^{-1}\|$ оценку нормы $\|x^{(n)}\|$. Тогда, если

$$\sqrt{N}\, \varepsilon_1 P_n \leqslant \beta, \quad 0 \leqslant \beta < 1, \tag{3.10.9}$$

то

$$\hat{\xi}_n \leqslant \frac{\varepsilon_1 \| A_n^{-1} \| (\sqrt{N} P_n + 1) \| f^{(n)} \|}{1 - \beta}. \qquad (3.10.10)$$

4°. Пусть теперь данные функции не полиномы или область Ω такова, что интеграл $\int_\Omega t^\alpha dt$ не вычисляется точно. На практике достаточно часто встречаются функции, которые выражаются точно через конечное число арифметических действий над независимыми переменными и над несколькими элементарными функциями, такими, как степенная, показательная, логарифмическая, тригонометрические и реже функции Бесселя, эллиптические интегралы и функции и некоторые другие. Имея это в виду, будем считать, что данные функции (коэффициенты и свободный член уравнения, координатные функции и т. п.) таковы, что с помощью специальных подпрограмм их значения могут быть вычислены, например, с точностью $O(\varepsilon_1^2)$. Имея значения подынтегральных функций, применяя квадратурные (или кубатурные) формулы и производя промежуточные вычисления с повышенной точностью, мы получим значения этих величин с погрешностями, которые можно оценить неравенствами (3.10.7), если к их правым частям прибавить оценки погрешности использованных кубатурных формул. После этого нетрудно написать формулы, аналогичные формулам (3.10.9), (3.10.10) и оценивающие нормы $\|\Gamma_n\|$ и $\|\delta^{(n)}\|$.

Глава 4

ПОГРЕШНОСТИ АЛГОРИТМА И ОКРУГЛЕНИЯ

§ 1. ЧИСЛО ОБУСЛОВЛЕННОСТИ

1°. Пусть A — ограниченный и ограниченно обратимый оператор, действующий из банахова пространства X в банахово пространство Y. Числом обусловленности оператора A называется произведение

$$P_A = \| A \| \cdot \| A^{-1} \|. \qquad (4.1.1)$$

Очевидно, что $P_A \geqslant 1$.

Если $X = Y = R_n$, то A есть числовая квадратная матрица порядка n. Если эта матрица положительно определенная, то $\| A \| = \lambda_n^{(n)}$, $\| A^{-1} \| = 1/\lambda_n^{(1)}$, где $\lambda_n^{(1)}$ и $\lambda_n^{(n)}$ суть соответственно наименьшее и наибольшее собственные числа матрицы A; в этом случае число обусловленности равно

$$P_A = \lambda_n^{(n)} / \lambda_n^{(1)}. \qquad (4.1.2)$$

Если A — произвольная неособенная квадратная матрица в R_n, то матрица A^*A положительно определенная. Пусть $s_n^{(n)}$ и $s_n^{(1)}$ — наибольшее и наименьшее собственные числа последней матрицы. Тогда, очевидно,

$$P_A = \sqrt{s_n^{(n)}/s_n^{(1)}}. \tag{4.1.3}$$

2°. Число обусловленности оператора играет большую роль во многих вопросах, касающихся оценок погрешностей. Особо важную роль это понятие играет в оценках погрешности решений линейных алгебраических систем (см., например, [25, 39, 202, 138], а также раздел II настоящей книги). Укажем еще один важный случай. Пусть A — ограниченный и ограниченно обратимый оператор, действующий в сепарабельном гильбертовом пространстве H. Уравнение

$$Ax = f \tag{4.1.4}$$

будем приближенно решать по методу наименьших квадратов: каждому натуральному n сопоставим (вообще говоря, другое) натуральное N и выделим из H подпространство N измерений H_n. Найдем элемент $x_*^{(n)} \in H_n$, такой, чтобы $\| Ax_*^{(n)} - f \| = \min$; существование и единственность такого элемента очевидны. Оценим [118] погрешность аппроксимации ρ_n элемента $x_*^{(n)}$, рассматриваемого как приближенное решение уравнения (4.1.4). Имеем $x_* - x_*^{(n)} = A^{-1}A(x_* - x_*^{(n)}) = A^{-1}(f - Ax_*^{(n)})$ и, следовательно, $\rho \leqslant \| A^{-1} \| \cdot \| f - Ax_*^{(n)} \|$. Если $x^{(n)}$ — произвольный элемент из H_n, то

$$\| f - Ax_*^{(n)} \| \leqslant \| f - Ax^{(n)} \| = \| A(x_* - x^{(n)}) \| \leqslant \| A \| \cdot \| x_* - x^{(n)} \|;$$

$$\rho_n \leqslant P_A \| x_* - x^{(n)} \|.$$

Взяв справа нижнюю грань по всевозможным $x^{(n)} \in H_n$ и обозначив

$$\forall x \in H, \ \mathscr{E}_n(x) = \inf_{x^{(n)} \in H_n} \| x - x^{(n)} \|,$$

получим интересующую нас оценку:

$$\rho_n \leqslant P_A \mathscr{E}_n(x_*). \tag{4.1.5}$$

§ 2. ПОГРЕШНОСТЬ АЛГОРИТМА В ИТЕРАЦИОННОМ ПРОЦЕССЕ

1°. Как хорошо известно, если A — самосопряженный оператор в гильбертовом пространстве H и его верхняя и нижняя грани удовлетворяют неравенствам $0 < m \leqslant M < \infty$, то задачу (4.1.4) можно преобразовать так, чтобы она решалась

итерациями. Достаточно заменить уравнение (4.1.4) таким, ему равносильным:

$$(I - B)x = 2f/(M + m),\ B = I - 2A/(M + m). \qquad (4.2.1)$$

Нижняя и верхняя грани оператора B равны соответственно $\mp(M - m)/(M + m)$, поэтому

$$\|B\| = (M - m)/(M + m) = (P_A - 1)/(P_A + 1) < 1 \qquad (4.2.2)$$

и простые итерации для уравнения (4.2.1) сходятся как геометрическая прогрессия со знаменателем $(P_A - 1)/(P_A + 1)$. Ясно, что сходимость тем быстрее, чем меньше число обусловленности; если же P_A достаточно велико, то сходимость может быть сколь угодно медленной.

2°. Соображения п. 1° применим к классическому методу Ритца. В сепарабельном гильбертовом пространстве H_0 рассмотрим задачу о минимуме функционала $|x|^2 - 2lx$, где $|\cdot|$ — норма в H_0, а l — линейный функционал, определенный и ограниченный на всем пространстве H_0. Выберем в H_0 координатную систему $\{\varphi_n\}$ и обозначим через $H_0^{(n)}$ подпространство пространства H_0 с базисом $(\varphi_1, \varphi_2, \ldots, \varphi_n)$. Приближенное решение нашей задачи построим как элемент подпространства $H_0^{(n)}$, реализующий минимум упомянутого функционала на этом подпространстве. При этом $x_*^{(n)}$ определяется выражением (2.1.3), а коэффициенты $a_k^{(n)}$ определяются из алгебраической системы (2.1.4), которую запишем в виде

$$M_n a^{(n)} = f^{(n)}; \qquad (4.2.3)$$

смысл обозначений очевиден. Обозначим через $\lambda_n^{(1)}$ и $\lambda_n^{(n)}$ соответственно наименьшее и наибольшее собственное число матрицы Ритца M_n и положим $B_n = I_n - 2M_n/(\lambda_n^{(1)} + \lambda_n^{(n)})$. Уравнение (4.2.3) заменим следующим:

$$a^{(n)} - B_n a^{(n)} = 2f^{(n)}/(\lambda_n^{(1)} + \lambda_n^{(n)}), \qquad (4.2.4)$$

которое разрешимо итерациями, так как $\|B_n\| = (P_n - 1)/(P_n + 1) < 1$; $P_n = P_{M_n}$. Если начальное приближение принять равным нулю и если ограничиться N итерациями, то соответствующая погрешность алгоритма для уравнения (4.2.4) имеет оценку

$$\frac{2\|f^{(n)}\|}{\lambda_n^{(1)} + \lambda_n^{(n)}} \frac{\|B\|^{N+1}}{1 - \|B\|} = \frac{\|f^{(n)}\|}{\lambda_n^{(1)}} \|B\|^{N+1}. \qquad (4.2.5)$$

Если $\lambda_n^{(1)} \xrightarrow[n\to\infty]{} 0$ или $\lambda_n^{(n)} \xrightarrow[n\to\infty]{} \infty$, то $\|B_n\| \xrightarrow[n\to\infty]{} 1$, и оценка погрешности алгоритма ухудшается. Поэтому желательно использовать такую координатную систему, для которой $\lambda_0 \leqslant \lambda_n^{(1)} \leqslant \lambda_n^{(n)} \leqslant \Lambda_0$, где λ_0 и Λ_0 — положительные постоянные; иначе говоря, целесообразно в качестве координатной брать почти

ортонормированную в H_0 систему. В этом случае можно определить матрицу B_n проще: можно положить $B_n = I_n - 2M_n/(\Lambda_0 + \lambda_0)$ и вместо (4.2.4) составить уравнение

$$a^{(n)} - B_n a^{(n)} = 2f^{(n)}/(\Lambda_0 + \lambda_0), \tag{4.2.6}$$

также равносильное уравнение (4.2.3). При таком определении нижняя и верхняя грани матрицы B_n соответственно равны

$$1 - \frac{2\lambda_n^{(n)}}{\Lambda_0 + \lambda_0} \geqslant 1 - \frac{2\Lambda_0}{\Lambda_0 + \lambda_0} = -\frac{\Lambda_0 - \lambda_0}{\Lambda_0 + \lambda_0},$$

$$1 - \frac{2\lambda_n^{(1)}}{\Lambda_0 + \lambda_0} \geqslant 1 - \frac{2\lambda_0}{\Lambda_0 + \lambda_0} = \frac{\Lambda_0 - \lambda_0}{\Lambda_0 + \lambda_0}.$$

Отсюда следует, что $\|B_n\| \leqslant (\Lambda_0 - \lambda_0)/(\Lambda_0 + \lambda_0)$; правая часть этого неравенства постоянна и меньше единицы. Для погрешности алгоритма после N итераций получаем оценку

$$\frac{\|f^{(n)}\|}{\lambda_0}\left(\frac{\Lambda_0 - \lambda_0}{\Lambda_0 + \lambda_0}\right)^{N+1}. \tag{4.2.7}$$

Докажем, что $\|f^{(n)}\| \leqslant C$, где C не зависит от n. Почти ортонормированная система сильно минимальна; в [90] доказано, что если координатная система сильно минимальна, то последовательность векторов $a^{(n)}$, дополненных нулями, имеет предел в l_2. Отсюда вытекает, что последовательность $\|a^{(n)}\|_{R_n}$ ограничена. Пусть $\|a^{(n)}\|_{R_n} \leqslant c_0 = \text{const}$. Тогда из уравнения (4.2.3) находим

$$\|f^{(n)}\| \leqslant \|M_n\| \cdot \|a^{(n)}\| \leqslant \lambda_n^{(n)} \|a^{(n)}\| \leqslant \Lambda_0 c_0,$$

и из (4.2.7) следует оценка погрешности алгоритма, не зависящая от n:

$$\frac{\Lambda_0 c_0}{\lambda_0}\left(\frac{\Lambda_0 - \lambda_0}{\Lambda_0 + \lambda_0}\right)^{N+1}. \tag{4.2.8}$$

3°. Соображения и результаты настоящего параграфа очевидным образом переносятся на обобщенный метод Ритца.

§ 3. ПОГРЕШНОСТЬ ОКРУГЛЕНИЯ В ИТЕРАЦИОННОМ ПРОЦЕССЕ

1°. Обратимся к системе (3.6.2); будем считать, что она может быть как конечной, так и бесконечной. Для дальнейшего важно, что операторы $A_{nk} + \Gamma_{nk}$ и свободные члены $f^{(n)} + \delta^{(n)}$ известны точно. Для краткости обозначим

$$A_{nk} + \Gamma_{nk} = \tilde{A}_{nk}, \quad f^{(n)} + \delta^{(n)} = \tilde{f}^{(n)}; \tag{4.3.1}$$

тогда система (3.6.2) запишется в виде

$$\forall n \geqslant 0, \quad \sum_{k=0}^{n} \tilde{A}_{nk} z^{(n)} = \tilde{f}^{(n)}. \tag{4.3.2}$$

Нетрудно видеть, что если нормы $\| \tilde{A}_{nn}^{-1}\Gamma_{nk}\|$ достаточно малы, то операторы $\tilde{A}_{nn}$ ограниченно обратимы, а произведения $\tilde{A}_{nn}^{-1}\tilde{A}_{nk}$ ограничены, и что если справедливо неравенство (3.7.4), то справедливо и аналогичное неравенство

$$n \geqslant 0, \ \|z^{(n)}\| \leqslant \tilde{C}_0 \max_{l \leqslant n} \|\tilde{A}_{ll}^{-1} f^{(l)}\|, \ \tilde{C}_0 = \text{const}. \tag{4.3.3}$$

Точное решение системы (4.3.2) есть

$$n \geqslant 0, \ z^{(n)} = -\sum_{k=0}^{n-1} \tilde{A}_{nn}^{-1}\tilde{A}_{nk} z^{(k)} + \tilde{A}_{nn}^{-1}\tilde{f}^{(n)}. \tag{4.3.4}$$

Вычисление правой части равенства (4.3.4) обычно связано с некоторой погрешностью. Пусть $v^{(1)}, v^{(2)}, \ldots, v^{(n-1)}$ — некоторые точно известные элементы и пусть $\bar{z}^{(n)}$ — точное значение выражения

$$-\sum_{k=0}^{n-1} \tilde{A}_{nn}^{-1}\tilde{A}_{nk} v^{(n)} + \tilde{A}_{nn}^{-1}\tilde{f}^{(n)}; \tag{4.3.5}$$

примем, что, вычисляя величину (4.3.5), мы в результате ошибок округления получим не элемент $\bar{z}^{(n)}$, а некоторый другой элемент $v^{(n)} = \bar{z}^{(n)} + \alpha^{(n)}$. Здесь $\alpha^{(n)}$ — результирующая ошибка округления; допустим, что она удовлетворяет неравенству

$$\|\alpha^{(n)}\| \leqslant \varepsilon \|v^{(n)}\|, \tag{4.3.6}$$

где ε — заранее заданное положительное число, которое может быть сколь угодно малым. Следовательно, мы принимаем, что элемент $\bar{z}^{(n)}$ вычисляется с относительной погрешностью, не превосходящей ε. Если величина (4.3.5) вычисляется с повышенной точностью, то можно в соответствии с формулой (1.1.9) принять $\varepsilon = \varepsilon_1$.

2°. Обозначим $-\tilde{A}_{nn}^{-1}\tilde{A}_{nk} = a_{nk}$, $\tilde{A}_{nn}^{-1}\tilde{f}^{(n)} = b^{(n)}$. В этих обозначениях система (4.3.4) и неравенство (4.3.3) примут вид

$$n \geqslant 0, \ z^{(n)} = \sum_{k=0}^{n-1} a_{nk} z^{(k)} + b^{(n)} \tag{4.3.7}$$

и

$$n \geqslant 0, \ \|z^{(n)}\| \leqslant \tilde{C}_0 \max_{l \leqslant n} \|b^{(l)}\|. \tag{4.3.8}$$

Решая систему (4.3.7) и учитывая ошибки округления, получаем последовательность элементов

$$v^{(0)} = z^{(0)} + \alpha^{(0)} := z^{(0)} + \gamma^{(0)},$$

$$v^{(1)} = a_{10}v^{(0)} + b^{(1)} + \alpha^{(1)} = z^{(1)} + a_{10}\alpha^{(0)} + \alpha^{(1)} := z^{(1)} + \gamma^{(1)}$$

и т. д. Общие формулы имеют вид

$$v^{(n)} = \sum_{k=0}^{n-1} a_{nk} v^{(k)} + b^{(n)} + \alpha^{(n)}; \tag{4.3.9}$$

$$v^{(n)} = z^{(n)} + \gamma^{(n)}; \tag{4.3.10}$$

$$\gamma^{(n)} = \sum_{k=0}^{n-1} a_{nk}\gamma^{(k)} + \alpha^{(n)}. \tag{4.3.11}$$

Система (4.3.11) с точностью до обозначений совпадает с системой (4.3.7); из (4.3.8) следует теперь, что

$$\|\gamma^{(n)}\| \leqslant \tilde{C}_0 \max_{j \leqslant n} \|\alpha^{(j)}\| \leqslant \tilde{C}_0 \varepsilon \max_{j \leqslant n} \|v^{(j)}\| \leqslant$$

$$\leqslant \tilde{C}_0 \varepsilon \left[\max_{j \leqslant n} \|z^{(j)}\| + \max_{j \leqslant n} \|\gamma^{(j)}\| \right]. \tag{4.3.12}$$

Допустим, что нормы $\|z^{(n)}\|$ ограничены независимо от n; это будет, например, если выполнены условия теоремы 3.7.1, а погрешности Γ_{nk} и $\delta^{(n)}$ достаточно малы. Пусть $\|z^{(n)}\| \leqslant \tilde{C}_1$, тогда

$$\|\gamma^{(n)}\| \leqslant \tilde{C}_0 \tilde{C}_1 \varepsilon + C_0 \varepsilon \max_{j \leqslant n} \|\gamma^{(j)}\|. \tag{4.3.13}$$

Если $k \leqslant n$, то из (4.3.13) следует

$$\|\gamma^{(k)}\| \leqslant \tilde{C}_0 \tilde{C}_1 \varepsilon + \tilde{C}_0 \varepsilon \max_{j \leqslant k} \|\gamma^{(j)}\| \leqslant \tilde{C}_0 \tilde{C}_1 \varepsilon + \tilde{C}_0 \varepsilon \max_{j \leqslant n} \|\gamma^{(j)}\|.$$

Допустим, что $\max \|\gamma^{(j)}\|$ достигается при $j = l$. Тогда $\|\gamma^{(l)}\| \leqslant \tilde{C}_0 \tilde{C}_1 \varepsilon + \tilde{C}_0 \varepsilon \|\gamma^{(l)}\|$ или $\|\gamma^{(j)}\| \leqslant \tilde{C}_0 \tilde{C}_1 \varepsilon / (1 - \tilde{C}_0 \varepsilon)$ и тем более

$$\|\gamma^{(n)}\| \leqslant \tilde{C}_0 \tilde{C}_1 \varepsilon / (1 - C_0 \varepsilon). \tag{4.3.14}$$

Таким образом, при сделанных выше предположениях ошибки округления ограничены и стремятся к нулю вместе с ε равномерно по n.

3°. Близкие результаты получаются и в том случае, когда вычисления ведутся с заданной абсолютной погрешностью. По-прежнему считаем, что при вычислении элемента (4.3.5) получается элемент $v^{(n)} = z^{(n)} + \alpha^{(n)}$, но на этот раз $\|\alpha^{(n)}\| \leqslant \varepsilon$; число ε по-прежнему считаем сколь угодно малым. Соотношения (4.3.7)—(4.3.11) остаются в силе; цепочка неравенств (4.3.12) упрощается и приводит к нужному результату:

$$\|\gamma^{(n)}\| \leqslant \tilde{C}_0 \max_{j \leqslant n} \|\alpha^{(j)}\| \leqslant \tilde{C}_0 \varepsilon. \tag{4.3.15}$$

§ 4. ПОГРЕШНОСТЬ ОКРУГЛЕНИЯ МЕТОДА НАИСКОРЕЙШЕГО СПУСКА

1°. Как было отмечено в § 8 главы 3, метод наискорейшего спуска сводится к рекуррентному вычислительному процессу; входящие в описание этого процесса операторы B_n и свободные члены $f^{(n)}$ определяются формулами (3.8.5) и (3.8.2).

Погрешность округления в общем случае определяется рекуррентным соотношением (4.3.11). Для определенности допустим, что вычисления производятся с заданной абсолютной погрешностью, так что $\|\alpha^{(n)}\| \leqslant \varepsilon$. Для метода наискорейшего спуска

$$A_{nn} = I; \quad A_{n,n-1} = -B = -(I - \sigma_{n-1} A); \quad A_{nk} = 0, \quad k \leqslant n-2;$$

$$\Gamma_{nk} = 0, \quad k \neq n-1; \quad \Gamma_{n,n-1} = -v_n A$$

(§ 8 главы 3). Отсюда

$$\tilde{A}_{nn} = I;\ \tilde{A}_{n,\,n-1} = -I + (\sigma_{n-1} + \upsilon_n) A;\ \tilde{A}_{nk} = 0,\ k \leqslant n-2;$$
$$a_{n,\,n-1} = I - (\sigma_{n-1} + \upsilon_n) A;\ a_{nk} = 0,\ k \neq n-1.$$

Уравнение (4.3.11) в данном случае имеет вид

$$\forall n \geqslant 1,\ \gamma^{(n)} = [I - (\sigma_{n-1} + \upsilon_n) A]\,\gamma^{(n-1)} + \alpha^{(n)}. \qquad (4.4.1)$$

Оценим величину $\tau_n := \|a_{n,\,n-1}\|$. Допустим, что σ_{n-1} при любом n вычисляется с малой относительной погрешностью ε', так что $|\vartheta_n| \leqslant \varepsilon' \sigma_{n-1}$. Оператор $a_{n,\,n-1}$ — ограниченный самосопряженный; его нижняя и верхняя грани суть соответственно

$$1 - (\sigma_{n-1} + \upsilon_n) M \geqslant 1 - (1/\mu + \varepsilon'/\mu) M = -(M - \mu)/\mu - \varepsilon' M/\mu$$

и

$$1 - (\sigma_{n-1} + \upsilon_n)\,\mu \leqslant 1 - (1/M - \varepsilon'/\mu)\,\mu = (M - \mu)/M + \varepsilon'\mu/M.$$

Отсюда

$$\tau_n \leqslant (M - \mu)/\mu + \varepsilon' M/\mu := \tau. \qquad (4.4.2)$$

Из этого соотношения вытекает неравенство

$$\forall n \geqslant 1,\ \|\gamma^{(n)}\| \leqslant \tau \|\gamma^{(n-1)}\| + \varepsilon'; \qquad (4.4.3)$$

сразу отметим, что $\gamma^{(0)} = 0$, потому что $\gamma^{(0)}$ есть погрешность точно заданного элемента $x^{(0)}$. Решение системы неравенств (4.4.3) есть

$$\forall n \geqslant 1,\ \|\gamma^{(n)}\| \leqslant (\tau^{(n)} - 1)\,\varepsilon'/(\tau - 1). \qquad (4.4.4)$$

Если $M < 2\mu$, то при достаточно малом ε' будет $\tau < 1$ и

$$\gamma^{(n)} < \varepsilon'/(1 - \tau). \qquad (4.4.5)$$

В этом случае погрешность округления ограничена независимо от n и имеет порядок $O(\varepsilon)$. Если же $M \geqslant 2\mu$, то не исключено, что с возрастанием n погрешность округления может неограниченно возрастать. Тот же по существу результат получается, если округление производится с заданной относительной погрешностью.

2°. **Пример.** В пространстве $L_2(0, 1)$ рассмотрим уравнение

$$(at + 1)\,x(t) = 1,\ a = \text{const} > 0, \qquad (4.4.6)$$

с очевидным решением $x = (at + 1)^{-1}$. Оператор умножения на функцию $at + 1$ — положительно определенный и ограниченный, и уравнение (4.4.6) можно решать методом наискорейшего спуска. За начальное приближение возьмем $x_0 = 0$. Вычисления произведем для значений $a = 1/2,\ 1,\ 3,\ 10$. В нашем примере $\mu = 1$, $M = a + 1$ и, следовательно, $(M - \mu)/(M + \mu) = a/(a + 2)$ и $(M - \mu)/\mu = a$. Наименьшее число итераций, достаточное для получения погрешности алгоритма, меньшей,

чем 10^{-4}, было подсчитано по формуле (3.8.3); значения этого числа приведены в табл. 12. В табл. 13 приведены точные и приближенные значения искомой функции. Как и следовало ожидать, приближенные значения вполне удовлетворительны для значения $a = 1/2$, при котором $M < 2\mu$; они оказались удовлетворительными также при $a = 1$, когда $M = 2\mu$. Большие значения a приходят к следующему результату: при не очень большом числе итераций точность приближенных значений функции $x(t)$ повышается, затем эта точность резко падает.

Таблица 12

$a = \frac{M-\mu}{\mu}$	1/2	1	3	10
$\frac{M-\mu}{M+\mu}$	1/5	1/3	1/2	5/6
Число итераций	6	9	19	51

Таблица 13

Узлы	$a=1/2$		$a=1$		$a=3$			$a=10$		
	Точное решение	Результат 6 итераций	Точное решение	Результат 9 итераций	Точное решение	Результат 10 итераций	Результат 15 итераций	Точное решение	Результат 8 итераций	Результат 10 итераций
0,0	1,0000	0,9920	1,0000	0,9966	1,0000	0,9525	—21 609	1,0000	0,9238	3983
0,1	0,9524	0,9484	0,9091	0,9081	0,7692	0,7601	—16 654	0,5000	0,5016	—337,8
0,2	0,9091	0,9075	0,8333	0,8331	0,6250	0,6240	—5187	0,3333	0,3333	23,97
0,3	0,8696	0,8691	0,7692	0,7692	0,5263	0,5264	689,9	0,2500	0,2556	—79,02
0,4	0,8333	0,8333	0,7143	0,7143	0,4545	0,4546	440,7	0,2000	0,2052	—7152
0,5	0,8000	0,8000	0,6667	0,6667	0,4000	0,4000	33,34	0,1667	0,1667	—13,63
0,6	0,7692	0,7692	0,6250	0,6250	0,3571	0,3571	—48,32	0,1429	0,1398	8091
0,7	0,7407	0,7410	0,5882	0,5882	0,3226	0,3226	—1792	0,1250	0,1236	4940
0,8	0,7143	0,7152	0,5556	0,5556	0,2942	0,2942	17 528	0,1111	0,1119	—3715
0,9	0,6897	0,6919	0,5263	0,5267	0,2703	0,2707	135 369	0,1000	0,0995	2581
1,0	0,6667	0,6711	0,5000	0,5010	0,2500	0,2527	455 927	0,0909	0,0930	—14 112

Далее наступает переполнение и вычисление последующих итераций становится невозможным. Эти явления находятся в соответствии с результатами данного параграфа и § 8 главы 3.

Раздел II

ЛИНЕЙНЫЕ АЛГЕБРАИЧЕСКИЕ СИСТЕМЫ

Прикладное значение теории линейных алгебраических систем хорошо известно. Наиболее распространенные в настоящее время методы приближенного решения задач механики и математической физики, такие, как методы Ритца, Бубнова — Галеркина, наименьших квадратов, конечных элементов, сеток, интегральных уравнений и др., будучи применены к линейным задачам, сразу приводят к линейным алгебраическим системам. Если же перечисленные методы применяются к нелинейным задачам, то они приводят к нелинейным системам с конечным числом числовых неизвестных, которые, в свою очередь, могут быть сведены (например, по методу Ньютона — Канторовича) к линейным алгебраическим системам. Естественным поэтому является то большое внимание, которое в последние десятилетия уделяется улучшению старых и разработке новых методов численного решения линейных алгебраических систем и оценке погрешностей этих методов. Первой в мировой литературе книгой, в которой были систематизированы важнейшие методы решения задач, связанных с линейными алгебраическими системами, была монография В. Н. Фаддеевой [141]; в значительно расширенном и переработанном виде эта монография вышла под именами Д. К. Фаддева и В. Н. Фаддеевой [139]. Позднее появились также завоевавшие широкую известность монографии Дж. Х. Уилкинсона [202, 138], В. В. Воеводина [25] и некоторых других авторов, а также большое количество журнальных статей; в этих работах наряду с описанием самих методов решения задач, связанных с линейными алгебраическими системами, содержится много интересных и важных результатов, относящихся к погрешностям этих методов.

Автору кажется, что при всей значительности результатов упомянутых здесь работ проблема погрешностей линейно-алгеб-

раических методов в этих работах не исчерпана до конца. По этой причине автор считает возможным изложить некоторые результаты своих исследований названной здесь проблемы. Эти исследования опираются на данную автором классификацию погрешностей (см. главу 1) и на содержание предшествующих глав данной книги; они частично опубликованы в [114—117]. Самые результаты исследований автора частично известны, частично, по-видимому, являются новыми.

Глава 5

МЕТОД ГАУССА

§ 1. СХЕМА ВЫБОРА НАИБОЛЬШЕГО ЭЛЕМЕНТА. ИСКАЖЕНИЕ МАТРИЦЫ И СТОЛБЦА СВОБОДНЫХ ЧЛЕНОВ

1°. Рассмотрим линейную алгебраическую систему

$$Ax = f \tag{5.1.1}$$

или, в более подробной записи,

$$\sum_{j=1}^{m} a_{ij}x_j = f_i,\ i = 1,\ 2,\ \ldots,\ m,$$

так что $A = [a_{ij}]_{i,\,j=1}^{m}$, $f = (f_1, f_2, \ldots, f_m)'$, $x = (x_1, x_2, \ldots, x_m)'$.

Будем считать, что матрица A — неособенная. Это предположение, а также введенные здесь обозначения мы сохраним на протяжении всего второго раздела.

Решать систему (5.1.1) будем по методу Гаусса, который подробно изложен в [139]. Более определенно мы воспользуемся схемой метода Гаусса, связанной с выбором наибольшего элемента. Этот выбор можно выполнить двояко: можно выбрать наибольший элемент рассматриваемой матрицы и соответственно переставить ее строки и столбцы; можно также выбрать элемент, наибольший в рассматриваемом столбце — при этом придется переставлять только строки. Для последующего это различие не играет роли; можно только заметить, что первый способ приводит, вообще говоря, к меньшим оценкам погрешности.

Указанные здесь перестановки приводят к появлению некоторой новой матрицы $\bar{A}$ и новых векторов $\bar{x}$, $\bar{f}$, связанных уравнением

$$\bar{A}\bar{x} = \bar{f}. \tag{5.1.2}$$

Нетрудно видеть, что в R_m

$$\|\bar{A}\| = \|A\|,\ \|\bar{A}^{-1}\| = \|A^{-1}\|,\ \|\bar{x}\| = \|x\|,\ \|\bar{f}\| = \|f\|. \tag{5.1.3}$$

2°. По схеме выбора наибольшего элемента матрица $\bar{A}$ разлагается в произведение $\bar{A} = L^{-1}U$, где L и U — треугольные

матрицы, соответственно нижняя и верхняя (по терминологии книги [139] — левая и правая). Матрица L есть произведение нижних треугольных матриц L_{ij} вида [35]

$$L_{ij}=\begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & \ddots & & & \\ & & & & 1 & & \\ & & & & \ddots & \ddots & \\ & & & & \alpha_{ij}^{(0)} & & \ddots \\ & & & & \vdots & & & 1 \end{bmatrix}, \quad \begin{array}{l} i \geqslant 2, \\ 1 \leqslant j \leqslant m-1, \\ i > j, \end{array}$$

каждая из которых имеет единицы на главной диагонали и единственный отличный от нуля элемент $\alpha_{ij}^{(0)}$ ниже главной диагонали на пересечении i-й строки и j-го столбца. Существенно, что

$$|\alpha_{ij}^{(0)}| \leqslant 1. \tag{5.1.4}$$

При фиксированном j матрицы L_{ij} коммутируют. Если обозначить через L_j произведение матриц L_{ij}, $i = j+1, j+2, \ldots, m$, то [35]

$$L = L_{m-1}L_{m-2} \ldots L_2L_1. \tag{5.1.5}$$

L_j есть нижняя треугольная матрица с единицами на главной диагонали. Элементы j-го столбца, расположенные под главной диагональю, суть $\alpha_{j+1,j}^{(0)}$, $\alpha_{j+2,j}^{(0)}$, $\ldots$, $\alpha_{mj}^{(0)}$; остальные элементы этой матрицы равны нулю. Таким образом,

$$L_j=\begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & \ddots & & & \\ & & & & 1 & & \\ & & & & \ddots & \ddots & \\ & & & & \alpha_{j+1,j}^{(0)} & & \ddots \\ & & & & \alpha_{j+2,j}^{(0)} & 1 & \\ & & & & \vdots & & \ddots \\ & & & & \alpha_{mj}^{(0)} & & 1 \end{bmatrix}. \tag{5.1.6}$$

3°. Прямым ходом метода Гаусса система (5.1.1) преобразуется в рекуррентную систему

$$U\bar{x} = L\bar{f}. \tag{5.1.7}$$

обратный ход состоит в решении этой системы.

Матрица U и вектор $L\bar{f}$ вычисляются с некоторыми ошибками, которые приводят к искажению системы (5.1.7). Пусть Γ — матрица искажения матрицы U, а $\bar{\delta}$ — вектор искажения вектора $L\bar{f}$. Вместо системы (5.1.7) мы на самом деле построим и будем решать искаженную систему

$$(U + \Gamma)\bar{z} = L\bar{f} + \bar{\delta}. \tag{5.1.8}$$

Наша ближайшая цель — оценить норму $\|\Gamma\|$. Попутно будет оценена и норма $\|\bar{\delta}\|$.

§ 2. МАТРИЦА L И ЕЕ ПОГРЕШНОСТИ

1°. Вычислим произведение $L^{(j)} = L_j L_{j-1} \ldots L_2 L_1$. Начнем со случая $j = 2$:

$$L_2 L_1 = \begin{bmatrix} 1 & & & & & & & & \\ 0 & 1 & & & & & & & \\ 0 & \alpha_{32}^{(0)} & 1 & & & & & & \\ 0 & \alpha_{42}^{(0)} & 0 & 1 & & & & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & & & & \\ \cdot & \cdot & \cdot & \cdot & & \cdot & & & \\ \cdot & \cdot & \cdot & \cdot & & & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & & & & 1 & \\ \cdot & \cdot & \cdot & \cdot & & & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & & & & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & & & & & & & 1 \\ 0 & \alpha_{m2}^{(0)} & 0 & 0 & \cdots & & & & & & 1 \end{bmatrix} \times$$

$$\times \begin{bmatrix} 1 & & & & & & & & \\ \alpha_{21}^{(0)} & 1 & & & & & & & \\ \alpha_{31}^{(0)} & 0 & 1 & & & & & & \\ \cdot & \cdot & & \cdot & & & & & \\ \cdot & \cdot & & & \cdot & & & & \\ \cdot & \cdot & & & & \cdot & & & \\ \cdot & \cdot & & & & & 1 & & \\ \cdot & \cdot & & & & & & \cdot & \\ \cdot & \cdot & & & & & & & \cdot \\ \cdot & \cdot & & & & & & & & 1 \\ \alpha_{m1}^{(0)} & 0 & 0 & \cdots & & & & & & 1 \end{bmatrix}.$$

Легко убедиться, что произведение есть треугольная нижняя матрица с единицами на главной диагонали и что под главной диагональю в произведении ненулевые элементы располагаются только в двух первых столбцах, причем произведение имеет вид

$$L^{(2)}=\begin{bmatrix} 1 & & & & & & & & & & \\ \alpha_{21}^{(0)} & 1 & & & & & & & & & \\ \alpha_{31}^{(1)} & \alpha_{32}^{(0)} & 1 & & & & & & & & \\ \alpha_{41}^{(1)} & \alpha_{42}^{(0)} & 0 & 1 & & & & & & & \\ \cdot & \cdot & \cdot & & \cdot & & & & & & \\ \cdot & \cdot & \cdot & & & \cdot & & & & & \\ \cdot & \cdot & \cdot & & & & \cdot & & & & \\ \cdot & \cdot & \cdot & & & & & 1 & & & \\ \cdot & \cdot & \cdot & & & & & & \cdot & & \\ \cdot & \cdot & \cdot & & & & & & & \cdot & \\ \cdot & \cdot & \cdot & & & & & & & & \cdot \\ \alpha_{m1}^{(1)} & \alpha_{m2}^{(0)} & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \end{bmatrix}; \qquad (5.2.1)$$

элементы $\alpha_{i1}^{(1)}$ вычисляются по формуле

$$i \geqslant 2, \quad \alpha_{i1}^{(1)} = \alpha_{i2}^{(0)}\alpha_{21}^{(0)} + \alpha_{i1}^{(0)}. \qquad (5.2.2)$$

2°. Ясно, что для $k=2$ справедливо представление

$$L^{(k)} =$$

$$=\begin{bmatrix} 1 & & & & & & & & & \\ \alpha_{21}^{(0)} & 1 & & & & & & & & \\ \alpha_{31}^{(1)} & \alpha_{32}^{(0)} & 1 & & & & & & & \\ \alpha_{41}^{(2)} & \alpha_{42}^{(1)} & \alpha_{43}^{(0)} & 1 & & & & & & \\ \cdot & \cdot & \cdot & \cdot & & & & & & \\ \alpha_{k1}^{(k-2)} & \alpha_{k2}^{(k-3)} & \cdot & \cdot & \alpha_{k,k-1}^{(0)} & 1 & & & & \\ \alpha_{k+1,1}^{(k-1)} & \alpha_{k+1,2}^{(k-2)} & \cdot & \cdot & \alpha_{k+1,k-1}^{(1)} & \alpha_{k+1,k}^{(0)} & 1 & & & \\ \alpha_{k+2,1}^{(k-1)} & \alpha_{k+2,2}^{(k-2)} & \cdot & \cdot & \alpha_{k+2,k-1}^{(1)} & \alpha_{k+2,k}^{(0)} & 0 & 1 & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \alpha_{m1}^{(k-1)} & \alpha_{m2}^{(k-2)} & & & \alpha_{m,k-1}^{(1)} & \alpha_{mk}^{(0)} & 0 & \cdot & \cdot & 1 \end{bmatrix}, \qquad (5.2.3)$$

где элементы $\alpha_{ij}^{(\varkappa)}$, $\varkappa \leqslant k-1$, вычисляются по формуле

$$\alpha_{ij}^{(\varkappa)} = \alpha_{i,j+\varkappa}^{(0)}\alpha_{j+\varkappa,j}^{(\varkappa-1)} + \alpha_{ij}^{(\varkappa-1)}. \qquad (5.2.4)$$

Докажем, что если представление (5.2.3), (5.2.4) верно для некоторого k, то оно верно и для $k+1$. Умножим $L^{(k)}$ слева на

$$L_{k+1}=\begin{bmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ & & \ddots & & & & & & \\ & & & 1 & & & & & \\ & & & \alpha^{(0)}_{k+1,\,k+1} & 1 & & & & \\ & & & \alpha^{(0)}_{k+2,\,k+1} & & \ddots & & & \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & \alpha^{(0)}_{m,\,k+1} & & & & & 1 \end{bmatrix}.$$

Используя правило умножения матриц, нетрудно убедиться, что произведение $L^{(k+1)}=L_{k+1}L^{(k)}$ получается из $L^{(k)}$ заменой k на $k+1$, причем элементы $\alpha^{(\varkappa)}_{ij}$, $\varkappa\leqslant k$, вычисляются по формуле (5.2.4). Тем самым формула (5.2.3) доказана для всех $k\leqslant m-1$; при $k=m-1$ имеем $L^{(m)}=L$ и, следовательно,

$$L=\begin{bmatrix} 1 & & & & & \\ \alpha^{(0)}_{21} & 1 & & & & \\ \alpha^{(1)}_{31} & \alpha^{(0)}_{32} & 1 & & & \\ \alpha^{(2)}_{41} & \alpha^{(1)}_{42} & \alpha^{(0)}_{43} & 1 & & \\ \cdots & \cdots & \cdots & \cdots & \ddots & \\ \alpha^{(m-2)}_{m1} & \alpha^{(m-3)}_{m2} & \alpha^{(m-4)}_{m3} & \cdots & \alpha^{(0)}_{m,\,m-1} & 1 \end{bmatrix}. \qquad (5.2.5)$$

3°. Оценим норму матрицы L. Ее можно представить в виде

$$L=\sum_{k=0}^{m-1}M_k,\quad M_0=I;$$

$$k>0,\ M_k=\begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & \ddots & & & \\ & & & \ddots & & \\ \alpha^{(k-1)}_{k+1,\,1} & & & & \ddots & \\ & \alpha^{(k-1)}_{k+2,\,2} & & & & \ddots \\ & & \ddots & & & \\ & & & \ddots & & \\ & & & & \alpha^{(k-1)}_{m,\,m-k} & 0 \end{bmatrix}; \qquad (5.2.6)$$

отличные от нуля элементы матрицы M_k расположены только на $(k+1)$-й нижней диагонали. Имеем

$$\|L\| \leqslant \sum_{k=0}^{m-1} \|M_k\| = 1 + \sum_{k=1}^{m-1} \|M_k\|.$$

Так как $|\alpha_{ij}^{(0)}| \leqslant 1$, то, как это следует из формулы (5.2.4),

$$\max_{i,j} |\alpha_{ij}^{(k)}| \leqslant 2 \max_{i,j} |\alpha_{ij}^{(k-1)}|,$$

откуда вытекает неравенство

$$\max_{i,j} |\alpha_{ij}^{(k)}| \leqslant 2^k. \tag{5.2.7}$$

Матрица M_k осуществляет следующее преобразование $R_m \to R_m$: если $x=(x_1, x_2, \ldots, x_m)$ и $y=(y_1, y_2, \ldots, y_m)=M_k x$, то $y_1 = y_2 = \ldots = y_k = 0$, $j > k$, $y_j = \alpha_{j,\,j-k}^{(k-1)} x_{j-k}$. Отсюда $\|M_k x\| \leqslant 2^{k-1}\|x\|$, $\|M_k\| \leqslant 2^{k-1}$ и

$$\|L\| \leqslant 1 + \sum_{k=1}^{m-1} 2^{k-1} = 2^{m-1}. \tag{5.2.8}$$

Оценку (5.2.8) можно в некоторых случаях улучшить. Если

$$|\alpha_{ij}^{(0)}| \leqslant \alpha \leqslant 1, \tag{5.2.9}$$

то по формуле (5.2.4)

$$\max_{i,j} |\alpha_{ij}^{(k)}| \leqslant (1+\alpha) \max_{i,j} |\alpha_{ij}^{(k-1)}|.$$

Отсюда

$$\max_{i,j} |\alpha_{ij}^{(k)}| \leqslant (1+\alpha)^k \max_{i,j} |\alpha_{ij}^{(0)}| \leqslant \alpha(1+\alpha)^k. \tag{5.2.10}$$

Теперь $\|M_k\| \leqslant \alpha(1+\alpha)^{k-1}$ и

$$\|L\| \leqslant 1 + \alpha \sum_{k=1}^{m-1} (1+\alpha)^{k-1} = (1+\alpha)^{m-1}. \tag{2.5.11}$$

Замечание 1. Формула (5.2.11) точна по порядку. Чтобы убедиться в этом, рассмотрим случай $\alpha_{ij}^{(0)} = \alpha = \text{const}$, $0 < \alpha \leqslant 1$. В этом случае $\alpha_{ij}^{(k)} = \alpha(1+\alpha)^k$. Пусть $x \in R_m$ имеет только одну составляющую, отличную от нуля, $x_1 > 0$. Тогда $\|x\| = x_1$ и

$$Lx = (x_1, \alpha x_1, \alpha(1+\alpha)x_1, \ldots, \alpha(1+\alpha)^{m-2} x_1).$$

Отсюда

$$\|Lx\|^2 = \left[1 + \alpha^2 \sum_{k=0}^{m-2} (1+\alpha)^{2k}\right] x_1^2 =$$

$$= \left\{1 + \frac{\alpha[(1+\alpha)^{2m-2} - 1]}{2+\alpha}\right\} \|x\|^2 \geqslant \frac{\alpha}{2+\alpha} (1+\alpha)^{2m-2} \|x\|^2$$

и, следовательно,

$$\|L\| \geqslant \sqrt{\alpha/(2+\alpha)}\,(1+\alpha)^{m-1}. \tag{5.2.12}$$

Замечание 2. Из вывода формулы (5.2.11) ясно, что

$$\|\hat{L}\| \leqslant (1+\alpha)^{m-1} \tag{5.2.13}$$

(определение матрицы $\hat{L}$ см. в главе 1, § 2, п. 1°).

4°. Неравенство (5.2.12) показывает, что оценка (5.2.11) достигается (по крайней мере по порядку), если числа $\alpha_{ij}^{(0)}$ положительны и постоянны. В общем же случае может оказаться, что оценка (5.2.11) завышена, что видно из следующего примера. Пусть $\alpha_{ij}^{(0)} = -\alpha = \text{const}$, $0 < \alpha < 1$. Из формулы (5.2.4) легко вытекает, что в данном случае $\alpha_{ij}^{(k)} = -\alpha(1-\alpha)^k$, и ненулевые элементы матрицы M_k (формула (5.2.7)) равны $-\alpha(1-\alpha)^k$. Отсюда

$$\|M_k x\|^2 = \alpha^2 (1-\alpha)^{2k} \sum_{j=1}^{m-k} x_j^2 \leqslant \alpha^2 (1-\alpha)^{2k} \|x\|^2;$$

$$\|M_k\| \leqslant \alpha(1-\alpha)^k; \quad \|L\| \leqslant 1 + \alpha \sum_{k=1}^{m-1} (1-\alpha)^k = 2 - (1-\alpha)^m < 2,$$

и в нашем примере норма $\|L\|$ ограничена независимо от m.

В более общем случае, если бы оказалось, что $-\alpha \leqslant \alpha_{ij}^{(0)} < 0$, $0 < \alpha < 1$, и что при любом фиксированном $k \leqslant m-2$ числа $\alpha_{ij}^{(k)}$ все имеют один и тот же знак, то слагаемые в (5.2.4) имели бы разные знаки, и было бы справедливо одно из двух неравенств:

$$\max_{i,j} |\alpha_{ij}^{(k)}| \leqslant \max_{i,j} |\alpha_{ij}^{(0)}| \cdot \max_{i,j} |\alpha_{ij}^{(k-1)}|; \quad \max_{i,j} |\alpha_{ij}^{(k)}| \leqslant \max_{i,j} |\alpha_{ij}^{(k-1)}|.$$

Заметим, что из первого неравенства вытекает и второе:

$$\max_{i,j} |\alpha_{ij}^{(k)}| \leqslant \alpha \max_{i,j} |\alpha_{ij}^{(k-1)}| < \max_{i,j} |\alpha_{ij}^{(k-1)}|.$$

В свою очередь, отсюда следует, что в обоих случаях $\max\limits_{ij} |\alpha_{ij}^{(k)}| \leqslant \alpha$. Теперь

$$\forall x \in R_m, \quad \|Lx\|^2 \leqslant \|x\|^2 + \alpha^2 \sum_{i=2}^{m} \Big(\sum_{j=1}^{i-1} |x_j|\Big)^2 \leqslant$$
$$\leqslant \|x\|^2 [1 + \alpha^2 m(m-1)/2]$$

и, следовательно,

$$\|L\| \leqslant \sqrt{1 + \alpha^2 (m-1) m/2}. \tag{5.2.14}$$

Замечание 3. Формулы (5.2.8), (5.2.11) — (5.2.14), дающие оценки нормы $\|L\|$ в различных условиях, очевидно, остаются верными и для нормы $\|\hat{L}\|$.

5°. Обозначим через Λ погрешность матрицы L и через $\lambda_{ij}^{(k)}$ — погрешность числа $\alpha_{ij}^{(k)}$, так что

$$\Lambda = [\lambda_{ij}^{(k)}]_{i,j=1}^{m}, \quad k = i-j-1; \quad (\alpha_{ij}^{(k)})_{\text{м}} = \alpha_{ij}^{(k)} + \lambda_{ij}^{(k)}.$$

По формуле (1.1.9) находим, что с точностью до членов высшего порядка малости $|\lambda_{ij}^{(k)}| \leqslant \varepsilon_1 |a_{ij}^{(k)}|$. Отсюда следует, что $\Lambda \varpi \varepsilon_1 \hat{L}$ и

$$\|\Lambda\| \leqslant \varepsilon_1 \|\hat{L}\|. \tag{5.2.15}$$

§ 3. ОЦЕНКА ПОГРЕШНОСТИ ИСКАЖЕНИЯ

1°. Пусть, как об этом сказано в § 1, $\Gamma = \delta(U)$. Тогда

$$U + \Gamma = (L + \Lambda)\bar{A} + \Lambda', \tag{5.3.1}$$

где Λ' — погрешность вычисления произведения $(L + \Lambda)\bar{A} := D$. По формуле (1.2.4) $\|\Lambda'\| = \|\delta(D)\| \leqslant \varepsilon_1 \|\hat{D}\|$, или, если отбросить члены высшего порядка малости, $\|\Lambda'\| \leqslant \varepsilon_1 \|\widehat{L\bar{A}}\| \leqslant \varepsilon_1 \|\hat{L}\| \cdot \|A\| \sqrt{m}$. Из (5.3.1) следует, что $\Gamma = \Lambda\bar{A} + \Lambda'$ и

$$\|\Gamma\| \leqslant \|\Lambda\bar{A}\| + \|\Lambda'\| \leqslant \|\Lambda\| \cdot \|A\| + \|\Lambda'\|. \tag{5.3.2}$$

Используя полученные выше оценки норм Λ и Λ', найдем

$$\|\Gamma\| \leqslant \varepsilon_1 (1 + \sqrt{m}) \|\bar{L}\| \cdot \|A\|. \tag{5.3.3}$$

2°. Положим $\bar{\delta} = \delta(L\bar{f})$ и оценим норму $\|\bar{\delta}\|$. Имеем $\bar{\delta} = \Lambda\bar{f} + \delta'$, где δ' — погрешность вычисления произведения $(L + \Lambda)\bar{f}$, равная с точностью до малых высшего порядка погрешности произведения $L\bar{f}$. Из (1.1.9) видно, что $\|\delta'\| \leqslant \varepsilon_1 \|\hat{L}\| \cdot \|f\|$. Соотношение (1.2.7) показывает, что та же величина оценивает норму $\|\Lambda\bar{f}\|$, и получается оценка

$$\|\bar{\delta}\| \leqslant 2\varepsilon_1 \|\hat{L}\| \cdot \|f\|. \tag{5.3.4}$$

3°. Воспользуемся формулой (3.1.6); она верна, если $\|A_n^{-1}\Gamma_n\| \leqslant \beta$, $0 \leqslant \beta < 1$. В данном случае надо считать, что величины, входящие в формулу (3.1.6), не зависят от n, и ее можно написать так: если $\|A^{-1}\Gamma\| \leqslant \beta$, $0 \leqslant \beta < 1$, то

$$\xi := \|z - x\| \leqslant (1 - \beta)^{-1} [\|A^{-1}\Gamma x\| + \|A^{-1}\delta\|]. \tag{5.3.5}$$

Из (5.3.5) вытекает утверждение, которое легче проверять: если $\|A^{-1}\| \cdot \|\Gamma\| \leqslant \beta$, $0 \leqslant \beta < 1$, то

$$\xi \leqslant \frac{\|A^{-1}\|}{1 - \beta} [\|\Gamma\| \cdot \|A^{-1}\| \cdot \|f\| + \|\delta\|]. \tag{5.3.6}$$

В нашей задаче следует заменить A на U; так как $U = L\bar{A}$, то $U^{-1} = \bar{A}^{-1}L^{-1}$, $\|U^{-1}\| \leqslant \|A^{-1}\| \cdot \|L^{-1}\|$, и мы приходим к такой оценке: если

$$\|A^{-1}\| \cdot \|L^{-1}\| \cdot \|\Gamma\| \leqslant \beta, \qquad 0 \leqslant \beta < 1, \tag{5.3.7}$$

то

$$\xi = \|z - x\| \leqslant \frac{\|A^{-1}\| \cdot \|L^{-1}\|}{1 - \beta} [\|\Gamma\| \cdot \|A^{-1}\| \cdot \|f\| + \|\delta\|]. \tag{5.3.8}$$

Величины $\|\Gamma\|$ и $\|\delta\|$ оцениваются по формулам п. 2°, и задача сводится к оценке нормы $\|L^{-1}\|$.

4°. Из (5.1.5) следует, что

$$L^{-1} = L_1^{-1} L_2^{-1} \dots L_{m-1}^{-1}; \tag{5.3.9}$$

матрицы L_k определяются формулой (5.1.6). Вычислим L_k^{-1}. Пусть

$$\begin{gathered} x \in R_m, \quad y = L_k x; \quad y_1 = x_1, \quad y_2 = x_2, \dots, y_k = x_k, \\ y_{k+1} = \alpha_{k+1,\,k}^{(0)} x_k + x_{k+1}, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots \\ y_m = \alpha_{mk}^{(0)} x_k + x_m. \end{gathered} \tag{5.3.10}$$

Решая систему (5.3.10) относительно $x_1, x_2, \dots, x_m$, находим

$$\begin{gathered} x_1 = y_1, \quad x_2 = y_2, \dots, x_k = y_k, \\ x_{k+1} = -\alpha_{k+1,\,k}^{(0)} y_k + y_{k+1}, \\ \dots\dots\dots\dots\dots\dots \\ x_m = -\alpha_{mk}^{(0)} y_k + y_m. \end{gathered}$$

Отсюда видно, что L_k^{-1} получается из L_k заменой α_{ik} на $-\alpha_{ik}$:

$$L_k^{-1} = \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & -\alpha_{k+1,\,k}^{(0)} & & & \\ & & & -\alpha_{k+2,\,k}^{(0)} & \ddots & & \\ & & & \vdots & & \ddots & \\ & & & -\alpha_{mk}^{(0)} & & & 1 \end{bmatrix}. \tag{5.3.11}$$

5°. Докажем тождество

$$\begin{bmatrix} 1 & & & & & & \\ \alpha_{21} & 1 & & & & & \\ \alpha_{31} & \alpha_{32} & 1 & & & & \\ \dots & \dots & \dots & \ddots & & & \\ \alpha_{k1} & \alpha_{k2} & \dots & \alpha_{k,\,k-1} & 1 & & \\ \alpha_{k+1,1} & \alpha_{k+1,2} & \dots & \alpha_{k+1,\,k-1} & 0 & 1 & \\ \dots & \dots & \dots & \dots & \dots & \dots & \ddots \\ \alpha_{m1} & \alpha_{m2} & \dots & \alpha_{m,\,k-1} & 0 & & 1 \end{bmatrix} \times$$

$$\times\begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & \alpha_{k+1,k} & \ddots & \\ & & & \alpha_{k+2,k} & & \ddots \\ & & & \vdots & & \\ & & & \alpha_{mk} & \cdots\cdots & 1 \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & & & & & & \\ \alpha_{21} & 1 & & & & & \\ \alpha_{31} & \alpha_{32} & 1 & & & & \\ \cdots & \cdots & \cdots & \ddots & & & \\ \alpha_{k1} & \alpha_{k2} & \cdots & \alpha_{k,k-1} & 1 & & \\ \alpha_{k+1,1} & \alpha_{k+1,2} & \cdots & \alpha_{k+1,k-1} & \alpha_{k+1,k} & 1 & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{m,k-1} & \alpha_{mk} & 0 \ldots & 1 \end{bmatrix}. \quad (5.3.12)$$

Введем обозначения ρ_{ij}, σ_{ij}, τ_{ij} соответственно для элементов левого сомножителя, правого сомножителя и произведения. Очевидно,

$$\rho_{ij} = \sigma_{ij} = 0,\ i < j;\quad \rho_{ii} = \sigma_{ii} = 1;$$
$$\rho_{ij} = \alpha_{ij},\ i > j;\quad \rho_{ij} = 0,\ k \leqslant j < i;$$
$$\sigma_{ij} = 0,\ j < i,\ j \neq k;\quad \sigma_{ik} = \alpha_{ik},\ i > k.$$

Произведение в (5.3.12) — нижняя треугольная матрица с единицами на главной диагонали, и достаточно вычислить элементы τ_{ij} при значениях индексов $i \geqslant 2$ и $j < i$.

Пусть $2 \leqslant i \leqslant k-1$ и $j < i$. Воспользуемся обычной формулой умножения матриц:

$$\tau_{ij} = \sum_{s=1}^{m} \rho_{is}\sigma_{sj}.$$

В данном случае $\rho_{is} = 0$, $s > i$, и потому

$$\tau_{ij} = \sum_{s=1}^{i} \rho_{is}\sigma_{sj}. \quad (5.3.13)$$

Но $j \leqslant k-1$, и если $s \neq j$, то $\sigma_{sj} = 0$, а при $s = j$ будет $\sigma_{sj} = 1$. В результате получаем $\tau_{ij} = \rho_{ij} = \alpha_{ij}$. Пусть теперь $i \geqslant k$. В этом случае

$$\rho_{is} = \begin{cases} \alpha_{is}, & s \leqslant k < 1, \\ 1, & s = i, \\ 0, & s \neq i,\ s > k, \end{cases}$$

и сумма (5.3.13) сводится к такой:

$$\tau_{ij} = \rho_{ii}\sigma_{ij} + \sum_{s=1}^{k-1} \alpha_{is}\sigma_{sj} = \sigma_{ij} + \sum_{s=1}^{k-1} \alpha_{is}\sigma_{sj}.$$

Если $j \leqslant k-1$, то $\sigma_{jj} = 1$, остальные $\sigma_{sj} = 0$, и тогда $\tau_{ij} = \alpha_{ij}$. Далее, если $j = k$, то $\sigma_{ik} = \alpha_{ik}$ и $\sigma_{sk} = 0$, $s \leqslant k-1$; в результате $\tau_{ik} = \alpha_{ik}$. Наконец, если $i > j > k$, то $\sigma_{ij} = 0$ и $\sigma_{sj} = 0$, так что $\tau_{ij} = 0$. Формула (5.3.12) доказана.

6°. Теперь нетрудно получить выражение для L^{-1}. По формулам (5.3.11), (5.3.12) находим

$$L_1^{-1}L_2^{-1} =$$

$$= \begin{bmatrix} 1 & & & & \\ -\alpha_{21}^{(0)} & 1 & & & \\ -\alpha_{31}^{(0)} & 0 & 1 & & \\ \cdots & \cdots & \cdots & \cdots & \\ -\alpha_{m1}^{(0)} & 0 & \ldots & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & -\alpha_{32}^{(0)} & 1 & & \\ \cdots & \cdots & \cdots & \cdots & \\ 0 & -\alpha_{m2}^{(0)} & 0 & \ldots & 0 \quad 1 \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & & & & \\ -\alpha_{21}^{(0)} & 1 & & & \\ -\alpha_{31}^{(0)} & -\alpha_{32}^{(0)} & 1 & & \\ \cdots & \cdots & \cdots & \cdots & \\ -\alpha_{m1}^{(0)} & -\alpha_{m2}^{(0)} & 0 & \ldots & 0 \quad 1 \end{bmatrix}.$$

Умножая последнее равенство справа на $L_3^{-1}, \ldots, L_{m-1}^{-1}$ и используя каждый раз формулу (5.3.12), получаем окончательно

$$L^{-1} = \begin{bmatrix} 1 & & & & \\ -\alpha_{21}^{(0)} & 1 & & & \\ -\alpha_{31}^{(0)} & -\alpha_{32}^{(0)} & 1 & & \\ \cdots & \cdots & \cdots & \cdots & \\ -\alpha_{m1}^{(0)} & -\alpha_{m2}^{(0)} & \ldots & -\alpha_{m,\,m-1}^{(0)} & 1 \end{bmatrix}. \qquad (5.3.14)$$

7°. Оценим норму $\|L^{-1}\|$. Пусть $x \in R_m$, $y = L^{-1}x$:

$$y_1 = x_1,$$
$$y_2 = -\alpha_{21}^{(0)}x_1 + x_2,$$
$$\cdots\cdots\cdots$$
$$y_m = -\alpha_{m1}^{(0)}x_1 - \alpha_{m2}^{(0)}x_2 - \ldots - \alpha_{m,\,m-1}^{(0)}x_{m-1} + x_m.$$

Так как $|\alpha_{ij}^{(0)}| \leqslant 1$, то

$$\|y\|^2 = \sum_{i=1}^{m}\left[-\sum_{j=1}^{i-1}\alpha_{ij}^{(0)}x_j + x_i\right]^2 \leqslant \sum_{i=1}^{i}\left[\sum_{j=1}^{i}|x_j|\right]^2 \leqslant$$
$$\leqslant \|x\|^2 \sum_{i=1}^{m} i = \frac{m(m+1)}{2}\|x\|^2$$

и, следовательно,

$$\|L^{-1}\| \leqslant \sqrt{m(m+1)/2} = m/\sqrt{2} + O(1). \qquad (5.3.15)$$

Формула (5.3.15) точна по порядку. Действительно, если положить $x_j = \bar{x} = \text{const}$ и $\alpha_{ij}^{(0)} = 1$, то

$$\|y\|^2 = \bar{x}^2 \sum_{j=1}^{m} (j-2)^2 = \bar{x}^2 \left[1 + \sum_{j=1}^{m-2} j^2\right] =$$

$$= \frac{\|x\|^2}{m}\left[1 + \frac{(m-2)(m-1)(2m-3)}{6}\right] = \|x\|^2 \frac{2m^2 - 9m + 13}{6}.$$

Отсюда следует, что

$$\|L^{-1}\| \geqslant \sqrt{(2m^2 - 9m + 13)/6} > 3^{-1/2}(m - 3/4). \qquad (5.3.16)$$

8°. Вернемся к погрешности искажения. Из формул (5.3.8), (5.3.9), (5.3.15) вытекает следующая оценка этой погрешности: если

$$\sqrt{m(m+1)/2}\,\|A^{-1}\| \cdot \|\Gamma\| \leqslant \beta, \quad 0 \leqslant \beta < 1, \qquad (5.3.17)$$

то

$$\xi = \|z - x\| \leqslant (1-\beta)^{-1} \sqrt{\frac{m(m+1)}{2}}\,\|A^{-1}\|[\|\Gamma\| \cdot \|x\| + \|\delta\|]. \qquad (5.3.18)$$

§ 4. ОЦЕНКА ПОГРЕШНОСТИ ОКРУГЛЕНИЯ

Погрешность округления в методе Гаусса связана с обратным ходом. Последний состоит в решении треугольной системы (5.1.7) или, точнее, (5.1.8), поскольку матрица U и свободный член Lf остаются неизвестными: известны на самом деле $\tilde{U} = U + \Gamma$ и $\tilde{f} = Lf + \delta$. Нам предстоит решить рекуррентную систему

$$\begin{aligned} &b_{mm} z_m = \tilde{f}_m, \\ &b_{m-1,\,m-1} z_{m-1} + b_{m-1,\,m} z_m = \tilde{f}_{m-1}, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ &b_{11} z_1 + b_{12} z_2 + \ldots + b_{1m} z_m = \tilde{f}_1; \end{aligned} \qquad (5.4.1)$$

смысл обозначений очевиден. Для того чтобы получить решение системы (5.4.1), необходимо выполнить некоторые арифметические действия, которые сопряжены с погрешностями округления. В результате вычислений будут получены не величины z_k, а некоторые другие величины v_k. Пусть величины $v_m, v_{m-1}, \ldots, v_{k+1}$ уже вычислены. Каждая из них дает значение соответствующей величины $z_m, z_{m-1}, \ldots, z_{k+1}$ со своей погреш-

ностью, так что $v_i = z_i + \gamma_i$, $k+1 \leqslant i \leqslant m$. Следующее уравнение системы (5.4.1) даст нам величину

$$v_k = \left[\tilde{f}_k - \sum_{i=k+1}^{m} b_{ki} v_i\right]\left(\frac{1}{b_{kk}}\right) + \alpha_k / b_{kk}; \qquad (5.4.2)$$

здесь α_k / b_{kk} — погрешность, допущенная при решении указанного уравнения. Из (5.4.1) следует, что $v_k = z_k + \gamma_k$, где

$$\gamma_k = \left[\alpha_k - \sum_{i=k+1}^{m} b_{ki}\gamma_i\right](1/b_{kk}),$$

или, короче,

$$\sum_{i=k}^{m} b_{ki}\gamma_i = \alpha_k. \qquad (5.4.3)$$

Обозначая $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$, имеем $(U + \Gamma)\gamma = \alpha$, или $\gamma = (U + \Gamma)^{-1}\alpha$. Далее,

$$(U + \Gamma)^{-1} = (L\bar{A} + \Gamma)^{-1} = (I + \bar{A}^{-1}L^{-1}\Gamma)^{-1}\bar{A}^{-1}L^{-1}.$$

Мы предполагаем, что Γ удовлетворяет неравенству (5.3.8), а тогда из последнего неравенства вытекает, что

$$\|\gamma\| \leqslant \frac{\|A^{-1}\|}{1-\beta}\sqrt{\frac{m(m+1)}{2}}\,\|\alpha\|.$$

Допустим, что вычисления производятся с повышенной точностью и с последующим отбрасыванием лишних знаков. Тогда, по неравенству (1.1.9), $|\alpha_k b_{kk}| \leqslant \varepsilon_1 |v_k|$ и $\|\alpha\| \leqslant \varepsilon_1 b_0 \|v\|$, $b_0 = \max |b_{kk}|$. Отсюда следует, что с точностью до малых высших порядков $\|\alpha\| \leqslant \varepsilon_1 b_0 \|x\|$ и окончательно

$$\|\gamma\| \leqslant \varepsilon_1 \frac{\|A^{-1}\|}{1-\beta}\sqrt{\frac{m(m+1)}{2}}\, b_0 \|x\|. \qquad (5.4.4)$$

Глава 6

ДРУГИЕ ТОЧНЫЕ МЕТОДЫ

§ 1. МЕТОД ПРОГОНКИ

1°. Метод прогонки представляет собой разновидность метода Гаусса, приспособленную для систем с ленточными матрицами. Он изложен, в частности, в монографии [129], из которой мы заимствуем приводимые ниже, в п. 2°, сведения об этом методе.

2°. Ограничимся случаем, когда система (5.1.1) — трехдиагональная. В подробной записи она имеет вид

$$\begin{gathered} c_0 x_0 - b_0 x_1 = f_0, \\ -a_i x_{i-1} + c_i x_i - b_i x_{i+1} = f_i, \quad 1 \leqslant i \leqslant m-1, \\ -a_m x_{m-1} + c_m x_m = f_m. \end{gathered} \qquad (6.1.1)$$

Процесс исключения (прямой ход по Гауссу) приводит к рекуррентной системе

$$x_m = \beta_{m+1},$$
$$x_i = \alpha_{i+1} x_{i+1} + \beta_{i+1}, \quad m-1 \geqslant i \geqslant 0; \tag{6.1.2}$$

коэффициенты α_i и β_i определяются следующими формулами:

$$\alpha_1 = \frac{b_0}{c_0}, \quad \alpha_{i+1} = \frac{b_i}{c_i - a_i\alpha_i}, \quad 1 \leqslant i \leqslant m-1, \tag{6.1.3}$$

$$\beta_1 = \frac{f_0}{c_0}, \quad \beta_{i+1} = \frac{f_i + a_i\beta_i}{c_i - a_i\alpha_i}, \quad 1 \leqslant i \leqslant m; \tag{6.1.4}$$

обратный ход, как обычно, заключается в решении системы (6.1.2).

Доказывается следующее утверждение.

Теорема 6.1.1. *Пусть коэффициенты системы* (6.1.1) *удовлетворяют условиям*

$$\begin{aligned} &|c_0| > 0, \ |c_m| > 0; \\ &|a_i| > 0, \ |b_i| > 0, \quad 1 \leqslant i \leqslant m-1; \\ &|c_i| \geqslant |a_i| + |b_i|, \quad 1 \leqslant i \leqslant m-1; \quad |c_0| \geqslant |b_0|, \ |c_m| \geqslant |a_m|, \end{aligned} \tag{6.1.5}$$

причем хотя бы в одном из соотношений третьей строки имеет место строгое неравенство. Тогда верны неравенства

$$|c_i - a_i\alpha_i| \geqslant b := \min |b_i|, \quad |\alpha_i| \leqslant 1. \tag{6.1.6}$$

Наметим доказательство этой теоремы. По условиям (6.1.5) $|\alpha_1| = |b_0|/|c_0| \leqslant 1$. Если для некоторого $i \leqslant m-1$ верно неравенство $|\alpha_i| \leqslant 1$, то $|c_i - a_i\alpha_i| \geqslant |c_i| - |a_i| \geqslant |b_i| \geqslant b$. Далее,

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

Наконец, доказывается, что $c_m - a_m\alpha_m \neq 0$.

Таким образом, условия (6.1.5) достаточны для выполнимости метода прогонки. Доказывается также, что если еще $|c_0| > |b_0|$, то $|\alpha_i| < 1$, $1 \leqslant i \leqslant m$. Поэтому будем считать, что $|\alpha_i| \leqslant \alpha = \text{const}$, $0 < \alpha < 1$.

3°. Выполнение прямого хода приводит к погрешности искажения; оценим ее. Данные задачи (6.1.1) будем считать точными.

Вычисления по формулам (6.1.3), (6.1.4) выполним с повышенной точностью и с последующим — по окончании всех вычислений по этим формулам — отбрасыванием лишних знаков. По формуле (1.1.9) получим

$$|\delta(\alpha_j)| \leqslant \varepsilon_1 |\alpha_j|; \quad |\delta(\beta_j)| \leqslant \varepsilon_1 |\beta_j|. \tag{6.1.7}$$

Но $|\alpha_j| \leqslant \alpha$, поэтому

$$|\delta(\alpha_j)| \leqslant \varepsilon_1 \alpha. \tag{6.1.8}$$

Далее оценим $|\beta_j|$. Обозначим $F = \max\limits_j |f_j|$. По формулам (6.1.4) имеем $|\beta_1| \leqslant F/|c_0|$ и $|\beta_{j+1}| \leqslant b_j^{-1}(F + a_j|\beta_j|)$. Пусть τ_i — решение разностного уравнения

$$\tau_{i+1} = F/b + \sigma\tau_i, \quad \tau_1 = F/|c_0|; \quad \sigma = \max_i |a_i/b_i|. \tag{6.1.9}$$

Очевидно, что $|\beta_i| \leqslant \tau_i$ и, тем более, $|\beta_i| \leqslant \max\limits_i \tau_i$. Легко находим, что

$$\tau_i = \begin{cases} (i-1)\dfrac{F}{b} + \dfrac{F}{|c_0|}, & \sigma = 1; \\ \dfrac{F}{b}\dfrac{\sigma^i - 1}{\sigma - 1} + \dfrac{F}{|c_0|}\sigma^i, & \sigma \neq 1. \end{cases} \tag{6.1.10}$$

Обозначая через τ максимум правой части формулы (6.1.10) при $1 \leqslant i \leqslant m$, получаем

$$|\delta(\beta_j)| \leqslant \varepsilon_1\tau. \tag{6.1.11}$$

4°. Теперь оценим погрешность обратного хода и одновременно погрешность приближенного решения, получаемого методом прогонки. Из формулы (6.1.2) находим, что с точностью до малых высших порядков

$$\delta(x_j) = \alpha_{i+1}\delta(x_{i+1}) + x_{i+1}\delta(\alpha_{i+1}) + \delta(\beta_{i+1}) \tag{6.1.12}$$

и в силу формул (6.1.8), (6.1.11)

$$|\delta(x_i)| \leqslant \alpha|\delta(x_{i+1})| + \varepsilon_1\alpha|x_{i+1}| + \tau\varepsilon_1. \tag{6.1.13}$$

Найдем оценку для $|x_{i+1}|$. Из уравнений (6.1.2) следует

$$|x_i| \leqslant \alpha|x_{i+1}| + \beta, \quad \beta = \max_i |\beta_i|.$$

Если η_i удовлетворяет разностному уравнению

$$\eta_i = \alpha\eta_{i+1} + \beta, \quad \eta_m = \beta, \tag{6.1.14}$$

то $|x_i| \leqslant \eta_i$. Решение уравнения (6.1.14) дается формулой

$$\eta_i = \frac{\beta}{1-\alpha}(1 - \alpha^{m-i+1}) < \frac{\beta}{1-\alpha},$$

и, следовательно, $|x_i| < \beta/(1-\alpha)$. Неравенство (6.1.13) заменяется более простым:

$$\begin{gathered} |\delta(x_i)| < \alpha|\delta(x_{i+1})| + \varepsilon_1(\alpha\beta/(1-\alpha) + \tau); \\ |\delta(x_m + 1)| \leqslant \varepsilon_1\tau. \end{gathered} \tag{6.1.15}$$

Обычным способом решим последнее неравенство. Обозначим $\alpha\beta/(1-\alpha) + \tau = \gamma$ и рассмотрим разностное уравнение

$$y_i = \alpha y_{i+1} + \varepsilon_1\gamma, \quad 1 \leqslant i \leqslant m; \quad y_m = \varepsilon_1\tau. \tag{6.1.16}$$

Ясно, что $|\delta(x_i)| < y_i$. Решение уравнения (6.1.16) есть

$$y_i = \frac{\varepsilon_1 \gamma}{1-\alpha} - \varepsilon_1 \left(\frac{\gamma}{1-\alpha} - \tau\right) \alpha^{m-i+1}.$$

Вычитаемое положительно, так как

$$\frac{\gamma}{1-\alpha} - \tau = \frac{\alpha\beta}{(1-\alpha)^2} + \frac{\tau\alpha}{1-\alpha} = \frac{\alpha}{1-\sigma}\left(\frac{\beta}{1-\alpha} + \tau\right),$$

поэтому $y_i < \varepsilon_1 \gamma/(1-\alpha)$ и окончательно

$$|\delta(x_i)| < \varepsilon_1 \frac{\alpha\beta + 2(1-\alpha)}{(1-\alpha)^2}. \tag{6.1.17}$$

§ 2. МЕТОД ХОЛЕЦКОГО

1°. Этот метод достаточно полно изложен в [139] под названием метода квадратных корней, поэтому в данном пункте мы приведем только самые важные факты, необходимые для понимания дальнейшего.

Пусть в уравнении (5.1.1) матрица A — положительно определенная. Тогда ее можно представить в виде

$$A = S^* S, \tag{6.2.1}$$

где S и S^* — сопряженные треугольные матрицы, соответственно верхняя и нижняя. Пусть $A = [a_{ij}]_{i,\,j=1}^{m}$, $S = [s_{ij}]_{i,\,j=1}^{m}$ и $s_{ij} = 0$, $i < j$. По правилу умножения матриц

$$a_{ij} = \sum_{k=1}^{\min(i,\,j)} s_{ki} s_{kj}. \tag{6.2.2}$$

Отсюда

$$s_{11} = \sqrt{a_{11}}, \quad s_{1j} = a_{1j}/s_{11}, \quad j > 1;$$

$$i \geqslant 2, \quad s_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} s_{ki}^2}; \quad j > i, \quad s_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} s_{ki} s_{kj}}{s_{ii}}. \tag{6.2.3}$$

После того как матрицы S и S^* построены, задача сводится к решению двух рекуррентных систем

$$S^* f' = f, \quad Sx = f'. \tag{6.2.4}$$

2°. В данном случае мы имеем дело с тремя источниками погрешностей: с погрешностью искажения матрицы S и с погрешностями округления, возникающими при решении систем (6.2.4).

3°. Оценим погрешность матрицы S. Будем считать, что вычисления по формулам (6.2.3) производятся с повышенной точностью и с последующим отбрасыванием лишних знаков. Тогда

по неравенству (1.1.9) $|\delta(s_{ij})| \leqslant \varepsilon_1 |s_{ij}|$; обозначая через Λ погрешность матрицы S, имеем, очевидно,

$$\| \Lambda \| \leqslant \varepsilon_1 \| \hat{S} \|. \tag{6.2.5}$$

Тому же неравенству (6.2.5) удовлетворяет и сопряженная матрица Λ^*, которая является погрешностью матрицы S^*.

4°. В результате искажения матриц S и S^* мы на самом деле будем решать не системы (6.2.4), а системы

$$(S^* + \Lambda^*) y = f, \quad (S + \Lambda) z = y. \tag{6.2.6}$$

Рассмотрим первую из них, которую представим в виде $\bar{S} y = f$, $\bar{S} = S^* + \Lambda^*$. Матрица $\bar{S}$ — нижняя треугольная, и рассматриваемую систему можно записать в виде

$$\begin{aligned} &c_{11} y_1 = f_1, \\ &c_{21} y_1 + c_{22} y_2 = f_2, \\ &\cdots\cdots\cdots \\ &c_{m1} y_1 + c_{m2} y_2 + \ldots + c_{mm} y_m = f_m; \\ &c_{ij} = s_{ij} + \delta(s_{ij}). \end{aligned} \tag{6.2.7}$$

С точностью до обозначений система (6.2.7) совпадает с системой (5.4.1), и можно воспользоваться рассуждениями § 4 главы 5. Пусть, решая уравнения (6.2.7), мы вместо y_k, $k = 1, 2, \ldots, m$, получим величины $v_k = y_k + \gamma_k$; при этом

$$v_k = \frac{1}{c_{kk}} \Big[f_k - \sum\nolimits_{i=1}^{k-1} c_{ki} v_i \Big] + \frac{\alpha_k}{c_{kk}}, \tag{6.2.8}$$

где α_k / c_{kk} есть погрешность, допущенная при решении k-го из уравнений (6.2.7). Заменив в уравнении (6.2.8) v_j на $y_j + \gamma_j$, найдем, что γ_k удовлетворяет системе

$$\sum\nolimits_{k=1}^{i} c_{ik} \gamma_k = \alpha_i,$$

или, короче, $\bar{S} \gamma = \alpha$. Имеем

$$\gamma = \bar{S}^{-1} \alpha = (S^* + \Lambda^*)^{-1} \alpha = [I + (S^*)^{-1} \Lambda^*]^{-1} (S^*)^{-1} \alpha, \tag{6.2.9}$$

откуда следует

$$\| \gamma \| \leqslant \| [I + (S^*)^{-1} \Lambda^*]^{-1} \| \cdot \| A^{-1} \|^{1/2} \cdot \| \alpha \|.$$

Далее, $\| (S^*)^{-1} \Lambda^* \| \leqslant \| A^{-1} \|^{1/2} \cdot \| \Lambda^* \|$, и по неравенству (6.2.5)

$$\| (S^*)^{-1} \Lambda^* \| \leqslant \varepsilon_1 \| A^{-1} \|^{1/2} \cdot \| A \|^{1/2} = \varepsilon_1 \sqrt{P_A} := \beta, \tag{6.2.10}$$

где P_A — число обусловленности матрицы A. Потребуем, чтобы было $\beta < 1$. Тогда $\| [I + (S^*)^{-1} \Lambda^*]^{-1} \| \leqslant (1 - \beta)^{-1}$, и погрешность округления вектора y имеет оценку

$$\| \gamma \| \leqslant (1 - \beta)^{-1} \sqrt{\| A^{-1} \|} \cdot \| \alpha \|.$$

Как и в § 4 главы 5, найдем, что

$$\|\alpha\| \leqslant \varepsilon_1 c_0 \|v\|, \quad c_0 = \max |c_{kk}|,$$

и окончательно, с точностью до малых высших порядков, получим

$$\|v - y\| = \|\gamma\| \leqslant \varepsilon_1 (1-\beta)^{-1} c_0 \|x\| \sqrt{\|A^{-1}\|}. \qquad (6.2.11)$$

5°. Обратимся ко второй системе (6.2.6), в которой, очевидно, следует заменить y на $y+\gamma$. Для матрицы $S^{-1}\Lambda$ верна та же оценка, что и для $(S^*)^{-1}\Lambda^*$, именно $\|S^{-1}\Lambda\| \leqslant \beta$, и если $\beta < 1$, то $\|(S+\Lambda)^{-1}\| \leqslant (1-\beta)^{-1}\sqrt{\|A^{-1}\|}$. Обозначим через w приближенное решение системы

$$(S+\Lambda)w = v = y + \delta, \qquad (6.2.12)$$

полученное в результате ошибок округления. Эта система также треугольная, ее точное решение равно $z+\tilde{\gamma}$, где $\tilde{\gamma} = (S+\Lambda)^{-1}\gamma$. Для разности $w-(z+\tilde{\gamma})$ при условии $\beta < 1$ верна оценка (6.2.11):

$$\|w - z - \tilde{\gamma}\| \leqslant \varepsilon_1 (1-\beta)^{-1} c_0 \|x\| \sqrt{\|A^{-1}\|},$$

или

$$\|w - z\| \leqslant \varepsilon_1 (1-\beta)^{-1} c_0 \|x\| \sqrt{\|A^{-1}\|} + \|\tilde{\gamma}\|.$$

Но

$$\|\tilde{\gamma}\| \leqslant \|(S+\Lambda)^{-1}\| \cdot \|\gamma\| \leqslant \varepsilon_1 (1-\beta)^{-2} c_0 \|x\| \sqrt{\|A^{-1}\|},$$

и

$$\|w - z\| \leqslant \varepsilon_1 (2-\beta)(1-\beta)^{-2} c_0 \|A^{-1}\|^{3/2} \|f\|. \qquad (6.2.13)$$

6°. Остается оценить разность $z-x$. Из (6.2.6) находим $(S^*+\Lambda^*)(S+\Lambda)z = f = Ax$; если раскрыть скобки и отбросить член второго порядка малости $\Lambda^*\Lambda$, то получится $A(z-x) + (S^*\Lambda + \Lambda S^*)z = 0$. Отсюда $z - x = -A^{-1}(S^*\Lambda + \Lambda^* S)z$, и по неравенству (6.2.5)

$$\|z - x\| \leqslant 2\varepsilon_1 \sqrt{m\|A\|} \cdot \|A^{-1}\| \cdot \|z\|.$$

Таким образом, $\|z-x\|$ есть величина порядка ε_1. Заменив справа $\|z\|$ на $\|x\|$, получим оценку, верную с точностью до величин высшего порядка малости:

$$\|z - x\| \leqslant 2\varepsilon_1 \|A^{-1}\|^2 \sqrt{m\|A\|} \cdot \|f\|. \qquad (6.2.14)$$

Сопоставив это с (6.2.13), придем к окончательной оценке погрешности метода Холецкого:

$$\|w - x\| \leqslant \varepsilon_1 \|A^{-1}\|^{3/2} \cdot \|f\| \cdot \left[2\sqrt{mP_A} + \frac{2-\beta}{(1-\beta)^2} c_0\right]. \qquad (6.2.15)$$

7°. В качестве примера рассмотрим систему, приведенную ниже, в п. 11° § 4 данной главы. Эта система имеет вид $Ax = f$,

где матрица A и вектор f заданы формулами (6.4.20). Матрица A — положительно определенная, и к ней можно применить метод Холецкого. По формуле (6.2.15) оценим погрешность этого метода для рассматриваемой системы. Из результатов п. 12° § 4 данной главы (см. ниже) следует, что

$$\|A\| = \frac{1}{2(1-\cos(2\pi/25))} < 16,$$

$$\|A^{-1}\| = 2(1+\cos(\pi/25)) < 4,$$

$$P_A = \|A\| \cdot \|A^{-1}\| < 64, \quad m = 12, \quad \sqrt{mP_A} < 32,$$

$$\|f\| = 1, \quad \beta = \varepsilon_1 \sqrt{P_A} < 2 \cdot 10^{-11}, \quad c_0 < 4.$$

Мы принимаем, что вычисления производились на ЭВМ БЭСМ-6 и $\varepsilon_1 < 2 \cdot 10^{-12}$. Теперь

$$\|w - x\| \leqslant 2 \cdot 10^{-12} \cdot 4^{3/2} \cdot 1 \left[32 + \frac{2 \cdot 2 \cdot 10^{-11}}{(1 - 2 \cdot 10^{-11})^2} \cdot 4\right] < 10^{-9}.$$

§ 3. МЕТОД ОКАЙМЛЕНИЯ

1°. Метод окаймления, предложенный Моррисом (см. [25]), обстоятельно изложен в [139]. Как отмечено в этой книге, данный метод целесообразно применять для систем Ритца и Бубнова — Галеркина, если надо уточнить уже построенное решение.

2°. Существо метода таково. В уравнении (5.1.1) матрицу $A = [a_{ij}]_{i,\,j=1}^{m}$ можно получить из матрицы $\tilde{A} = [a_{ij}]_{i,\,j=1}^{m-1}$ порядка $m-1$ окаймлением, т. е. присоединением m-й строки и m-го столбца. Это может быть записано так:

$$A = \begin{bmatrix} \tilde{A} & u \\ v & a_{mm} \end{bmatrix}. \tag{6.3.1}$$

Здесь u и v суть вектор-столбец и вектор-строка:

$$u = (a_{1m}, a_{2m}, \ldots, a_{m-1,\,m})', \quad v = (a_{m1}, a_{m2}, \ldots, a_{m,\,m-1}).$$

Допустим, что обе матрицы A и $\tilde{A}$ — неособенные и что матрица $\tilde{A}^{-1}$ известна. Доказывается, что матрицу A^{-1} можно представить в виде

$$A^{-1} = \begin{bmatrix} P & r \\ q & 1/\alpha \end{bmatrix}, \tag{6.3.2}$$

где P — квадратная матрица порядка $m-1$; q — вектор-строка; r — вектор-столбец, оба размерностью $m-1$, и α — число. Для

этих четырех объектов получаются следующие формулы:

$$\alpha = a_{mm} - (v,\ \tilde{A}^{-1}u), \tag{6.3.3}$$

$$r = -\alpha^{-1}\tilde{A}^{-1}u, \tag{6.3.4}$$

$$q = -\alpha^{-1}v\tilde{A}^{-1}, \tag{6.3.5}$$

$$P = \tilde{A}^{-1} + \alpha^{-1}\tilde{A}^{-1}uv\tilde{A}^{-1}. \tag{6.3.6}$$

Запишем еще решение системы (5.1.1), полученное методом окаймления. Если обозначить свободный член f через $(\tilde{f},\ f_m)'$, где $\tilde{f} = (f_1,\ f_2,\ \ldots,\ f_{m-1})$, а решения систем $\tilde{A}x^* = \tilde{f}$ и $\tilde{A}y^* + u = 0$ соответственно через x^* и y^*, то

$$x = A^{-1}f = (x^*,\ 0)' + \frac{f_m - (v,\ x^*)}{a_{mm} + (v,\ y^*)}(y^*,\ 1)'. \tag{6.3.7}$$

3°. Оценим норму $\|A^{-1}\|$. Пусть x — произвольный вектор из R_m, $x = (x_1,\ x_2,\ \ldots,\ x_m)'$. По формуле (6.3.2)

$$A^{-1}x = \Big(\sum_{k=1}^{m-1} p_{1k}x_k + r_1x_m,\quad \sum_{k=1}^{m-1} p_{2k}x_k + r_2x_m,\ \ldots$$
$$\ldots,\ \sum_{k=1}^{m-1} p_{m-1,\,k}x_k + r_{m-1}x_m,\ \sum_{k=1}^{m-1} q_kx_k + (1/\alpha)x_m\Big)'; \tag{6.3.8}$$

смысл обозначений очевиден. Отсюда

$$\|A^{-1}x\|^2 = \sum_{j=1}^{m-1}\Big(\sum_{k=1}^{m-1} p_{jk}x_k + r_jx_m\Big)^2 + \Big(\sum_{k=1}^{m-1} q_kx_k + \alpha^{-1}x_m\Big)^2 \leqslant$$
$$\leqslant \Big[\sum_{j,\,k=1}^{m-1} p_{jk}^2 + \sum_{j=1}^{m-1}(r_j^2 + q_j^2) + \alpha^{-1}\Big]\cdot\|x\|^2$$

и, как следствие,

$$\|A^{-1}\| \leqslant \big[\|P\|_E^2 + \|r\|^2 + \|q\|^2 + \alpha^{-2}\big]^{1/2}; \tag{6.3.9}$$

$\|P\|_E$ — евклидова норма матрицы P.

Можно вывести оценку, в некоторых случаях более точную, чем оценка (6.3.9). Обозначим $\check{x} = (x_1,\ x_2,\ \ldots,\ x_{m-1},\ 0)'$, $\bar{x} = (0,\ 0,\ \ldots,\ 0,\ x_m)'$, так что $x = \check{x} + \bar{x}$. Имеем $A^{-1}x = A^{-1}\check{x} + A^{-1}\bar{x}$, откуда $\|A^{-1}x\|^2 \leqslant 2\|A^{-1}\check{x}\|^2 + 2\|A^{-1}\bar{x}\|^2$. По формуле (6.3.8)

$$\|A^{-1}\check{x}\|^2 = \|P\check{x}\|^2_{R_{m-1}} + \big|(q,\ \check{x})_{R_{m-1}}\big|^2 \leqslant \big[\|P\|^2 + \|q\|^2\big]\|\check{x}\|^2,$$
$$\|A^{-1}\bar{x}\|^2 = \big[\|r\|^2 + \alpha^{-2}\big]x_m^2.$$

Из последних неравенств следует, что

$$\|A^{-1}\| \leqslant \sqrt{2}\max\Big[\sqrt{\|P\|^2 + \|q\|^2},\quad \sqrt{\|r_m\|^2 + \alpha^{-2}}\Big]. \tag{6.3.10}$$

Оценки величин, входящих в правые части неравенств (6.3.8) — (6.3.10), сразу вытекают из формул (6.3.3) — (6.3.6):

$$\begin{aligned} &\|P\| \leqslant \|\tilde{A}^{-1}\| \cdot [1 + \|u\| \cdot \|v\| \cdot \|\tilde{A}^{-1}\| \cdot |\alpha|^{-1}], \\ &\|q\| \leqslant |\alpha^{-1}| \cdot \|\tilde{A}^{-1}\| \cdot \|v\|, \\ &\|r\| \leqslant |\alpha^{-1}| \cdot \|\tilde{A}^{-1}\| \cdot \|u\|, \\ &|\alpha| = |a_{mm} - vA^{-1}u|. \end{aligned} \tag{6.3.11}$$

4°. Переходим к оценке погрешностей. Прежде всего примем, что величины a_{jk} — элементы матрицы A — заданы неточно, причем мы допускаем, что λ_{jk} — погрешности чисел a_{jk} — имеют известную нам оценку: $|\lambda_{jk}| \leqslant \lambda = \text{const}$; примем еще, что нам известна оценка сверху для нормы $\|\tilde{A}^{-1}\|$. Обозначим через $\tilde{\Gamma}$ погрешность матрицы $\tilde{A}$; очевидно, что $\|\tilde{\Gamma}\| \leqslant \leqslant \lambda(m-1)$. Пусть λ достаточно мало, тогда

$$\|\tilde{A} + \tilde{\Gamma}\| \approx \|\tilde{A}\|, \quad \|(\tilde{A} + \tilde{\Gamma})^{-1} - \tilde{A}^{+1}\| \approx \|\tilde{A}^{-1}\|^2 \cdot \|\tilde{\Gamma}\|. \tag{6.3.12}$$

Займемся оценкой погрешности объектов (6.3.11). Обозначим через λ_α, λ_u, λ_v погрешности числа α и векторов u, v соответственно. Сразу же заметим, что $\|\lambda_u\|$, $\|\lambda_v\| \leqslant \sqrt{m-1}\,\lambda$. По формуле (6.3.3)

$$\alpha + \lambda_\alpha = a_{mm} + \lambda_{mm} - (v + \lambda_v)(\tilde{A} + \tilde{\Gamma})^{-1}(u + \lambda_u) + \delta_м(\alpha); \tag{6.3.13}$$

$\delta_м(\alpha)$ — суммарная погрешность машинных арифметических действий, необходимых для вычисления величины α. Из (6.3.13) следует равенство, верное с точностью до слагаемых более высокого порядка малости:

$$\lambda_\alpha = \lambda_{mm} - \lambda_v \tilde{A}^{-1}u - v\tilde{A}^{-1}\lambda_v + v\tilde{A}^{-1}\tilde{\Gamma}\tilde{A}^{-1}u + \delta_м(\alpha).$$

Отсюда

$$\begin{aligned} |\lambda_\alpha| \leqslant \lambda + \lambda\sqrt{m-1}\,\|\tilde{A}^{-1}\| \cdot \|u\| + \lambda\sqrt{m-1}\,\|\tilde{A}^{-1}\| \cdot \|v\| + \\ + \lambda(m-1)\|\tilde{A}^{-1}\|^2 \cdot \|u\| \cdot \|v\| + |\delta_м(\alpha)| = \\ = \lambda\left[1 + \sqrt{m-1}\,\|\tilde{A}^{-1}\| \cdot \|u\|\right]\left[1 + \sqrt{m-1}\,\|\tilde{A}^{-1}\| \cdot \|v\|\right] + |\delta_м(\alpha)|; \end{aligned}$$

мы здесь воспользовались соотношениями (6.3.12).

Чтобы оценить $\delta_м(\alpha)$, воспользуемся неравенствами § 1, 2 главы 1. Имеем

$$\delta_м(v\tilde{A}^{-1}u) = \delta_м(v, \tilde{A}^{-1}u) = (v, (\tilde{A}^{-1}u)_м)_м - (v, \tilde{A}^{-1}u).$$

Обозначим $\tilde{A}^{-1}u = z$, тогда $(\tilde{A}^{-1}u)_м = z + \delta_м(z)$; по неравенству (1.2.7) $\|\delta_м(z)\| \leqslant \varepsilon_1\sqrt{m-1}\,\|\tilde{A}^{-1}\| \cdot \|u\|$. Далее,

$$\begin{aligned} |\delta_м(v, \tilde{A}^{-1}u)| = [(v, z + \delta_м(z))_м - (v, z + \delta_м(z))] + \\ + [(v, z + \delta_м(z)) - (v, z)|. \end{aligned} \tag{*}$$

Первая квадратная скобка есть машинная погрешность скалярного произведения $(v, z+\delta_м(z))$. По формуле (1.1.9) она имеет оценку $\varepsilon_1\|v\|\cdot\|z+\delta_м(z)\|$. Отбросив члены высшего порядка малости, найдем, что первая квадратная скобка в (*) оценивается величиной $\varepsilon_1\|v\|\cdot\|z\|\leqslant\varepsilon_1\|\tilde{A}^{-1}\|\cdot\|u\|\cdot\|v\|$. Вторая скобка оценится так:

$$|(v, z+\delta_м(z))-(v, z)|=|(v, \delta_м(z))|\leqslant\|v\|\cdot\|\delta_м(z)\|\leqslant$$
$$\leqslant\varepsilon_1\sqrt{m-1}\,\|\tilde{A}^{-1}\|\cdot\|u\|\cdot\|v\|,$$

и окончательно $|\delta_м(v, \tilde{A}^{-1}u)|\leqslant\varepsilon_1(1+\sqrt{m-1})\|\tilde{A}^{-1}\|\cdot\|u\|\cdot\|v\|$.

Очевидно, что при добавлении числа a_{mm} к величине $-v\tilde{A}^{-1}u$ погрешность увеличится не более чем на $\varepsilon_1|a_{mm}|$, и мы приходим к искомой оценке:

$$|\delta_м\alpha|\leqslant\varepsilon_1[|a_{mm}|+(1+\sqrt{m-1})\|\tilde{A}^{-1}\|\cdot\|u\|\cdot\|v\|].$$

Теперь из (6.3.13) следует, что

$$|\lambda_\alpha|\leqslant\lambda[1+\sqrt{m-1}\,\|\tilde{A}^{-1}\|\cdot\|u\|][1+\sqrt{m-1}\,\|\tilde{A}^{-1}\|\cdot\|v\|]+$$
$$+\varepsilon_1[a_{mm}+(1+\sqrt{m-1})\|\tilde{A}^{-1}\|\cdot\|u\|\cdot\|v\|]. \qquad (6.3.14)$$

Займемся оценкой величины $\|\lambda_r\|$. Вычисляя по формуле (6.3.4), мы на самом деле получим вектор

$$r+\lambda_r=-(\alpha+\lambda_\alpha)^{-1}(\tilde{A}+\tilde{\Gamma})^{-1}(u+\lambda_u)+\delta_м(r).$$

Так же, как при оценке величины $\delta_м(\alpha)$, мы найдем, что с точностью до членов высшего порядка малости $\|\delta_м(r)\|\leqslant\varepsilon_1|\alpha|^{-1}\times$ $\times\|\tilde{A}^{-1}\|(1+\sqrt{m-1})$ и

$$\|\lambda_r\|\leqslant|\lambda_\alpha|\cdot\|\tilde{A}^{-1}\|\cdot\|u\|+\lambda|\alpha|^{-1}\times$$
$$\times\|\tilde{A}^{-1}\|\sqrt{m-1}\{1+\sqrt{m-1}\,\|\tilde{A}^{-1}\|\cdot\|u\|\}+$$
$$+\varepsilon_1|\alpha|^{-1}\|\tilde{A}^{-1}\|\cdot\|u\|(1+\sqrt{m-1}). \qquad (6.3.15)$$

Аналогично получается оценка вектора λ_q:

$$\|\lambda_q\|\leqslant|\lambda_q|\cdot\|\tilde{A}^{-1}\|\cdot\|v\|+\lambda|\alpha|^{-1}\|\tilde{A}^{-1}\|\sqrt{m-1}\times$$
$$\times\{1+\sqrt{m-1}\,\|\tilde{A}^{-1}\|\cdot\|v\|\}+\varepsilon_1\|\tilde{A}^{-1}\|\cdot\|v\|(1+\sqrt{m-1}). \qquad (6.3.16)$$

Остается оценить погрешность Λ_P матрицы P. По формуле (6.3.6)

$$P+\Lambda_P=(\tilde{A}+\tilde{\Gamma})^{-1}+(\alpha+\lambda_\alpha)^{-1}(\tilde{A}+\tilde{\Gamma})^{-1}(u+\lambda_u)(v+\lambda_v)\times$$
$$\times(\tilde{A}+\tilde{\Gamma})^{-1}+\delta_м(P). \qquad (6.3.17)$$

Оценим погрешность $\delta_м(P)$. При ее вычислении числа α и элементы матрицы $\tilde{A}^{-1}$ и векторов u и v считаются заданными

точно. Вычисления начинаются с произведения $uv = [u_i v_j]_{i,j=1}^{m}$ По формуле (1.1.1) $|\delta(u_i v_j)| \leqslant \varepsilon_1 |u_i v_j|$; это значит, что $\delta(uv) \Subset \varepsilon_1 \widehat{uv}$ и потому

$$\|\delta(uv)\| \leqslant \varepsilon_1 \|\widehat{uv}\| \leqslant \varepsilon_1 \left[\sum_{i,j=1}^{m} u_i^2 v_j^2\right]^{1/2} = \varepsilon_1 \|u\| \cdot \|v\|.$$

Заметим, что та же оценка верна и для $\|\delta(\widehat{uv})\|$.

Теперь рассмотрим погрешность произведения $\tilde{A}^{-1}uv\tilde{A}^{-1}$. Имеем

$$\delta_{\text{м}}(\tilde{A}^{-1}uv\tilde{A}^{-1}) = (\tilde{A}^{-1}(uv)_{\text{м}}\tilde{A}^{-1})_{\text{м}} - \tilde{A}^{-1}uv\tilde{A}^{-1} =$$
$$= [(\tilde{A}^{-1}(uv)_{\text{м}}\tilde{A}^{-1})_{\text{м}} - \tilde{A}^{-1}(uv)_{\text{м}}\tilde{A}^{-1}] + [\tilde{A}^{-1}(uv)_{\text{м}}\tilde{A}^{-1} - \tilde{A}^{-1}uv\tilde{A}^{-1}]. \tag{6.3.18}$$

Первая квадратная скобка оценивается по неравенству (1.2.6): ее норма имеет оценку

$$2\varepsilon_1 \|\tilde{A}^{-1}\|^2 \|(\widehat{uv})_{\text{м}}\| \leqslant 2\varepsilon_1 (m-1) \|\tilde{A}^{-1}\|^2 (\|\widehat{uv}\| + \|\delta(\widehat{uv})\|).$$

Но $\|\widehat{uv}\| \leqslant \|uv\|_E = \|u\| \cdot \|v\|$, а $\|\delta(\widehat{uv})\| = O(\varepsilon_1)$. Отбросив в последнем неравенстве члены второго порядка малости, получим для первой квадратной скобки более простую оценку $2\varepsilon_1(m-1) \times \|\tilde{A}^{-1}\|^2 \|u\| \cdot \|v\|$.

Вторая квадратная скобка в (6.3.18) равна $\tilde{A}^{-1}\delta(uv)\tilde{A}^{-1}$; ее норма не превосходит $\varepsilon_1 \|\tilde{A}^{-1}\|^2 \|u\| \cdot \|v\|$. Собрав найденные оценки, получим

$$\|\delta_{\text{м}}(\tilde{A}^{-1}uv\tilde{A}^{-1})\| \leqslant \varepsilon_1 (2m-1) \|\tilde{A}^{-1}\|^2 \|u\| \cdot \|v\|.$$

Умножение матрицы $\tilde{A}^{-1}uv\tilde{A}^{-1}$ на число α^{-1} приведет по формуле (1.2.4а) к умножению оценки ее погрешности на $|\alpha|^{-1}\sqrt{m-1}$:

$$\|\delta_{\text{м}}(\alpha^{-1}\tilde{A}^{-1}uv\tilde{A}^{-1})\| \leqslant \varepsilon_1 |\alpha|^{-1} \sqrt{m-1}\,(2m-1) \|\tilde{A}^{-1}\| \cdot \|u\| \cdot \|v\|.$$

Теперь рассмотрим величину $\delta_{\text{м}}(P)$. Временно обозначим $\tilde{A}^{-1} = K$, $\alpha^{-1}\tilde{A}^{-1}uv\tilde{A}^{-1} = L$; тогда $\|\delta_{\text{м}}(P)\| = \|\delta_{\text{м}}(K+L)\|$. Далее, $\delta_{\text{м}}(K+L) = (K+L)_{\text{м}} - (K+L)$ и $(K+L)_{\text{м}} = (K + (L_{\text{м}}))_{\text{м}}$. Отсюда

$$\delta_{\text{м}}(K+L) = (K + (L_{\text{м}}))_{\text{м}} - (K+L) = [(K + (L)_{\text{м}})_{\text{м}} - (K + (L_{\text{м}}))] +$$
$$+ [K + (L_{\text{м}}) - (K+L)] = \delta_{\text{м}}(K + (L)_{\text{м}}) + \delta_{\text{м}}(L).$$

Второе слагаемое справа мы оценили выше, а первое оценивается по неравенству (1.2.8):

$$\delta\|(K + (L)_{\text{м}})\| \leqslant \varepsilon_1 \sqrt{m-1}\,(\|K\| + \|(L)_{\text{м}}\|) \approx$$
$$\approx \varepsilon_1 \sqrt{m-1}\,(\|K\| + \|L\|) \leqslant \varepsilon_1 \sqrt{m-1}\,[\|\tilde{A}^{-1}\| +$$
$$+ |\alpha|^{-1} \|\tilde{A}^{-1}\|^2 \cdot \|u\| \cdot \|v\|].$$

Сложив это с оценкой для $\delta_м(L)$, найдем

$$\|\delta_м(P)\| \leqslant \varepsilon_1 \sqrt{m-1}\,\|\tilde{A}^{-1}\|\,[1 + 2m|\alpha|^{-1}\cdot\|\tilde{A}^{-1}\|\cdot\|u\|\cdot\|v\|]. \tag{6.3.18а}$$

Вернемся к формуле (6.3.17). Отбросив в ней члены высшего порядка малости, получим

$$\Lambda_P = \tilde{A}^{-1}[\tilde{\Gamma} - \alpha^{-2}\lambda_a uv - \alpha^{-1}\tilde{\Gamma}\tilde{A}^{-1}uv + \alpha^{-1}\lambda_u v + \alpha^{-1}u\lambda_v - \alpha^{-1}uv\tilde{A}^{-1}\tilde{\Gamma}]\,\tilde{A}^{-1} + \delta_м(P)$$

или, если воспользоваться неравенством (6.3.18а),

$$\|\Lambda_P\| \leqslant \|\tilde{A}^{-1}\|^2\{\lambda(m-1) + \alpha^{-2}\lambda\|u\|\cdot\|v\| + 2|\alpha^{-1}|\lambda\sqrt{m-1}\,(1 + \sqrt{m-1}\,\|\tilde{A}^{-1}\|)\} + \varepsilon_1\sqrt{m-1}\,\|\tilde{A}^{-1}\|\,[1 + 2m|\alpha|^{-1}\cdot\|\tilde{A}^{-1}\|\cdot\|u\|\cdot\|v\|]. \tag{6.3.19}$$

С помощью формулы (6.3.7) теперь нетрудно найти оценку погрешности решения системы (5.1.1) по методу окаймления.

§ 4. МЕТОД СОПРЯЖЕННЫХ НАПРАВЛЕНИЙ

1°. Приведем некоторые сведения о методе сопряженных направлений, который с достаточной полнотой изложен в [139]. Пусть A — неособенная матрица порядка m, а матрицы B и C того же порядка m таковы, что матрица

$$K = CAB \tag{6.4.1}$$

— симметричная и положительно определенная. Так, если A — положительно определенная матрица, то можно принять $B = C = I$; в общем случае можно положить $B = A^*$, $C = I$ или $B = I$, $C = A^*$. Построим систему K-ортонормированных векторов $\omega_1, \omega_2, \ldots, \omega_m$; это значит, что

$$(K\omega_i, \omega_j) = \delta_{ij}. \tag{6.4.2}$$

Решение x системы (5.1.1) будем строить в виде

$$x = \sum_{i=1}^{m} a_i B\omega_i. \tag{6.4.3}$$

Тогда

$$Ax = \sum_{i=1}^{m} a_i AB\omega_i = f,$$

или, если умножить это слева на C, $\sum_{i=1}^{m} a_i K\omega_i = Cf$. В силу соотношений (6.4.2) получаем искомое решение

$$x = \sum_{i=1}^{m} (Cf, \omega_i)\, B\omega_i.$$

Ограничимся одной из простейших возможностей и положим $B = A^*$, $C = I$. Тогда $K = AA^*$, и решение системы (5.1.1)

дается формулой

$$x = \sum_{i=1}^{m} (f, \omega_i) A^{*}\omega_i. \qquad (6.4.4)$$

Если матрица A симметричная и положительно определенная, то можно считать $K = A$. В этом случае формула для решения упрощается:

$$x = \sum_{i=1}^{m} (f, \omega_i)\,\omega_i.$$

2°. Векторы ω_i обычно строят, применяя известный прием ортогонализации в метрике, порожденной скалярным произведением $[x, y] = (Kx, y)$ и нормой $|x| = \sqrt{[x, x]} = \sqrt{(Kx, x)}$, к произвольно выбранному базису в R_m. Известно, что процесс ортогонализации может оказаться численно неустойчивым, а тогда применение формулы (6.4.4) может привести к большим погрешностям. Приведем некоторые относящиеся к этому вопросу соображения.

Пусть дана система m линейно независимых векторов φ_1, φ_2, ..., φ_m. В соответствии с процессом ортогонализации положим

$$\psi_1 = \varphi_1, \qquad \omega_1 = \psi_1/|\psi_1|, \qquad (6.4.5)$$

$$2 \leqslant k \leqslant m, \qquad \psi_k = \varphi_k - \sum_{j=1}^{k-1} [\varphi_k, \omega_j]\,\omega_j, \qquad \omega_k = \psi_k/|\psi_k|.$$

Как легко видеть, $|\psi_k| = |\varphi_k| \sin \alpha_k$, где α_k — угол в K-метрике между вектором φ_k и подпространством векторов φ_1, φ_2, ..., φ_{k-1}. Если векторы φ_1, φ_2, ..., φ_k близки к линейно зависимым, то $\min\limits_{2 \leqslant j \leqslant k} \sin \alpha_j$ будет малой величиной; можно принять указанный минимум за меру линейной независимости названных векторов. Мы покажем, что в метрике $|\cdot|$ эта мера может быть сколь угодно малой, даже если ортогонализируемые в этой метрике векторы ортогональны в R_m.

3°. Рассмотрим пример. Пусть $m = 2$ и

$$K = A = \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix}, \qquad a > 0.$$

Если $x = (x_1, x_2)$, $y = (y_1, y_2)$, то $[x, y] = x_1y_1 + ax_2y_2$, $|x|^2 = x_1^2 + ax_2^2$. Положим $\bar{x} = (\cos\theta, \sin\theta)'$, $\bar{y} = (-\sin\theta, \cos\theta)'$. Эти векторы ортогональны и нормированы в R_2. Оценим угол между этими векторами в метрике $[\ ,\ ]$. Имеем

$$\cos\alpha = \frac{[\bar{x}, \bar{y}]}{|\bar{x}| \cdot |\bar{y}|} = \frac{a - 1}{\sqrt{1 + a(\operatorname{tg}^2\theta + \operatorname{ctg}^2\theta) + a^2}}. \qquad (6.4.6)$$

Угол α будет минимальным, если $\theta = \pi/4$. В этом случае $\sin\alpha = 2\sqrt{a}/(a + \ldots)$. Если a достаточно велико или достаточно мало (в данном случае и то и другое означает, что число обу-

словленности матрицы $K = A$ достаточно велико), то угол между $\bar{x}$ и $\bar{y}$, который в метрике R_2 равен $\pi/2$, в метрике $[\ ,\ ]$ будет сколь угодно малым, и в этой метрике указанные векторы будут «почти линейно зависимыми». Этого не будет, если число обусловленности матрицы K будет не очень велико.

4°. Докажем следующее, более общее утверждение: если число обусловленности матрицы K произвольной размерности достаточно велико, то найдутся ортогональные в R_m векторы, угол между которыми в K-метрике сколь угодно мал. Ортогональным в R_m преобразованием L можно перевести матрицу K в диагональную. Обозначим последнюю через K', так что $K' = L^{-1}KL$. Если $(\bar{x}, \bar{y}) = 0$, то и $(L\bar{x}, L\bar{y}) = 0$. Кроме того, $(K'\bar{x}, \bar{y}) = (L^{-1}KL\bar{x}, \bar{y}) = (KL\bar{x}, L\bar{y})$; отсюда следует, что угол между векторами $L\bar{x}$ и $L\bar{y}$ в K-метрике совпадает с углом между $\bar{x}$ и $\bar{y}$ в K'-метрике. Следовательно, достаточно рассмотреть угол в K'-метрике между векторами, ортогональными в R_m. Имеем $K' = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$, где λ_i — собственные числа матрицы K; их можно расположить в порядке возрастания. Пусть $\bar{x} = (\xi_1, 0, \ldots, 0, \xi_m)$, $\bar{y} = (\bar{\eta}_1, 0, \ldots, 0, \bar{\eta}_m)$; так как эти векторы ортонормированы в R_m, то можно положить $\xi_1 = \eta_m = \cos\theta$, $\xi_m = -\eta_1 = \sin\theta$. Угол α между этими векторами в K'-метрике определяется формулой

$$\cos\alpha = \frac{(\lambda_m - \lambda_1)\cos\theta\sin\theta}{[(\lambda_1\cos^2\theta + \lambda_m\sin^2\theta)(\lambda_1\sin^2\theta + \lambda_m\cos^2\theta)]^{1/2}} =$$

$$= \frac{a-1}{[1 + a(\operatorname{tg}^2\theta + \operatorname{ctg}^2\theta) + a^2]^{1/2}};$$

здесь $a = \lambda_m/\lambda_1$ — число обусловленности матрицы K. Последняя формула совпадает с формулой (6.4.6), и для рассматриваемого случая верен вывод п. 3°: если число обусловленности матрицы K достаточно велико, то угол α будет сколь угодно мал.

5. Справедливо утверждение, в некотором смысле обратное только что доказанному: если векторы $\varphi_1, \varphi_2, \ldots, \varphi_m$ ортогональны в R_m, то

$$\sin\alpha_k \geqslant 1/\sqrt{a}, \tag{6.4.7}$$

где a — число обусловленности матрицы K. Обозначим через λ_1 и λ_m наименьшее и наибольшее собственные числа матрицы K. Тогда $a = \lambda_m/\lambda_1$, и если x — произвольный вектор из R_m, то

$$\sqrt{\lambda_1}\,\|x\|_{R_m} \leqslant |x| \leqslant \sqrt{\lambda_m}\,\|x\|_{R_m}. \tag{6.4.8}$$

Как сказано в п. 2° данного параграфа,

$$\sin\alpha_k = \frac{|\psi_k|}{|\varphi_k|} = \frac{1}{|\varphi_k|}\left|\varphi_k - \sum_{j=1}^{k-1}[\varphi_k, \omega_j]\,\omega_j\right|.$$

Сумма под знаком нормы есть некоторая линейная комбинация векторов $\varphi_1, \varphi_2, \ldots, \varphi_{k-1}$:

$$\sum_{j=1}^{k-1} [\varphi_k, \omega_j]\omega_j = \sum_{j=1}^{k-1} b_j\varphi_j.$$

Отсюда

$$\sin\alpha_k = \frac{1}{|\varphi_k|}\cdot\left|\varphi_k - \sum_{j=1}^{k-1} b_j\varphi_j\right| \geqslant \frac{\sqrt{\lambda_1}}{|\varphi_k|}\left\|\varphi_k - \sum_{j=1}^{k-1} b_j\varphi_j\right\| \geqslant$$

$$\geqslant \frac{\sqrt{\lambda_1}}{|\varphi_k|}\cdot\|\varphi_k\| \geqslant \frac{1}{\sqrt{a}},$$

что и требовалось доказать.

6°. Выберем векторы φ_j следующим образом: примем, что $\varphi_j = e_j/|e_j|$, где e_j — координатные векторы в R_m, так что $e_{jk} = \delta_{jk}$, $k = 1, 2, \ldots, m$. Для простоты допустим, что числа $|e_j|$ и векторы φ_j вычислены с достаточной точностью, так что их погрешностями можно пренебречь. Допустим, что для некоторого j оценки погрешностей $\delta(\omega_k)$, $k < j$, уже получены и нормы указанных погрешностей малы. Это допущение справедливо для $j = 2$, потому что тогда $\omega_k = \omega_1 = \varphi_1$, и по сделанному выше допущению норма $\|\delta(\varphi_1)\|$ пренебрежимо мала.

Приняв последнее допущение, займемся оценкой погрешностей величин $|\psi_j|$ и $1/|\psi_j|$. По формулам (6.4.5)

$$|\psi_j|^2 = |\varphi_j|^2 - \sum_{k=1}^{j-1} [\varphi_j, \omega_k]^2.$$

При вычислении величин $[\varphi_j, \omega_k]^2$ будут допущены малые погрешности $\delta([\varphi_j, \omega_k]^2)$. Последняя формула заменится такой:

$$(|\psi_j|^2)_м = |\psi_j|^2 + \delta(|\psi_j|^2) = 1 - \sum_{k=1}^{j-1}([\varphi_k, \omega_j]^2 + \delta([\varphi_k, \omega_j]^2)),$$

откуда

$$\delta(|\psi_j|^2) = -\delta\left(\sum_{k=1}^{j-1}[\varphi_j, \omega_k]^2\right).$$

Вычислим правую часть последнего равенства. Временно обозначим $[\varphi_j, \omega_k]^2 = \sigma_{jk}$. Имеем

$$\delta\left(\sum_{k=1}^{j-1}\sigma_{jk}\right) = \left(\sum_{k=1}^{j-1}(\sigma_{jk})_м\right)_м - \sum_{k=1}^{j-1}\sigma_{jk} =$$

$$= \left[\sum_{k=1}^{j-1}(\sigma_{jk})_м\right)_м - \sum_{k=1}^{j-1}(\sigma_{jk})_м\Big] + \Big[\sum_{k=1}^{j-1}(\sigma_{jk})_м - \sum_{k=1}^{j-1}\sigma_{jk}\Big].$$

Первая квадратная скобка справа есть погрешность суммы; если вычисления производятся с повышенной точностью, то по формуле (1.1.9) эта погрешность получит оценку

$$\left|\delta\left(\sum_{k=1}^{j-1}(\sigma_{jk})_м\right)\right| \leqslant \varepsilon_1\sum_{k=1}^{j-1}(\sigma_{jk})_м \approx \varepsilon_1\sum_{k=1}^{j-1}\sigma_{jk} =$$

$$= \varepsilon_1\sum_{k=1}^{j-1}[\varphi_j, \omega_k]^2 \leqslant \varepsilon_1|\varphi_j|^2 = \varepsilon_1.$$

Отметим, что при вычислении с обычной точностью получится оценка

$$\left|\delta\left(\sum_{k=1}^{j-1}(\sigma_{jk})_{\text{м}}\right)\right| \leqslant (j-2)\varepsilon_1.$$

Оценим вторую квадратную скобку. Обозначая $\tau_{jk} = [\varphi_j, \omega_k]$, имеем

$$\left|\sum_{k=1}^{j-1}(\sigma_{jk})_{\text{м}} - \sum_{k=1}^{j-1}\sigma_{jk}\right| \leqslant \sum_{k=1}^{j-1}|\delta(\tau_{jk}^2)|$$

и по формуле (1.1.1) $\delta(\tau_{jk}^2) \leqslant \varepsilon_1\tau_{jk}^2 = \varepsilon_1[\varphi_j, \omega_k]^2$. В результате вторая квадратная скобка получает оценку

$$2\varepsilon_1\sum_{k=1}^{j-1}[\varphi_j, \omega_k]^2 \leqslant 2\varepsilon_1|\varphi_j|^2 = 2\varepsilon_1.$$

Сложив оценки для обеих квадратных скобок, найдем $|\delta|\psi_j|^2| \leqslant 3\varepsilon_1$; если вычисления производятся с обычной точностью, то $|\delta(|\psi|)_j|^2 \leqslant j\varepsilon_1$. Чтобы учесть обе возможные оценки, обозначим через j величину, которая может равняться либо 3, либо j, и запишем общую оценку

$$|\delta(|\psi_j|^2)| \leqslant \hat{j}\varepsilon_1. \qquad (6.4.8а)$$

Отметим следствие из последней формулы:

$$(|\psi_j|^2)_{\text{м}} = |\psi_j|^2 + \theta\hat{j}\varepsilon_1, \quad |\theta| \leqslant 1.$$

Вычисляя величину $|\psi_j|$, мы будем извлекать корень не из точного значения величины $|\psi_j|^2$, а из ее машинного значения. Полученное таким образом значение корня обозначим через $|\psi_j|_{\text{м}}$, так что $|\psi_j|_{\text{м}} = (\sqrt{(|\psi_j|^2)_{\text{м}}})_{\text{м}}$. По неравенству (1.1.12)

$$|\delta(|\psi_j|_{\text{м}})| \leqslant \varepsilon_1\sqrt{(|\psi_j|^2)_{\text{м}}} = \varepsilon_1\sqrt{|\psi_j|^2 + \theta\hat{j}\varepsilon_1} \approx \varepsilon_1|\psi_j|.$$

Далее,

$$||\psi_j|_{\text{м}} - |\psi_j|| \leqslant |(\sqrt{(|\psi_j|^2)_{\text{м}}})_{\text{м}} - \sqrt{(|\psi_j|^2)_{\text{м}}}| + |\sqrt{(|\psi_j|^2)_{\text{м}}} - |\psi_j|| =$$

$$= |\delta(\sqrt{(|\psi_j|^2)_{\text{м}}})| + \frac{||\psi_j|^2_{\text{м}} - |\psi_j|^2|}{\sqrt{(|\psi_j|^2)_{\text{м}}} + |\psi_j|} \leqslant \varepsilon_1|\psi_j|_{\text{м}} +$$

$$+ \frac{\varepsilon_1\hat{j}}{\sqrt{|\psi_j|^2(1+\varepsilon_1/\theta)} + |\psi_j|} \approx \varepsilon_1\left(|\psi_j| + \frac{\hat{j}}{2|\psi_j|}\right) \leqslant \varepsilon_1\left(|\psi_j| + \frac{\hat{j}}{2}\sqrt{a}\right) \leqslant$$

$$\leqslant \varepsilon_1\left(|\psi_j| + \frac{\hat{j}}{2}P_A\right).$$

Отсюда, между прочим, следует, что

$$|\psi_j|_{\text{м}} = |\psi_j| + \varepsilon_1\theta(|\psi_j| + 2^{-1}\hat{j}P_A). \qquad (6.4.9)$$

Теперь можно оценить величину

$$|\delta(|\psi_j|^{-1})| = \left|\left(\frac{1}{|\psi_j|_{\text{м}}}\right)_{\text{м}} - \frac{1}{|\psi_j|}\right| \leqslant \left|\left(\frac{1}{|\psi_j|_{\text{м}}}\right)_{\text{м}} - \frac{1}{|\psi_j|_{\text{м}}}\right| + \left|\frac{1}{|\psi_j|_{\text{м}}} - \frac{1}{|\psi_j|}\right|.$$

Первое слагаемое справа оценивается по формуле (1.1.1): оно не превосходит величины $\varepsilon_1/|\psi_j|_{\text{м}} \approx \varepsilon_1/|\psi_j| \leqslant \varepsilon_1 P_A$. Далее, по формуле (6.4.9)

$$\left|\frac{1}{|\psi_j|_{\text{м}}} - \frac{1}{|\psi_j|}\right| \leqslant \varepsilon_1 \frac{|\psi_j| + 2^{-1}\hat{j} P_A}{|\psi_j| \cdot |\psi_j|_{\text{м}}} \approx \varepsilon_1 \left(\frac{1}{|\psi_j|} + \frac{2^{-1}\hat{j} P_A}{|\psi_j|^2}\right)$$

и окончательно

$$|\delta(1/|\psi_j|)| \leqslant \varepsilon_1 P_A (2 + 2^{-1}\hat{j} P_A^2). \tag{6.4.10}$$

Таким образом, погрешности величин $1/|\psi_j|$ малы, если число обусловленности матрицы A не очень велико, например, если $P_A \leqslant \varepsilon_1^{\alpha - 1/3}$, $\alpha > 0$.

7°. Решим вспомогательную задачу: оценим погрешность вектора

$$x = \sum_{k=1}^{j} \gamma_k \xi^{(k)}, \tag{6.4.11}$$

где γ_k — числа и $\xi^{(k)} = (\xi_i^{(k)})_{i=1}^m$ — векторы с m составляющими. Числа γ_k и векторы $\xi^{(k)}$ будем считать заданными точно. Положим

$$x_i = \sum_{k=1}^{j} \gamma_k \xi_i^{(k)}.$$

По формуле (1.1.9)

$$|\delta(x_i)| \leqslant \varepsilon_1 \sum_{k=1}^{j} |\gamma_k \xi_i^{(k)}| \leqslant \varepsilon_1 \|\gamma\| \left[\sum_{k=1}^{j} (\xi_i^{(k)})^2\right]^{1/2};$$

$$\|\gamma\|^2 = \sum_{k=1}^{j} \gamma_k^2,$$

откуда

$$\|\delta(x)\| = \left[\sum_{i=1}^{m} \left|\delta\left(\sum_{k=1}^{j} \gamma_k \xi_i^{(k)}\right)\right|^2\right]^{1/2} \leqslant \varepsilon_1 \|\gamma\| \left[\sum_{k=1}^{j} \|\xi^{(k)}\|^2\right]^{1/2}. \tag{6.4.12}$$

Оценим еще величину $|\delta(x)|$. Воспользуемся известным неравенством $\sqrt{\sigma_1}\,\|x\| \leqslant |x| \leqslant \sqrt{\sigma_m}\,\|x\|$, в котором σ_m и σ_1 суть соответственно наибольшее и наименьшее собственные числа матрицы K. Неравенство (6.4.12) умножим на $\sqrt{\sigma_m}$, возведем в квадрат и заменим левую часть меньшей величиной $|\delta(x)|$. В итоге получим

$$|\delta(x)|^2 \leqslant \varepsilon_1^2 \sigma_m \|\gamma\|^2 \sum_{k=1}^{j} \|\xi^{(k)}\|^2 \leqslant \varepsilon_1^2 \frac{\sigma_m}{\sigma_1} \|\gamma\|^2 \sum_{k=1}^{j} |\xi^{(k)}|^2 =$$

$$= \varepsilon_1^2 P_A^2 \|\gamma\|^2 \sum_{k=1}^{j} |\xi^{(k)}|^2,$$

или

$$|\delta(x)| \leqslant \varepsilon_1 P_A \|\gamma\| \left[\sum_{k=1}^{j} |\xi^{(k)}|\right]^{1/2}. \tag{6.4.13}$$

В частности, если $|\xi^{(k)}| = 1$, то

$$|\delta(x)| \leqslant \varepsilon_1 P_A \|\gamma\| \sqrt{j}. \qquad (6.4.14)$$

8°. Используя формулу (6.4.14), оценим погрешности векторов ω_j. По формуле (6.4.5)

$$\omega_j = \frac{\psi_j}{|\psi_j|} = \frac{\varphi_j}{|\psi_j|} - \sum_{\nu=1}^{j-1} \frac{[\varphi_j, \omega_\nu]}{|\psi_j|} \omega_\nu. \qquad (6.4.15)$$

Выражение (6.4.15) подходит под формулу (6.4.11), если заменить x на ω_j и положить $\gamma_1 = 1/|\psi_1|$, $\gamma_{\nu+1} = [\varphi_j, \omega_\nu]/|\psi_j|$, $\nu = 1, 2, \ldots, j-1$; $\xi^{(1)} = \varphi_j$, $\xi^{(\nu+1)} = \omega_\nu$, $\nu = 1, 2, \ldots, j-1$. Непосредственно применить формулу (6.4.14) нельзя, так как точные значения γ_k и $\xi^{(k)}$ неизвестны. Нетрудно, однако, видеть, что замена всех величин в правой части (6.4.15) их машинными значениями, которые известны точно, введет в оценку для $|\delta(\omega_j)|$ слагаемые порядка $\varepsilon_1^2 P_A^3$, которыми можно пренебречь, если P_A не очень велико. Но тогда можно воспользоваться формулой (6.4.14). Заметим, что в данном случае

$$\|\gamma\|^2 = |\psi_j|^{-2}\left[1 + \sum_{\nu=1}^{j-1} [\varphi_j, \omega_\nu]^2\right] \leqslant 2|\psi_j|^{-2} \leqslant 2P_A^2,$$

и формула (6.4.14) дает оценку

$$|\delta(\omega_j)| \leqslant \varepsilon_1 \sqrt{2m}\, P_A^2. \qquad (6.4.16)$$

9°. Теперь можно оценить погрешность формулы (6.4.4), дающей решение линейной алгебраической системы. В соответствии с этой формулой

$$x + \delta(x) = \left(\sum_{j=1}^{m} (f, \omega_j + \delta(\omega_j))\, A^*(\omega_j + \delta(\omega_j))\right)_{\text{м}}. \qquad (6.4.17)$$

Выражение под знаком $(\ldots)_{\text{м}}$ справа в (6.4.17) подходит под формулу (6.4.11), если положить $\gamma_j = (f, \omega_j + \delta(\omega_j))$, $\xi^{(j)} = A^*(\omega_j + \delta(\omega_j))$. По неравенству (1.1.9)

$$\|\delta(x)\| \leqslant \varepsilon_1 \left[\sum_{j=1}^{m} (f, \omega_j + \delta(\omega_j))^2\right]^{1/2} \cdot \left[\sum_{j=1}^{m} \|A^*(\omega_j + \delta(\omega_j))\|^2\right]^{1/2}.$$

Слагаемые $\delta(\omega_j)$ справа дадут в оценке погрешности члены порядка $\varepsilon_1^2 P_A^2$. Если P_A не очень велико, то их можно отбросить, что приведет нас к более простому неравенству

$$\|\delta(x)\| \leqslant \varepsilon_1 \|f\| \cdot \|A\| \left[\sum_{j=1}^{m} \|\omega_j\|^2\right]^{1/2}.$$

Далее,

$$\|\omega_j\|^2 \leqslant \sigma_1^{-1} |\omega_j|^2 = \|(AA^*)^{-1}\| = \|A^{-1}\|^2,$$

поэтому

$$\|\delta(x)\| \leqslant \varepsilon_1 \sqrt{m}\, P_A \|f\|. \qquad (6.4.18)$$

Из формул (6.4.10), (6.4.16) видно, что методом сопряженных направлений можно пользоваться для решения линейной алгебраической системы, и формула (6.4.18) верна с точностью до членов порядка малости более высокого, чем ε_1, если $P_A^3 \ll \varepsilon_1$. Достаточно, например, чтобы $P_A < \varepsilon_1^{-1/3+\alpha}$, $\alpha > 0$,

10°. Если вычисления производятся с обычной точностью, то погрешность суммы придется оценивать по формуле (1.1.6). В конечном счете это приведет к оценке

$$\|\delta(x)\| \leqslant \varepsilon_1 \sqrt{m}\,(m-1)\,P_A \|f\|. \tag{6.4.19}$$

Последняя оценка применима, если, например, $P_A \leqslant (m^{-1}\varepsilon_1)^{-1/3+\alpha}$, $\alpha > 0$.

11°. Для иллюстрации оценок, полученных в данном параграфе, была решена по методу сопряженных направлений система 12-го порядка $Ax = f$, где

$$A = \begin{bmatrix} 12 & 11 & 10 & 9 & \dots & 2 & 1 \\ 11 & 11 & 10 & 9 & \dots & 2 & 1 \\ 10 & 10 & 10 & 9 & \dots & 2 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 2 & 2 & 2 & 2 & \dots & 2 & 1 \\ 1 & 1 & 1 & 1 & \dots & 1 & 1 \end{bmatrix}, \qquad f = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 0 \end{bmatrix}. \tag{6.4.20}$$

Матрица A приведена в [26]; там же приведена и обратная матрица

$$A^{-1} = \begin{bmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & & \cdot & & & & \\ & & & \cdot & & & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 2 \end{bmatrix}. \tag{6.4.21}$$

Точное решение системы (6.4.20) очевидно: $x = (-1, 2, -1, 0, \dots, 0)'$. Система (6.4.20) решалась на ЭВМ класса ЕС с ординарной ($\varepsilon_1 \approx 10^{-7}$) и двойной ($\varepsilon_1 \approx 10^{-16}$) точностью, без использования регистра младших разрядов. Погрешности решения приведены в табл. 14. Здесь x_i^* — координаты точного решения, x_i — координаты приближенного решения. По данным таблицы вычислены нормы векторов погрешности:

$$\varepsilon_1 \approx 10^{-7}, \qquad \|\delta(x)\| = 1{,}07 \cdot 10^{-2};$$
$$\varepsilon_1 \approx 10^{-16}, \qquad \|\delta(x)\| = 4{,}9 \cdot 10^{-12}.$$

12° Проанализируем результаты вычислений. В данном случае $\|f\| = 1$, $\sqrt{m}\,(m-1) = 11\sqrt{12} < 22 \cdot 1{,}74 = 38{,}28$. Чтобы применить формулу (6.4 19), остается вычислить P_A. Этим мы и займемся. Достаточно найти наибольшее и наименьшее собственные числа матрицы (6.4.21); их отношение и даст нам число P_A. Собственные числа упомянутой матрицы суть корни уравнения

$$\mathrm{Det}\,(A^{-1} - \lambda I) = 0. \tag{6.4.22}$$

Таблица 14

t	$\varepsilon_1 \approx 10^{-7}$ $x_t - x_x^*$	$\varepsilon_1 \approx 10^{-16}$ $x_t - x_t^*$
1	$2{,}4 \cdot 10^{-3}$	$2{,}2 \cdot 10^{-13}$
2	$9{,}3 \cdot 10^{-4}$	$9{,}2 \cdot 10^{-13}$
3	$-5{,}7 \cdot 10^{-3}$	$-2{,}4 \cdot 10^{-12}$
4	$-3{,}5 \cdot 10^{-3}$	$-8{,}3 \cdot 10^{-13}$
5	$-5{,}9 \cdot 10^{-4}$	$2{,}4 \cdot 10^{-12}$
6	$1{,}5 \cdot 10^{-4}$	$-4{,}8 \cdot 10^{-13}$
7	$-9{,}5 \cdot 10^{-4}$	$1{,}8 \cdot 10^{-13}$
8	$-2{,}2 \cdot 10^{-3}$	$-1{,}9 \cdot 10^{-13}$
9	$4{,}5 \cdot 10^{-4}$	$2{,}4 \cdot 10^{-13}$
10	$-1{,}3 \cdot 10^{-3}$	$-1{,}6 \cdot 10^{-14}$
11	$-1{,}1 \cdot 10^{-3}$	$-6{,}3 \cdot 10^{-14}$
12	$-2{,}8 \cdot 10^{-4}$	$-2{,}8 \cdot 10^{-14}$

Это уравнение легко приводится к виду (полагаем $-\lambda + 2 = 2\cos\theta$) $\Delta_{12} - \Delta_{11} = 0$, где

$$\Delta_n = \begin{vmatrix} 2\cos\theta & 1 & & & & \\ 1 & 2\cos\theta & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & 2\cos\theta & 1 & \\ & & & 1 & 2\cos\theta & \end{vmatrix}_n ;$$

индекс внизу означает порядок определителя. Как известно (см [140], а также [84, 1-е изд., § 97, с. 462]), $\Delta_n = \sin(n-1)\theta/\sin\theta$, и уравнение (6.4.22) приводится к такому:

$$\sin(25\theta/2)/\sin(\theta/2) = 0, \tag{6.4.23}$$

его решения суть $\theta = 2k\pi/25$, $k = 1, 2, \ldots, 12$. Экстремальные значения λ равны

$$\lambda_{\min} = 2(1 - \cos(2\pi/25)), \quad \lambda_{\max} = 2(1 - \cos(24\pi/25)) = 2(1 + \cos(\pi/25)).$$

Отсюда

$$P_A = \frac{1 + \cos(\pi/25)}{1 - \cos(2\pi/25)} < 64.$$

Теперь, по формуле (6.4.19)

$$\|\delta(x)\| \leqslant \varepsilon_1 \cdot 38{,}28 \cdot 64 < 2456\varepsilon_1. \tag{6.4.24}$$

В соответствии с формулой (6.4.24) можно ожидать потери точности не более чем в четыре десятичных знака, между тем,

как мы видим, при точности $\varepsilon_1 \approx 10^{-7}$ теряются пять десятичных знаков. Это, видимо, объясняется тем, что в данном случае P_A^3 есть величина того же порядка, что и $(\varepsilon_1/m)^{-1/3}$, и формула (6.4.19) становится недостаточно надежной. Как показывает табл. 14, при $\varepsilon_1 \approx 10^{-16}$ теряются четыре десятичных знака, что согласуется с формулой (6.4.24).

§ 5. МЕТОД МАТРИЦ ОТРАЖЕНИЯ

1°. Этот метод был предложен Ф. Хаусхолдером; подробнее см. [139]. С. К. Годунов [39] дал такое видоизменение метода Хаусхолдера, которое сводит данную систему не к общей рекуррентной, а к двухдиагональной системе. В той же книге [39] дана оценка погрешности видоизмененного метода матриц отражения. Здесь метод матриц отражения исследуется в той форме, в какой он изложен в [139].

Существо метода таково. Пусть дана система (5.1.1). Строится ортогональная матрица E_1 порядка m, которая переводит первый столбец данной матрицы в новый столбец, у которого все элементы, начиная со второго, равны нулю. Первый элемент нового столбца отличен от нуля — в противном случае данная матрица была бы особенной. Далее строится вторая ортогональная матрица E_2, переводящая второй столбец матрицы E_1A в столбец, у которого все элементы, начиная с третьего, равна нулю. Продолжая этот процесс, мы получим ортогональную матрицу $E = E_{m-1}E_{m-2} \ldots E_1$, такую, что матрица $U = EA$ оказывается верхней треугольной, у которой все элементы главной диагонали отличны от нуля.

2°. За E_k принимаются некоторые матрицы отражения. Скажем об этом несколько подробнее. Пусть w — единичный вектор в R_m и Q — ортогональная к нему $(m-1)$-мерная плоскость, проходящая через начало координат. Матрица E_0 отражения относительно плоскости Q имеет вид

$$E_0 = I - 2ww', \tag{6.5.1}$$

где ww' есть матрица порядка m и ранга единица, полученная умножением столбца w на строку w':

$$ww' = \begin{bmatrix} w_1^2 & w_1w_2 & \ldots & w_1w_m \\ w_2w_1 & w_2 & \ldots & w_2w_m \\ \ldots & \ldots & \ldots & \ldots \\ w_mw_1 & w_mw_2 & \ldots & w_m^2 \end{bmatrix}.$$

Вектор w можно выбрать так, чтобы преобразование E_0 переводило любой заданный вектор s в вектор, параллельный заданному координатному вектору e_i; для этого достаточно

взять

$$w = \rho^{-1}(s - \alpha e_i), \qquad \alpha = \pm \|s\|, \qquad \rho = \|s - \alpha e_i\| = \\ = \sqrt{2\alpha^2 - 2\alpha(s, e_i)}. \tag{6.5.2}$$

Знак α целесообразно выбрать так, чтобы ρ имело большее значение: $\operatorname{sign}\alpha = -\operatorname{sign}(s, e_i)$.

За E_1 возьмем матрицу отражения, которая переводит первый столбец матрицы A в координатный вектор $(1, 0, \ldots, 0)$, умноженный на постоянную. Матрица E_1A имеет вид

$$E_1A = \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ \cdot & \cdot & \cdot \\ 0 & \times & \times \end{bmatrix}; \tag{6.5.3}$$

крестиками обозначены элементы, которые могут быть отличны от нуля (эту символику мы заимствовали из [39]). Обозначим через A_2 матрицу, полученную из матрицы (6.5.3) вычеркиванием первой строки и первого столбца. Построим матрицу отражения $\bar{E}_2$ порядка $m - 1$, такую, что первый столбец матрицы $\bar{E}_2A_2$ имеет только один ненулевой элемент и этот элемент расположен на первом месте. Матрица

$$E_2 = \begin{bmatrix} 1 & 0 \\ 0 & \bar{E}_2 \end{bmatrix}$$

— ортогональная в R_m; легко убедиться, что

$$E_2E_1A = \begin{bmatrix} \times & \times & \times & \ldots & \times \\ 0 & \times & \times & \ldots & \times \\ 0 & 0 & \times & \ldots & \times \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \times & \ldots & \times \end{bmatrix}.$$

Описанный процесс продолжаем до исчерпания. Нетрудно видеть, что связь между E_k и $\bar{E}_k$ дается формулой

$$E_k = \begin{bmatrix} I_{k-1} & 0 \\ 0 & \bar{E}_k \end{bmatrix},$$

в которой I_{k-1} — единичная матрица порядка $k - 1$, и что

$$EA = E_{m-1}E_{m-2} \ldots E_1A = \begin{bmatrix} \times & \times & \times & \ldots & \times \\ 0 & \times & \times & \ldots & \times \\ 0 & 0 & \times & \ldots & \times \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & \times \end{bmatrix}. \tag{6.5.4}$$

3°. В данном пункте мы изложим важный, на наш взгляд, результат С. К. Годунова [39] о погрешности произведения нескольких матриц, из которых одна произвольная, а остальные — ортогональные. Ограничимся случаем произведения (6.5.4).

Итак, пусть в произведении $B = E_{m-1}E_{m-2} \dots E_1 A$ матрица A — произвольная, а матрицы E_i — ортогональные. Положим

$$B_1 = E_1 A, \quad B_2 = E_2 B_1, \dots, \quad B_{m-1} = B = E_{m-1} B_{m-2}. \tag{6.5.5}$$

Умножения (6.5.5) совершаются с погрешностями, и вместо матриц B_i мы будем получать искаженные матрицы

$$\tilde{B}_i = E_i \tilde{B}_{i-1} + \Gamma_i, \quad i = 1, 2, \dots, m-1; \quad B_0 = A. \tag{6.5.6}$$

Справедливо следующее утверждение: если Q — ортогональная, а D — произвольная матрица, то

$$\| \delta(QD) \| \leqslant \eta \| D \|, \tag{6.5.7}$$

где η не зависит от D. В [39] выводится некоторая оценка числа η.

Из (6.5.6) вытекает представление

$$\tilde{B}_{m-1} = E_{m-1}E_{m-2} \dots E_1 A + \Gamma,$$

где

$$\Gamma = \Gamma_{m-1} + E_{m-1}\Gamma_{m-2} + E_{m-1}E_{m-2}\Gamma_{m-3} + \dots + E_{m-1}E_{m-2} \dots E_2 \Gamma_1.$$

При умножении на ортогональную матрицу норма матрицы не меняется, поэтому

$$\| \Gamma \| \leqslant \sum_{i=1}^{m-1} \| \Gamma_i \|. \tag{6.5.8}$$

По формуле (6.5.7) находим последовательно

$$\| \Gamma_1 \| \leqslant \eta \| A \|, \quad \| \tilde{B}_1 \| \leqslant (1+\eta) \| A \|;$$

$$\| \Gamma_2 \| \leqslant \eta \| \tilde{B}_1 \| \leqslant \eta(1+\eta) \| A \|; \quad \| \tilde{B}_2 \| \leqslant (1+\eta)^2 \| A \|;$$

. . . . . .

$$\| \Gamma_{m-1} \| \leqslant \eta (1+\eta)^{m-2} \| A \|.$$

Теперь

$$\| \Gamma \| \leqslant \eta \| A \| \sum_{i=0}^{m-2} (1+\eta)^i \leqslant \eta (1+\eta)^{m-2} (m-1) \| A \|.$$

С точностью до малых высших порядков можно принять $(1+\eta)^{m-2} = 1$ и, следовательно,

$$\| \Gamma \| \leqslant \eta (m-1) \| A \|. \tag{6.5.9}$$

4°. Простую оценку величины η дает формула (1.2.4). Для любой ортогональной матрицы Q, как легко видеть, $\| Q \|_E = \sqrt{m}$ и указанная формула дает

$$\eta \leqslant \varepsilon_1 \sqrt{m}. \tag{6.5.10}$$

Отсюда вытекает следующая оценка погрешности искажения матрицы EA:

$$\|\delta(EA)\| \leqslant \varepsilon_1 (m-1)\sqrt{m}\,\|A\|. \qquad (6.5.11)$$

5°. Повторяя рассуждения п. 3° и используя опять неравенство (5.5.10), находим искажение вектора Ef:

$$\|\delta\| := \|\delta(Ef)\| \leqslant \varepsilon_1 (m-1)\sqrt{m}\,\|f\|. \qquad (6.5.12)$$

По методу Хаусхолдера система (5.1.1) сводится к системе $EAx = Ef$, или, если учесть искажения матрицы EA и столбца Ef, к искаженной системе

$$(U+\Gamma)z = Ef + \delta, \quad U = EA. \qquad (6.5.13)$$

Если m не очень велико, то $\varepsilon_1 (m-1)\sqrt{m}$ есть величина малая, и можно воспользоваться формулами (5.3.6), (5.3.7), в которых следует заменить A на $U = EA$. Так как матрица E — ортогональная, то $\|U^{-1}\| = \|A^{-1}\|$, и можно воспользоваться упомянутыми формулами (5.3.6), (5.3.7), не меняя множителя $\|A^{-1}\|$ и заменяя $\|\Gamma\|$ и их оценками (6.5.11), (6.5.12). В результате получится следующая оценка погрешности искажения метода Хаусхолдера: если P_A — число обусловленности матрицы A и

$$\varepsilon_1 (m-1)\sqrt{m}\,P_A \leqslant \beta, \quad 0 \leqslant \beta < 1, \qquad (6.5.14)$$

то

$$\xi = \|z - x\| \leqslant \varepsilon_1 (1-\beta)^{-1}(m-1)\sqrt{m}. \qquad (6.5.15)$$

6°. Оценим теперь погрешность округления, возникающую при обратном ходе, т. е. при решении рекуррентной системы (6.5.13). Как и в § 4 главы 5, обозначим $U+\Gamma = \bar{U}$, $Ef+\delta = \bar{f}$ и запишем систему (6.5.13) в виде (5.4.1). Это приведет нас к соотношениям

$$v_k = z_k + \gamma_k, \quad (U+\Gamma)\gamma = \alpha, \qquad (6.5.16)$$

где α — вектор погрешностей, определенных соотношением (5.4.2). По аналогии с § 4 главы 5 имеем далее

$$\gamma = (U+\Gamma)^{-1}\alpha = (EA+\Gamma)^{-1}\alpha = (I + A^{-1}E^{-1}\Gamma)^{-1} A^{-1}E^{-1}\alpha.$$

Мы принимаем, что верны неравенства (6.5.11), (6.5.14), поэтому

$$\|A^{-1}E^{-1}\Gamma\| \leqslant \|A^{-1}\|\cdot\|\Gamma\| \leqslant \beta, \quad \|(I - A^{-1}E^{-1}\Gamma)^{-1}\| \leqslant (1-\beta)^{-1}$$

и $\|\gamma\| \leqslant (1-\beta)^{-1}\|A^{-1}\|\cdot\|\alpha\|$. Допуская, что $|\alpha_k| < \varepsilon$ и, следовательно, $\|\alpha\| \leqslant \varepsilon\sqrt{m}$, получаем оценку погрешности округления в методе Хаусхолдера:

$$\|\gamma\| \leqslant \|A^{-1}\|(1-\beta)^{-1}\sqrt{m}\,\varepsilon. \qquad (6.5.17)$$

§ 6. ОЦЕНКА НОРМЫ МАТРИЦЫ И ЕЕ ОБРАТНОЙ

1°. Как видно из содержания глав 5 и 6, оценки погрешностей решений линейных алгебраических систем весьма часто зависят от норм $\|A\|$ и $\|A^{-1}\|$, где A — матрица данной системы. Ниже, в § 4 главы 8, описан способ получения оценок сверху для этих норм, основанный на вычислении характеристического полинома матрицы A^*A, однако этот способ требует большого (порядка $O(m^4)$) числа арифметических действий. Здесь мы изложим некоторые соображения о том, как получить нужные оценки с помощью меньшего числа действий.

Просто находятся границы нормы $\|A\|$. Так, например, вычисление нормы $\|A\|_E$ требует $2m^2$ сложений и умножений плюс одно извлечение квадратного корня, и для операторной нормы справедлива двусторонняя оценка (см. [39], а также формулу (1.2.2) настоящей книги)

$$m^{-1/2}\|A\|_E \leqslant \|A\| \leqslant \|A\|_E. \tag{6.6.1}$$

2°. К сожалению, этот способ, видимо, нецелесообразен, когда речь идет о норме обратной матрицы, так как ее вычисление, например, по методу Леверье — Фаддеева (см. § 3 главы 8) или через решение систем с матрицей A и с координатными векторами в правых частях требует $O(m^4)$ арифметических действий. Границы для нормы $\|A^{-1}\|$ может дать следующий прием. Составим матрицу $B = A^*A$. Если A — неособенная, то B — положительно определенная матрица; обозначая через $\sigma_1 \leqslant \sigma_2 \leqslant \ldots \leqslant \sigma_m$ собственные числа, имеем

$$\|A\| = \sqrt{\sigma_m} = [\sup_{\|x\|=1}(Bx, x)]^{1/2} \qquad \|A^{-1}\| = \sigma_1^{-1/2} = [\inf_{\|x\|=1}(Bx, x)]^{-1/2}. \tag{6.6.2}$$

Заметим, что вычисление матрицы B требует $2m^3$ арифметических действий.

Пусть $B = [b_{ij}]_{i,j=1}^m$, тогда $(Bx, x) = \sum_{i,j=1}^m b_{ij}x_ix_j$ причем $b_{ii} > 0$. Положим $\bar{x} = (0, \ldots, 1, 0, \ldots, 0)$ с единицей на k-м месте. Тогда $b_{kk} = (B\bar{x}, \bar{x}) \geqslant \inf_{\|x\|=1}(Bx, x) = \sigma_1$. Отсюда

$$\|A^{-1}\| = 1/\sigma_1 \geqslant 1/\sqrt{b_{kk}}, \qquad \|A^{-1}\| \geqslant (\min b_{kk})^{-1}, \tag{6.6.3}$$

и мы получаем нижнюю границу нормы $\|A^{-1}\|$. Пусть K — какая-нибудь верхняя граница нормы $\|B\| = \|A\|^2$; можно, например, положить $K = \|A\|_E^2$. Составим матрицу $D = [d_{ij}]_{i,j=1}^m = KI - B$; ее наибольшее собственное число, которое мы обозначим через $\bar{\sigma}_m$, равно $K - \sigma_1$. Это число вычислим по итерационному методу [139, § 53], и пусть K_1 — полученное этим

методом достаточно точное значение $\bar{\sigma}_m$, так что можно считать $K_1 \geqslant K - \sigma_1 - \varepsilon$, где ε — достаточно малое число. Тогда $\sigma_1 \geqslant K - K_1 - \varepsilon$, и получается оценка нормы $\|A^{-1}\|$ сверху:

$$\|A^{-1}\| \leqslant 1/\sqrt{K - K_1 - \varepsilon}. \tag{6.6.4}$$

Глава 7

ИТЕРАЦИОННЫЕ МЕТОДЫ

§ 1. МЕТОД ПОСЛЕДОВАТЕЛЬНЫХ ПРИБЛИЖЕНИЙ

1°. В простейшем случае метод последовательных приближений осуществляется так: если в уравнении (5.1.1) $A = I - B$, так что этому уравнению можно придать вид $x = Bx + f$, то выбираем произвольный вектор x_0 и полагаем затем

$$x_{n+1} = Bx_n + f. \tag{7.1.1}$$

Для сходимости процесса (7.1.1) при произвольном x_0 необходимо и достаточно, чтобы было $\rho = \rho(B) < 1$; $\rho(B)$ — спектральный радиус оператора B:

$$\rho(B) = \lim_{n\to\infty} \|B^n\|^{1/n}. \tag{7.1.2}$$

Для матриц это условие равносильно такому: все собственные числа матрицы B меньше единицы по модулю. В частности, процесс (7.1.1) сходится, если $\|B\| < 1$. По поводу высказанных здесь утверждений см. [139].

2°. С процессом последовательных приближений связаны погрешности алгоритма и округления; погрешности аппроксимации и искажения отсутствуют. Погрешность алгоритма оценивается просто: полагая для простоты $x_0 = f$, получаем

$$x_* = \sum_{n=0}^{\infty} B^n f. \tag{7.1.3}$$

Отсюда

$$\zeta_n = \|x_* - x_n\| = \Big\|\sum_{k=n+1}^{\infty} B^k f\Big\| \leqslant \|f\| \sum_{k-n+1}^{\infty} \|B^k\|.$$

Выберем такое число $\varepsilon > 0$, чтобы было $\rho + \varepsilon < 1$; при достаточно больших k будет $\|B^k\| \leqslant (\rho + \varepsilon)^k$ и

$$\zeta_n \leqslant \|f\| \frac{(\rho+\varepsilon)^{n+1}}{1-\rho-\varepsilon}. \tag{7.1.4}$$

В частности, если $\|B\| < 1$, то

$$\zeta_n \leqslant \|f\| \frac{\|B\|^{n+1}}{1-\|B\|}. \tag{7.1.5}$$

3°. Рассмотрим погрешность округления для процесса (7.1.1). Для общих рекуррентных процессов оценка погрешности округления дана в главе 4. В случае процесса (7.1.1) вывод этой оценки можно упростить и уточнить Приведем этот вывод.

Прежде всего заметим, что

$$\| x_n \| \leqslant \sum_{k=0}^{\infty} \| B^k \| \cdot \| f \|$$

— это сразу вытекает из формулы $x_n = \sum_{k=0}^{n} B^k f$ и из сходимости ряда $\sum_{k=0}^{\infty} \| B^k \|$.

Пусть теперь $v_0 = x_0 = f$, v_1, v_2, ..., v_{k-1} — точно известные векторы. Обозначим через $\bar{x}_k$ точное значение вектора

$$\bar{x}_k = B v_{k-1} + f \tag{7.1.6}$$

и через v_k — машинное значение правой части последнего равенства. Пусть еще $\alpha_k = v_k - \bar{x}_k$. Величины справа в (7.1.6) все заданы точно, и из формулы (1.1.9) легко вытекает, что

$$\| \alpha_k \| \leqslant \varepsilon_1 (\| B \| \cdot \| v_{k-1} \| + \| f \|). \tag{7.1.7}$$

Найдем связь между v_l и x_l. Прежде всего элемент $x_0 = f$ задан точно, поэтому $v_0 = \bar{x}_0 = x_0$. Далее,

$$v_1 = \bar{x}_1 + \alpha_1 = B v_0 + f + \alpha_1 = x_1 + \alpha_1;$$

$$v_2 = \bar{x}_2 + \alpha_2 = B v_1 + f + \alpha_2 = B x_1 + f + B\alpha_2 + \alpha_2 = x_2 + B\alpha_1 + \alpha_2;$$

по индукции

$$v_l = x_l + \sum_{i=1}^{l} B^{l-i} \alpha_i. \tag{7.1.8}$$

Подставим (7.1.8) в (7.1.7):

$$\| \alpha_k \| \leqslant \varepsilon_1 \left\{ \| B \| \left[\| x_{k-1} \| + \max_{i \leqslant k-1} \| \alpha_i \| \sum_{i=1}^{k-1} \| B^{k-1-i} \| \right] + \| f \| \right\} \leqslant$$

$$\leqslant \varepsilon_1 (C_1 + C_2 \max_{i \leqslant k} \| \alpha_i \|). \tag{7.1.9}$$

Очевидно, можно положить

$$C_1 = \left[\| B \| \sum_{k=0}^{\infty} \| B^k \| + 1 \right] \| f \|, \quad C_2 = \sum_{k=0}^{\infty} \| B^k \|. \tag{7.1.10}$$

В частности, если $\| B \| = q < 1$, то

$$C_1 = [\| B \| / (1 - q) + 1] \| f \|, \quad C_2 = 1/(1 - q). \tag{7.1.11}$$

Допустим, что $\varepsilon_1 C_2 < 1$, и пусть $\max\limits_{i \leqslant k} \| \alpha_i \| = \alpha_l$, $l \leqslant k$. Если $l = k$, то сразу получается оценка погрешности округления:

$$\| \alpha_k \| \leqslant \varepsilon_1 C_1 / (1 - \varepsilon_1 C_2). \tag{7.1.12}$$

Если же $l < k$, то, заменив в (7.1.9) k на l, мы увидим, что α_l удовлетворяет неравенству (7.1.12). При этом, если $i \leqslant k$, то $\|\alpha_i\| \leqslant \|\alpha_l\|$, и α_k удовлетворяет тому же неравенству, которое таким образом установлено для всех k.

Если C_2 не очень велико, то погрешности округления удовлетворяют с точностью до величин высшего порядка малости более простому неравенству

$$\|\alpha_k\| \leqslant C_1 \varepsilon_1. \tag{7.1.13}$$

§ 2. ПРЕОБРАЗОВАНИЕ СИСТЕМЫ

1°. Во многих случаях можно применять метод последовательных приближений не к данной системе, а к системе, полученной из данной некоторым преобразованием. Чаще всего используют такое преобразование: система (5.1.1.) заменяется системой, ей эквивалентной и, вообще говоря, зависящей от целочисленного параметра n:

$$x = x + H_n(f - Ax), \tag{7.2.1}$$

где H_n — неособенная матрица. Последовательные приближения определяются по формуле

$$x_{n+1} = x_n + H_n(f - Ax_n), \tag{7.2.2}$$

вектор x_0 выбирается произвольно. Для сходимости процесса (7.2.2) необходимо и достаточно, чтобы

$$(I - H_k A)(I - H_{k-1}A) \ldots (I - H_1 A) \xrightarrow[k\to\infty]{} 0. \tag{7.2.3}$$

В частности, матрицы H_n могут быть шаровыми, тогда их можно рассматривать как скалярные множители. Другой интересный случай — это случай постоянной матрицы $H_n = H$. Несколько примеров такого рода будут рассмотрены в настоящем параграфе ниже.

Вопросы, затронутые в настоящем пункте, подробно освещены в [139].

2°. Пусть матрица A в уравнении (5.1.1) — положительно определенная, λ_1 и λ_m — ее собственные числа, соответственно наименьшее и наибольшее. Положим [124] $H_n = H = 2I(\lambda_1 + \lambda_m)^{-1}$, так что система (7.2.1) принимает вид

$$x = [I - 2(\lambda_1 + \lambda_m)^{-1} A]\, x + 2(\lambda_1 + \lambda_m)^{-1} f. \tag{7.2.4}$$

Верхняя и нижняя грани оператора $I - 2(\lambda_1 + \lambda_m)^{-1}A$ равны соответственно

$$1 - \frac{2\lambda_1}{\lambda_1 + \lambda_m} = \frac{\lambda_m - \lambda_1}{\lambda_1 + \lambda_m};\quad 1 - \frac{2\lambda_m}{\lambda_1 + \lambda_m} = -\frac{\lambda_m - \lambda_1}{\lambda_1 + \lambda_m}.$$

Отсюда следует, что

$$\left\| I - \frac{2A}{\lambda_1 + \lambda_m} \right\| = \frac{\lambda_m - \lambda_1}{\lambda_1 + \lambda_m} = \frac{P_A - 1}{P_A + 1} := q < 1; \tag{7.2.5}$$

здесь $P_A = \lambda_m/\lambda_1$ — число обусловленности матрицы A. Из неравенства (7.2.5) вытекает, что последовательные приближения для системы (7.2.4) сходятся как прогрессия с знаменателем q. Найдем выражения коэффициентов, входящих в оценки погрешностей алгоритма и округления. Формула (7.1.5) сразу дает

$$\zeta_n \leqslant \frac{2q^{n-1}}{1-q}\,\frac{\|f\|}{\lambda_1+\lambda_m}. \tag{7.2.6}$$

Далее, по формулам (7.1.10), (7.1.11) имеем

$$C_2 = \frac{1}{1-q}\,\frac{2\|f\|}{\lambda_1+\lambda_m},\quad B = I - \frac{2A}{\lambda_1+\lambda_m},$$

$$C_1 = \|\hat{B}\|\max_k \|x_k\| + C_2 \leqslant \left[\frac{\|\hat{B}\|}{1-q}+1\right]\frac{2\|f\|}{\lambda_1+\lambda_m}.$$

При увеличении C_1 и C_2 оценки погрешности остаются верными поэтому можно положить

$$C_1 = \left[\frac{\|\hat{B}\|}{1-q}+1\right]\frac{2\|f\|}{\lambda_1+\lambda_m},\quad C_2 = \frac{1}{1-q}\,\frac{2\|f\|}{\lambda_1+\lambda_m}. \tag{7.2.7}$$

Соотношения (7.2.7) вместе с неравенствами (7.1.12), (7.1.13) оценивают погрешность округления в рассматриваемом случае.

Иногда (например, при использовании метода Ритца) приходится сталкиваться с такой ситуацией, когда числа λ_1 и λ_m неизвестны, но известны числа $0 < \lambda_0 < \lambda_1$ и $\Lambda_0 > \lambda_m$. В этом случае можно заменить систему (7.2.4) следующей:

$$x = \left[I - \frac{2A}{\lambda_0+\Lambda_0}\right]x + \frac{2}{\lambda_0+\Lambda_0}f. \tag{7.2.8}$$

Легко убедиться, что

$$\left\|I - \frac{2A}{\lambda_1+\Lambda_0}\right\| \leqslant q_0 := \frac{\Lambda_0-\lambda_0}{\Lambda_0+\lambda_0} < 1,$$

и последовательные приближения сходятся. Легко также получить в этом случае оценки погрешностей алгоритма и округления.

Преобразование (7.2.4) было предложено И. П. Натансоном [124] применительно к произвольным ограниченным самосопряженным и положительно определенным операторам в произвольном гильбертовом пространстве.

3°. Одношаговый циклический процесс, или процесс Зайделя, подробно рассмотрен в [139], и мы здесь ограничимся самыми общими указаниями. Одношаговый циклический процесс есть итерационный процесс, сформулированный для систем вида $x = Bx + f$. Пусть $B = [b_{ij}]_{i,j=1}^{n}$; тогда упомянутый

процесс определяется формулой

$$x_i^{(n)} = \sum_{j=1}^{i-1} b_{ij} x_j^{(n)} + \sum_{j=i}^{m} b_{ij} x_j^{(n-1)} f_i. \qquad (7.2.9)$$

Положим

$$M = \begin{bmatrix} 0 & & & \\ b_{21} & 0 & & \\ \cdots & \cdots & \cdots & \cdots \\ b_{m1} & b_{m2} & \ldots & b_{m,\,m-1}\ 0 \end{bmatrix}, \quad N = \begin{bmatrix} b_{11} & b_{12} & \ldots & b_{1m} \\ & b_{22} & \ldots & b_{2m} \\ & \cdots & \cdots & \cdots \\ & & & b_{mm} \end{bmatrix},$$

тогда соотношение (7.2.9) можно записать в виде $x_n = Mx_n + Nx_{n-1} + f$, или, что равносильно,

$$x_n = (I - M)^{-1} N x_{n-1} + (I + M)^{-1} f.$$

Таким образом, одношаговый циклический процесс есть обычный процесс последовательных приближений для системы

$$x = (I - M)^{-1} N x + (I - M)^{-1} f,$$

равносильной системе $x = Bx + f$. Необходимым и достаточным условием сходимости этого процесса является условие, что спектральный радиус оператора $(I - M)^{-1}N$ меньше единицы: оно равносильно условию, чтобы модули всех корней уравнения

$$\begin{vmatrix} b_{11} - t & b_{12} & b_{13} & \ldots & b_{1m} \\ b_{21}t & b_{22} - t & b_{23} & \ldots & b_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ b_{m1}t & b_{m2}t & b_{m3}t & \ldots & b_{mm} - t \end{vmatrix} = 0$$

были меньше единицы.

Погрешности алгоритма и округления для одношагового циклического процесса оцениваются по формулам (7.1.10)—(7.1.13), в которых соответствующие константы должны быть вычислены для столбца свободных членов $(I - M)^{-1}f$ и матрицы $(I - M)^{-1}N$.

В [139] дано достаточное условие сходимости одношагового циклического процесса, значительно проще проверяемое, чем сформулированное выше условие.

4°. В статье [175] рассмотрен класс уравнений вида (5.1.1), обладающих следующими свойствами:

$$M_1 \|x\|^2 \leqslant (Ax, x) \leqslant M_2 \|x\|^2, \quad 0 < M_1 \leqslant M_2 < \infty, \qquad (7.2.10)$$

$$\|A - A^*\| \leqslant 2k, \quad k \geqslant 0, \qquad (7.2.11)$$

где числа M_1, M_2, k фиксированы. Неравенство (7.2.10) означает, что симметричная часть матрицы A, равная $(A + A^*)/2$, — положительно определенная. Система (5.1.1), матрица которой удовлетворяет условиям (7.2.10), (7.2.11), решается мето-

дом последовательных приближений, которые строятся по формуле

$$x_{n+1} = x_n - p(Ax_n - f), \tag{7.2.12}$$

где p — некоторое число. Доказывается, что при подходящем выборе этого числа приближения (7.1.12) сходятся как прогрессия с знаменателем $q < 1$; формулы для p и q таковы:

$$p = \begin{cases} M_1 (M_1^2 + k^2)^{-1}, & 2k^2 \geqslant (M_2 - M_1) M_1, \\ 2 (M_1 + M_2)^{-1}, & 2k^2 < (M_2 - M_1) M_1; \end{cases} \tag{7.2.13}$$

$$q = \begin{cases} k (M_1^2 + k^2)^{-1/2}, & 2k^2 \geqslant (M_2 - M_1) M_1, \\ [4k^2 + (M_2 - M_1)^2]^{1/2} (M_1 + M_2)^{-1}, & 2k^2 < (M_2 - M_1) M_1. \end{cases} \tag{7.2.14}$$

Отметим, что при $k = 0$, т. е. когда матрица A — положительно определенная, формулы (7.2.13), (7.2.14) совпадают с точностью до обозначений с соответствующими формулами п. 2°, так что результаты статьи [175] можно рассматривать как обобщение результатов Натансона на случай не положительно определенных матриц. Вместе с тем необходимо сказать, что это обобщение нетривиально.

Оценки для погрешностей алгоритма и округления сразу получаются из формул (7.1.5), (7.1.12), (7.1.13), если заменить в них B и f на $I - pA$ и pf.

5°. Коротко остановимся на методе релаксации. Пусть A — положительно определенная матрица и x — произвольный вектор, который мы будем рассматривать как приближенное решение системы (5.1.1); ее точное решение пусть будет x_*. Погрешность приближенного решения x можно оценивать так называемой функцией ошибок $F(x) = (A(x - x_*), x - x_*)$. Поставим задачу: найти вектор x', который отличался бы от вектора x только i-й составляющей, так, чтобы $F(x') < F(x)$. Положим $x' = x + \alpha e_i$; e_i — i-й координатный вектор. Легко доказать, что

$$F(x') = F(x) + a_{ii}^{-1} (\alpha a_{ii} - r_i)^3 - r_i^2 a_{ii}^{-1}; \tag{7.2.15}$$

здесь a_{ii} — i-й элемент главной диагонали матрицы A и r_i — i-я составляющая невязки вектора x: $r_i = (Ax - f, e_i)$. Мы хотим, чтобы сумма второго и третьего членов справа в (7.2.15) была отрицательной; это будет, если $|\alpha a_{ii} - r_i| < |r_i|$, или

$$\alpha = q r_i / a_{ii}, \quad 0 < q < 2. \tag{7.2.16}$$

При этом изменение функции ошибок равно

$$F(x') - F(x) = -q (2 - q) r_i^2 a_{ii}^{-1}. \tag{7.2.17}$$

Выбор q в указанных пределах произволен. Если $q = 1$, то говорят о полной релаксации, так как в этом случае убыва-

ние функции ошибок максимально. При $q \neq 1$ говорят о неполной релаксации; если $q < 1$, то релаксацию называют нижней, если $q > 1$, то — верхней.

Если процесс релаксации циклический, так что значение q_i множителя релаксации q, соответствующее значению i, $1 \leqslant i \leqslant m$, при дальнейшем возрастании i циклически повторяется, то процесс релаксации оказывается частным случаем процесса последовательных приближений. Именно, если D и Q — диагональные матрицы, $D = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{mm})$ и $Q = \operatorname{diag}(q_1, q_2, \ldots, q_m)$, то процесс циклической релаксации есть процесс последовательных приближений для системы

$$x = (I - QD^{-1}A)x + QD^{-1}f, \qquad (7.2.18)$$

равносильной системе (5.1.1). Если упомянутый процесс сходится (в [139] сходимость доказана для нижней релаксации), то погрешности алгоритма и округления можно оценить по формулам (7.1.5) и (7.1.12), (7.1.13), заменив в них B на $QD^{-1}A$ и f на $QD^{-1}f$.

§ 3. МЕТОДЫ НАИСКОРЕЙШЕГО СПУСКА И МИНИМАЛЬНЫХ НЕВЯЗОК

1°. Погрешности метода наискорейшего спуска исследованы в главе 3, § 8, и в главе 4, § 4 для случая, когда в уравнении (5.1.1) A есть ограниченный положительно определенный оператор в произвольном гильбертовом пространстве. Следовательно, этот метод применим и к линейным алгебраическим системам с положительно определенными матрицами. Все выводы, сделанные в указанных параграфах, для таких систем остаются в силе.

К сказанному можно добавить следующее. В теории линейных алгебраических систем известен метод сопряженных градиентов (метод Ланцоша); этот метод подробно изложен в учебнике [6]. По методу Ланцоша последующее приближение x_{k+1} к решению системы (5.1.1) с положительно определенной матрицей A строится по предшествующему приближению следующим образом. Выбирается число $n < m$ (m — порядок матрицы A) и за x_{k+1} принимается выражение

$$x_{k+1} = x_k + \sum_{i=0}^{n-1} \alpha_i^{(k)} A^i r_k, \quad r_k = Ax_k - f,$$

в котором коэффициенты $\alpha_i^{(k)}$ определяются из условия $(Ax_{k+1}, x_{k+1}) - 2(f, x_{k+1}) = \min$. При $n = 1$ эти формулы тождественны с соответствующими формулами метода наискорейшего спуска, который, следовательно, можно рассматривать как частный случай метода сопряженных градиентов.

Для линейных алгебраических систем метод Ланцоша — точный: если матрица A имеет $q \leqslant m$ различных собственных

чисел, то q итераций приводят к точному решению (разумеется, если все итерации выполнены без погрешностей); подробнее об этом см. [6]. Во всяком случае достаточно выполнить m итераций. Это, конечно, верно и для метода наискорейшего спуска. Но тогда погрешности искажения и округления для этого метода можно оценивать по формулам (3.8.7) и (4.4.4), если заменить в них n на m.

2°. Метод минимальных невязок был предложен М. А. Красносельским и С. Г. Крейном в 1952 г. для систем с положительно определенными матрицами. Ю. А. Кузнецов указал на применимость этого метода к некоторым классам несимметричных матриц. А. А. Самарский доказал сходимость метода. В. С. Козякин и М. А. Красносельский изложили в [175] метод минимальных невязок в достаточно общем виде, дали новое доказательство сходимости, пригодное для соответствующих классов операторов в произвольном гильбертовом пространстве; в [175] дана и библиография вопроса.

Пусть система (5.1.1) удовлетворяет условиям (7.2.10), (7.2.11). По методу минимальных невязок для решения упомянутой системы строится итерационный процесс

$$x_{n+1} = x_n - p_n(Ax_n - f); \tag{7.3.1}$$

численный множитель p_n выбирается так, чтобы норма невязки $r_{n+1} = Ax_{n+1} - f$ была минимальной. Из (7.3.1) находим

$$r_{n+1} = Ax_n - f - p_nA(Ax_n - f) = r_n - p_nAr_n.$$

Отсюда

$$\|r_{n+1}\|^2 = \|r_n\|^2 - 2p_n(Ar_n, r_n) + p_n^2\|Ar_n\|^2,$$

и норма $\|r_{n+1}\|$ будет минимальной, если

$$p_n = (Ar_n, r_n)/\|Ar_n\|^2. \tag{7.3.2}$$

Доказывается, что если $m \geqslant 2$, то

$$\|r_{n+1}\| \leqslant q\|r_n\|, \tag{7.3.3}$$

где

$$q = \frac{M_2^2 - M_1^2 + 4k\left(k^2 + M_1M_2\right)^{1/2}}{(M_1 + M_2)^2 + 4k^2}; \tag{7.3.4}$$

существенно, что $q < 1$. Действительно, разность между знаменателем и числителем оценивается так:

$$(M_1 + M_2)^2 + 4k^2 - M_2^2 + M_1^2 - 4k\sqrt{k^2 + M_1M_2} =$$
$$= 2M_1(M_1 + M_2) - 4k\left(\sqrt{k^2 + M_1M_2} - k\right) =$$
$$= 2M_1(M_1 + M_2) - \frac{4kM_1M_2}{\sqrt{4k^2 + M_1M_2} + k} >$$
$$> 2M_1(M_1 + M_2) - 2M_1M_2 = 2M_1^2 > 0.$$

Отсюда следует, что метод минимальных невязок сходится для любой матрицы, удовлетворяющей условиям (7.2.10), (7.2.11), точнее, для любой матрицы с положительно определенной симметричной частью.

3°. Покажем, что если $k = 0$, так что матрица A — положительно определенная, то метод минимальных невязок для уравнения (5.1.1) совпадает с методом наискорейшего спуска для равносильного уравнения $A^{3/2}x = A^{1/2}f$. Если $k = 0$, то

$$p_n = \frac{(Ar_n, r_n)}{(A^2r_n, r_n)} = \frac{\| A^{1/2}r_n \|^2}{(AA^{1/2}r_n, A^{1/2}r_n)}.$$

Обозначим $A^{1/2}x = \tilde{x}$, $A^{1/2}f = \tilde{f}$, $A^{1/2}r_n = \tilde{r}_n$. Уравнение $A^{3/2}x = A^{1/2}f$ переходит в уравнение $A\tilde{x} = \tilde{f}$. Отметим, что $\tilde{r}_n$ есть невязка уравнения $A\tilde{x} = \tilde{f}$ при подстановке x_n вместо x. В новых обозначениях

$$p_n = \| \tilde{r}_n \|^2 / (A\tilde{r}_n, \tilde{r}_n),$$

что с точностью до обозначений совпадает с σ_n (формула (3.8.2)), и уравнение (7.3.1) переходит в уравнение, определяющее процесс приближений для метода наискорейшего спуска.

Отметим еще, что в общем случае, когда $k \geqslant 0$, число p_n при больших n есть частное двух малых чисел, которое, следовательно, надо вычислять с большой точностью.

Очевидно, что если при любом n будет $\|I - p_nA\| \leqslant \bar{q} < 1$, то погрешности искажения и округления процесса (7.3.1) будут ограниченными; этого может не быть, если последнее неравенство неверно.

§ 4. МЕТОД РИЧАРДСОНА

В настоящем параграфе мы кратко излагаем содержание части статьи [172] с точки зрения результатов главы 4.

В пространстве R_m рассмотрим векторное уравнение

$$(I - B)x = f. \tag{7.4.1}$$

Здесь I — единичная матрица в R_m, B — симметричная матрица порядка $m \times m$, спектральный радиус которой $\rho < 1$; следовательно, существует ограниченный оператор $(I - B)^{-1}$. Примем, что матрица B и вектор f заданы точно, без погрешности. По методу Ричардсона решение уравнения (7.4.1) строится как предел векторов $x^{(n)}$, которые вычисляются по двухшаговому рекуррентному процессу

$$x^{(n)} - \omega Bx^{(n-1)} + (\omega - 1)x^{(n-2)} = \omega f, \tag{7.4.2}$$

$$\omega = 2/(1 + \sqrt{1 - \rho^2}), \tag{7.4.3}$$

начальные приближения $x^{(0)}$ и $x^{(1)}$ задаются произвольно.

Рассмотрим несколько более общий рекуррентный процесс

$$t^{(n)} - \omega B t^{(n-1)} + (\omega - 1) t^{(n-2)} = h^{(n)} \qquad (7.4.4)$$

и докажем, что для этого процесса справедливо соотношение (3.7.4). С этой целью построим решение конечной системы

$$k \leqslant n, \quad t^{(k)} - \omega B t^{(k-1)} + (\omega - 1) t^{(k-2)} = h^{(k)}, \qquad (7.4.5)$$

$t^{(0)}$ и $t^{(1)}$ считаем заданными произвольно.

Матрицу B можно представить в виде $B = Q\Lambda Q^*$, где $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ — диагональная матрица собственных чисел матрицы B, Q — ортогональная матрица, составленная из собственных векторов матрицы B, Q^* — сопряженная с Q матрица. В [172] доказывается, что решение системы (7.4.5) можно представить в виде

$$k \leqslant n, \quad t^{(k)} = \sum_{j=1}^{k-1} Q S_{k-j} Q^* h^{(j)}, \qquad (7.4.6)$$

где S_ν — диагональные матрицы, элементы которых, лежащие на главной диагонали, определяются так: если $\theta_j = \arccos(\lambda_j/\rho)$ и $\omega - 1 = r^2$, то

$$(S_\nu)_{jj} = \begin{cases} r^{\nu-1} \dfrac{\sin \nu\theta_j}{\sin \theta_j}, & \theta_j \neq 0, \pi, \\ \nu r^{\nu-1}, & \theta_j = 0, \\ (-1)^{\nu-1} \nu r^{\nu-1}, & \theta_j = \pi. \end{cases}$$

Отсюда $\|QS_{k-j}Q^*h^{(j)}\| \leqslant (k-j) r^{k-j-1} \|h^{(j)}\|$, и по формуле (7.4.6)

$$k \leqslant n, \quad \|t^{(k)}\| \leqslant \max_{j \leqslant n} \|h^{(j)}\| \sum_{j=1}^{k} (k-j) r^{k-j-1}. \qquad (7.4.7)$$

Так как $0 < r < 1$, то сумма в (7.4.7) ограничена. Но тогда условие (3.7.4) выполнено, и можно оценить погрешность округления для процесса Ричардсона (7.4.2).

Решая бесконечную систему (7.4.2), мы из-за погрешностей округления вместо последовательности $\{x^{(n)}\}$ получаем некоторую другую последовательность $\{v^{(n)}\}$, которая точно удовлетворяет другой бесконечной системе

$$v^{(n)} - \omega B v^{(n-1)} + (\omega - 1) v^{(n-2)} = \omega f + \delta^{(n)} \qquad (7.4.8)$$

Допуская, например, что вычисления ведутся с фиксированной абсолютной погрешностью ε, имеем $\|\delta^{(n)}\| \leqslant \varepsilon$. Из результатов § 3 главы 4 вытекает, что

$$\|\gamma^{(n)}\| := \|v^{(n)} - x^{(n)}\| \leqslant C\varepsilon, \quad C = \text{const}. \qquad (7.4.9)$$

Постоянную C легко определить. Вычитаем (7.4.2) из (7.4.8):

$$\gamma^{(n)} - \omega B \gamma^{(n-1)} + (\omega - 1) \gamma^{(n-2)} = \delta^{(n)}.$$

Сравнивая это с (7.4.5), находим по неравенству (7.4.7)

$$\|\gamma^{(n)}\| \leqslant \max_{j \leqslant n} \|\delta^{(j)}\| \sum_{j=1}^{n} (n-j) r^{n-j-1} \leqslant$$

$$\leqslant \left[\frac{1-r^n}{(1-r)^2} - \frac{r^{n-1}}{1-r} \right] \max_{j \leqslant n} \|\delta^{(j)}\| \leqslant C\varepsilon, \qquad (7.4.10)$$

$$C = \frac{1-r^n}{(1-r)^2} - \frac{r^{n-1}}{1-r};$$

несколько увеличив выражение в квадратной скобке, придем к неравенству

$$\|\gamma^{(n)}\| \leqslant \varepsilon/(1-r)^2. \qquad (7.4.11)$$

Аналогичные оценки получаются в предположении, что вычисления выполняются с заданной относительной погрешностью.

Глава 8

ХАРАКТЕРИСТИЧЕСКИЙ ПОЛИНОМ МАТРИЦЫ

§ 1. МЕТОД А. Н. КРЫЛОВА

1°. Собственные числа квадратной матрицы A суть корни уравнения

$$\operatorname{Det}(A - \lambda I) = 0; \qquad (8.1.1)$$

полином в левой части этого уравнения есть характеристический полином матрицы A. Метод Крылова (подробнее см. [139]) позволяет при известных условиях сравнительно просто вычислить этот полином. Упомянутый метод состоит в следующем.

Пусть $A = [a_{ij}]_{i,j=1}^{m}$. Составим однородную систему $(A-\lambda I)\,x = 0$, или, более подробно,

$$\lambda x_i = \sum_{j=1}^{m} a_{ij} x_j, \quad i = 1, 2, \ldots, m; \qquad (8.1.2)$$

эта система имеет нетривиальное решение тогда и только тогда, когда λ есть собственное число матрицы A.

Первое из уравнений (8.1.2) умножим последовательно на $\lambda, \lambda^2, \ldots, \lambda^{m-1}$ и каждый раз будем заменять в правых частях произведения $\lambda x_1, \lambda x_2, \ldots, \lambda x_m$ соответствующими правыми частями уравнений (8.1.2). В результате получится новая система

$$\lambda^k x_1 = \sum_{j=1}^{m} b_{jk} x_j, \quad k = 1, 2, \ldots, m; \qquad (8.1.3)$$

$$b_{1j} = a_{1j}; \; k > 1, \; b_{kj} = \sum_{i=1}^{m} b_{k-1,\,i} a_{ij}.$$

Условие нетривиальной разрешимости новой системы есть

$$\begin{vmatrix} b_{11}-\lambda & b_{12} & \dots & b_{1m} \\ b_{21}-\lambda^2 & b_{22} & \dots & b_{2m} \\ \dots & \dots & \dots & \dots \\ b_{m1}-\lambda_m & b_{m2} & \dots & b_{mm} \end{vmatrix} = 0. \qquad (8.1.4)$$

Определитель (8.1.4) сравнительно нетрудно вычислить, разложив его по элементам первого столбца, — для этого достаточно вычислить m определителей порядка $m-1$ с числовыми элементами.

Если система (8.1.2) нетривиально разрешима, то система (8.1.3) также нетривиально разрешима. Отсюда следует, что все корни уравнения (8.1.1) удовлетворяют и уравнению (8.1.4), а тогда левая часть уравнения (8.1.4) делится на характеристический полином. В данном случае делимое и делитель суть полиномы одной и той же степени, поэтому частное есть постоянная. Сравнивая коэффициенты при λ^m, находим, что это частное равно

$$N = \begin{vmatrix} b_{12} & b_{13} & \dots & b_{1m} \\ b_{22} & b_{23} & \dots & b_{2m} \\ \dots & \dots & \dots & \dots \\ b_{m-1,2} & b_{m-1,3} & \dots & b_{m-1,m} \end{vmatrix}.$$

Может случиться, что $N=0$; тогда уравнение (8.1.4) превращается в бесполезное тождество. Будем считать, что $N \neq 0$. В этом случае, как только что было доказано, полином (8.1.4) сравнительно легко вычисляется, и дело сводится далее к нахождению корней известного полинома.

2°. Проанализируем погрешности метода Крылова. Матрицу A предполагаем заданной точно. Прежде всего оценим погрешности элементов b_{kj}. Введем обозначения

$$b_{1j}^{(0)} = |b_{1j}|;\ k>1,\ b_{kj}^{(\nu)} = \sum_{i=1}^{m} |b_{k-1,i}^{(\nu-1)} a_{ij}|.$$

По формулам (8.1.3) и (1.1.9) имеем $|\delta(b_{2j})| \leqslant \varepsilon_1 |b_{2j}^{(1)}|$. Далее, $b_{3j} = \sum_{i=1}^{m} b_{2i}a_{ij}$ и отсюда

$$\delta(b_{3j}) = \Big(\sum_{i=1}^{m} (b_{2i})_{\text{м}}\, a_{ij}\Big)_{\text{м}} - \sum_{i=1}^{m} b_{2i}a_{ij} =$$

$$= \Big(\sum_{i=1}^{m} (b_{2i})_{\text{м}}\, a_{ij}\Big)_{\text{м}} - \sum_{i=1}^{m} (b_{2i})_{\text{м}}\, a_{ij} +$$

$$+ \sum_{i=1}^{m} (b_{2i})_{\text{м}}\, a_{ij} - \sum_{i=1}^{m} b_{2i}a_{ij} = \delta\Big(\sum_{i=1}^{m} (b_{2i})_{\text{м}}\, a_{ij}\Big) +$$

$$+ \sum_{i=1}^{m} \delta(b_{2i})\, a_{ij};$$

$$|\delta(b_{3j})| \leqslant \varepsilon_1 \Big\{ \Big|\sum_{i=1}^{m} b_{2i}a_{ij}\Big| + \Big|\sum_{i=1}^{m} b_{2i}^{(1)} a_{ij}\Big| \Big\}.$$

Теперь $|\delta(b_{3l})| \leqslant \varepsilon_1 (|b_{3l}^{(1)}| + |b_{3l}^{(2)}|)$; по индукции найдем

$$|\delta(b_{il})| \leqslant \varepsilon_1 \sum_{\nu=1}^{i-1} |b_{il}^{(\nu)}|. \tag{8.1.5}$$

Раскрыв определитель (8.1.4), получим уравнение степени m

$$\sum_{k=0}^{m} b_k \lambda^k = 0. \tag{8.1.6}$$

Коэффициенты b_k на самом деле неизвестны, и вместо (8.1.6) получится искаженное уравнение

$$\sum_{k=1}^{m} (b_k + \delta(b_k)) \lambda^k = 0. \tag{8.1.7}$$

Из (8.1.5) следует, что $|\delta(b_k)| \leqslant \varepsilon_1 a$, где a — постоянная, зависящая от матрицы A. Полином (8.1.6) искажается на величину порядка

$$\varepsilon_1 a \sum_{k=0}^{m} |\lambda|^k = \varepsilon_1 a (|\lambda|^{m+1} - 1)(|\lambda| - 1)^{-1}.$$

Пусть число r_0 таково, что все корни полинома (8.1.6) содержатся в круге $|\lambda| < r_0$. Тогда в этом круге искажение названного полинома не превосходит числа

$$\varepsilon_1 a (r_0^{m+1} - 1)(r_0 - 1)^{-1}. \tag{8.1.8}$$

В силу известной теоремы алгебры полиномов за r_0 можно принять любое число, удовлетворяющее неравенству

$$r_0 > 1 + |b_m|^{-1} \max_{0 \leqslant k \leqslant m-1} |b_k|. \tag{8.1.9}$$

§ 2. ПОГРЕШНОСТИ ВЫЧИСЛЕНИЯ КОРНЕЙ ПОЛИНОМА

1°. Выясним, как искажаются корни полинома (8.1.6), если каждый его коэффициент испытывает искажение, не превосходящее величины $\varepsilon_1 a$, $a = \text{const}$. Обозначим этот полином через $f(\lambda)$, его искажение — через $\varphi(\lambda)$. Пусть α — корень полинома $f(\lambda)$ кратности $p \geqslant 1$. Если ε_1 достаточно мало, то можно найти такое число $\varepsilon > 0$, что

$$\min_{|\lambda - \alpha| = \varepsilon} |f(\lambda)| > \varepsilon_1 a (r_0^{m+2} - 1)(r_0 - 1)^{-1}. \tag{8.2.1}$$

2°. Выбор ε можно осуществить так. Разложим $f(\lambda)$ по степеням $\lambda - \alpha$

$$f(\lambda) = (\lambda - \alpha)^p \sum_{k=0}^{m-p} c_k (\lambda - \alpha)^k, \quad c_0 \neq 0.$$

Если $|\lambda - \alpha| = \varepsilon$, то

$$|f(\lambda)| \geqslant \varepsilon^p \Big[|c_0| - \sum_{k=1}^{m-p} |c_k| \varepsilon^k \Big].$$

Пусть ε столь мало, что

$$\sum_{k=1}^{m-p} |c_k| \varepsilon^k \leqslant |c_0|/2. \tag{8.2.2}$$

Тогда

$$|f(\lambda)| \geqslant \varepsilon^p |c_0|/2. \tag{8.2.3}$$

В силу формулы (8.1.8) теперь достаточно потребовать, чтобы

$$\varepsilon > \left[\frac{2\varepsilon_1 a}{|c_0|}\,\frac{r_0^{m+1}-1}{r_0-1}\right]; \tag{8.2.4}$$

если ε_1 достаточно мало, то соотношения (8.2.2) и (8.2.4) не противоречат друг другу.

По теореме Руше (см., например, [131, п. 22]) искаженный полином $f(\lambda)+\psi(\lambda)$ имеет в круге $|\lambda-\alpha|<\varepsilon$ ровно p корней $\alpha_1, \alpha_2, \ldots, \alpha_p$; это значит, что искажение корня α не превосходит числа ε.

3°. Применение результатов настоящего параграфа к методу Крылова очевидно. Столь же очевидно их применение и к методу следующего параграфа.

§ 3. МЕТОД ЛЕВЕРЬЕ — ФАДДЕЕВА

1°. Пусть дана квадратная матрица порядка m. Выполним построения, описанные следующими формулами:

$$A_1 = A,\ p_1 = \operatorname{Sp} A_1,\ B_1 = A_1 - p_1 I,$$

$$A_2 = AB_1,\ p_2 = 2^{-1}\operatorname{Sp} A_2,\ B_2 = A_2 - p_2 I,$$

. . . . . .

$$A_{m-1} = AB_{m-2},\ p_{m-1} = (m-1)^{-1}\operatorname{Sp} A_{m-1},\ B_{m-1} = A_{m-1} - p_{m-1}I,$$

$$A_m = AB_{m-1},\ p_m = m^{-1}\operatorname{Sp} A_m,\ B_m = A_m - p_m I. \tag{8.3.1}$$

Доказывается, что: а) $p_1, p_2, \ldots, p_m$ суть коэффициенты характеристического полинома матрицы A, записанного в форме

$$f(\lambda) = (-1)^m(\lambda^m - p_1\lambda^{m-1} - p_2\lambda^{m-2} - \ldots - p_m); \tag{8.3.2}$$

б) матрица B_m — нулевая; в) если матрица A — неособенная, то

$$A^{-1} = p_m^{-1} B_{m-1}. \tag{8.3.3}$$

В [139] приведены доказательства утверждений а) — в) и дана библиография вопроса.

2°. Оценим погрешность метода Леверье — Фаддеева. Элементы матрицы A предполагаются заданными точно, поэтому $\delta(a_{ij}^{(0)}) = 0$. По правилу умножения матриц и по формуле (1.1.9) получим при $k \geqslant 2$

$$|\delta(a_{ij}^{(k)})| \leqslant \varepsilon_1 |a_{ij}^{(k)}| = \varepsilon_1 \left|\sum\nolimits_{s=1}^{m} a_{is} b_{si}^{(k-1)}\right|;$$

$$|\delta(b_{ij}^{(k-1)})| \leqslant \varepsilon_1 |b_{ij}^{(k-1)}|,\ |\delta(p_k)| \leqslant \varepsilon_1 |p_k| = \frac{\varepsilon_1}{k}\left|\sum\nolimits_{i=1}^{m} a_{ii}^{(k)}\right|. \tag{8.3.4}$$

где $a_{ij}^{(k)}$ и $b_{ij}^{(k)}$ — элементы матриц A_k и B_k соответственно. Используя результаты § 2 главы 1, теперь легко оценить нормы $\|\delta(A_k)\|$ и $\|\delta(B_k)\|$, а также погрешности коэффициентов характеристического полинома; если воспользоваться еще и соотношением (8.3.3), то можно получить и оценку погрешности обратной матрицы A^{-1}.

§ 4. ОЦЕНКА ЧИСЛА ОБУСЛОВЛЕННОСТИ МАТРИЦЫ

1°. Как мы видели на протяжении всего раздела II, для оценок погрешностей различных методов решения задач линейной алгебры важную роль играют величины $\|A\|$, $\|A^{-1}\|$, P_A; их двусторонним оценкам посвящен § 6 главы 6. Здесь мы скажем о том, как можно оценить указанные величины сверху с помощью характеристического полинома.

Пусть A — неособенная квадратная матрица порядка m. Составим матрицу $K = AA^*$ и вычислим ее характеристический полином. Если $\sigma_1 \leqslant \sigma_2 \leqslant \ldots \leqslant \sigma_m$ — корни этого полинома, то, как известно,

$$\|A\| = \sqrt{\sigma_m},\quad \|A^{-1}\| = \sqrt{\sigma_1^{-1}},\quad P_A = \sqrt{\sigma_m/\sigma_1}. \tag{8.4.1}$$

Мы получим оценки для $\|A\|$, $\|A^{-1}\|$, P_A сверху, если найдем оценку сверху для σ_m и оценку снизу для σ_1. Пусть

$$F(\lambda) = \sum_{k=0}^{m} \alpha_k \lambda^k \tag{8.4.2}$$

— характеристический полином матрицы K. Как известно, верхняя граница корней полинома $F(\lambda)$ дается выражением

$$\beta = 1 + a/|\alpha_m|,\quad a = \max_{0 \leqslant i \leqslant m-1} |\alpha_i|; \tag{8.4.3}$$

это выражение, естественно, является верхней границей для σ_m.

Полином $\lambda^m F(\lambda^{-1})$ имеет корни σ_k^{-1}, $k = 1, 2, \ldots, m$. Применяя к этому полиному оценку (8.4.3), находим, что нижняя граница для σ_1 дается выражением

$$\alpha = |\alpha_0|/(|\alpha_0| + b),\quad b = \max_{1 \leqslant i \leqslant m} |\alpha_i|. \tag{8.4.4}$$

Теперь

$$\begin{gathered} \|A\| \leqslant \sqrt{\beta} = \sqrt{(|\alpha_m| + a)/|\alpha_m|}, \\ \|A^{-1}\| \leqslant \sqrt{\alpha^{-1}} = \sqrt{(|\alpha_0| + b)/|\alpha_0|}, \\ P_A \leqslant \sqrt{\frac{\beta}{\alpha}} = \sqrt{\frac{(|\alpha_m| + a)/(|\alpha_0| + b)}{|\alpha_0 \alpha_m|}}. \end{gathered} \tag{8.4.5}$$

2°. Оценки (8.4.5) могут оказаться грубыми. Для их улучшения достаточно уточнить оценки для σ_m и σ_1. Это можно сделать, например, с помощью теоремы Штурма об отделении корней полинома.

Раздел III

НЕКОТОРЫЕ КОНКРЕТНЫЕ ЛИНЕЙНЫЕ ПРОЦЕССЫ И ЗАДАЧИ

Главы 9 и 10 мы посвящаем методу конечных элементов (МКЭ). В главе 9 даются краткое описание МКЭ, анализ и оценки погрешностей, связанных с этим методом; в главе 10 изучается, постоянный множитель, входящий в оценку аппроксимации МКЭ. В обеих главах мы в основном ограничиваемся функциями одной независимой переменной и в соответствии с этим краевыми задачами для обыкновенных дифференциальных уравнений. Переход от одномерных задач к многомерным во многих случаях осуществляется сравнительно просто. В частности, в работах автора [96, 182], а также в некоторых его более ранних работах исследуется погрешность искажения МКЭ для широкого класса многомерных задач, приближенное решение которых строится на кубической сетке; в [47] исследуется погрешность вариационно-сеточной аппроксимации в соболевских пространствах $W_n^{(s)}$; для $p=2$ эта погрешность была ранее исследована Стрэнгом и Фиксом в [199]; имеется довольно обширная литература, в которой та же погрешность исследуется на симплициальных сетках, из последних работ этого направления отметим статьи [147, 148], в которых исследуется погрешность искажения МКЭ на треугольных сетках. В [97, 180] исследована константа в оценке вариационно-сеточной аппроксимации также и в случае многих переменных; сетка предполагается кубической.

В идейном плане исследование погрешностей на кубической сетке не очень отличается от того же исследования в одномерном случае, поэтому автор считает возможным ограничиться в основном изучением последнего случая. Лишь в конце каждой из глав 9 и 10 (§ 6 главы 9 и § 6 главы 10) рассматриваются функции многих переменных.

Глава 9

ПОГРЕШНОСТИ МКЭ

§ 1. ОПИСАНИЕ МЕТОДА ДЛЯ ОДНОМЕРНЫХ ЗАДАЧ

1° Пусть t — вещественная переменная и $x(t)$ — функция, определенная на промежутке $[0, 1]$ оси t. В основе метода конечных элементов лежит идея приближенного представления функций того или иного класса линейными агрегатами некоторых специальных функций. Ниже излагается один из многих вариантов построения таких функций. (Подробнее см. [96, 182].)

2°. Зададим натуральное число s и построим s функций $\omega_q(t)$, $0 \leqslant q \leqslant s-1$, обладающих следующими свойствами: а) функции $\omega_q(t)$ определены на всей оси t и $\omega_q \in W_p^{(s)}(R_1)$, где p, $1 < p < \infty$, — фиксированное число. В приложениях более интересен случай $p = 2$; б) $\operatorname{supp} \omega_q \subset [-1, +1]$. Из а) и б) вытекает, что $\omega_q \in C^{(s-1)}(R_1)$ и что

$$\omega_q^{(\alpha)}(\pm 1) = 0, \quad 0 \leqslant \alpha, \quad q \leqslant s-1; \tag{9.1.1}$$

в) функции ω_q удовлетворяют равенствам

$$\omega_q^{(\alpha)}(0) = \delta_{\alpha q}, \quad 0 \leqslant \alpha, \quad q \leqslant s-1. \tag{9.1.2}$$

Совокупность функций, удовлетворяющих условиям а) — в), назовем *системой исходных функций высоты* $s-1$.

Зададим натуральное число n и положим $h = 1/(2n)$. Координатные функции МКЭ определяются формулой

$$\varphi_{qjh} = \omega_q(t/h - j), \quad 0 \leqslant t \leqslant 1; \tag{9.1.3}$$

здесь j — целое число; оно должно быть таким, чтобы $\varphi_{qjh}(t) \not\equiv 0$. По условию б) это будет, если $-1 \leqslant t/h - j \leqslant 1$; если $t \in [0, 1]$, то j должно меняться в пределах $0 \leqslant j \leqslant 2n$.

Пусть $x \in W_p^{(s)}(0, 1)$. Рассмотрим подпространство функций вида

$$\sum_{q=0}^{s-1} \sum_{j=0}^{2n} a_{qj} \omega_q(t/h - j), \quad a_{qj} = \text{const}; \tag{9.1.4}$$

размерность подпространства (9.1.4) равна $N = s(2n+1)$. Если положить $a_{qj} = h^q x^{(q)}(jh)$, то соответствующая функция

$$x^h(t) = \sum_{q=0}^{s-1} \sum_{j=0}^{2n} h^q x^{(q)}(jh)\, \omega_q(t/h - j) \tag{9.1.5}$$

решает задачу эрмитовой интерполяции для функции $x(t)$. Именно

$$\frac{d^\alpha}{dt^\alpha} x^h(j_0 h) = \frac{d^\alpha}{dt^\alpha} x(j_0 h), \quad 0 \leqslant \alpha \leqslant s-1, \quad 0 \leqslant j_0 \leqslant 2n. \tag{9.1.6}$$

3°. Вопрос о сходимости $x^h(t) \xrightarrow[n\to\infty]{} x(t)$ решается следующей теоремой

Теорема 9.1.1. *Предельное соотношение*

$$\| x^h - x \|_{W_p^{(s)}} \xrightarrow[n\to\infty]{} 0 \tag{9.1.7}$$

имеет место для любой функции $x \in W_p^{(s)}(0, 1)$ *тогда и только тогда, когда исходные функции удовлетворяют соотношениям*

$$\sum_{q=0}^{s-1} \sum_{i=0}^{1} \frac{i^{\gamma-q}}{(\gamma-q)!}\, \omega_q(t-i) = \frac{t^\gamma}{\gamma!}, \quad 0 \leqslant \gamma \leqslant s, \quad 0 \leqslant t \leqslant 1. \tag{9.1.8}$$

Эти соотношения мы называем фундаментальными. Более подробно можно записать фундаментальные соотношения (9.1.8) так:

$$\omega_\gamma(t) + \sum_{q=0}^{\gamma} \frac{\omega_q(t-1)}{(\gamma-q)} = \frac{t^\gamma}{\gamma!}, \quad 0 \leqslant t \leqslant 1, \quad 0 \leqslant \gamma \leqslant s-1;$$

$$\sum_{q=0}^{s-1} \frac{\omega_q(t-1)}{(s-q)!} = \frac{t^s}{s!}, \quad 0 \leqslant t \leqslant 1. \tag{9.1.8а}$$

Простейшие исходные функции — это функции, совпадающие на каждом из промежутков $[-1, 0]$ и $[0, 1]$ с некоторыми полиномами степени $2s-1$:

$$-1 \leqslant t \leqslant 0, \quad \omega_q(t) = \frac{1}{q!}(1+t)^s t^q \left[a_0^{(s)} - a_1^{(s)} t + \ldots \right.$$

$$\left. \ldots + (-1)^{s-q-1} a_{s-q-1}^{(s)} t^{s-q-1}\right]; \tag{9.1.9}$$

$$0 \leqslant t \leqslant 1, \quad \omega_q(t) = \frac{1}{q!}(1-t)^s t^q \left[a_0^{(s)} + a_1^{(s)} t + \ldots + a_{s-q-1}^{(s)} t^{s-q-1}\right];$$

числа $a_k^{(s)}$ определяются однозначно: они суть коэффициенты ряда Маклорена для функции $(1-t)^{-s}$:

$$a_0^{(s)} = 1; \quad k > 0, \quad a_k^{(s)} = \frac{1}{k!} s(s+1)\ldots(s+k-1). \tag{9.1.10}$$

Отметим, что из (9.1.10) вытекает рекуррентная формула

$$a_k^{(s)} = a_{k-1}^{(s)} + a_k^{(s-1)}, \quad 1 \leqslant k \leqslant s-2. \tag{9.1.10а}$$

4°. Функции (9.1.9), (9.1.10) удовлетворяют «усиленным» фундаментальным соотношениям. Так мы называем соотношения (9.1.8), если они верны при значениях γ, $0 \leqslant \gamma \leqslant \tilde{s}$, где $\tilde{s} > s$. Приведем эти соотношения в подробной записи:

$$\omega_\gamma(t) + \sum_{q=0}^{\gamma} \frac{\omega_q(t-1)}{(\gamma-q)!} = \frac{t^\gamma}{\gamma!}, \quad 0 \leqslant t \leqslant 1, \quad 0 \leqslant \gamma \leqslant s-1; \tag{9.1.8б}$$

$$\sum_{q=0}^{s-1} \frac{\omega_q(t-1)}{(\gamma-q)!} = \frac{t^\gamma}{\gamma!}, \quad 0 \leqslant t \leqslant 1, \quad s \leqslant \gamma \leqslant \tilde{s}.$$

Для функций (9.1.9), (9.1.10) $\tilde{s} = 2s - 1$. Не существует одномерных исходных функций, соответствующих определению п. 2° и удовлетворяющих соотношениям (9.1.8б) при $\tilde{s} > 2s - 1$.

Кусочно-полиномиальные исходные функции дают (см. ниже, § 2) наилучший порядок по h погрешности аппроксимации сравнительно с любыми другими исходными функциями, удовлетворяющими условиям а)—в) настоящего параграфа и фундаментальным соотношениям (9.1.8). По этой причине мы назовем *оптимальными* полиномы, образующие правую часть второй формулы (9.1.9).

5°. Рассмотрим краевую задачу с естественными краевыми условиями для обыкновенного дифференциального уравнения

$$Ax := \sum_{k=0}^{s} \frac{d^k}{dt^k}\left(p_k(t)\frac{d^k x}{dt^k}\right) = f(t), \quad 0 < t < 1. \qquad (9.1.11)$$

Примем, что коэффициенты $p_k(t)$ измеримы и ограничены, оператор A задачи — положительно определенный в $L_2(0, 1)$, $p_s(t) \geqslant \text{const} > 0$. Тогда существует единственный элемент $x \in W_2^{(s)}(0, 1)$, который является обобщенным решением рассматриваемой задачи. Будем решать эту задачу по МКЭ: приближенное решение $x_h(t)$ будем искать в виде (9.1.4) и определим коэффициенты a_{qj} из условия, чтобы интеграл энергии

$$(Ax^h, x^h) - 2(f, x^h) \qquad (9.1.12)$$

был минимальным. Это условие приводит к линейной алгебраической системе для неизвестных коэффициентов, которые мы здесь и далее будем обозначать через $a_{qj}^{(h)}$.

$$\sum_{q=0}^{s-1}\sum_{i=0}^{2n}[\varphi_{qjh}, \varphi_{q'j'h}]a_{qj}^{(h)} = (f, \varphi_{q'j'h}), \quad 0 \leqslant q' \leqslant s-1,$$
$$0 \leqslant j' \leqslant 2n; \qquad (9.1.13)$$

квадратные скобки означают скалярное произведение в энергетической метрике, а $\varphi_{qjh} = \omega_q(t/h - j)$. Система (9.1.13) имеет одно и только одно решение.

Если x — произвольный элемент энергетического пространства, то [84]

$$(Ax, x) - (f, x) = |x - x_*|^2 - |x_*|^2. \qquad (9.1.14)$$

здесь x_* — точное решение задачи, $|\cdot|$ — норма в энергетической метрике. Отсюда вытекает следующее: если x^h — произвольный элемент подпространства (9.1.4), то

$$|x_h - x_*| \leqslant |x^h - x_*|. \qquad (9.1.15)$$

6°. Отметим еще следующее: если уравнение (9.1.11) решается при условиях первой краевой задачи

$$x^{(\alpha)}(0) = x^{(\alpha)}(1) = 0, \quad 0 \leqslant \alpha \leqslant s-1; \qquad (9.1.16)$$

то подпространство (9.1.4) можно заменить несколько более простым подпространством функций

$$x^h(t) = \sum_{q=0}^{s-1} \sum_{j=1}^{2n-1} a_{qj} \omega_q(t/h - j). \qquad (9.1.17)$$

§ 2. ПОГРЕШНОСТЬ АППРОКСИМАЦИИ

1°. Пусть $x \in W_2^{(s+1)}(0, 1)$, тогда $x \in C^{(s)}(0, 1)$ и можно построить функцию $x^h(t)$ по формуле (3.1.5). Оценим норму разности $x - x^h$ в метрике $W_2^{(\bar{s})}(0, 1)$, $0 \leqslant \bar{s} \leqslant s$; эту метрику можно задать, например, формулой

$$(x, y)_{W_2^{(\bar{s})}} = \sum_{\alpha=0}^{\bar{s}} \int_0^1 x^{\alpha}(t)\, y^{\alpha}(t)\, dt.$$

Рассмотрим интеграл

$$\int_0^1 \left[D^{\alpha} x(t) - D^{\alpha} x^h(t)\right]^2 dt =$$

$$= \sum_{j_0=0}^{2n-1} \int_{j_0 h}^{(j_0+1)h} \left[D^{\alpha} x(t) - D^{\alpha} x^h(t)\right]^2 dt, \quad 0 \leqslant \alpha \leqslant \bar{s},$$

$$D^{\alpha} = \frac{d^{\alpha}}{dt^{\alpha}}.$$

Выделим из последней суммы одно слагаемое:

$$J_{j_0} = \int_{j_0 h}^{(j_0+1)h} \left[D^{\alpha} x(t) - D^{\alpha} x^h(t)\right]^2 dt =$$

$$= \int_{j_0 h}^{(j_0+1)h} \left[D^{\alpha} x(t) - \sum_{q=0}^{s-1} \sum_{j=0}^{2n} x^{(q)}(jh)\, h^{-\alpha} \omega_q(t/h - j)\right]^2 dt.$$

Обозначим $j_0 h = t_0$ и сделаем подстановку $t = t_0 + \tau h$, $0 \leqslant \tau \leqslant 1$. Слагаемые в последней сумме отличны от тождественного нуля, если $-1 < j_0 + \tau - j < 1$, т. е. если $j = j_0$ или $j = j_0 + 1$. Теперь

$$J_{j_0} = h \int_0^1 \left[x^{(\alpha)}(t_0 + \tau h) - h^{-\alpha} \sum_{q=0}^{s-1} \sum_{i=0}^{1} x^{(q)}(t_0 + ih)\, \omega_q^{(\alpha)}(\tau - i)\right]^2 d\tau.$$

Разложим $x^{(\alpha)}(t_0 + \tau h)$ по формуле Тейлора по степеням h до членов порядка $s - \alpha$, а $x^{(q)}(t_0 + ih)$ — до членов порядка $s - q$; в обоих случаях остаточные члены будут содержать производную $x^{(s+1)}(t)$. Легко проверить, что в силу фундаментальных соотношений главные члены формул Тейлора исчезнут, а остаточные члены дадут оценку

$$J_{j_0} \leqslant C h^{2s-2\alpha+2} \left\| x^{(s+1)} \right\|^2_{L_2(t_0,\, t_0+h)}$$

Суммируя по j_0, а затем по α, получаем

$$\|x - x^h\|_{W_2^{(\bar{s})}} \leqslant Ch^{s-\bar{s}+1} \|x^{(s+1)}\|_{L_2(0,\,1)}. \qquad (9.2.1)$$

2°. Лучшая оценка получается, если исходные функции удовлетворяют усиленным фундаментальным соотношениям, а функция $x(t)$ обладает соответственно большей гладкостью. Пусть фундаментальные соотношения удовлетворяются для $\gamma = 2s - 1$, а $x \in W_2^{(2s)}(0, 1)$. Тогда рассуждения, аналогичные предшествующим, приводят к оценке

$$\|x - x^h\|_{W_2^{(\bar{s})}} \leqslant Ch^{2s-\bar{s}} \|x^{(2s)}\|_{L_2(0,\,1)}, \quad 0 \leqslant \bar{s} \leqslant 2s - 1. \qquad (9.2.2)$$

3°. Обратимся к задаче п. 5° § 1. Отметим, что для этого случая энергетическая норма эквивалентна норме $W_2^{(s)}(0, 1)$. Допустим, что решение рассматриваемой задачи $x_* \in W_2^{(2s)}(0, 1)$, для этого достаточно, чтобы коэффициенты уравнения (9.1.11) и его свободный член обладали некоторой определенной гладкостью. В качестве исходных функций возьмем кусочно-полиномиальные функции (9.1.9). По формулам (9.2.2) и (9.1.15) находим оценку погрешности аппроксимации МКЭ

$$x_* - x_h \leqslant Ch^{2s-2} \|x_*^{(2s)}\|_{L_2(0,\,1)}. \qquad (9.2.3)$$

§ 3. ПОГРЕШНОСТЬ ИСКАЖЕНИЯ

1°. При исследовании погрешности искажения для нас будет существенно, что коэффициенты a_{qj} в приближенном решении x_h зависят от h, поэтому будем обозначать их через $a_{qj}^{(h)}$, так что

$$x_h(t) = \sum_{q=0}^{s-1} \sum_{j=0}^{2n} a_{qj}^{(h)} \omega_q(t/h - j). \qquad (9.3.1)$$

Совокупность коэффициентов $a_{qj}^{(h)}$ при фиксированном h будем рассматривать как вектор $a^{(h)}$ в пространстве R_N, $N = s(2n + 1)$, а совокупность правых частей системы (9.1.13) $F_{qj}^{(h)} := (f, \varphi_{qjh})$ — как вектор $F^{(h)}$ в том же пространстве R_N.

2°. Решим вспомогательную задачу: выведем неравенство, связывающее норму функции x^h вида (9.1.15) в $L_2(0, 1)$ с нормой соответствующего вектора коэффициентов a, составленного из коэффициентов a_{qj}, в R_N. Имеем

$$\|x^{(h)}\|^2 := \|x^{(h)}\|^2_{L_2(0,\,1)} = \sum_{l=0}^{2n-1} \int_{lh}^{(l+1)h} [x^{(h)}(t)]^2\, dt.$$

В промежутке $(lh, (l + 1)h)$ отличны от тождественного нуля только координатные функции $\omega_q(t/h - l)$ и $\omega_q(t/h - l - 1)$;

отсюда

$$\int_{lh}^{(l+1)h}[x^h(t)]^2\,dt=\int_{lh}^{(l+1)h}\Big[\sum_{q=0}^{s-1}(a_{ql}\omega_q(t/h-l)+$$

$$+a_{q,\,l+1}\omega_q(t/h-l-1))\Big]^2\,dt=h\int_0^1\Big[\sum_{q=0}^{s-1}(a_{ql}\omega_q(\tau)+$$

$$+a_{q,\,l+1}\omega_q(\tau-1))\Big]^2\,d\tau. \qquad (9.3.2)$$

Последний интеграл есть неотрицательная квадратичная форма от $2s$ переменных $a_{ql}, a_{q,l+1}$, $q=0, 1, \ldots, s-1$, с коэффициентами не зависящими от h. Нетрудно видеть, что эта форма — положительно определенная. Действительно, пусть ее значение равно нулю. Тогда

$$\sum_{q=0}^{s-1}(a_{ql}\omega_q(\tau)+a_{q,\,l+1}\omega_q(\tau-1))\equiv 0,\quad 0\leqslant\tau\leqslant 1.$$

Дифференцируя это тождество α раз, $\alpha=0, 1, \ldots, s-1$, полагая затем $\tau=1$ и $\tau=0$ и пользуясь соотношениями (9.1.1), (9.1.2), находим, что все коэффициенты $a_{ql}=a_{q,\,l+1}=0$. Этим доказана положительная определенность рассматриваемой формы. Отсюда следует существование таких постоянных c_1, c_2, положительных и не зависящих от h и l, что

$$c_1^2h\sum_{j=l}^{l+1}\sum_{q=0}^{s-1}a_{ql}^2\leqslant\int_{lh}^{(l+1)h}[x^h(t)]^2\,dt\leqslant 2^{-1}c_2^2h\sum_{j=l}^{l+1}a_{ql}^2.$$

Суммируя последнее по l, получаем искомые неравенства

$$c_1\sqrt{h}\,\|a\|_{R_N}\leqslant\|x^h\|_{L_2(0,\,1)}\leqslant c_2\sqrt{h}\,\|a\|_{R_N}. \qquad (9.3.3)$$

3°. Введем два новых пространства X_N и Y_N, состоящие, как и R_N, из N-компонентных векторов, с нормами

$$\|a\|_{X_N}=\sqrt{h}\,\|a\|_{R_N},\quad \|a\|_{Y_N}=h^{-1/2}\|a\|_{R_N}. \qquad (9.3.4)$$

Теперь неравенствам (9.3.3) можно придать вид

$$c_1\|a\|_{X_N}\leqslant\|x^h\|_{L_2(0,\,1)}\leqslant c_2\|a\|_{X_N}. \qquad (9.3.5)$$

Пусть M_h — матрица системы (9.1.6) и A_h — оператор, порожденный этой матрицей и действующий из X_N в Y_N. Векторы $a^{(h)}$ и $F^{(h)}$ будем рассматривать как элементы пространств X_N и Y_N соответственно. Систему (9.1.6) можно записать в виде

$$A_ha^{(h)}=F^{(h)}. \qquad (9.3.6)$$

Теорема 9.3.1. *Вычислительный процесс* (9.3.6) (*процесс вычисления коэффициентов* $a_{ql}^{(h)}$) *устойчив в последовательности пар пространств* (X_N, Y_N) *в смысле второго определения* (§ 2 *главы* 3).

В силу следствия 3.2.1 достаточно установить, что величины $\|A_h^{-1}\|$ и $\|a^{(h)}\|_{X_N}$ ограничены независимо от h.

Оценим норму $\|A_h^{-1}\|$:

$$\|A_h^{-1}\| = \sup_{b \in R_N} \frac{\|A_h^{-1}\|_{X_N}}{\|b\|_{Y_N}} = h \sup_{b \in R_N} \frac{\|A_h^{-1} b\|_{R_N}}{\|b\|_{R_N}} = \frac{h}{\mu_1^{(n)}}, \quad (9.3.7)$$

где $\mu_1^{(n)}$ — наименьшее собственное число матрицы M_n.

Пусть λ_0 — нижняя грань оператора A: $\lambda_0 = \inf\limits_{v \in H_A} |v|^2/\|v\|^2$, $|\cdot|$ — энергетическая норма оператора A. Этот оператор — положительно определенный, поэтому $\lambda_0 > 0$. Пусть

$$v^h(t) = \sum_{q=0}^{s-1} \sum_{j=0}^{2n} b_{qj} \omega_q (t/h - j)$$

и пусть $b \in R_N$ вектор с составляющий b_{qj}. Очевидно, что $v^h \in H_A$, поэтому $\lambda_0 \leqslant |v^h|^2/\|v^h\|^2$. Далее

$$|v^h|^2 = \Big| \sum_{q=0}^{s-1} \sum_{j=0}^{2n} b_{qj} \varphi_{qjh} \Big|^2 =$$

$$= \sum_{q, q'=0}^{s-1} \sum_{j, j'=0}^{2n} [\varphi_{qjh}, \varphi_{q'j'h}] b_{qj} b_{q'j'} = (M_h b, b)_{R_N}.$$

По формуле (9.3.9) $\|v^h\|^2 \geqslant c_1^2 h \|b\|_{R_N}^2$ и, следовательно,

$$\lambda_0 \leqslant \frac{1}{c_1^2 h} \frac{(M_h b, b)_{R_N}}{\|b\|_{R_N}^2}.$$

Правую часть можно заменить ее нижней гранью по всем возможным векторам $b \in R_N$. Это приведет нас к неравенству $\lambda_0 \leqslant \mu_1^{(n)}/c_1^2 h$, или $\mu_1^{(n)} \geqslant \lambda_0 c_1^2 h$. Подставив последнее неравенство в (9.3.7), найдем, что $\|A_h^{-1}\| \leqslant (\lambda_0 c_1^2)^{-1} = \text{const}$.

Как хорошо известно, для рассматриваемой нами задачи процесс МКЭ сходится в H_A. Тем более он сходится в $L_2(0, 1)$. Отсюда следует, что L_2-нормы приближенных решений ограничены: $\|x_h\| \leqslant c_3 = \text{const}$. По неравенству (9.3.3) нормы $\|a^{(h)}\|_{X_N}$ также ограничены: $\|a^{(h)}\|_{X_N} \leqslant c_3/c_1$. Теорема 9.3.1 доказана.

4°. Устойчивость процесса (9.3.6) означает следующее. Пусть матрица M_h вычислена с погрешностью Γ_h, а вектор F^h — с погрешностью $\delta^{(h)}$ и пусть $b^{(h)}$ — решение искаженной системы МКЭ

$$(A_h + \tilde{\Gamma}_h) b^{(h)} = F^h + \delta^{(h)}, \quad (9.3.8)$$

где $\tilde{\Gamma}_h$ — оператор, порожденный матрицей Γ_h и действующий из X_N в Y_N. Существуют такие положительные постоянные p,

q, r, что при любом h справедливы оценки погрешности искажения: если $\|\tilde{\Gamma}_h\| \leqslant r$, то

$$\|b^{(h)} - a^{(h)}\|_{X_N} \leqslant p\|\tilde{\Gamma}_h\| + q\|\delta^{(h)}\|_{Y_N}. \tag{9.3.9}$$

5°. Нетрудно описать вычислительный процесс, дающий последовательность приближенных решений МКЭ. Функции (9.1.5) образуют в H_A подпространство, которое мы обозначим через $H_A^{(h)}$. Формула

$$x^{(h)}(t) = \sum_{q=0}^{s-1} \sum_{j=0}^{2n} a_{qj}^{(h)} \omega_q(t/h - j)$$

определяет оператор, действующий из X_N в $H_A^{(h)}$ и сопоставляющий вектору коэффициентов $a^{(h)}$, удовлетворяющему уравнению (9.3.6), приближенное решение $x_h(t)$. Обозначим этот оператор через Π_h. Он очевидно обратим, так что $x_h = \Pi_h a^{(h)}$ и $a^{(h)} = \Pi_h^{-1} x_h$. Подставив последнее выражение в (9.3.6), получим вычислительный процесс, определяющий приближенное решение x_h

$$A_h \Pi_h^{-1} x_h = F^{(h)}. \tag{9.3.10}$$

Теорема 9.3.2. *Вычислительный процесс* (9.3.10) *устойчив в смысле второго определения в последовательности пар пространств* $(H_A^{(h)}, Y_N)$.

В результате искажения мы вместо уравнения (9.3.10) получим искаженное уравнение

$$(A_h + \tilde{\Gamma}_h)\Pi^{-1} z_h = F^{(h)} + \delta^{(h)} \tag{9.3.11}$$

и искаженное приближенное решение

$$z_h = \sum_{q=0}^{s-1} \sum_{j=0}^{2n} b_{qj}^{(h)} \varphi_{qjh}(t), \tag{9.3.12}$$

$b_{qj}^{(h)}$ — составляющие вектора $b^{(h)}$. Очевидно, что

$$|z_h - x_h|^2 = (M_h(b^{(h)} - a^{(h)}),\ b^{(h)} - a^{(h)})_{R_N} \leqslant$$
$$\leqslant \|A_h(b^{(h)} - a^{(h)})\|_{Y_N} \cdot \|b^{(h)} - a^{(h)}\|_{X_N} \tag{9.3.13}$$

Второй множитель справа оценивается по неравенству (9.3.9), остается оценить первый множитель. Из уравнений (9.3.6), (9.3.8) находим

$$(A_h + \tilde{\Gamma}_h)(b^{(h)} - a^{(h)}) = \delta^{(h)} - \Gamma_h a^{(h)},$$

что приводится к виду

$$(I + \tilde{\Gamma}_h A_h^{-1}) A_h (b^{(h)} - a^{(h)}) = \delta^{(h)} - \tilde{\Gamma}_h a^{(h)}.$$

Как было показано выше, $\|A_h^{-1}\| \leqslant (c_1^2 \lambda_0)^{-1}$. Потребуем, чтобы $\|\Gamma_h\| \leqslant \beta c_1^2 \lambda_0$, где β — какое-нибудь число из интервала $(0, 1)$.

Тогда $\left\|\left(I+\tilde{\Gamma}_h A_h^{-1}\right)^{-1}\right\| \leqslant (1-\beta)^{-1}$ и

$$\|A_h(b^{(h)}-a^{(h)})\|_{Y_N} \leqslant (1-\beta)^{-1}\left[\|\delta^{(h)}\|_{Y_N}+\beta c_2^1\lambda_0\|\tilde{\Gamma}_h\|\right].$$

Теперь

$$|z_h-x_h| \leqslant \{(1-\beta)^{-1}\left[p\|\tilde{\Gamma}_h\|+q\|\delta^{(h)}\|_{Y_N}\right]\times$$

$$\times\left[\beta c_1^2\lambda_0\|\tilde{\Gamma}_h\|+\|\delta^{(h)}\|_{Y_N}\right]\}^{1/2} \leqslant p_1\|\tilde{\Gamma}_h\|+q_1\|\delta^{(h)}\|_{Y_N}; \quad (9.3.14)$$

последнее неравенство верно, если $\|\tilde{\Gamma}_h\| \leqslant r_1 = \min(r,\ \beta c_1^2\lambda_0)$.

Теоремы 9.3.1, 9.3.2 остаются в силе, если x — векторная функция. Изменится только значение N, именно будет $N = ks(2n+1)$, где k — число составляющих вектора x.

§ 4. ЧИСЛО ОБУСЛОВЛЕННОСТИ МАТРИЦЫ МКЭ

1°. Число обусловленности матрицы МКЭ M_h равно отношению $P_{M_h} = \mu_N^{(n)}/\mu_1^{(n)}$, где $\mu_N^{(n)}$ и $\mu_1^{(n)}$ суть соответственно наибольшее и наименьшее собственные числа матрицы M_h:

$$\mu_1^{(n)} = \min_{g\in R_N\setminus\{0\}} \frac{\left|\sum_{q=0}^{s-1}\sum_{j=0}^{2n} g_{ij}\varphi_{qjh}\right|^2}{\sum_{q=0}^{s-1}\sum_{j=0}^{2n} g_{ij}^2}, \quad (9.4.1)$$

$$\mu_N^{(n)} = \max_{g\in R_N\setminus\{0\}} \frac{\left|\sum_{q=0}^{s-1}\sum_{j=0}^{2n} g_{ij}\varphi_{gjh}\right|^2}{\sum_{q=0}^{s-1}\sum_{j=0}^{2n} g_{ij}^2}.$$

Обозначим через $g_{ij}^{(1)}$ и $g_{ij}^{(N)}$ значения коэффициентов g_{ij}, при которых достигаются соответственно минимум и максимум дроби (9.4.1). Пусть

$$w_1 = \sum_{q=0}^{s-1}\sum_{j=0}^{2n} g_{ij}^{(1)}\varphi_{qih}, \quad w_N = \sum_{q=0}^{s-1}\sum_{j=0}^{2n} g_{ij}^{(N)}\varphi_{qjh},$$

так что

$$\mu_1^{(n)} = \frac{|w_1|^2}{\sum_{q=0}^{s-1}\sum_{j=0}^{2n}\left[g_{ij}^{(1)}\right]^2}, \quad \mu_N^{(n)} = \frac{|w_N|^2}{\sum_{q=0}^{s-1}\sum_{j=0}^{2n}\left[g_{ij}^{(N)}\right]^2}.$$

Формулу (9.3.3) можно представить в виде

$$c_1'\|x^h\|^2 \leqslant h\sum_{q=0}^{s-1}\sum_{j=0}^{2n} a_{qj}^2 \leqslant c_2'\|x^h\|^2, \quad c_1',\ c_2' = \text{const} > 0. \quad (9.4.2)$$

Отсюда вытекают неравенства

$$\mu_1^{(n)} \geqslant \frac{h}{c_2'}\frac{|w_1|^2}{\|w_1\|^2} \geqslant \min_{x^h\in H_A^{(N)}\setminus\{0\}} \frac{|x^h|^2}{\|x^h\|^2}\frac{h}{c_2'} = \frac{h}{c_2'}\lambda_1^{(n)}, \quad (9.4.3)$$

$$\mu_N^{(n)} \leqslant \frac{h}{c_1'}\frac{|w_N|^2}{\|w_N\|^2} \leqslant \max_{x^h\in H_A^{(N)}\setminus\{0\}} \frac{|x^h|^2}{\|x^h\|^2}\frac{h}{c_1'} = \frac{h}{c_1'}\lambda_N^{(n)};$$

здесь $\lambda_1^{(n)}$ и $\lambda_N^{(n)}$ — приближения к первому и к N-му собственным числам оператора (9.1.11), построенные по МКЭ Для собственных чисел процесс МКЭ сходится, поэтому $\lambda_1^{(n)} \geqslant c_3' = \text{const} > 0$; кроме того, справедливо соотношение $\lambda_N^{(n)} \leqslant c_4' h^{-2s}$, $c_4' = \text{const}$ (см. [96] и [182], а также [122]). Теперь из неравенств (9.4 3) вытекает, что

$$\mu_1^{(n)} \geqslant c'h, \quad \mu_N^{(n)} \leqslant c''h^{1-2s}, \quad c', \ c'' = \text{const}, \tag{9.4.4}$$

и

$$P_{M_h} \leqslant ch^{-2s}, \quad c = \text{const}. \tag{9.4.5}$$

2°. Докажем, что при достаточно большом N верны и неравенства противоположного смысла

$$\mu_1^{(n)} \leqslant c_0'h, \quad \mu_N^{(n)} \geqslant c_0''h^{1-2s}, \quad P_{M_h} \geqslant c_0h^{-2s}, \quad c_0', \ c_0'', \ c_0 = \text{const} > 0. \tag{9.4.6}$$

Из (9.4.2) вытекает соотношение

$$\forall x^h \in H_A^{(N)} \setminus \{0\}, \quad \frac{|x^h|^2}{\sum_{q=0}^{s-1}\sum_{j=1}^{2n} a_{qj}^2} \leqslant \frac{h}{c_1'}\frac{|x^h|^2}{\|x^h\|^2}.$$

Выберем x^h так, чтобы правая часть последнего неравенства приняла свое наименьшее значение, равное $h\lambda_1^{(n)}/c_1'$. Одновременно заменим левую часть ее наименьшим значением $\mu_1^{(n)}$. В результате получим $\mu_1^{(n)} \leqslant h\lambda_1^{(n)}/c_1'$. Так как $\lambda_1^{(n)} \xrightarrow[N\to\infty]{} \lambda_1$, то при достаточно большом N будет $\lambda_1^{(n)} \leqslant 2\lambda_1$ и $\mu_1^{(n)} \leqslant 2h\lambda_1/c_1'$. Аналогично получается неравенство $\mu_N^{(n)} \geqslant \lambda_N^{(n)}/c_2'$; так как $\lambda_N^{(n)} \geqslant c_4h^{-2s}$ [122], то $\mu_N^{(n)} > c_4'h^{-2s}$ и окончательно $P_{M_h} \leqslant c_4'h^{-2s}/2\lambda_1c_1'$.

§ 5. ПОГРЕШНОСТИ АЛГОРИТМА И ОКРУГЛЕНИЯ

1°. Систему (9.3.8) можно записать в виде

$$(M_h + \Gamma_h)b^{(h)} = F^{(h)} + \delta^{(h)}, \tag{9.5.1}$$

при этом нужно рассматривать $b^{(h)}$, $F^{(h)}$ и $\delta^{(h)}$ как элементы пространства R_N. Матрица M_h — симметричная и положительно определенная в R_N; допустим, что $\|\tilde{\Gamma}\| \leqslant r$ (см. § 3, п. 4°), как это требуется соответствующей теоремой об устойчивости. Покажем, что при достаточно малом h матрица $\bar{M}_h = M_h + \Gamma_h$ будет положительно определенной. Прежде всего Γ_h можно считать симметричной, а тогда симметричной будет и $\bar{M}_h$. Далее, $\|M_h\| = \mu_N^{(n)}$ и, следовательно, $c_0''h^{1-2s} \leqslant \|M_h\| \leqslant c''h^{-2s}$. Теперь

$$\|A_h\| = \sup_{x^h \in R_N} \frac{\|A_hx^h\|_{Y_N}}{\|x^h\|_{X_N}} = \frac{1}{h}\|M_h\| \geqslant c_0''h^{-2s}.$$

Точно также найдем, что $\|\hat{\Gamma}_h\| = h^{-1}\|\Gamma_h\|$, и потому $\|\Gamma_h\| \leqslant rh$. Выбрав r достаточно малым, можно добиться того, чтобы нижняя грань матрицы M_h, которая больше, чем $\mu_1^{(N)} - rh$, была больше, чем $c'h/2$, и матрица $\bar{M}_h$ будет положительно определенной. Верхняя грань этой матрицы не превосходит величины $c''h^{1-2s} + rh$, и ее число обусловленности

$$P_{\bar{M}_h} \leqslant \frac{2c''}{c'} h^{-2s} + \frac{2r}{c'}.$$

Второй член справа мал по сравнению с первым; мы им пренебрежем и будем считать, что

$$P_{\bar{M}_h} \leqslant 2c''h^{-2s}/c'. \tag{9.5.2}$$

Заметим, впрочем, что неравенство (9.5.2) можно сделать точным, если несколько увеличить число c''.

К системе (9.5.1) с положительно определенной матрицей $\bar{M}_h$ можно применить прием Натансона (глава 7, § 2, п. 2°): система (9.5.1) заменяется системой (7.2.7), в которой следует положить

$$f = F^{(h)} + \delta^{(h)}, \quad \lambda_0 = 2^{-1}c'h, \quad \Lambda_0 = c''h^{1-2s}$$

(слагаемым rh в последнем равенстве пренебрегаем); система (7.2.7) принимает вид

$$b^{(h)} - \left[I - \frac{2\bar{M}_h}{2^{-1}c'h + c''h^{1-2s}}\right] b^{(h)} = \frac{2\,(F^{(h)} + \delta^{(h)})}{2^{-1}c'h + c''h^{1-2s}}. \tag{9.5.3}$$

При этом

$$\rho := \left\| I - \frac{2\bar{M}_h}{2^{-1}c'h + c''h^{1-2s}} \right\| \leqslant \frac{c''h^{1-2s} - 2^{-1}c'h}{c''h^{1-2s} + 2^{-1}c'h} = \frac{1 - c'h^{2s}/(2c'')}{1 + c'h^{2s}/(2c'')} < 1,$$

и если систему (9.5.3) решать по методу простой итерации, то эти итерации будут сходиться как прогрессии с знаменателем ρ. Погрешность после L итераций оценивается величиной

$$\frac{\rho^{L+1}}{1-\rho}\|F^{(h)} + \delta^{(h)}\| \approx \rho^{L+1}\frac{c''}{c'}h^{-2s}\|F^{(h} + \delta^{(h)}\|. \tag{9.5.4}$$

Потребуем, чтобы множитель при $\|F^{(h)} + \delta^{(h)}\|$ не превосходил заданного малого числа ε:

$$(L+1)\ln(1/\rho) - 2s\ln(1/h) - \ln(c''/c') \geqslant \ln(1/\varepsilon).$$

Заменим здесь $\ln(1/\rho)$ его приближенным значением $\frac{c'}{c''}h^{2s}$. Тогда последнее неравенство дает

$$L \geqslant \frac{c''}{c'}h^{-2s}\left(2s\ln\frac{1}{h} + \ln\frac{c''}{c'\varepsilon}\right) - 1.$$

2°. Займемся оценкой погрешности округления. Для краткости положим

$$I - 2\left(2^{-1}c'h + c''h^{1-2s}\right)^{-1}\bar{M}_h = U,$$
$$2\left(2^{-1}c'h + c''h^{1-2s}\right)^{-1}\left(F^{(h)} + \delta^{(h)}\right) = \bar{F}, \tag{9.5.5}$$

так что систему (9.5.3) можно записать в более компактной форме

$$b^{(h)} - Ub^{(h)} = \bar{F}; \tag{9.5.6}$$

при этом $\|U\| = \rho < 1$. Решение $b^{(h)}$ можно строить в виде

$$b^{(h)} = \sum_{\nu=0}^{\infty} b_\nu, \tag{9.5.7}$$

где $b_0 = F$, $b_\nu = Ub_{\nu-1}$, $\nu \geqslant 1$. Пусть векторы ξ и η связаны соотношением $\eta = U\xi$, так что

$$\eta_j = \sum_{k=1}^{N} U_{jk}\xi_k, \qquad j = 1, 2, \ldots, N.$$

Допустим, что вычисления производятся с повышенной точностью, тогда по формуле (1.1.9)

$$|\delta(\eta_j)| \leqslant \varepsilon_1 \sum_{k=1}^{N} |U_{jk}\xi_k|$$

и, следовательно,

$$\|\delta(\eta)\| \leqslant \varepsilon_1 \|U\|_E \cdot \|\xi\|. \tag{9.5.8}$$

Ниже будем обозначать $\|U\|_E = \rho_1$.

Из формул (9.5.7), (9.5.8) следует, что при $\nu \geqslant 1$

$$b_\nu + \delta(b_\nu) = U(b_{\nu-1} + \delta(b_{\nu-1})) + \theta\varepsilon_1\rho_1\|b_{\nu-1} + \delta(b_{\nu-1})\|.$$

Отсюда находим, что с точностью до членов высшего порядка малости относительно ε_1 будет

$$\nu \geqslant 1, \quad \delta(b_\nu) = U\delta(b_{\nu-1}) + \theta\varepsilon_1\rho_1\|b_{\nu-1}\|$$

и, следовательно,

$$\|\delta(b_\nu)\| \leqslant \rho\|\delta(b_{\nu-1})\| + \varepsilon_1\rho_1\|b_{\nu-1}\|,$$

или, обозначая $\delta(b_\nu) = \varepsilon_1\varkappa_\nu$,

$$\varkappa_\nu \leqslant \rho\varkappa_{\nu-1} + \rho\|b_{\nu-1}\|. \tag{9.5.9}$$

Заметим, что $\varkappa_0 = 0$, так как величина $b_0 = F$ известна точно.

По формуле (9.5.9) получаем $\varkappa_{\nu-1} \leqslant \rho\varkappa_{\nu-2} + \rho_1\|b_{\nu-2}\|$. Подставив это в (9.5.9), находим

$$\varkappa_\nu \leqslant \rho^2\varkappa_{\nu-2} + \rho\rho_1\|b_{\nu-2}\| + \rho_1\|b_{\nu-2}\|.$$

Продолжая таким же образом, получим в конечном счете

$$\varkappa_\nu \leqslant \rho^{\nu-1}\varkappa_1 + \rho_1\sum_{k=1}^{\nu-1}\|b_k\|\rho^{\nu-k-1}.$$

Но $\varkappa_1 \leqslant \rho_1 \| b_0 \| = \rho_1 \| \bar{F} \|$, $\| b_k \| = \| U^k b_0 \| \leqslant \rho^k \| \bar{F} \|$, поэтому

$$\varkappa_\nu \leqslant \rho_1 \| \bar{F} \| \nu \rho^{\nu-1}, \quad \| \delta(b_\nu) \| \leqslant \varepsilon_1 \rho_1 \| \bar{F} \| \nu \rho^{\nu-1}. \qquad (9.5.10)$$

Приближенное решение $b^{(h)}$, полученное из (9.5.6) в результате L итераций, имеет вид $b^{(h)} = \sum_{\nu=0}^{L} b_\nu$; составляющие $b_k^{(h)}$ вектора $b^{(h)}$ выражаются через составляющие $b_{k\nu}$ векторов b_ν по формуле $b_k^{(h)} = \sum_{\nu=0}^{L} b_{k\nu}$. Учитывая формулу (1.1.9), можно написать

$$b^{(h)} + \delta(b^{(h)}) = \sum_{\nu=0}^{L} (b_\nu + \delta(b_\nu)) + \varepsilon_1 \bar{b},$$

где $\bar{b}$ — вектор с составляющими

$$\theta_1 \sum_{\nu=0}^{L} b_{1,\nu}, \quad \theta_2 \sum_{\nu=0}^{L} b_{2,\nu}, \ \ldots, \ \theta_N \sum_{\nu=0}^{L} b_{N,\nu}; \quad \| \theta_l \| \leqslant 1;$$

в правой части последнего равенства отброшены члены более высокого порядка малости. Из упомянутого равенства находим

$$\| \delta(b^{(h)}) \| \leqslant \sum_{\nu=0}^{\infty} \| \delta(b_\nu) \| + \varepsilon_1 \| \bar{b} \|,$$

а из формулы (9.5.10) —

$$\sum_{\nu=0}^{\infty} \| \delta(b_\nu) \| \leqslant \varepsilon_1 \rho_1 \| \bar{F} \| \sum_{\nu=1}^{\infty} \nu \rho^{\nu-1} = (1-\rho)^{-2} \varepsilon_1 \rho_1 \| \bar{F} \|. \qquad (9.5.11)$$

Далее,

$$\| \bar{b} \|^2 = \sum_{k=1}^{N} \theta_k^2 \Big(\sum_{\nu=0}^{L} b_{k,\nu} \Big)^2 \leqslant \sum_{k=1}^{N} \Big(\sum_{\nu=0}^{L} | b_{k,\nu} | \Big)^2 = \Big\| \sum_{\nu=0}^{L} \bar{b}_\nu \Big\|^2,$$

здесь $\bar{b}_\nu$ — вектор с составляющими $| b_{k,\nu} |$. Теперь

$$\| \bar{b} \| \leqslant \Big\| \sum_{\nu=0}^{L} \bar{b}_\nu \Big\| \leqslant \sum_{\nu=0}^{L} \| \bar{b}_\nu \| \leqslant \sum_{\nu=0}^{L} \| b_\nu \| \leqslant \| \bar{F} \| \sum_{\nu=0}^{\infty} \rho^n =$$

$$= \| \bar{F} \| / (1-\rho). \qquad (9.5.12)$$

Теперь по формулам (9.5.11), (9.5.12)

$$\| \delta(b^{(h)}) \| \leqslant \varepsilon_1 \| \bar{F} \| [\rho_1/(1-\rho)^2 + 1/(1-\rho)].$$

Так как ρ довольно близко к единице, то вторым слагаемым в квадратной скобке можно пренебречь, и окончательно получается следующая оценка погрешности округления:

$$\| \delta(b^{(h)}) \| \leqslant \varepsilon_1 \| \bar{F} \| \rho_1 (1-\rho)^{-2} \leqslant \varepsilon_1 \| F \| \sqrt{N}\, \rho (1-\rho)^{-2} \leqslant$$

$$\leqslant \varepsilon_1 \| \bar{F} \| \sqrt{(2n+1)s}\, (1-\rho)^{-2} \qquad (9.5.13)$$

Но $1 - \rho = O(h^{2s})$ и $(2n+1)s = O(h^{-1})$, поэтому

$$\| \delta(b^{(h)}) \| = \varepsilon_1 \| \bar{F} \| O(h^{-4s-1/2}). \qquad (9.5.14)$$

3°. По-видимому, целесообразно выполнять такое количество итераций, чтобы погрешности алгоритма и округления,

т. е. правые части формул (9.5.4) и (9.5.14), были величинами одного порядка малости. Потребуем, чтобы $\rho^{L+1}h^{-2s} = \varepsilon_1 h^{-4s-1/2}$, или $\rho^{L+1} = \varepsilon_1 h^{-2s-1/2}$. Отсюда, так как $\ln \rho^{-1} \approx c' h_{2s}/c''$, то

$$L + 1 = (c''/c')\, h^{-2s} [\ln (1/\varepsilon_1) + (2s + 1/2) \ln (1/h)].$$

Обычно h существенно больше, чем ε_1, поэтому вторым слагаемым в квадратной скобке можно пренебречь и положить

$$L + 1 = (c'' h^{-2s}/c') \ln \varepsilon^{-1} \tag{9.5.15}$$

Результаты настоящего параграфа распространяются на положительно определенные задачи для уравнений в частных производных порядка $2s$ в кубе евклидова пространства R_m. Различие проявится только в том, что в данном случае $N = O(h^{-m})$, поэтому в формуле (9.5.14) справа появится величина $O(h^{-4s-m/2})$; формула (9.5.15) остается неизменной.

4°. Систему (9.5.1) с положительно определенной матрицей $\overline{M}_h$ можно решать по методу Холецкого; в этом случае погрешности оцениваются так, как об этом сказано в главе 6, § 2. Далее, в случае обыкновенного дифференциального уравнения второго порядка ($m = s = 1$) матрица $\overline{M}_h$ — трехдиагональная, и можно воспользоваться методом прогонки; для оценки погрешностей можно воспользоваться тогда результатами главы 6, § 1.

§ 6. ПОГРЕШНОСТИ МКЭ ДЛЯ МНОГОМЕРНЫХ ЗАДАЧ

1°. Для простоты ограничимся рассмотрением функций, заданных в замкнутом единичном кубе Q пространства R_m; о применении МКЭ к более общим областям см., например, [94, 96]. Выберем натуральное n, положим $h = 1/(2n)$ и построим в R_m кубическую сетку с шагом h. При этом куб Q будет точно покрыт кубами сетки. В качестве исходных функций возьмем произведения

$$\tilde{\omega}_q(t) = \prod_{k=1}^{m} \omega_{q_k}(t_k). \tag{9.6.1}$$

Здесь $t = (t_1, t_2, \ldots, t_m)$ — вектор в R_m, $q = (q_1, q_2, \ldots, q_m)$ — мультииндекс размерности m, ω_{q_k} — функции (9.1.9), (9.1.10). Напомним принятое нами обозначение: если a — вещественное число, то $\underline{a}$ — вектор, все m составляющих которого равны a. Далее, если $t = (t_1, t_2, \ldots, t_m)$, $\tau = (\tau_1, \tau_2, \ldots, \tau_m)$ — векторы или мультииндексы в R_m, то будем писать $t \leqslant \tau$, если $t_k \leqslant \tau_k$, $k = 1, 2, \ldots, m$; если при этом хотя бы для одного значения k будет $t_k < \tau_k$, то будем писать $t < \tau$. В этих обозначениях $Q = \{t : \underline{0} \leqslant t \leqslant \underline{1}\}$, мультииндекс q в (9.6.1) подчинен неравенствам $\underline{0} \leqslant q \leqslant \underline{s} - \underline{1}$, а радиус-вектор любого узла упомянутой выше сетки равен jh, где j — целочисленный вектор.

Из определения (9.6.1) вытекают очевидные свойства функций $\tilde{\omega}_q(t)$: а) $\operatorname{supp}\tilde{\omega}_q = Q_1 := \{t: -\underline{1} \leqslant t \leqslant \underline{1}\}$; б) $D^\alpha\tilde{\omega}_q(\underline{0}) = \delta_{\alpha q}$, $\underline{0} \leqslant \alpha$, $q \leqslant \underline{s-1}$; в) если $t \in \partial Q$, то $D^\alpha\tilde{\omega}_q(t) = 0$, $\underline{0} \leqslant \alpha$, $q \leqslant \underline{s-1}$.

Координатные функции МКЭ выражаются через исходные функции (9.6.1) по формуле

$$\varphi_{qjh}(t) = \tilde{\omega}_q(t/h - j), \tag{9.6.2}$$

j — любой целочисленный вектор.

2°. Если $x \in C^{(s-1)}(Q)$, то функция

$$x^{(h)}(t) = \sum_{\underline{0} \leqslant q \leqslant \underline{s-1}} \sum_{\underline{0} \leqslant j \leqslant \underline{2n}} h^{|q|} x^{(q)}(jh)\,\tilde{\omega}_q(t/h - j) \tag{9.6.3}$$

принадлежит тому же классу $C^{(s-1)}(Q)$ и удовлетворяет условиям

$$D^\alpha x^h(j_0 h) = D^\alpha x(j_0 h), \qquad \underline{0} \leqslant j_0 \leqslant \underline{2n},\ \underline{0} \leqslant \alpha \leqslant \underline{s-1}; \tag{9.6.4}$$

функция (9.6.3), следовательно, осуществляет эрмитову интерполяцию функции $x(t)$ по заданным значениям этой последней и ее производных до порядка $s-1$ в узлах сетки, принадлежащих кубу Q.

Мы будем рассматривать также функции $x(t)$ класса $W_p^{(\tilde{s})}(Q)$, $\tilde{s} \geqslant s$, $1 \leqslant p \leqslant \infty$. Для таких функций мы определим $x^h(t)$ более сложной формулой

$$x^h(t) = \sum_{\underline{0} \leqslant q \leqslant \underline{s-1}} \sum_{\underline{0} \leqslant j \leqslant \underline{2n}} h^{|q|} \bar{x}^{(q)}(jh)\,\tilde{\omega}_q(t/h - j), \tag{9.6.5}$$

где $\bar{x}(t)$ — усреднение функции $x(t)$ ранга, не меньшего s, по Ильину — Головкину [59, 47]. Отметим следующее свойство такого усреднения: если его ранг равен $s_0 \leqslant \tilde{s}$, то можно выбрать такой достаточно малый радиус усреднения, чтобы

$$\|x - \bar{x}\|_{W_p^{(s_0)}} \leqslant \varepsilon, \tag{9.6.6}$$

ε — произвольно заданное положительное число.

Опираясь на усиленные фундаментальные соотношения, которыми удовлетворяют одномерные исходные функции $\omega_{q_k}(t_k)$ (см. § 1), можно доказать, что для функций класса $C^{(s)}(Q)$ погрешность аппроксимации стремится к нулю, когда $h \to 0$, а для класса $C^{(\tilde{s})}(Q)$, $s+1 \leqslant \tilde{s} \leqslant 2s$, эта погрешность имеет оценку

$$\|x - x^h\|_{C^{(s)}} \leqslant C h^{\tilde{s}-1} \max_{|\alpha|=\tilde{s}} \|x^{(\alpha)}\|_C. \tag{9.6.7}$$

Если $x \in W_p^{(\tilde{s})}(Q)$, $s+1 \leqslant \tilde{s} \leqslant 2s$, то положим в (9.6.6) $\varepsilon = Ch^{\tilde{s}-s} \max_{|\alpha|=\tilde{s}} \|x^{(\alpha)}\|_{L_p}$ и $s_0 = s$. Далее, построим x^h по формуле

9.6.5). По неравенству (9.6.7) получим

$$\|\bar{x} - x^h\|_{W_p^{(s)}} \leqslant Ch^{\bar{s}-s} \max_{|\alpha|=\bar{s}} \|x^{(\alpha)}\|_{L_p}$$

и, следовательно,

$$\|x - x^h\|_{W_p^{(s)}} \leqslant Ch^{\bar{s}-s} \max_{|\alpha|=\bar{s}} \|x^{(\alpha)}\|_{L_p}. \qquad (9.6.8)$$

Постоянные C в последних формулах, вообще говоря, различны.

3°. В кубе Q рассмотрим формально самосопряженный эллиптический дифференциальный оператор и соответствующее уравнение в частных производных

$$\sum_{|\alpha|,\ |\beta|=0}^{s} (-1)^{|\alpha|} D^{\alpha}(A_{\alpha\beta} D^{\beta} x) = f(t); \qquad (9.6.9)$$

коэффициенты $A_{\alpha\beta}$ будем считать достаточно гладкими, а краевые условия такими, что оператор задачи — положительно определенный в $L_2(Q)$. Допустим, что решение нашей задачи $x \in W_2^{(\bar{s})}(Q)$, $s + 1 \leqslant \bar{s} \leqslant 2s$. Будем решать эту задачу методом конечных элементов с координатными функциями (9.6.2); приближенное решение $x_h(t)$ находится из условия $|x_* - x_h|^2 = \min$, где x_* — точное решение задачи, а $|\cdot|$ означает энергетическую норму. В таком случае $|x_* - x_h| \leqslant |x_* - x^h|$, где x^h строится по формуле (9.6.5) с заменой $x(t)$ на $x_*(t)$. Заметим, что в данном случае энергетическая норма оценивается через норму $W_2^{(s)}$, и формула (9.6.8) дает оценку погрешности аппроксимации точного решения x_* приближенным решением x_h

$$|x_* - x_h| \leqslant Ch^{\bar{s}-s} \max_{|\alpha|=\bar{s}} \|x^{(\alpha)}\|_{L_2}. \qquad (9.6.10)$$

4°. Погрешность искажения для задачи (9.6.9) оценивается в общем по той же схеме, что и в одномерном случае (§ 3). Приближенное решение имеет вид

$$x_h(t) = \sum_{0 \leqslant q \leqslant \underline{s}-1} \sum_{0 \leqslant j \leqslant 2n} a_{qj}^{(h)} \tilde{\omega}_q(t/h - j). \qquad (9.6.11)$$

Совокупность коэффициентов $a_{qj}^{(h)}$ при фиксированном h, занумерованных в каком-нибудь определенном порядке, обозначим через $a^{(h)}$ и будем рассматривать как вектор в евклидовом пространстве R_N, где на этот раз $N = s^m(2n+1)^m$. Как и в § 3, обозначим матрицу МКЭ через M_h, а совокупность свободных членов системы МКЭ — через $F^{(h)}$. Эту совокупность также будем рассматривать как вектор в R_N. Неискаженную систему МКЭ можно записать в виде

$$M_h a^{(h)} = F^{(h)}. \qquad (9.6.12)$$

Введем N-мерные гильбертовы пространства X_N, Y_N с нормами

$$\|a\|_{X_N} = h^{m/2}\|a\|_{R_N}, \quad \|a\|_{Y_N} = h^{-m/2}\|a\|_{R_N}. \tag{9.6.13}$$

Если обозначить через a_{qj} составляющие произвольного вектора $a \in R_N$ и положить

$$x^h(t) = \sum_{0 \leqslant q \leqslant s-1} \sum_{0 \leqslant j \leqslant 2n} a_{qj} \tilde{\omega}_q(t/h - j),$$

то можно доказать неравенство, аналогичное неравенству (9.3.5):

$$c_1 \|a\|_{X_N} \leqslant \|x^h\|_{L_2} \leqslant c_2 \|a\|_{X_N}. \tag{9.6.14}$$

Если рассматривать векторы $a^{(h)}$ и $F^{(h)}$ как элементы пространств X_N и Y_N соответственно и обозначить через A_h оператор, порожденный матрицей M_h и действующий из X_N в Y_N, то уравнение (9.6.14) запишется в виде (9.3.6). Остаются верными теоремы 9.3.1 и 9.3.2; сохраняют силу с некоторыми очевидными изменениями и их доказательства. Сохраняет силу и формула (9.5.2), оценивающая число обусловленности матрицы МКЭ, как искаженной, так и неискаженной. Несколько изменяются оценки норм матрицы M_h и ее обратной: они имеют порядок $O(h^{m-2s})$ и $O(h^{-m})$ соответственно.

Поскольку остается в силе оценка числа обусловленности матрицы МКЭ, то сохраняются с небольшими изменениями оценки погрешностей алгоритма и округления, если систему МКЭ решать по методу простой итерации. В частности, главная по порядку часть оценки погрешности округления имеет вид

$$\|\delta(b^{(h)})\| \leqslant \varepsilon_1 \|\bar{F}\| \cdot O(h^{-4s-m/2}) \tag{9.6.15}$$

§ 7. О ПОГРЕШНОСТЯХ СОСТАВЛЕНИЯ СИСТЕМ МКЭ

1°. Аналогичный вопрос для систем методов Ритца и Бубнова — Галеркина был рассмотрен в § 10 главы 3.

С целью упростить изложение ограничимся рассмотрением простейшей задачи

$$\frac{d}{dt}\left(p(t)\frac{dx}{dt}\right) + q(t)x = f(t), \qquad 0 < t < 1, \tag{9.7.1}$$

$$x(0) = x(1) = 0. \tag{9.7.2}$$

Переход к произвольному числу координат и к произвольному порядку уравнений (или их систем) не привнесет принципиальных трудностей.

Допустим, что p, q, f — функции, достаточно гладкие при $0 \leqslant t \leqslant 1$, причем $p(t)$ — строго положительная функция, а

$q(t)$ — неотрицательная функция. Положим

$$\omega(t)=\begin{cases} t+1, & -1\leqslant t\leqslant 0;\\ 1-t, & 0\leqslant t\leqslant 1;\\ 0, & t\notin[-1,\ 1].\end{cases}$$

Выберем натуральное число n и определим координатные функции МКЭ по формуле

$$\varphi_{nj}(t)=\omega(t/h-j),\qquad h=1/2n.$$

Достаточно считать, что $j=1,\ 2,\ \ldots,\ 2n-2$; тогда координатные функции отличны от тождественного нуля при $0\leqslant t\leqslant 1$; при этом они удовлетворяют краевым условиям (9.7.2). Оператор задачи (9.7.1), (9.7.2) — положительно определенный в $L_2(0,\ 1)$, его энергетические произведение и норма определяются формулами

$$[x,\ y]=\int_0^1 [p(t)\,x'(t)\,y'(t)+q(t)\,x(t)\,y(t)]\,dt,\quad |x|^2=[x,\ x]. \tag{9.7.3}$$

Приближенное по МКЭ решение задачи (9.7.1), (9.7.2) имеет вид

$$x_h(t)=\sum_{k=1}^{2n-2} a_k^{(n)}\omega(t/h-k)=\sum_{k=1}^{2n-2} a_k^{(n)}\varphi_{nk}(t);$$

коэффициенты $a_k^{(n)}$ удовлетворяют алгебраической системе МКЭ

$$\sum_{k=1}^{2n-2} a_k^{(n)}[\varphi_{nk},\ \varphi_{nj}]=(f,\ \varphi_{nj}),\qquad j=1,\ 2,\ \ldots,\ 2n-2. \tag{9.7.4}$$

2°. Рассмотрим погрешности свободных членов. Для начала разберем случай, когда $f(t)$ есть полином. Имеем

$$(f,\ \varphi_{nj})=\int_0^1 f(t)\,\varphi_{nj}(t)\,dt=\int_0^1 f(t)\,\omega(t/h-j)\,dt.$$

Так как $1\leqslant j\leqslant 2n-2$, то

$$(f,\ \varphi_{nj})=\int_{(j-1)h}^{(j+1)h} f(t)\,\omega(t/h-j)\,dt=h\int_{-1}^{+1} f(jh+\tau h)\,\omega(\tau)\,d\tau=$$

$$=h\int_0^1 [f(jh-\tau h)+f(jh+\tau h)]\,(1-\tau)\,d\tau, \tag{9.7.5}$$

и дело сводится к вычислению интеграла от полинома, т. е. к конечному числу арифметических действий. Если последние выполняются с повышенной точностью, то с точностью до величин порядка $o(\varepsilon_1)$ будет $|\delta((f,\varphi_{nj}))\leqslant\varepsilon_1|(f,\varphi_{nj})|$. Отсюда легко получить оценку для нормы $\|\delta^{(n)}\|_{R_N}$, $N=2n-2$, где

$\delta^{(n)}$ — погрешность столбца свободных членов системы (9.7.4). Именно

$$\|\delta^{(n)}\|^2_{R_N} \leqslant \varepsilon_1^2 \sum_{k=1}^{2n-2} (f,\ \varphi_{nk})^2.$$

Далее,

$$(f,\ \varphi_{nk})^2 = \left[\int_{(k-1)h}^{(k+1)h} f(t)\,\omega(t/h-k)\,dt\right]^2 \leqslant$$

$$\leqslant \int_{(k-1)h}^{(k+1)h} f^2(t)\,dt \int_{(k-1)h}^{(k+1)h} \omega^2(t/h-k)\,dt = (2/3)\,h \int_{(k-1)h}^{(k+1)h} f^2(t)\,dt.$$

Суммируя по k, получаем

$$\|\delta^{(n)}\|_{R_N} \leqslant (2/\sqrt{3})\,h^{1/2}\varepsilon_1 \|f\|_{L_2}, \tag{9.7.6}$$

Больший интерес представляет оценка в метрике Y_N:

$$\|\delta^{(n)}\|_{Y_N} = h^{-1/2}\|\delta^{(n)}\|_{R_N} \leqslant (2/\sqrt{3})\,\varepsilon_1 \|f\|_{L_2}. \tag{9.7.7}$$

Если $f(t)$ не полином, то допустим, как и в § 10 главы 3, что значения $f(t)$ и некоторых ее производных могут быть вычислены с погрешностью $o(\varepsilon_1)$. Пусть $f \in C^{(s)}$. Выражение в квадратных скобках в (9.7.5) разложим по формуле Тейлора:

$$f(jh-\tau h)+f(jh+\tau h) = 2\sum_{\sigma=0}^{[(s-1)/2]} \frac{(\tau h)^{2\sigma}}{(2\sigma)!} f^{(2\sigma)}(jh) + r_s,$$

$$|r_s| \leqslant \frac{2h^s}{s!}\tau^s \max|f^{(s)}(t)|.$$

Если отбросить остаточный член, то остается интеграл от полинома, коэффициенты которого известны с точностью $o(\varepsilon_1)$. Продолжая вычислять с повышенной точностью интересующий нас интеграл и отбрасывая лишние знаки в его окончательном значении, мы получим значение интеграла (f, φ_{nj}) с погрешностью, которая складывается из двух частей. Одна из них, связанная с вычислением интеграла от полинома, оценивается с точностью до малых высших порядков величиной

$$\varepsilon_1 |(f,\ \varphi_{nj})|\,dt. \tag{9.7.8}$$

Другая погрешность, связанная с отбрасыванием остаточного члена формулы Тейлора, имеет оценку

$$\frac{2h^s}{(s+1)!} \max|f^{(s)}(t)|. \tag{9.7.9}$$

Из двух последних оценок нетрудно получить оценку для $\delta^{(n)}$ — погрешности вектора свободных членов системы МКЭ (9.7.4).

3°. Займемся оценкой погрешности матрицы системы (9.7.4). Имеем

$$[\varphi_{nk},\ \varphi_{nj}] = \int_0^1 p(t)\,\varphi'_{nk}(t)\,\varphi'_{nj}(t)\,dt + \int_0^1 q(t)\varphi_{nk}(t)\varphi_{nj}(t)dt := \alpha_{jk}^{(n)} + \beta_{jk}^{(n)}.$$

Величины $\alpha_{jk}^{(n)}$ и $\beta_{jk}^{(n)}$ могут быть отличны от нуля только тогда, когда носители функций φ_{nk} и φ_{nj} пересекаются. Это будет, если $j-k=-1, 0, +1$; так как матрица элементов $[\varphi_{nk}, \varphi_{nj}]$ симметрична, то достаточно рассмотреть случаи $j-k=0$ и $j-k=1$. Имеем

$$\alpha_{jj}^{(n)} = h^{-2}\int_0^1 p(t)\,\omega'^2(t/h-j)\,dt = h^{-2}\int_{(j-1)h}^{(j+1)h} p(t)\,dt =$$

$$= h^{-1}\int_0^1 [p(jh-\tau h)+p(jh+\tau h)]\,d\tau,$$

и если $p\in C^{(s)}[0, 1]$, то

$$\alpha_{jj}^{(n)} = \frac{2}{h}\sum_{\sigma=0}^{[(s-1)/2]} \frac{h^{2\sigma}}{(2\sigma+1)!}\,p^{(2\sigma)}(jh)+\bar{r}_{jj}, \qquad (9.7.10)$$

$$|\bar{r}_{jj}| \leqslant \frac{2h^s}{(s+1)!}\max|p^{(s)}(t)|.$$

Отсюда легко видеть $(\gamma_{jj}^{(n)} = \delta(\alpha_{jj}^{(n)}))$, что

$$|\gamma_{jj}^{(n)}| \leqslant \frac{2\varepsilon_1}{h}\left|\sum_{\sigma=0}^{[(s-1)/2]} \frac{h^{2\sigma}}{(2\sigma+1)!}\,p^{(2\sigma)}(jh)\right| + \frac{2h^s}{(s+1)!}\max|p^{(s)}(t)|. \qquad (9.7.11)$$

Если в формуле (9.7.11) ограничиться слагаемым с индексом нуль, то с точностью до величин высшего порядка малости по ε_1 и по h получится

$$|\gamma_{jj}^{(n)}| \leqslant 2\varepsilon_1 h^{-1}\max p(t) + h\varepsilon_1\cdot 3^{-1}\max|p''(t)|.$$

При малых h второй член справа меньше первого. Пренебрегая им, получим оценку

$$|\gamma_{jj}^{(n)}| \leqslant 2\varepsilon_1 h^{-1}\max p(t). \qquad (9.7.12)$$

Рассмотрим теперь величину

$$\alpha_{j,\,j-1}^{(n)} = h^{-2}\int_0^1 \omega'(t/h-j)\,\omega'(t/h-j+1)\,p(t)\,dt.$$

Носители сомножителей под интегралом пересекаются по промежутку $[(j-1)h,\ jh]$; на этом промежутке $\omega'(t/h-j)=1$, $\omega'(t/h-j+1)=-1$, так что

$$\alpha_{j,\,j-1}^{(n)} = -h^{-2}\int_{(j-1)h}^{jh} p(t)\,dt = -h^{-1}\int_0^1 p((j-1)h+\tau h)\,d\tau.$$

Так как $p\in C^{(s)}[0, 1]$, то по формуле Тейлора

$$p((j-1)h+\tau h) = \sum_{\sigma=0}^{s-1}\frac{\tau^\sigma h^\sigma}{\sigma!}\,p^{(\sigma)}((j-1)h)+\bar{r}_{j,\,j-1},$$

$$|\bar{r}_{j,\,j-1}| \leqslant \frac{\tau^s h^s}{s!}\max|p^{(s)}(t)|;$$

отсюда

$$\alpha^{(n)}_{j,\,j-1} = -\frac{1}{h}\sum_{\sigma=0}^{s-1}\frac{h^\sigma}{(\sigma+1)!}\,p^{(\sigma)}((j-1)h) + \theta\frac{h^{s-1}}{(s+1)!}\max|p^{(s)}(t)|,$$

$$|\theta| \leqslant 1, \qquad (9.7.13)$$

и, следовательно,

$$\gamma^{(n)}_{j,\,j-1} := \delta\left(\alpha^{(n)}_{j,\,j-1}\right) = \frac{\varepsilon_1\theta'}{h}\sum_{\sigma=0}^{s-1}\frac{h^\sigma}{(\sigma+1)!}\,p^{(\sigma)}((j-1)h) +$$

$$+\theta'\varepsilon_1\frac{h^{s-1}}{(s+1)!}\max|p^{(s)}(t)|, \quad |\theta'| \leqslant 1; \qquad (9.7.14)$$

из последней формулы следует с точностью до величин более высокого порядка малости

$$|\gamma^{(n)}_{j,\,j-1}| \leqslant \varepsilon_1 h^{-1}\max p(t). \qquad (9.7.15)$$

Аналогично можно вычислить $\beta^{(n)}_{jk}$ и оценить погрешность $\delta(\beta^{(n)}_{jk})$. Нетрудно видеть, что ($k = j-1$, j)

$$|\delta(\beta_{jk})| \leqslant 2\varepsilon_1 h\max q(t) + (\varepsilon_1 h^3/12)\max|q''(t)|;$$

при малых h эта величина мала по сравнению с правой частью формулы (9.7.15). Это дает основание пренебречь погрешностью $\delta(\beta^{(n)}_{jk})$ и позволяет считать, что (с точностью до малых высших порядков)

$$\delta([\varphi_{nj},\ \varphi_{nk}]) = \gamma^{(n)}_{jk}, \qquad j-k = 0,\ 1.$$

Пусть $\Gamma_n = \left[\gamma^{(n)}_{jk}\right]^N_{j,\,k=1}$ — погрешность матрицы МКЭ. По формулам (9.7.12), (9.7.15) элементы матрицы Γ_n не превосходят по модулю элементов матрицы

$$\varepsilon_1 h^{-1}\max p(t)\,\Delta_N, \qquad (9.7.16)$$

где

$$\Delta_N = \begin{bmatrix} 2 & 1 & & & & & \\ 1 & 2 & 1 & & & & \\ & 1 & 2 & 1 & & & \\ & & & \cdot & & & \\ & & & & \cdot & & \\ & & & & & \cdot & \\ & & & & & 1\ 2\ 1 & \\ & & & & & 1\ 2 & \end{bmatrix}_N.$$

Оценим норму $\|\Delta_N\|$ как оператора в R_N: она равна наибольшему корню уравнения $\mathrm{Det}[\Delta_N - \lambda I] = 0$, I — единичная матрица порядка N. Положив $2 - \lambda = 2\cos\vartheta$, преобразуем последнее уравнение к виду (см. главу 5, § 4)

$$\sin(N+1)\vartheta/\sin\vartheta = 0;$$

в промежутке $0 \leqslant \vartheta \leqslant 2\pi$ лежат корни этого уравнения $\vartheta_k = k\pi/(N+1)$, $k = 1, 2, \ldots, N$; соответствующие значения λ^k суть

$$\lambda_k = 2 - 2\cos\frac{k\pi}{N+1} = 4\sin^2\frac{k\pi}{2(N+1)}.$$

Наибольший из корней λ_k равен $\lambda_N = 4\sin^2\frac{N\pi}{2(N+1)} < 4$, и, следовательно, $\|\Delta_N\|_{R_N} < 4$. Отсюда и из формулы (9.7.16) вытекает, что с точностью до величин, малых относительно h и ε_1,

$$\|\Gamma_n\|_{R_N} \leqslant 4\varepsilon_1 h^{-1} \max p(t). \qquad (9.7.17)$$

Глава 10

КОНСТАНТЫ В ОЦЕНКАХ АППРОКСИМАЦИИ МКЭ

Задача настоящей главы — оценить постоянную C в неравенстве (9.2.2), которое, в свою очередь, дает оценку погрешности аппроксимации МКЭ при использовании кусочно-полиномиальных исходных функций, удовлетворяющих усиленным фундаментальным соотношениям; самые исходные функции определяются формулами (9.1.9), (9.1.10). Для дальнейшего важно подчеркнуть, что исходные функции зависят от s, а постоянная C зависит от s и $\bar{s}$, поэтому мы введем обозначения $\omega_{qs}(t)$ для исходных функций и $C(s, \bar{s})$ для постоянной C из формулы (9.2.2).

Результаты настоящей главы содержатся в [97, 180]

§ 1. ОПТИМАЛЬНЫЕ ПОЛИНОМЫ

1°. Полином во второй строке справа в формуле (9.1.9) обозначим через $\sigma_{qs}(t)$:

$$\sigma_{qs}(t) = \frac{1}{q!}(1-t)^s t^q \left(a_0^{(s)} + a_1^{(s)} t + \ldots + a_{s-q-1}^{(s)} t^{s-q-1}\right), \qquad (10.1.1)$$

так что

$$\begin{aligned} &\forall t, \quad 0 \leqslant t \leqslant 1, \quad \omega_{qs}(t) = \sigma_{qs}(t); \qquad (10.1.2)\\ &-1 \leqslant t \leqslant 0, \quad \omega_{qs}(t) = (-1)^q \sigma_{qs}(-t). \end{aligned}$$

Полиномы $\sigma_{qs}(t)$ мы назвали оптимальными (см. главу 9, § 1), имея в виду, что они приводят к наилучшей по порядку оценке погрешности аппроксимации МКЭ.

В § 1 главы 9 было отмечено, что $a_k^{(s)}$ суть коэффициенты ряда Маклорена функции $(1-t)^{-s}$ и что все эти коэффициенты

неотрицательны. Отсюда следует, что на отрезке $[0, 1]$

$$0 \leqslant \sigma_{qs}(t) \leqslant 1/q! \tag{10.1.3}$$

Известное неравенство А. А. Маркова позволяет сразу, написать оценку для производных

$$|\sigma'_{qs}(t)| \leqslant 2(2s-1)^2/q! \tag{10.1.4}$$

и аналогичные оценки для старших производных. Как известно, в классе всех полиномов оценка А. А. Маркова точна, однако для весьма специальных оптимальных полиномов марковские оценки оказываются завышенными. В последующем будут получены более точные оценки производных порядка $\bar{s} \leqslant s$ (именно эти производные и представляют интерес для МКЭ) от оптимальных полиномов.

2°. Начнем с производных от полиномов $\sigma_{0s}(t)$. Имеем $\sigma_{0s}(t) = (1-t)^s \sum_{k=0}^{s-1} a_k^{(s)} t^k$. Отсюда

$$\sigma'_{0s}(t) = -s(1-t)^{s-1} \sum_{k=0}^{s-1} a_k^{(s)} t^k + (1-t)^s \frac{d}{dt} \sum_{k=0}^{s-1} a_k^{(s)} t^k \tag{10.1.5}$$

Производную, написанную справа, можно получить, взяв соответствующий отрезок предварительно продифференцированного ряда Маклорена функции $(1-t)^s$. Но

$$\frac{d}{dt} \sum_{k=0}^{\infty} a_k^{(s)} t^k = s(1-t)^{-(s+1)} = s \sum_{k=0}^{\infty} a_k^{(s+1)} t^k,$$

поэтому

$$\frac{d}{dt} \sum_{k=0}^{s-1} a_k^{(s)} t^k = s \sum_{k=0}^{s-2} a_k^{(s+1)} t^k.$$

Подставив это в (10.1.5) и воспользовавшись формулами (9.1.10), (9.1.10а), получим

$$\sigma'_{0s}(t) = -s \binom{2s-1}{s-1} [t(1-t)]^{s-1}. \tag{10.1.6}$$

По формуле Стирлинга

$$\binom{2s-1}{s-1} < \frac{e^{7/12}}{\sqrt{\pi}} \frac{2^{2s-1}}{\sqrt{s-1}} \approx 2{,}03 \frac{2^{2s-2}}{\sqrt{s-1}}.$$

Далее, $\max t(1-t) = 1/4$, и мы приходим к следующей оценке производной:

$$|\sigma'_{0s}(t)| \leqslant 2{,}03/\sqrt{s-1}$$

Заметим, что формула (10.1.4) приводит к менее точному неравенству $\sigma'_{0s}(t) \leqslant 2(2s-1)^2$.

Дифференцируя формулу (10.1.6), получаем

$$\sigma''_{0s}(t) = -s(s-1)\binom{2s-1}{s-1}[t(1-t)]^2(1-2t). \quad (10.1.7)$$

3°. Рассмотрим теперь производную $\sigma^{(\bar{s})}_{0s}(t)$, $2 \leqslant \bar{s} \leqslant s$. К тождеству (10.1.6) применим формулу Лейбница:

$$\sigma^{(\bar{s})}_{0s}(t) = s\binom{2s-1}{s-1}\sum_{j=0}^{\bar{s}-1}(-1)^{\bar{s}-j}\binom{s-1}{j}(s-1)\dots(s-j)\times$$

$$\times(s-1)\dots(s-\bar{s}+j+1)t^{s-j-1}(1-t)^{s-\bar{s}+j}. \quad (10.1.8)$$

Число множителей вида $s-l$ в формуле (10.1.8) равно $\bar{s}-1$. Пусть, например, $j \leqslant \bar{s}-j-1$, т. е. $j \leqslant (\bar{s}-1)/2$. Произведение названных множителей равно

$$(s-1)^2\dots(s-j)^2(s-j-1)\dots(s-(\bar{s}-j-1)). \quad (10.1.9)$$

В произведении $(s-j-1)\dots(s-(\bar{s}-j-1))$ множители расположены в порядке убывания. Заменим в этом произведении последнюю скобку первой, предпоследнюю — второй и т. д. Интересующее нас произведение оказывается меньше величины

$$(s-j-1)^2(s-j-2)^2\dots(s-(\bar{s}-1)/2)^2,\ \bar{s}\text{ — нечетное};$$

$$(s-j-1)^2(s-j-2)^2\dots(s-\bar{s}/2+1)^2(s-\bar{s}/2),\ \bar{s}\text{ — четное},$$

а произведение (10.1.9) меньше величины

$$\mu(s,\bar{s}) := \begin{cases} (s-1)^2(s-2)^2\dots(s-(\bar{s}-1)/2)^2, & \bar{s}\text{ — нечетное}, \\ (s-1)^2(s-2)^2\dots(s-\bar{s}/2+1)^2(s-\bar{s}/2), & \bar{s}\text{ — четное}; \end{cases} \quad (10.1.10)$$

в формуле (10.1.10) $\bar{s} > 2$. Тот же результат получается, если $j > \bar{s}-j-1$. Теперь из (10.1.9) вытекает неравенство

$$|\sigma^{(\bar{s})}_{0s}(t)| \leqslant s\binom{2s-1}{s-1}\mu(s,\bar{s})\sum_{j=0}^{\bar{s}-1}\binom{\bar{s}-1}{j}t^{s-j-1}(1-t)^{s-\bar{s}+j} =$$

$$= s\binom{2s-1}{s-1}\mu(s,\bar{s})[t(1-t)]^{s-\bar{s}}\sum_{j=0}^{\bar{s}-1}\binom{\bar{s}-1}{j}t^{\bar{s}-j-1}(1-t)^j.$$

Последняя сумма равна единице, и окончательно

$$|\sigma^{(\bar{s})}_{0s}(t)| \leqslant s\binom{2s-1}{s-1}\mu(s,\bar{s})[t(1-t)]^{s-\bar{s}}. \quad (10.1.11)$$

Заметим, что неравенство (10.1.11) будет верным и для $\bar{s} = 1, 2$, если положить

$$\mu(s,1) = 1, \quad \mu(s,2) = s-1. \quad (10.1.12)$$

В табл. 15 приводятся значения функции $\mu(s, \bar{s})$.

Таблица 15

s \ $\bar{s}$	1	2	3	4	5	6
3	1	2	4			
4	1	3	9	18		
5	1	4	16	48	144	
6	1	5	25	100	400	1200

4°. Ниже будут получены оценки производных оптимальных полиномов с индексами $q > 0$. Отдельно рассмотрим производные первого и второго порядков. Обозначая

$$(1-t)^s \sum_{k=0}^{s-q-1} a_k^{(s)} t^k = \psi_{qs}(t), \tag{10.1.13}$$

имеем $\sigma_{qs}(t) = t^q \psi_{qs}(t)/q!$. Отсюда

$$\sigma'_{qs}(t) = \frac{t^{q-1}}{(q-1)!} \psi_{qs}(t) + \frac{t^q}{q!} \psi'_{qs}(t).$$

Аналогично

$$\psi'_{qs}(t) = -s \binom{2s-q-1}{s-q-1} t^{s-q-1}(1-t)^{s-1} \tag{10.1.14}$$

и, следовательно,

$$\sigma'_{qs}(t) = \frac{t^{q-1}}{(q-1)!} \psi_{qs}(t) - \frac{s}{q!} \binom{2s-q-1}{s-q-1} [t(1-t)]^{s-1}. \tag{10.1.15}$$

Дифференцируя формулу (10.1.15), получаем

$$\sigma''_{qs}(t) = \frac{t^{q-2}}{(q-2)!} \psi_{qs}(t) + \frac{t^{q-1}}{(q-1)!} \psi'_{qs}(t) -$$

$$- \frac{s(s-1)}{q!} \binom{2s-q-1}{s-q-1} [t(t-1)]^{s-2}(1-2t),$$

или по формуле (10.1.14)

$$\sigma''_{qt}(t) = \frac{t^{q-2}}{(q-2)!} \psi_{qs}(t) - \frac{s}{q!} \binom{2s-q-1}{s-q-1} [t(1-t)]^{s-2} \times$$

$$\times (q+s-1-t(q+2s-2)). \tag{10.1.16}$$

Очевидно, что $0 \leqslant \psi_{qs}(t) \leqslant 1$; приняв еще во внимание, что линейная на сегменте функция достигает максимума модуля на одном из концов сегмента, получаем следующие оценки:

$$|\sigma'_{qs}(t)| \leqslant \frac{t^{q-1}}{(q-1)!} + \frac{s}{q!} \binom{2s-q-1}{s-q-1} [t(1-t)]^{s-1}; \tag{10.1.17}$$

$$|\sigma''_{qs}(t)| \leqslant \frac{t^{q-2}}{(q-2)!} + \frac{s(s+q-1)}{q!} \binom{2s-q-1}{s-q-1} [t(1-t)]^{s-2}. \tag{10.1.18}$$

5°. Переходим к общему случаю $2 < \bar{s} \leqslant s$. Имеем

$$\sigma_{qs}(t) = \frac{t}{q}\frac{t^{q-1}}{(q-1)!}(1-t)^s\left(a_0^{(s)} + a_1^{(s)}t + \ldots + a_{s-q-1}^{s}t^{s-q-1}\right) =$$

$$= \frac{t}{q}\sigma_{q-1,\,s}(t) - \frac{a_{s-q}^{(s)}}{q!}[t(1-t)]^s. \qquad (10.1.19)$$

Отсюда

$$\sigma'_{qs}(t) = \frac{1}{q}\sigma_{q-1,\,s}(t) + \frac{t}{q}\sigma'_{q-1,\,s}(t) - \frac{s}{q!}a_{s-q}^{(s)}[t(1-t)]^{s-1}(1-2t). \qquad (10.1.20)$$

Продифференцируем это $\bar{s}-1$ раз по формуле Лейбница:

$$\sigma_{qs}^{(\bar{s})}(t) = \frac{1}{q}\sigma_{q-1,\,s}^{(\bar{s})}(t) + \frac{t}{q}\sigma_{q-1,\,s}^{(\bar{s})}(t) + \frac{\bar{s}-1}{q}\sigma_{q-1,\,s}^{(\bar{s})}(t) -$$

$$- \frac{s}{q!}a_{s-q}^{(s)}(1-2t)\frac{d^{\bar{s}-1}}{dt^{\bar{s}-1}}[t(1-t)]^{s-1} + \frac{s(s-1)}{q!}a_{s-q}^{(s)}\frac{d^{\bar{s}-2}}{dt^{\bar{s}-2}}[t(1-t)]^{\bar{s}-1}.$$

Из формулы (10.1.6) получаем

$$s\frac{d^{\bar{s}-1}}{dt^{\bar{s}-1}}[t(1-t)]^{s-1} = -\binom{2s-1}{s-1}^{-1}\sigma_{0s}^{(\bar{s})}(t),$$

$$s\frac{d^{\bar{s}-2}}{dt^{\bar{s}-2}}[t(1-t)]^{s-1} = -\binom{2s-1}{s-1}^{-1}\sigma_{0s}^{(\bar{s}-1)}(t).$$

Это приводит к следующей рекуррентной формуле:

$$\sigma_{qs}^{(\bar{s})} = \frac{s}{q}\sigma_{q-1,\,s}^{(\bar{s}-1)}(t) + \frac{t}{q}\sigma_{q-1,\,s}^{(\bar{s})}(t) +$$

$$+ \binom{2s-q-1}{s-q-1}\left[q!\binom{2s-1}{s-1}\right]^{-1}(1-2t)\sigma_{0s}^{(\bar{s})}(t) -$$

$$- 2(s-1)\binom{2s-q-1}{s-q-1}\left[q!\binom{2s-1}{s-1}\right]^{-1}\sigma_{0s}^{(\bar{s}-1)}(t). \qquad (10.1.21)$$

§ 2. НОРМЫ ПРОИЗВОДНЫХ ОПТИМАЛЬНЫХ ПОЛИНОМОВ

1°. Норму производной $\sigma_{qs}^{(\bar{s})}(t)$ в пространстве $L_p(0, 1)$ будем обозначать через $A_{qsp}^{(\bar{s})}$; вместо $A_{qs\infty}^{(\bar{s})}$ будем писать $A_{qs}^{(\bar{s})}$. Отдельно рассмотрим случаи $\bar{s} = 0$, 1, 2 и $2 < \bar{s} \leqslant s$; сначала выведем оценки для $A_{qs}^{(\bar{s})}$, а затем для $A_{qsp}^{(\bar{s})}$.

Формула (10.1.3) означает, что

$$A_{qs}^{(0)} \leqslant 1/q! \qquad q = 0, 1, \ldots, s-1. \qquad (10.2.1)$$

Если $t \in [0, 1]$, то $t(1-t) \leqslant 1/4$, и из формул (10.1.6), (10.1.7) (10.1.11) вытекают оценки

$$A_{0s}^{(1)} \leqslant s\binom{2s-1}{s-1}2^{-2s+2}, \quad A_{0s}^{(2)} = s(s-1)\binom{2s-1}{s-1}2^{-2s+4}; \tag{10.2.2}$$

$$2 < \bar{s} \leqslant s, \quad A_{0s}^{(\bar{s})} \leqslant s\binom{2s-1}{s-1}\mu(s, \bar{s})\,2^{-2(s-\bar{s})}. \tag{10.2.3}$$

Неравенство (10.2.3) справедливо и для значений $\bar{s} = 1, 2$, если для этих значений определить μ по формулам (10.1.12).

Из формул (10.1.17), (10.1.18) следует далее

$$A_{qs}^{1} \leqslant \frac{1}{(q-1)!} + \frac{s}{q}\binom{2s-q-1}{s-q-1}2^{-2s+2}; \tag{10.2.4}$$

$$A_{qs}^{(2)} \leqslant \frac{1}{(q-2)!} + \frac{s(s+q-1)}{q}\binom{2s-q-1}{s-q-1}2^{-2s+4}, \quad q > 0, \tag{10.2.5}$$

а из формулы (10.1.21) вытекает неравенство

$$|\sigma_{qs}^{(\bar{s})}(t)| \leqslant \frac{s}{q}|\sigma_{q-1,s}^{(\bar{s}-1)}(t)| + \frac{1}{q}|\sigma_{q-1,s}^{(\bar{s})}(t)| +$$

$$+\binom{2s-q-1}{s-q-1}\left[q!\binom{2s-1}{s-1}\right]^{-1}\left[|\sigma_{0s}^{(\bar{s})}(t)| + 2(s-1)|\sigma_{0s}^{(\bar{s}-1)}(t)|\right].$$

Пусть $\|\cdot\|$ — произвольная монотонная норма — это значит, что $\|u\| \leqslant \|v\|$, если $|u(t)| \leqslant |v(t)|$, $0 \leqslant t \leqslant 1$. Тогда из последнего неравенства следует

$$\|\sigma_{qs}^{(\bar{s})}\| \leqslant \frac{1}{q}\left[s\|\sigma_{q-1,s}^{(\bar{s}-1)}\| + \|\sigma_{q-1,s}^{(\bar{s})}\| +\right.$$

$$\left.+\binom{2s-q-1}{s-q-1}\left[q!\binom{2s-1}{s-1}\right]^{-1}\left[2(s-1)\|\sigma_{0s}^{(\bar{s}-1)}\| + \|\sigma_{0s}^{(\bar{s})}\|\right]\right. \tag{10.2.6}$$

В частности,

$$A_{qp}^{(\bar{s})} \leqslant \frac{1}{q}(sA_{q-1,s}^{(\bar{s}-1)} + A_{q-1,s}^{(\bar{s})}) + \binom{2s-q-1}{s-q-1} \times$$

$$\times\left[q!\binom{2s-1}{s-1}\right]^{-1}\left[2(s-1)A_{0s}^{(\bar{s}-1)} + A_{0s}^{(\bar{s})}\right]. \tag{10.2.7}$$

2°. Переходим к неравенствам в метриках $L_p(0, 1)$. Очевидно, что $A_{qsp}^{(\bar{s})} \leqslant A_{qs}^{(\bar{s})}$. Уточним эту оценку. Прежде всего

$$A_{qsp}^{(\bar{s})} = \left[\int_0^1 \sigma_{qs}^{p}(t)\,qt\right]^{1/p} < \frac{1}{q!}\left[\int_0^1 t^{pq}\,dt\right]^{1/p} = \frac{1}{q!\,(pq+1)^{1/p}}. \tag{10.2.8}$$

Теперь оценим производные от оптимальных полиномов. По формуле (10.1.11), полагая $q = 0$, получаем

$$A_{0sp}^{(\bar{s})} \leqslant s\binom{2s-1}{s-1}\mu(s, \bar{s})\left[\int_0^1 [t(1-t)]^{p(s-\bar{s})}\,dt\right]^{1/p} =$$

$$= s\binom{2s-1}{s-1}\mu(s, \bar{s})\,\mathrm{B}^{1/p}(p(s-\bar{s})+1,\ p(s-\bar{s})+1) =$$

$$= s\binom{2s-1}{s-1}\mu(s, \bar{s})\frac{\Gamma^{2/p}(p(s-\bar{s})+1)}{\Gamma^{1/p}(2p(s-\bar{s})+2)}.$$

Если $s = \bar{s}$, то

$$A_{0sp}^{(s)} \leqslant s\binom{2s-1}{s-1}\mu(s, \bar{s}).$$

В общем случае, когда $\bar{s} < s$, воспользуемся формулой Стирлинга, записанной в виде

$$z > 0, \qquad \Gamma(z+1) = \sqrt{2\pi z}\,(z/e)^z e^{\theta/12z}, \qquad 0 < \theta < 1.$$

Из этой формулы следует

$$\Gamma^{2/p}(p(s-\bar{s})+1) < [2\pi p(s-\bar{s})]^{1/p}\left[\frac{p(s-\bar{s})}{e}\right]^{2(s-\bar{s})} e^{1/6p^2(s-\bar{s})};$$

$$\Gamma^{1/p}(2p(s-\bar{s})+2) > [2p(s-\bar{s})]^{1/p}\,\Gamma^{1/p}(2p(s-\bar{s})+1) >$$

$$> (2\pi)^{1/p}[2p(s-\bar{s})]^{3/2p}\left[\frac{2p(s-\bar{s})}{e}\right]^{2(s-\bar{s})}.$$

Отсюда

$$\frac{\Gamma^{2/p}(p(s-\bar{s})+1)}{\Gamma^{1/p}(2p(s-\bar{s})+2)} < 2^{-2(s-\bar{s})}\left[\frac{\pi}{4p(s-\bar{s})}\right]^{1/2p} e^{1/6p^2(s-\bar{s})}, \quad (10.2.9)$$

и, следовательно,

$$A_{0ps}^{(\bar{s})} \leqslant s\binom{2s-s}{s-\bar{s}}\mu(s, \bar{s})\,2^{-2(s-\bar{s})}\left[\frac{\pi}{4p(s-\bar{s})}\right]^{1/2p} e^{1/6p^2(s-\bar{s})}. \quad (10.2.10)$$

3°. Пусть теперь $q > 0$. Если $2 < \bar{s} \leqslant s$, то верна рекуррентная формула, вытекающая из (10.2.6) и аналогичная формуле (10.2.7):

$$A_{qsp}^{(\bar{s})} \leqslant \frac{1}{q}\left(sA_{q-1,\,sp}^{(\bar{s}-1)} + A_{q-1,\,sp}^{(\bar{s})}\right) +$$

$$+\binom{2s-q-1}{s-q-1}\left[q!\binom{2s-1}{s-1}\right]^{-1}\left(2(s-1)A_{0sp}^{(\bar{s}-1)} + A_{0sp}^{(\bar{s})}\right). \quad (10.2.11)$$

Особо рассмотрим случаи $s=1$ и $\bar{s}=2$. По формуле (10.1.17)

$$A_{qsp}^{(1)} \leqslant \frac{1}{(q-1)!\,[p(q-1)+1]^{1/p}} + \frac{s}{q!}\binom{2s-q-1}{s-q-1}\frac{\Gamma^{2/p}(p(s-1)+1)}{\Gamma^{1/p}(2p(s-1)+2)}.$$

Аналогично по формуле (10.1.18)

$$A_{qsp}^{(2)} \leqslant \frac{1}{(q-2)!\,[p(q-2)+1]^{1/p}} +$$

$$+ \frac{s(s+q-1)}{q!}\binom{2s-q-1}{s-q-1}\frac{\Gamma^{2/p}(p(s-2)+1)}{\Gamma^{1/p}(2p(s-2)+2)}.$$

По неравенству (10.2.9) получаем

$$A_{qsp}^{(1)} \leqslant \frac{1}{(q-1)!\,[p(q-1)+1]^{1/p}} +$$

$$+ \frac{s}{q!}\binom{2s-q-1}{s-q-1}2^{-2(s-1)}\left[\frac{\pi}{4p(s-1)}\right]^{1/2p} e^{1/6p^2(s-1)}; \quad (10.2.12)$$

$$\cdot A_{qsp}^{(2)} \leqslant \frac{1}{(q-2)!\,[p(q-2)+1]^{1/p}} +$$

$$+ \frac{s(s+q-1)}{q!}\binom{2s-q+1}{s-q-1}2^{-2(s-2)}\left[\frac{\pi}{4p(s-2)}\right]^{1/p} e^{1/6p^2(s-2)}.$$

§ 3. О РОСТЕ ПРОИЗВОДНЫХ ОПТИМАЛЬНЫХ ПОЛИНОМОВ

1°. Оценим порядок роста производных оптимальных полиномов в зависимости от роста s и $\bar{s}$. Прежде всего отметим, что $\binom{2s-1}{s-1}$ есть один из коэффициентов разложения бинома Ньютона с показателем $2s-1$, поэтому

$$\binom{2s-1}{s-1} < \frac{1}{2}\cdot 2^{2s-1} = 2^{2s-2} \quad (10.3.1)$$

Теперь по формуле (10.2.3) получаем

$$A_{0s}^{(\bar{s})} < s\mu(s, \bar{s})\,2^{2\bar{s}-1} \quad (10.3.2)$$

Сравним оценку (10.3.2) с оценкой, вытекающей из неравенства А. А. Маркова (сравнение с более точной формулой В. А. Маркова приводит качественно к тому же результату):

$$A_{0s}^{(\bar{s})} \leqslant 2^{\bar{s}}(2s-1)^2(2s-2)^2\ldots(2s-\bar{s})^2. \quad (10.3.3)$$

Пусть, например, $\bar{s}$ — нечетное. В силу формулы (10.1.10)

$$A_{0s}^{(\bar{s})} < 2^{2\bar{s}-2}s(s-1)^2(s-2)^2\ldots(s-(\bar{s}-1)/2)^2. \quad (10.3.4)$$

Правую часть формулы (10.3.3) запишем в виде

$$2^{3\bar{s}}(s-1/2)^2(s-1)^2\ldots(s-\bar{s}/2)^2=2^{3\bar{s}}(s-1/2)^2(s-3/2)^2\ldots$$
$$\ldots(s-\bar{s}/2)^2(s-1)^2(s-2)^2\ldots(s-(\bar{s}-1)/2)^2. \quad (10.3.5)$$

Отношение правых частей формул (10.3.4) и (10.3.5) равно

$$2^{-\bar{s}-2}s[(s-1/2)(s-3/2)\ldots(s-\bar{s}/2)]^2.$$

Отсюда видно, что для оптимальных полиномов оценка (10.3.2) существенно лучше общей оценки А. А. Маркова. Тот же результат получается и для четных $\bar{s}$.

2°. Оценку (10.3.2) можно несколько улучшить по порядку, если оценить величину $\binom{2s-1}{s-1}$ по формуле Стирлинга. Имеем

$$\binom{2s-1}{s-1}=\frac{(2s-1)!}{(s-1)!s!}=\sqrt{\frac{2s-1}{2\pi s(s-1)}}\,\frac{(2s-1)^{2s-1}}{(s-1)^{s-1}s^s}\times$$
$$\times\exp\left[\frac{1}{12}\left(\frac{\theta}{2s-1}-\frac{\theta'}{s-1}-\frac{\theta''}{s}\right)\right],\quad 0<\theta,\ \theta',\ \theta''<1.$$

Отсюда

$$\binom{2s-1}{s-1}<\frac{1}{\sqrt{2\pi}}\sqrt{\frac{1}{s}+\frac{1}{s-1}}\left(\frac{2s-1}{s-1}\right)^{s-1}\left(\frac{2s-1}{s}\right)^s e^{1/12}<$$
$$<\frac{e^{1/12}}{\sqrt{\pi(s-1)}}2^s\left(\frac{2s-1}{s-1}\right)^{s-1}.$$

Далее, $\left(\frac{2s-1}{s-1}\right)^{s-1}\leqslant 2^{s-1}\left(1+\frac{1}{2(s-1)}\right)^{s-1}<2^{s-1}\sqrt{e}$ и окончательно

$$\binom{2s-1}{s-1}<\frac{e^{7/12}}{\sqrt{\pi(s-1)}}2^{2s-1}<2{,}03\,\frac{2^{2s-2}}{\sqrt{s-1}}. \quad (10.3.6)$$

Теперь можно заменить оценку (10.3.2) следующей, более точной по порядку:

$$A_{0s}^{(\bar{s})}<\frac{2e^{7/12}}{\sqrt{\pi}}\,\frac{1}{\sqrt{s-1}}\,\mu(s,\bar{s})\,2^{2\bar{s}-2}<2{,}03\cdot 2^{(\bar{s}-1)}\,\frac{1}{\sqrt{s-1}}\,\mu(s,\bar{s}).$$
(10.3.7)

Отметим, что при $s>5$ оценка (10.3.7) точнее оценки (10.3.2) и численно. Ниже мы все же будем пользоваться более простыми оценками (10.3.1), (10.3.2).

3°. Выясним порядок роста правой части формулы (10.3.2) при возрастании $\bar{s}$. Отношение этих правых частей, соответ-

ствующих значениям $\bar{s}$ и $\bar{s}-1$, равно $4\mu(s, \bar{s})/\mu(s, \bar{s}-1)$. Если $\bar{s}$ — четное, то

$$\mu(s, \bar{s}) = (s-1)^2 \ldots (s-\bar{s}/2+1)^2(s-\bar{s}/2),$$

$$\mu(s, \bar{s}-1) = (s-1)^2 \ldots (s-\bar{s}/2-1)^2,$$

и интересующее нас отношение равно $2(2s-\bar{s})$. Если же $\bar{s}$ — нечетное, то указанное отношение равно $2(2s-\bar{s}+1)$.

4°. Допустим, что для некоторого $q-1$ верна оценка

$$A^{(\bar{s})}_{q-1,\,s} \leqslant \frac{C_{q-1}}{(q-1)!}\, 2^{2\bar{s}-2} s\mu(s, \bar{s}), \tag{10.3.8}$$

в которой C_{q-1} — некоторая постоянная. Оценим величину $A^{(\bar{s})}_{qs}$. Заметим прежде всего, что в силу формулы (10.3.2) оценка (10.3.8) верна для индекса $q-1=0$, причем можно положить $C_0 = 1$. Из формул (10.3.8) и (10.2.7) следует

$$A^{(\bar{s})}_{qs} \leqslant \frac{1}{q!}\Big\{ C_{q-1}\left[2^{2\bar{s}-4}s^2\mu(s, \bar{s}-1) + 2^{2\bar{s}-2}s\mu(s, \bar{s})\right] +$$

$$+ \binom{2s-q-1}{s-q-1}\binom{2s-1}{s-1}^{-1}\left[2^{2\bar{s}-3}s(s-1)\mu(s, \bar{s}-1) + \right.$$

$$\left. + 2^{2\bar{s}-2}s\mu(s, \bar{s})\right]\Big\}. \tag{10.3.9}$$

Оценку (10.3.9) упростим. Прежде всего оценим коэффициент перед второй квадратной скобкой. Имеем

$$\binom{2s-q-1}{s-q-1}\binom{2s-1}{s-1}^{-1} = \frac{(2s-q-1)!}{(s-q-1)!\,s!}\,\frac{s!\,(s-1)!}{(2s-1)!} =$$

$$= \frac{s-q}{2s-q}\cdot\frac{s-q+1}{2s-q+1}\cdot\ldots\cdot\frac{s-1}{2s-1}.$$

Дроби справа имеют вид $(s-k)/(2s-k)$, $1 \leqslant k \leqslant q$; так как $2(s-k) < 2s-k$, то каждая из этих дробей меньше $1/2$ и рассматриваемый коэффициент меньше, чем 2^{-q}.

Теперь оценим произведение $s\mu(s, \bar{s}-1)$. Пусть $\bar{s}$ — нечетное, тогда

$$s\mu(s, \bar{s}-1) = s(s-1)^2 \ldots (s-(\bar{s}-3)/2)^2(s-(\bar{s}-1)/2).$$

Так как $\bar{s} \leqslant s$, то $s = s-(\bar{s}-1)/2+(\bar{s}-1)/2 < s-(\bar{s}-1)/2+s/2$. Отсюда $s < 2(s-(\bar{s}-1)/2)$ и $s\mu(s, \bar{s}-1) < 2\mu(s, \bar{s})$. То же получается и при $\bar{s}$ четном. В этом случае $s\mu(s, \bar{s}-1) = s(s-1)^2 \ldots (s-(\bar{s}-2)/2)^2$. Далее, $s = s-\bar{s}/2+\bar{s}/2 < s-\bar{s}/2+s/2$ и $s \leqslant 2(s-\bar{s}/2)$. Отсюда $s\mu(s, s-1) < 2\mu(s, \bar{s})$. Полученные результаты подставим в (10.3.9). Заметим при этом, что $(s-1)\mu(s, \bar{s}-1) < s\mu(s, \bar{s}-1) <$

$< 2\mu(s, \bar{s})$ и потому

$$A_{qs}^{(\bar{s})} \leqslant (C_q/q!)\, s\mu(s, \bar{s})\, 2^{2\bar{s}-2}, \tag{10.3.10}$$

где

$$C_q = \frac{3}{2} C_{q-1} + \frac{1}{2^{q-1}}. \tag{10.3.11}$$

Решение разностного уравнения (10.3.11) имеет вид $C_q = C(3/2)^q - 1/2^q$. При $q = 0$ имеем $C_0 = 1$, отсюда $C = 2$ и $C_q = 2(3/2)^q - 1/2^q$. Окончательно

$$A_{qs}^{(\bar{s})} \leqslant \frac{1}{q!}\left[2\left(\frac{3}{2}\right)^q - \frac{1}{2^q}\right] s\mu(s, \bar{s})\, 2^{2\bar{s}-2}. \tag{10.3.12}$$

5°. Оценка (10.2.10) отличается от оценки (10.2.3) лишь наличием множителя $\left[\frac{\pi}{4p(s-\bar{s})}\right]^{1/2} e^{1/6p^2(s-\bar{s})}$, а рекуррентные оценки (10.2.11) и (10.2.7) тождественны. Отсюда нетрудно заключить, что для произвольного p, $1 \leqslant p \leqslant \infty$, справедливо неравенство

$$A_{qsp}^{(\bar{s})} \leqslant \frac{1}{q!}\left[2\left(\frac{3}{2}\right)^q - \frac{1}{2^q}\right]\left[\frac{\pi}{4p(s-\bar{s})}\right]^{1/2p} e^{1/6p^2(s-\bar{s})} \mu(s, \bar{s})\, 2^{2s-2}. \tag{10.3.13}$$

Заметим еще, что если оценивать величину $\binom{2s-1}{s-1}$ не как коэффициент биноминального разложения, а по формуле Стирлинга, то можно получить для $A_{qs}^{(s)}$ оценку, несколько более точную по порядку, чем оценка (10.3.12):

$$A_{qs}^{(\bar{s})} \leqslant \frac{2e^{7/12}}{\sqrt{\pi}\, q!}\left[2\left(\frac{3}{2}\right)^q - \frac{1}{2^q}\right] \frac{s}{\sqrt{s-1}} \mu(s, \bar{s})\, 2^{2\bar{s}-2} <$$
$$< \frac{2{,}03}{q!}\left[2\left(\frac{3}{2}\right)^q - \frac{1}{2^q}\right] \frac{s}{\sqrt{s-1}} \mu(s, \bar{s})\, 2^{2\bar{s}-2}. \tag{10.3.14}$$

6°. Приведем некоторые численные значения оценок величин $A_{qs}^{(\bar{s})}$ (табл. 16). Эти оценки получены из формул (10.2.2)—(10.2.5) и (10.3.12). При этом использована табл. 15 значений функции $\mu(s, \bar{s})$ (см. конец § 2 настоящей главы). Оценки вычислены с избытком.

Таблица 16

$s = 2$

q \ $\bar{s}$	0	1	2
0	1	3/2	6
1	1	3/2	4

$s = 3$

q \ $\bar{s}$	0	1	2	3
0	1	15/8	15	120
1	1	7/4	9	480
2	1/2	41/32	5/2	408

$s = 4$

q \ $\bar{s}$	0	1	2	3	4
0	1	35/16	105/4	315	2 520
1	1	31/16	15	440	11 520
2	1/2	31/32	33/8	1224	10 368
3	1/6	49/36	5/4	636	5 088

$s = 5$

q \ $\bar{s}$	0	1	2	3	4	5
0	1	2,47	39,4	630	$756 \cdot 10$	$908 \cdot 10^2$
1	1	2,10	21,9	$320 \cdot 10$	$384 \cdot 10^2$	$461 \cdot 10^3$
2	1/2	1,21	5,93	$272 \cdot 10$	$327 \cdot 10^2$	$391 \cdot 10^3$
3	1/6	0,52	1,55	$142 \cdot 10$	$170 \cdot 10^2$	$263 \cdot 10^3$
4	1/24	0,17	0,52	537	$644 \cdot 10$	$733 \cdot 10^2$

Как видно из табл. 16, оценки величин $A_{qs}^{(\bar{s})}$ заметно возрастают при переходе от $\bar{s} = 2$ к $\bar{s} = 3$ и выше. Возможно, это означает, что оценка (10.3.12) завышена.

§ 4. АППРОКСИМАЦИЯ ФУНКЦИЙ В РАВНОМЕРНЫХ МЕТРИКАХ

1°. В данном параграфе будет получена оценка постоянной $C = C(2s, \bar{s})$, $\bar{s} \leqslant s$, в неравенстве

$$\|D^{\bar{s}}x - D^{\bar{s}}x^h\| \leqslant C(2s, \bar{s})\, h^{2s-\bar{s}} \|x^{(2s)}\|; \qquad (10.4.1)$$

знаком $\|\cdot\|$ в этом параграфе обозначается норма в $L_\infty(0, 1)$.

Пусть $t = t_0 + \tau h$, $t_0 = j_0 h$, j_0 — целое число, $0 \leqslant j_0 \leqslant 2n - 1$, $0 \leqslant \tau \leqslant 1$. Составим разность

$$D^{\bar{s}}x(t_0 + \tau h) - D^{\bar{s}}x^h(t_0 + \tau h) = D^s x(t_0 + \tau h) - \\ - \sum_{q=0}^{s-1} \sum_{j=0}^{2n} h^{q-\bar{s}} x^{(q)}(jh)\, \omega_{qs}^{(\bar{s})}(j_0 - j + \tau).$$

Функция $\omega_{qs}(j_0 - j + \tau)$ отлична от тождественного нуля только при $j = j_0$ и $j = j_0 - 1$. Это приводит к несколько более простому тождеству:

$$D^{\bar{s}}x(t_0 + \tau h) - D^{\bar{s}}x^h(t_0 + \tau h) = D^{\bar{s}}x(t_0 + \tau h) - \\ - \sum_{q=0}^{s-1} h^{q-\bar{s}} \left[x^{(q)}(t_0)\, \omega_{qs}^{(\bar{s})}(\tau) + x^{(q)}(t_0 + h)\, \omega_{qs}^{(\bar{s})}(\tau - 1)\right]. \qquad (10.4.2)$$

Далее, по формуле Тейлора

$$D^{\bar{s}}x(t_0+\tau h)=\sum_{r=0}^{2s-\bar{s}-1}\frac{x^{(\bar{s}+r)}(t_0)}{r!}h^r+\frac{(\tau h)^{2s-\bar{s}}}{(2s-\bar{s})!}\times$$
$$\times\int_0^1 x^{(2s}(t_0+\tau hz)(1-z)^{2s-\bar{s}-1}dz;$$

$$x^q(t_0+h)=\sum_{r=0}^{2s-q-1}\frac{x^{(q+r)}(t_0)}{r!}h^r+\frac{h^{2s-\bar{s}}}{(2s-q-1)!}\times$$
$$\times\int_0^1 x^{(2s)}(t_0+hz)(1-z)^{2s-q}dz\cdot\omega_{qs}^{(\bar{s})}(\tau-1).$$

Подставим это в (10.4.2). Вследствие фундаментальных соотношений главные члены формул Тейлора исчезнут, что даст новое соотношение

$$h^{-(2s-\bar{s})}[D^{\bar{s}}x(t)-D^{\bar{s}}x^h(t)]=\frac{\tau^{2s-\bar{s}}}{(2s-\bar{s}-1)!}\times$$
$$\times\int_0^1 x^{(2s)}(t_0+\tau hz)(1-z)^{2s-\bar{s}-1}dz-$$
$$-\sum_{q=0}^{s-1}\frac{\omega_{qs}^{(\bar{s})}(\tau-1)}{(2s-q-1)!}\int_0^1 x^{(2s)}(t_0+hz)(1-z)^{2s-q-1}dz. \qquad (10.4.3)$$

Первый член справа мажорируется величиной

$$\frac{\|x^{(2s)}\|}{(2s-\bar{s}-1)!}\int_0^1(1-z)^{2s-\bar{s}-1}dz=\frac{\|x^{(2s)}\|}{(2s-\bar{s})!},$$

а второй член — величиной

$$\|x^{(2s)}\|\sum_{q=0}^{s-1}\frac{A_{qs}^{(\bar{s})}}{(2s-q)!}\leqslant\frac{\|x^{(2s)}\|}{(s+1)!}\sum_{q=0}^{s-1}A_{qs}^{(\bar{s})}.$$

Отсюда

$$h^{-(2s-\bar{s})}\|D^{\bar{s}}x-D^{\bar{s}}x^h\|_{L_\infty}\leqslant\frac{1}{(2s-\bar{s})!}+\frac{1}{(s+1)!}\sum_{q=0}^{s-1}A_{qs}^{(\bar{s})}$$

и, следовательно,

$$C(2s,\ \bar{s})\leqslant\frac{1}{(2s-\bar{s})!}+\frac{1}{(s+1)!}\sum_{q=0}^{s-1}A_{qs}^{(\bar{s})}. \qquad (10.4.4)$$

Заметим, что первый член в оценке (10.4.4) существенно меньше второго.

Сумму в (10.4.4) можно оценить, исходя из формулы (10.3.12):

$$\sum_{q=0}^{s-1}A_{qs}^{(\bar{s})}<s\mu(s,\ \bar{s})\,2^{2\bar{s}-2}\sum_{q=0}^{\infty}\left[\frac{1}{q!}\left(2\left(\frac{3}{2}\right)^q-\frac{1}{2^q}\right)\right]=$$
$$=(2e^{3/2}-e^{1/2})\,s\mu(s,\ \bar{s})\,2^{2\bar{s}-2}<7{,}32s\mu(s,\ \bar{s})\,2^{2\bar{s}-2},$$

и для величины $C(2s,\ \bar{s})$ получается сравнительно простая оценка

$$C(2s,\ \bar{s})<\frac{1}{(2s-\bar{s})!}+\frac{7{,}32}{(s+1)!}s\mu(s,\ \bar{s})\,2^{2s-2}. \qquad (10.4.5)$$

2°. В табл. 17 приведены оценки величины $C(2s, \bar{s})$. Эти оценки вычислены по формуле (10.4.5), в которой отброшено первое слагаемое.

Таблица 17

q \ $\bar{s}$	1	2	3	4	5
1	3,66				
2	2,45	9,77			
3	0,92	7,92	23,3		
4	0,245	2,93	35,2	141	
5	0,0407	0,813	11,4	157	1880

§ 5. АППРОКСИМАЦИЯ ФУНКЦИЙ В ИНТЕГРАЛЬНЫХ МЕТРИКАХ

1°. Начнем с очевидной формулы

$$\| D^{\bar{s}}x - D^{\bar{s}}x^h \|_{L_p}^p = \sum_{j=0}^{2n-1} \int_{t_j}^{t_{j+1}} | D^{\bar{s}}x(t) - D^{\bar{s}}x^h(t) |^p \, dt,$$

$$t_j = jh, \quad h = 1/(2n). \tag{10.5.1}$$

В интеграле справа положим $t = t_j + \tau h$:

$$\int_{t_j}^{t_{j+1}} | D^{\bar{s}}x(t) - D^{\bar{s}}x^h(t) |^p \, dt = h \int_0^1 | D^{\bar{s}}x(t_j + \tau h) -$$

$$- D^{\bar{s}}x^h(t_j + \tau h) |^p \, d\tau. \tag{10.5.2}$$

В формуле (10.4.3) заменим t_0 на t_j, возведем модули обеих частей этой формулы в степень p и проинтегрируем по τ в пределах от нуля до единицы:

$$h^{-p(2s-\bar{s})} \int_{t_j}^{t_{j+1}} | D^{\bar{s}}x(t) - D^{\bar{s}}x^h(t) |^p \, dt =$$

$$= \int_0^1 \left| \frac{\tau^{2s-\bar{s}} \, d\tau}{(2s-\bar{s}-1)!} \int_0^1 x^{(2s)}(t_j + \tau h z)(1-z)^{2s-\bar{s}-1} \, dz - \right.$$

$$\left. - \sum_{q=0}^{s-1} \frac{\omega_{qs}^{(\bar{s})}(\tau - 1)}{(2s-q-1)!} \int_0^1 x^{(2s)}(t_j + hz)(1-z)^{2s-q-1} \, dz \right|^p d\tau. \tag{10.5.3}$$

В силу неравенства Гельдера для сумм правая часть формулы (10.5.3) не превосходит величины

$$2^{p-1} \int_0^1 d\tau \int_0^1 \left\{ \frac{\tau^{(2s-\bar{s})p}}{[(2s-\bar{s}-1)!]^p} \left| x^{(2s)}(t_j + \tau h z) \right|^p (1-z)^{(2s-\bar{s}-1)p} + \right.$$

$$\left. + | x^{(2s)}(t_j + hz) |^p \left[\sum_{q=0}^{s-1} \frac{(1-z)^{2s-q-1}}{(2s-q-1)!} \, | \omega_{qs}^{(\bar{s})}(\tau - 1) | \right]^p \right\} dz. \tag{10.5.4}$$

Двойной интеграл в (10.5.4) естественно распадается на два.

В первом заменим $1-z$ на единицу и затем положим $t_j + \tau hz = y$. В результате получится следующая оценка упомянутого интеграла:

$$\frac{1}{h\,[(2s-\bar{s}-1)!]^p}\int_0^1 \tau^{(2s-\bar{s})p}\,d\tau \int_{t_j}^{t_{j+1}} |\,x^{(2s)}(y)\,|^p\,dy = \\ = \frac{1}{h\,[(2s-\bar{s}-1)!]^p\,(2s-\bar{s})\,p}\int_{t_j}^{t_{j+1}} |\,x^{(2s)}(y)\,|^p\,dy. \quad (10.5.5)$$

Второй интеграл в (10.5.4) меньше, чем

$$\int_0^1 d\tau \int_0^1 |\,x^{(2s)}(t_j+hz)\,|^p \left[\sum_{q=0}^{s-1}\frac{|\,\omega_{qs}^{(\bar{s})}(\tau-1)\,|}{(2s-q-1)!}\right]^p dz = \\ = \frac{1}{h}\int_0^1 \left[\sum_{q=0}^{s-1}\frac{|\,\omega_{qs}^{(\bar{s})}(t-1)\,|}{(2s-q-1)!}\right]^p dt \int_{t_j}^{t_{j+1}} |\,x^{(2s)}(t)\,|^p\,dt.$$

По неравенству Гельдера

$$\left[\sum_{q=0}^{s-1}\frac{|\,\omega_{qs}^{(\bar{s})}(t-1)\,|}{(2s-q-1)!}\right]^p \leqslant \\ \leqslant \sum_{q=0}^{s-1} |\,\omega_{qs}^{(\bar{s})}(t-1)\,|^p \left\{\sum_{q=0}^{s-1}\frac{1}{[(2s-q-1)!]^{p'}}\right\}^{p/p'},$$

и второй интеграл в (10.5.4) получает оценку

$$\left\{\sum_{q=0}^{s-1}\frac{1}{[(2s-q-1)!]^{p'}}\right\}^{p/p'} \sum_{q=0}^{s-1} |\,A_{qsp}^{(\bar{s})}\,|^p \int_{t_j}^{t_{j+1}} |\,x^{(2s)}(t)\,|^p\,dt. \quad (10.5.6)$$

Из (10.5.5), (10.5.6) находим

$$h^{-(2s-\bar{s})p}\int_{x_j}^{x_{j+1}} |\,D^{\bar{s}}x(t) - D^{\bar{s}}x^h(t)\,|^p\,dt \leqslant \\ \leqslant 2^{p-1}\int_{t_j}^{t_{j+1}} |\,x^{(2s)}(t)\,|^p\,dt \left\{\frac{1}{[(2s-\bar{s}-1)!]^p\,[(2s-\bar{s})\,p+1]} + \right. \\ \left. + \left[\sum_{q=0}^{s-1}\frac{1}{[(2s-q-1)!]^{p'}}\right]^{p/p'} \sum_{q=0}^{s-1} |\,A_{qsp}^{(\bar{s})}\,|^p\right\}.$$

Просуммируем это по j и извлечем корень степени p, что приведет нас к окончательной оценке

$$h^{-(2s-\bar{s})}\,\|\,D^{\bar{s}}x - D^{\bar{s}}x^h\,\|_p \leqslant 2^{1/p'}\,\|\,x^{(2s)}\,\|_p \times \\ \times \left\{\frac{1}{[(2s-\bar{s}-1)!]^p\,[(2s-\bar{s})\,p+1]} + \left[\sum_{q=0}^{s-1}\frac{1}{[(2s-q-1)!]^{p'}}\right]^{p-1} \times \right. \\ \left. \times \sum_{q=0}^{s-1} |\,A_{qsp}^{(\bar{s})}\,|^p\right\}^{1/p'}. \quad (10.5.7)$$

2° Оценку (10.5.7) можно упростить, сделав ее, однако, несколько более грубой. Так как $\bar{s} \leqslant s$, то $2s-\bar{s}-1 \geqslant s-1$ и $2s-\bar{s} \geqslant s$, поэтому

$$\frac{1}{[(2s-\bar{s}-1)!]^p\,[(2s-\bar{s})\,p+1]} \leqslant \frac{1}{[(s-1)!]^p\,sp}$$

Далее, $q \leqslant s-1$. Отсюда $2s-q-1 \geqslant s$ и

$$\sum_{q=0}^{s-1} \frac{1}{[(2s-q-1)!]^{p'}} < \frac{s}{(s!)^{p'}}.$$

Выражение в фигурных скобках в формуле (10.5.7) мажорируется величиной

$$\frac{1}{[(s-1)!]^p\, s}\left[\frac{1}{p} + \sum_{q=0}^{s-1} (A_{qsp}^{(\bar{s})})^p\right].$$

Теперь

$$h^{-(2s-\bar{s})} \| D^s x - D^{\bar{s}} x^h \|_p \leqslant$$

$$\leqslant \frac{2^{1/p'} \| x^{(2s)} \|_p}{(s-1)!\, s^{1/p}} \left\{\frac{1}{p} + \sum_{q=0}^{s-1} (A_{qsp}^{(\bar{s})})^p\right\}^{1/p} \leqslant$$

$$\leqslant \frac{2^{1/p'} \| x^{(2s)} \|_p}{(s-1)!\, s^{1/p}} \left[p^{-1/p} + \sum_{q=0}^{s-1} A_{qsp}^{(\bar{s})}\right].$$

Последнюю сумму можно оценить так же, как в § 4, или, точнее по порядку, по формуле для $A_{qsp}^{(s)}$, аналогичной формуле (10.3.14):

$$\sum_{q=0}^{s-1} A_{qsp}^{(s)} < \frac{2e^{7/12}}{\sqrt{\pi}} (2e^{3/2} - e^{1/2}) \frac{s}{\sqrt{s-1}}\, \mu(s, \bar{s}) \times$$

$$\times 2^{2\bar{s}-2} \left[\frac{\pi}{4p\,(s-\bar{s})}\right]^{1/2p}.$$

Окончательно

$$h^{-(2s-\bar{s})} \| D^{\bar{s}} x - D^{\bar{s}} x^h \|_p \leqslant \| x^{(2s)} \|_p \frac{2e^{7/12}}{\sqrt{\pi}} (2e^{3/2} - e^{1/2}) \times$$

$$\times \frac{s}{\sqrt{s-1}} \frac{2^{1/p}}{(s-1)!\, s^{1/p}} \mu(s, \bar{s}) \left[\frac{\pi}{4p\,(s-\bar{s})}\right]^{1/2p} =$$

$$= \| x^{(2s)} \|_p \frac{4{,}80s}{\sqrt{s-1}} \frac{2^{1/p'}}{(s-1)!\, s^{1/p}} \mu(s, \bar{s}) \left[\frac{\pi}{4p\,(s-\bar{s})}\right]^{1/2p}. \quad (10.5.8)$$

Отметим особо важный для приложений случай $p=2$:

$$h^{-(2s-\bar{s})} \| D^{\bar{s}} x - D^{\bar{s}} x^h \|_2 \leqslant \| x^{(2s)} \|_2 \cdot 5{,}25 \sqrt{\frac{s}{s-1}} \frac{\mu(s, \bar{s})}{(s-1)!\,(s-\bar{s})}. \quad (10.5.9)$$

§ 6. АППРОКСИМАЦИЯ ФУНКЦИЙ МНОГИХ ПЕРЕМЕННЫХ

1°. Будем пользоваться кусочно-полиномиальными многомерными исходными функциями, введенными в § 6 главы 9 и полученными перемножением одномерных кусочно-полиномиальных функций, также введенных в главе 9 и изученных в

предшествующих параграфах данной главы. Введем обозначение

$$\tilde{A}^{(\alpha)}_{qsp} = \| \tilde{\omega}^{(\alpha)}_{qs} \|_{L_p(Q)}, \tag{10.6.1}$$

где $\tilde{\omega}_{qs}$ — упомянутые выше многомерные исходные функции, Q — куб, $\underline{0} \leqslant t \leqslant \underline{1}$. Положим еще

$$\tilde{A}^{(\bar{s})}_{qsp} = \max_{|\alpha|=\bar{s}} \tilde{A}^{(\alpha)}_{qsp}. \tag{10.6.2}$$

Если $p < \infty$, то (определение $A^{(\alpha_k)}_{q_k sp}$ см. § 2, п. 1°)

$$[\tilde{A}^{(\alpha)}_{qsp}]^p = \int_Q |\tilde{\omega}^{(\alpha)}_{qs}(t)|^p \, dt = \prod_{k=1}^{m} \int_0^1 \left| \omega^{(\alpha_k)}_{q_k s}(t_k) \right|^p dt_k = \prod_{k=1}^{m} \left(A^{(\alpha_k)}_{q_k sp} \right)^p.$$

Отсюда

$$\tilde{A}^{(\alpha)}_{qsp} = \prod_{k=1}^{m} A^{(\alpha_k)}_{q_k sp}. \tag{10.6.3}$$

Полагая $p \to \infty$, находим, что формула (10.6.3) верна и для $p = \infty$. Вместо $\tilde{A}^{(\alpha)}_{qs\infty}$ будем писать $\tilde{A}^{(\alpha)}_{qs}$, так что

$$\tilde{A}^{(\alpha)}_{qs} = \prod_{k=1}^{m} A^{(\alpha_k)}_{q_k s}. \tag{10.6.4}$$

2°. Пусть Ω — конечная область пространства R_m и пусть эта область такова, что функции пространства $C^{(2s)}(\Omega)$ допускают продолжение на все R_m с сохранением класса, причем носители всех продолженных функций лежат в одном и том же кубе Q. Выбрав подходящим образом оси координат и единицу длины, можно считать, что $Q = \{t : \underline{0} \leqslant t \leqslant \underline{1}\}$.

Выберем натуральное число n и построим кубическую сетку с шагом $h = 1/(2n)$; число n выберем столь большим, чтобы расстояние между границами $\partial\Omega$ и ∂Q было больше h.

Пусть функция $x \in C^{(2s)}(\Omega)$. Продолжим ее так, как указано выше; продолженную функцию обозначим тем же символом x. Построим аппроксимирующую функцию в виде

$$x^h(t) = \sum_{|q|=0}^{s-1} \sum_{j=\underline{0}}^{\underline{2n}} h^{|q|} x^{(q)}(jh) \, \omega_{qs}(t/h - j). \tag{10.6.5}$$

Примем, что $t \in \operatorname{supp} x$. Положим $t = t_0 + \tau h$, $t_0 = j_0 h$, где j_0 — целочисленный вектор и $0 < \tau < 1$. Легко показать, что в сумме (10.6.5) могут быть отличны от тождественного нуля только те слагаемые, для которых $j_0 = j_0 + i$, $i \in I$, где I — множество вершин куба Q; формулу (10.6.5) можно привести к виду

$$x^{(h)}(t) = \sum_{|q|=0}^{s-1} \sum_{i \in I} h^{|q|} x^{(q)}(t_0 + ih) \, \tilde{\omega}_{qs}(t/h - i).$$

Пусть α — мультииндекс и $|\alpha| = \bar{s} \leqslant s$. Рассмотрим разность

$$D^\alpha x(t) - D^\alpha x^h(t) = x^{(\alpha)}(t_0 + \tau h) - \\ - \sum_{q=0}^{s-1} \sum_{i \in I} h^{|q|} x^{(q)}(t_0 + ih)\, \tilde{\omega}_{qs}(t - i), \qquad (10.6.6)$$

Выражения $x^{(\alpha)}(t_0 + \tau h)$ и $x^{(q)}(t_0 + ih)$ разложим по формуле Тейлора вокруг точки t_0 до членов порядка $2s - 1$ и подставим эти разложения в (10.6.6). Главные члены формул Тейлора исчезнут потому, что исходные функции удовлетворяют усиленным фундаментальным соотношениям, и мы получим

$$D^\alpha x(t) - D^\alpha x^h(t) = (2s - \bar{s} + 1)\, h^{2s-\bar{s}} \int_0^1 \sum_{|\beta|=2s-\bar{s}} (\tau^\beta/\beta!) \times \\ \times x^{(\alpha+\beta)}(t_0 + \tau h z)(1 - z)^{2s-\bar{s}}\, dz - \\ - \sum_{|q|=0}^{s-1} (2s - |q| + 1)\, h^{2s-\bar{s}} \int_0^1 \sum_{|\beta|=2s-\bar{q}} \sum_{i \in I} (i^\beta/\beta!) \times \\ \times (1 - z)^{2s-|q|} x^{(\beta+q)}(t_0 + ihz)\, \tilde{\omega}_{qs}^{(\alpha)}(t - i)\, dz. \qquad (10.6.7)$$

Введем обозначение

$$\| x^{(k)} \|_C = \max_{|\alpha|=k} \| x^{(\alpha)} \|_C; \qquad (10.6.8)$$

аналогичными обозначениями будем пользоваться и для других метрик. Оценим правую часть формулы (10.6.7). Уменьшаемое мажорируется величиной

$$h^{2s-\bar{s}} \| x^{(2s)} \|_C \sum_{|\beta|=2s-\bar{s}} (1/\beta!). \qquad (10.6.9)$$

В равенстве

$$(x_1 + x_2 + \ldots + x_m)^k = \sum_{|\beta|=k} (k!/\beta!)\, x_1^{\beta_1} x_2^{\beta_2} \ldots x_m^{\beta_m}$$

положим $k = 2s - \bar{s}$, $x_1 = x_2 = \ldots = x_m = 1$. В результате получим

$$\sum_{|\beta|=2s-\bar{s}} 1/\beta! = m^{2s-\bar{s}}/(2s - \bar{s})!$$

Теперь из (10.6.9) следует оценка для уменьшаемого в (10.6.7): оно не превосходит величины

$$h^{2s-\bar{s}} \| x^{(2s)} \|_C\, m^{2s-\bar{s}}/(2s - \bar{s})!. \qquad (10.6.10)$$

Вычитаемое в (10.6.7) мажорируется величиной

$$h^{2s-\bar{s}} \| x^{(2s)} \|_C \sum_{|q|=0}^{2s-1} \tilde{A}_{qs}^{(\alpha)} \sum_{|\beta|=2s-|q|} (1/\beta!) = \\ = h^{2s-\bar{s}} \| x^{(2s)} \|_C \sum_{|q|=0}^{2s-1} \frac{m^{2s-|q|}}{(2s - |q|)!}\, \tilde{A}_{qs}^{(\alpha)}. \qquad (10.6.11)$$

Из (10.6.10), (10.6.11) вытекает искомая оценка

$$|D^{\alpha}x(t)-D^{\alpha}x^{h}(t)|\leqslant C(s,\ \bar{s},\ m)\,h^{2s-\bar{s}}\,\|x^{(2s)}\|_{C},\qquad (10.6.12)$$

где

$$C(s,\ \bar{s},\ m)=\frac{m^{2s-\bar{s}}}{(2s-\bar{s})!}+\sum_{|q|=0}^{2s-1}\frac{m^{2s-|q|}}{(2s-|q|)!}\,\tilde{A}_{qs}^{(\bar{s})}. \qquad (10.6.13)$$

3°. Рассмотрим функцию $x\in W_{\vec{p}}^{(s)}(\Omega)$. Продолжим ее с сохранением класса на R_m и обозначим продолженную функцию той же буквой x; будем считать, что такое продолжение возможно и что «постоянная продолжения» $\varkappa$, характеризуемая неравенством

$$\|x\|_{W_p^{(s)}(R_m)}\leqslant\varkappa\|x\|_{W_p^{(s)}(\Omega)},$$

известна или по крайней мере известна ее оценка сверху. Как и в пп. 1, 2°, будем считать, что носитель продолженной функции заключен внутри куба $Q=\{t:0\leqslant t\leqslant 1\}$. Построим кубическую сетку, такую же, как в п. 2°. В данном случае, когда $x\in W_p^{(s)}(R_m)$, вариационно-сеточная аппроксимация строится по формуле (9.6.5). Если Q_j — куб $\{t: jh\leqslant t\leqslant(j+1)h\}$, то справедливо неравенство, вытекающее из (10.6.12):

$$\begin{gathered}t\in Q_j,\quad |\alpha|=\bar{s}\leqslant s,\\ |D^{\alpha}\bar{x}_h(t)-D^{\alpha}x^{h}(t)|\leqslant h^{2s-\bar{s}}C(s,\ \bar{s},\ m)\,\|\bar{x}_h^{(2s)}\|_{C(Q_j)}.\end{gathered}\qquad (10.6.14)$$

Пусть $\omega_h(y)$ — усредняющее ядро Ильина — Головкина, равное нулю при $|y|\leqslant h$, тогда

$$\bar{x}_h(t)=\int_{R_m}\omega_h(t-t')\,x(t')\,dt'=\int_{|t-t'|<h}\omega_h(t-t')\,x(t')\,dt'.$$

Дифференцируя эту формулу, а затем интегрируя по частям и принимая во внимание, что при $|t-t'|\geqslant h$ само усредняющее ядро и все его производные равны нулю, получим

$$|\beta|=2s,\quad \bar{x}_h^{(\beta)}(t)=\int_{|t-t'|<h}x^{(\beta)}(t')\,\omega_h(t-t')\,dt'.$$

По неравенству Гельдера

$$\begin{aligned}|\bar{x}_h^{(\beta)}(t)|\leqslant&\Big[\int_{|t-t'|<h}|x^{(\beta)}(t')|^{p}\,\omega_h(t-t')\,|dt'\Big]^{1/p}\times\\ &\times\Big[\int_{|t-t'|\leqslant h}|\omega_h(t-t')\,|dt'\Big]^{1/p'}.\end{aligned}\qquad (10.6.15)$$

Для дальнейших оценок примем за $\omega_h(t-t')$ то конкретное усредняющее ядро, которое было построено в статье К. К. Головкина [52]. Приведем его описание. Пусть $z\in R_1$ и $\zeta(z)$ — четная неотрицательная функция класса $C_0^{\infty}(R_1)$, носитель ко-

торой заключен в отрезке $[-m^{-1/2}, m^{-1/2}]$. Пусть еще $\int_{-\infty}^{+\infty}\zeta(z)\,dz=1$. Положим

$$\nu_s(z)=\sum_{k-1}^{s}\frac{(-1)^{k-1}}{k}\binom{s}{k}\zeta\left(\frac{z}{k}\right),$$

$$\omega(t)=\prod_{j=1}^{m}\nu_s(t_j);\quad \omega_h(t)=h^{-m}\omega\left(\frac{t}{h}\right).$$

При таком определении

$$\int_{R_m}|\omega(t)|\,dt<\left[\sum_{k=1}^{s}\frac{1}{k}\binom{s}{k}\int_{-\infty}^{+\infty}\zeta\left(\frac{z}{k}\right)dz\right]^m=$$
$$=\left[\sum_{k=1}^{m}\binom{s}{k}\right]^m=(2^s-1)^m$$

и $$\int_{|t-t'|<h}|\omega_h(t-t')|\,dt'=\int_{R_m}|\omega(t)|\,dt<(2^s-1)^m.$$

Подставив это в (10.6.15), получим

$$|\bar{x}_h^{(\beta)}(t)|\leqslant(2^s-1)^{m/p'}\left[\int_{|t-t'|<h}|x^{(\beta)}(t')|^p\times|\omega_h(t-t')|\,dt'\right]^{1/p}. \tag{10.6.16}$$

Пусть j_0 — целочисленный вектор, удовлетворяющий неравенству $0\leqslant j_0\leqslant 2n-1$. Если $t\in Q_{j_0}$, то по неравенствам (10.6.14) и (10.6.16)

$$|D^\alpha\bar{x}_h(t)-D^\alpha x^h(t)|\leqslant 2^{s-\bar{s}}C(s,\bar{s},m)(2^s-1)^{m/p'}\times$$
$$\times\left[\sum_{|\beta|=2s}\int_{|t-t'|<h}|x^{(\beta)}(t')|^p\cdot|\omega_h(t-t')|\,dt'\right]^{1/p}.$$

Возведем это в степень p и проинтегрируем по Q_{j_0}:

$$\int_{Q_{j_0}}|D^\alpha\bar{x}_h(t)-D^\alpha x^h(t)|^p\,dt\leqslant h^{(2s-\bar{s})p}C^p(s,\bar{s},m)\times$$
$$\times(2^s-1)^{m(p-1)}\sum_{|\beta|=2s}\int_{Q_{j_0}}dt\int_{|t-t'|<h}|x^{(\beta)}(t')^p|\,\omega_h(t-t')\,dt'.$$

Справа изменим порядок интегрирования. Тогда t будет пробегать шар $|t-t'|<h$, а t' — некоторую область $Q_{j_0}^*$, заключенную в кубе $(j_0-2)h\leqslant t'\leqslant(j_0+1)h$. Получается неравенство

$$\int_{Q_{j_0}}|D^\alpha\bar{x}_h(t)-D^\alpha x^h(t)|^p\,dt\leqslant h^{(2s-\bar{s})p}C^p(s,\bar{s},m)\times$$
$$\times(2^s-1)^{m(p-1)}\sum_{|\beta|=2s}\int_{Q_{j_0}^*}|x^{(\beta)}(t')|^p\,dt'\int_{|t-t'|<h}|\omega_h(t-t')|\,dt\leqslant$$
$$\leqslant h^{(2s-\bar{s})p}(2^s-1)^p\sum_{|\beta|=2s}\int_{Q_{j_0}^*}|x^{(\beta)}(t)|^p\,dt.$$

Это неравенство просуммируем по целочисленным векторам j_0, удовлетворяющим неравенству $0 \leqslant j_0 \leqslant 2n-1$. Так как кратность наложения областей $Q_{l_0}^*$ не превосходит 4^m, то мы приходим к следующему неравенству:

$$\|D^{\alpha}x_h - D^{\alpha}x^h\|_{L_p(Q)} \leqslant h^{2s-\bar{s}}\,4^{m/p}\,(2^{2s}-1)^m\,C(s,\bar{s},m)\times$$
$$\times \sum_{|\beta|=2s} \|x^{(\beta)}\|_{L_p(Q)}.$$

Отсюда

$$\|D^{\alpha}\overline{x}_h - D^{\alpha}x^h\|_{L_p(Q)} \leqslant h^{2s-\bar{s}}\,4^{m/p}\,(2^{2s}-1)^m\,C(s,\bar{s},m)\times$$
$$\times \varkappa \sum_{|\beta|=2s} \|x^{(\beta)}\|_{L_p(Q)}; \qquad (10.6.17)$$

$\varkappa$ — постоянная продолжения. Можно доказать (подробнее см. [180, глава III, § 4])

$$\|x^{(\alpha)} - \overline{x}_h^{(\alpha)}\|_{p,\Omega} \leqslant h^{2s-\bar{s}}\,\frac{(2^{2s}-1)^m\,m^{2s-\bar{s}}}{(2s-\bar{s}-1)!\,[(2s-\bar{s}-1)\,p'+1]^{1/p'}}\,\|x^{(2s)}\|_{p,\Omega}. \qquad (10.6.18)$$

Из формул (10.6.14)—(10.6.18) вытекает искомая оценка:

$$\|x - x^h\|_{p,\bar{s},\Omega} \leqslant h^{2s-\bar{s}}C_p(s,\bar{s},m,\Omega)\sum_{|\beta|=2s}\|x^{(\beta)}\|_{p,\Omega}, \qquad (10.6.19)$$

где
$$C_p(s,\bar{s},m,\Omega) = \varkappa(2^{2s-\bar{s}}-1)^m \times$$
$$\times\left[4^{m/p}C(s,\bar{s},m) + \frac{m^{2s-\bar{s}}}{(2s-\bar{s}-1)!\,[(2s-\bar{s}-1)^{p'}+1]^{1/p'}}\right]. \qquad (10.6.20)$$

4°. Из оценок констант, полученных в пп. 1—3° данного параграфа, легко найти соответствующие оценки для констант в оценке погрешности аппроксимации МКЭ для многомерных задач. Подробнее см. [97, 180].

Глава 11

ПОГРЕШНОСТИ ПРИБЛИЖЕННЫХ РЕШЕНИЙ ИНТЕГРАЛЬНЫХ УРАВНЕНИЙ

Проблемы, сформулированные в названии данной главы, по-видимому, исследованы пока еще не вполне достаточно. Перечислим те результаты, которые кажутся нам наиболее важными. Л. В. Канторович [64] исследовал погрешность аппроксимации метода механических квадратур и методы замены ядра вырожденным для уравнений Фредгольма. Большое количество результатов, относящихся к приближенному решению уравнений Фредгольма, а также одномерных сингулярных урав-

нений, принадлежит Б. Г. Габдулхаеву (см. [30], где приведена достаточно полная библиография работ этого автора). Некоторые результаты по приближенному решению интегральных уравнений, фредгольмовых и одномерных сингулярных, приведены в [119, глава 6]. В [42] с большой полнотой изложено применение проекционных методов к уравнениям в свертках, которые являются обобщением известных уравнений Винера — Хопфа. Некоторые новые подходы к приближенному решению уравнений Фредгольма указаны в [156]. Приближенному решению одномерных сингулярных интегральных уравнений посвящена книга [192]. Две главы книги [183] посвящены приближенному решению сингулярных интегральных уравнений, одномерных и многомерных. Приближенному решению многомерных сингулярных уравнений посвящены работы [98] и [118]. Отметим еще сборники статей [162] и [155]. Особо отметим весьма интересную статью [173], в которой получено необходимое и достаточное условие сходимости одного варианта метода коллокации, связанного с МКЭ, и дана оценка соответствующей погрешности аппроксимации. В ряде работ автора данной книги развит метод приближенного вычисления резольвенты Фредгольма, основанный на конечноэлементной аппроксимации ядра. Результаты этих работ приведены вместе с необходимой библиографией в [184]; с некоторыми усилениями и дополнениями эти результаты даны ниже, в главе 11.

Для упрощения записей, а отчасти и выкладок, мы будем ниже, говоря об уравнении Фредгольма, иметь в виду только скалярные уравнения в R_1 и будем считать, что промежуток интегрирования есть отрезок $[0, 1]$. Распространение результатов на более общие уравнения Фредгольма очевидно.

§ 1. МЕТОД МЕХАНИЧЕСКИХ КВАДРАТУР ДЛЯ УРАВНЕНИЯ ФРЕДГОЛЬМА

1°. Рассмотрим уравнение

$$x(t) + \int_0^1 K(t, \tau)\, x(\tau)\, d\tau = f(t). \qquad (11.1.1)$$

Допустим, что $f \in C^{(s)}[0, 1]$, $K \in C^{(s, s)}([0, 1] \times [0, 1])$ и что уравнение (11.1.1) разрешимо единственным образом. Условие на ядро надо понимать так: в квадрате $[0, 1] \times [0, 1]$ существуют и непрерывны все производные вида

$$\frac{\partial^{k+l} K}{\partial t^k\, \partial \tau^l}, \quad 0 \leqslant k,\ l \leqslant s.$$

Из наших условий следует, что решение уравнения (11.1.1), которое мы обозначим через x_*, принадлежит к классу $C^{(s)}[0, 1]$.

Будем приближенно решать уравнение (11.1.1) по методу механических квадратур, заменяя интеграл по квадратурной формуле вида

$$\int_0^1 \varphi(\tau)\,d\tau = \sum_{k=1}^{n} c_k^{(n)} \varphi(\tau_k^{(n)}) + \rho_n. \qquad (11.1.2)$$

Примем, что эта формула обладает следующими свойствами:

$$c_k^{(n)} \geqslant 0, \quad \sum_{k=1}^{n} c_k^{(n)} = 1; \qquad (11.1.3)$$

если $\varphi \in C^{(s)}[0, 1]$, то

$$\rho_n \leqslant C_s h^s \| \varphi^{(s)} \|_C, \quad h = \max_k (\tau_{k+1}^{(n)} - \tau_k^{(n)}). \qquad (11.1.4)$$

Заменяя в уравнении (11.1.1) интеграл по формуле (11.1.2) и отбрасывая остаточный член, получаем новое уравнение

$$y^{(n)}(t) + \sum_{k=1}^{n} c_k^{(n)} K(t, \tau_k)\, y^{(n)}(\tau_k) = f(t), \quad \tau_k = \tau_k^{(n)}, \qquad (11.1.5)$$

решение $y^{(n)}(t)$ которого будем рассматривать как приближенное решение уравнения (11.1.1). Чтобы определить $y^{(n)}(t)$, положим в (11.1.5) $t = \tau_j$, $1 \leqslant j \leqslant n$. Это приведет нас к алгебраической системе

$$y_j^{(n)} + \sum_{k=1}^{n} c_k^{(n)} K(\tau_j, \tau_k)\, y_k^{(n)} = f_j^{(n)}, \quad 1 \leqslant j \leqslant n, \qquad (11.1.6)$$

в которой $y_j^{(n)} = y^{(n)}(\tau_j)$, $f_j^{(n)} = f(\tau_j)$. Если система (11.1.6) разрешима единственным образом, то тем же свойством обладает и уравнение (11.1.5). Очевидно и обратное; очевидно также, что любое решение уравнения (11.1.5) принадлежит к классу $C^{(s)}[0, 1]$.

Операторы

$$\int_0^1 K(t, \tau)\, x(\tau)\, d\tau, \qquad \sum_{k=1}^{n} c_k^{(n)} K(t, \tau_k)\, x(\tau_k) \qquad (11.1.7)$$

будем рассматривать как операторы в $C^{(s)}[0, 1]$; оценим норму их разности. Эта норма равна

$$\sup_{\|x\|_{C^{(s)}}=1} \sum_{j=0}^{s} \left\| \int_0^1 \frac{\partial^j K(t, \tau)}{\partial \tau^j} x(\tau)\, d\tau - \sum_{k=1}^{n} c_k^{(n)} \frac{\partial^j K(t, \tau_k)}{\partial t^j} x(\tau_k) \right\|_C;$$

по оценке (11.1.4) выражение под знаком sup не превосходит величины

$$C_s h^s \sum_{j=0}^{s} \left\| \frac{\partial^s}{\partial \tau^s} \left[\frac{\partial^j K(t, \tau)}{\partial t^j} x(\tau) \right] \right\|_C \leqslant C' h^s \| x \|_{C^{(s)}}, \quad C' = \text{const}. \qquad (*)$$

Таким образом, при достаточно малом h разность операторов (11.1.7) сколь угодно мала по норме $C^{(s)}$.

Интегральные операторы (11.1.7) соответственно обозначим через K и K_n, а их разность — через σ_n. Уравнение (11.1.5)

можно записать в виде

$$y^{(n)}(t) + [(Ky^{(n)}) - (\sigma_n y^{(n)})](t) = f(t).$$

Отсюда

$$y^{(n)}(t) - [(I + K)^{-1}\sigma_n(y)](t) = (I + K)^{-1} f(t) = x_*(t)$$

и, следовательно,

$$\|x_* - y_*^{(n)}\|_{C(s)} \leqslant \|(I + K)^{-1}\|_{C(s)} \cdot \|\sigma_n\|_{C(s)} \cdot \|y_*^{(n)}\|_{C(s)}.$$

При достаточно малом h будет $\|(I + K)^{-1}\|_{C(s)} \cdot \|\sigma_n\|_{C(s)} \leqslant 1/2$ и

$$\|y_*^{(n)}\|_{C(s)} - \|x_*\|_{C(s)} \leqslant 2^{-1}\|y_*^{(n)}\|_{C(s)}; \quad \|y_*^{(n)}\|_{C(s)} \leqslant 2\|x_*\|_{C(s)}.$$

Теперь по оценке (*)

$$\|x_* - y_*^{(n)}\|_{C(s)} \leqslant Ch^s \|x_*\|_{C(s)}. \tag{11.1.8}$$

Постоянную C можно уточнить:

$$C = 2C_s' \|(I + K)^{-1}\|_{C(s)}. \tag{11.1.9}$$

Формулы (11.1.6)—(11.1.9) дают оценку аппроксимации метода механических квадратур для уравнения Фредгольма, если его ядро и свободный член суть достаточно гладкие функции.

2°. Рассмотрим случай, когда ядро и свободный член уравнения (11.1.1) только непрерывны. Для произвольной функции определим ее модуль непрерывности $\omega(x, h)$ и допустим, что

$$|x(t') - x(t'')| \leqslant M_x \gamma(h),$$

где $|t' - t''| \leqslant h$ и $\gamma(h) \xrightarrow[h \to 0]{} 0$, причем M_x зависит от x, но не от h. Этим выделяется подкласс непрерывных функций, модули непрерывности которых имеют оценку $O(\gamma(h))$ с фиксированной функцией $\gamma(h)$. Наименьшее возможное значение M_x обозначим через $M(x)$; очевидно, что

$$\omega(x, h) \leqslant M(x)\gamma(h). \tag{11.1.10}$$

Аналогичный подкласс непрерывных функций можно выделить и в случае, когда рассматриваются функции, зависящие от нескольких переменных.

Допустим теперь, что модули непрерывности ядра $K(t, \tau)$ и свободного члена $f(t)$ удовлетворяют неравенству (11.1.10) с одной и той же функцией $\gamma(h)$. Тогда тому же неравенству удовлетворяет и решение $x_*(t)$ уравнения (11.1.1).

Введем пространство S функций, непрерывных на отрезке $[0, 1]$, с модулем непрерывности, удовлетворяющим неравенству (11.1.10) при фиксированной функции $\gamma(h)$; норму в этом пространстве зададим формулой

$$\|\cdot\|_S = \|\cdot\|_C + M(\cdot). \tag{11.1.11}$$

Очевидно, что $f,\ x_* \in S$. Докажем, что операторы (11.1.7) ограничены в S. Обозначим $(Kx)(t) = \varphi(t)$. Имеем $\|\varphi\|_C \leqslant \leqslant \|K\|_C \cdot \|x\|_C$. Далее, $\omega(\varphi, h) \leqslant \omega(K, h)\|x\|_C \leqslant M_t(K)\gamma(h)\|x\|_C$. Отсюда $M(\varphi) \leqslant M_t(K)\|x\|_C$ и

$$\|\varphi\|_S \leqslant [\|K\|_C + M_t(K)]\|x\|_C \leqslant [\|K\|_C + M_t(K)]\ \|x\|_S.$$

Таким образом, $\|K\|_{S\to S} \leqslant \|K\|_C + M_t(K)$. Теперь рассмотрим оператор K_n. Обозначим $\psi_n(t) = (K_n x)(t)$. Тогда

$$|\psi_n(t)| \leqslant \sum_{k=1}^{n} c_k^{(n)} \|K\|_C \cdot \|x\|_C = \|K\|_C \cdot \|x\|_C;$$

$$\omega(\psi_n, h) \leqslant \sum_{k=1}^{n} c_k^{(n)} \omega(K, h)\|x\|_C = \omega(K, h)\|x\|_C \leqslant$$
$$\leqslant M_t(K)\|x\|_C\,\gamma(h).$$

Отсюда $M(\psi_n) \leqslant M_t(K)\|x\|_C$ и $\|K_n\|_{S\to S} \leqslant \|K\|_C + M_t(K)$.

Допустим теперь, что для функций $\varphi(t)$, удовлетворяющих условию (11.1.10), остаточный член квадратурной формулы (11.1.4) имеет оценку

$$\rho_n \leqslant C_0\gamma(h)\|\varphi\|_S. \tag{11.1.12}$$

Дополнительно предположим, что

$$M_t(K) := \frac{1}{\gamma(h)} \sup_{\substack{|t'-t''|\leqslant h \\ 0\leqslant\tau\leqslant 1}} |K(t'', \tau) - K(t'\ \tau)| \leqslant C_1;$$

$$M_{t,\tau}(K) := \frac{1}{\gamma^2(h)} \sup_{\substack{|t'-t''|\leqslant h \\ |\tau'-\tau''|\leqslant h}} |K(t', \tau') - K(t'', \tau') -$$
$$- K(t', \tau'') + K(t'', \tau'')| \leqslant C_1,\ C_1 = \text{const}.$$

Оценим норму $\|\sigma_n\|_{S\to S}$. Имеем $\|\sigma_n x\|_S = \|\sigma_n x\|_C + M(\sigma_n x)$. По оценке (11.1.12)

$$\|\sigma_n x\|_C = \|Kx - K_n x\|_C \leqslant C_0\gamma(h)[\|K\|_C \cdot \|x\|_C +$$
$$+ M(Kx)] \leqslant 2C_0\|K\|_C \cdot \|x\|_C\,\gamma(h) = C_2\|x\|_C\,\gamma(h),\quad C_2 = \text{const}. \tag{11.1.13}$$

Далее, по той же оценке (11.1.12)

$$\omega(\sigma_n x, h) = \sup_{|t''-t'|\leqslant h} \left| \int_0^1 [K(t'', \tau) - K(t', \tau)]\, x(\tau)\, d\tau - \right.$$
$$\left. - \sum_{k=1}^{n} c_k^{(n)} [K(t'', \tau_k) - K(t', \tau_k)]\, x(\tau_k) \right| \leqslant$$
$$\leqslant C_0\gamma(h)[\max_{\tau} |K(t'', \tau) - K(t', \tau)| \cdot \|x\|_C +$$
$$+ M_t(K(t'', \tau) - K(t', \tau))\|x\|_C \leqslant 2C_0C_1\gamma^2(h)\|x\|_C.$$

Отсюда $M(\sigma_n x) \leqslant 2C_0C_1\gamma(h)\|x\|_C$ и

$$\|\sigma_n(x)\|_S \leqslant C\gamma(h)\|x\|_C \leqslant C\gamma(h)\|x\|_S,\quad C = \text{const}.$$

Таким образом, $\|\sigma_n\|_{S\to S}\leqslant C\gamma(h)$ и при h достаточно малом будет $\|(I-K)^{-1}\|_{S\to S}\cdot\|\sigma_n\|_{S\to S}<1/2$. Поступая далее, как в конце § 1, получим

$$\|x_*-y_*^{(n)}\|_S\leqslant C\gamma(h),\quad C=\text{const}. \tag{11.1.14}$$

3°. Рассмотрим погрешность искажения для уравнений Фредгольма. Обозначим через A_n матрицу системы (11.1.6): $A_n=I_n+K^{(n)}$, где $K^{(n)}$ — матрица элементов $c_k^{(n)}K(\tau_j,\ \tau_k)$, I_n — единичная матрица порядка n. Будем рассматривать $\bar y^{(n)}=(y_1^{(n)},\ y_2^{(n)},\ \ldots,\ y_n^{(n)})$ и $\bar f^{(n)}=(f_1^{(n)}\ f_2^{(n)},\ \ldots,\ f_n^{(n)})$ как векторы в пространстве Y_{n1} (см. главу 3, § 9, п. 1°). Пусть $s_n^{(1)}$ — наименьшее сингулярное число матрицы A_n. Так как $y_k^{(n)}$ суть значения $y_*^{(n)}$ — решения уравнения (11.1.5) — в точках τ_j, а $y_*^{(n)}\xrightarrow{C}x_*$ и тем более $y_*^{(n)}\xrightarrow{L_2}x_*$, то нормы $\|\bar y^{(n)}\|_{Y_{n1}}$ ограничены в совокупности; по следствию 3.2.2 для устойчивости процесса (11.1.6) в последовательности пар пространств $(Y_{n1},\ Y_{n1})$ необходимо и достаточно, чтобы $\|A_n^{-1}\|_{Y_{n1}\to Y_{n1}}\leqslant C=\text{const}$. Но

$$\|A_n^{-1}\|_{Y_{n1}\to Y_{n1}}=\|A_n^{-1}\|_{R_n\to R_n},$$

и для устойчивости процесса (11.1.6) оказывается необходимым и достаточным, чтобы $s_n^{(1)}$ было положительно ограничено снизу. Легко убедиться, что это же условие необходимо и достаточно для устойчивости процесса вычисления приближенного решения в соответствующей последовательности пар подпространств.

4°. Погрешности алгоритма и округления для метода механических квадратур можно оценить, используя соответствующие оценки для линейных алгебраических систем (см. главы 5—7).

§ 2. УРАВНЕНИЯ ФРЕДГОЛЬМА, РЕШАЕМЫЕ ИТЕРАЦИЯМИ

1°. Пусть ядро уравнения (11.1.1) удовлетворяет неравенству

$$\max_{0\leqslant t,\ \tau\leqslant 1}|K(t,\ \tau)|=q<1; \tag{11.2.1}$$

тогда решение названного уравнения можно построить по методу простой итерации

$$x_*=f(t)+\sum_{k=1}^{\infty}(-1)^k\int_0^1 K_n(t,\ \tau)f(\tau)\,d\tau, \tag{11.2.2}$$

K_n — итерированные ядра. Ряд (11.2.2) сходится равномерно на отрезке $[0,\ 1]$ как прогрессия с знаменателем q; точнее, n-й

член этого ряда не превосходит величины $q^n \int_0^1 |f(\tau)|\,d\tau$. Удобнее производить вычисления не по формуле (11.2.2), а по обычной схеме итераций:

$$x_0(t) = f(t); \quad n \geqslant 1, \; x_n(t) = f(t) - \int_0^1 K(t, \tau)\, x_{n-1}(\tau)\, d\tau. \qquad (11.2.3)$$

Однако, если ядро имеет более или менее сложную структуру, то вычисление интегралов, входящих в формулу (11.2.3), может оказаться затруднительным. Поэтому мы прибегнем к следующему приему [109]. Введем функцию

$$\omega(t) = \begin{cases} t, & 0 \leqslant t \leqslant 1, \\ 2 - t, & 1 < t \leqslant 2, \\ 0, & t \notin [0,2], \end{cases} \qquad (11.2.4)$$

и обозначим

$$K^h(t, \tau) = \sum_{j,\,k=-1}^{2n-1} K(t_{j+1}, t_{k+1})\, \omega(t/h - j)\, \omega(t/h - k). \qquad (11.2.5)$$

Здесь n — произвольно выбранное натуральное число, $h = 1/(2n)$, $t_j = (j+1)h$. Если $K \in C^{(2)}(Q)$, $Q = [0,1] \times [0,1]$, то по формулам (10.6.12), (10.6.13)

$$\|K - K^h\|_{C(Q)} \leqslant Ch^2 \max_{|\alpha| = 2} \|D^\alpha K\|_{C(Q)}. \qquad (11.2.6)$$

Возьмем n столь большим, чтобы правая часть последней формулы была меньше, чем $1 - q$. Тогда, очевидно, $\|K^h\|_{C(Q)} := q_1 < 1$. Заменим в уравнении (11.1.1) ядро K на K^h, а свободный член $f(t)$ — функцией $f^h(t) = \sum_{j=-1}^{2n-1} f(t_{j+1})\, \omega(t/h - j)$. Разность $f(t) - f^h(t)$ есть разность ординат кривой $y = f(t)$ и вписанной в нее ломаной $y = f^h(t)$. Если $f \in C^{(2)}[0,1]$, то эту разность можно вычислить по формуле

$$t_{j+1} \leqslant t \leqslant t_{j+2}, \; f(t) - f^h(t) = -\frac{1}{h} \int_{t_{j+1}}^{t_{j+2}} G(t - t_{j+1}, \tau - t_{j+1}) f''(\tau)\, d\tau,$$

где $G(t, \tau) = \begin{cases} t(h - \tau), & 0 \leqslant t \leqslant \tau, \\ \tau(h - t), & 0 \leqslant \tau \leqslant t. \end{cases}$

Отсюда

$$\|f - f^h\|_C \leqslant h^2 \|f''\|_C. \qquad (11.2.7)$$

2°. Оценим погрешность аппроксимации от описанной в п. 1° замены. Новое интегральное уравнение имеет вид

$$x^h(t) + \int_0^1 K^h(t, \tau)\, x^h(\tau)\, d\tau = f^h(t). \qquad (11.2.8)$$

Вычитая это из (11.1.1), получаем $(I + K)(x - x^h) = f - f^h + (K - K^h)\, x^h$, или

$$x - x^h = (I + K)^{-1} [f - f^h - (K - K^h)(x - x^h) + (K - K^h)\, x].$$

Отсюда

$$\|x - x^h\| \leqslant \frac{\|(I+K)^{-1}\| \cdot [\|f - f^h\| + \|K - K^h\| \cdot \|x\|]}{1 - \|(I+K)^{-1}\| \cdot \|K - K^h\|};$$

все нормы взяты в $C[0,1]$ или соответственно в $C([0,1]\times[0,1])$. Используя неравенства (11.2.6), (11.2.7), а также неравенство $\|(I+K)^{-1}\| \leqslant (1-q)^{-1}$, находим оценку погрешности аппроксимации:

$$\|x - x^h\| \leqslant [1 - C'q \max_{|\alpha|=2} \|D^\alpha K\| h^2]^{-1} [C'' \max_{|\alpha|=2} \|D^\alpha K\| \cdot \|x\| + \\ + \|f''\|] h^2; \; C', \; C'' = \text{const}; \qquad (11.2.9)$$

можно еще воспользоваться неравенством

$$\|x\| \leqslant \|f\|/(1-q). \qquad (11.2.10)$$

3°. О погрешности искажения в более общей ситуации сказано в § 1. Рассмотрим уравнение, полученное из (11.2.8) искажением; решим это искаженное уравнение по методу итераций и оценим погрешность алгоритма. В данном случае при достаточно малом h и при достаточно малой погрешности искажения итерации сходятся как прогрессия с знаменателем $q_2 < 1$, где $q_2 = \max |\tilde{K}^h(t,\tau)|$ и $\tilde{K}^h(t,\tau)$ есть искаженное значение ядра $K^h(t,\tau)$. Если выполнено точно N итераций, то

$$\|\tilde{x}^h - \tilde{x}^h_N\|_C \leqslant \sum_{k=N+1}^{\infty} \left\| \int_0^1 \tilde{K}^h_k(t,\tau) \tilde{f}^h(\tau)\, d\tau \right\|.$$

Здесь $\tilde{f}^h(t)$ — искаженное значение функции $f^h(t)$, $\tilde{x}^h(t)$ — точное решение уравнения, полученного из (11.2.8) искажением. Из последнего неравенства вытекает оценка погрешности алгоритма:

$$\|\tilde{x}^h - \tilde{x}^h_N\|_C \leqslant \|\tilde{f}^{(h)}\|_C \, q_2^{N+1}/(1-q_2). \qquad (11.2.11)$$

4°. Представляет интерес оценка погрешности алгоритма для коэффициентов последовательных приближений $\tilde{x}^h_k(t)$. Справедлива точная формула

$$\tilde{x}^h_k(t) = \tilde{f}^h(t) - \int_0^1 \tilde{K}^h(t,\tau) \tilde{x}^h_{k-1}(\tau)\, d\tau. \qquad (11.2.12)$$

Нетрудно убедиться, что

$$\tilde{x}^h_{k-1}(t) = \sum_{j=-1}^{2n-1} \tilde{a}^h_{j,\,k-1} \omega(t/h - j), \quad \tilde{a}^h_{j,\,k-1} = \text{const}.$$

Подставив это, а также значения $\tilde{f}^h(t)$ и $\tilde{K}^h(t,\tau)$ в (11.2.12) и приравняв коэффициенты при $\omega(t/h - j)$, получим

$$\tilde{a}^h_{jk} = \tilde{f}^h(t_{j+1}) - \sum_{l,m=-1}^{2n-1} \tilde{a}^h_{m,\,k-1} \tilde{K}^h(t_{j+1},\, t_{l+1}) \times \\ \times \int_0^1 \omega(\tau/h - l)\, \omega(\tau/h - m)\, d\tau.$$

Интеграл в последнем тождестве вычисляется просто. Обозначая его через hw_{lm}, имеем $w_{lm} = w_{ml}$ и

$$w_{lm} = \begin{cases} 0, & |m-l| \geqslant 2, \\ 1/6, & m-l = \pm 1, \\ 2/3, & m = l,\ 0 \leqslant m \leqslant 2n-2, \\ 1/3, & m = l = -1,\ m = l = 2n-1. \end{cases} \tag{11.2.13}$$

В результате мы приходим к системе рекуррентных соотношений для коэффициентов $\tilde{a}_{jk}$:

$$\tilde{a}^h_{jk} = \tilde{f}^h(t_{j+1}) - h \sum_{l,m=-1}^{2n-1} \tilde{a}^h_{n,k-1} \tilde{K}^h(t_{j+1},\ t_{l+1}) w_{lm}: \tag{11.2.14}$$

эту систему надо решить при начальном условии $\tilde{a}^h_{j0} = \tilde{f}^h(t_{j+1})$.

Если мы введем в пространстве $2n$-компонентных векторов норму по формуле $\|a\| = \max|a_i|$, то норма оператора в правой части равенства (11.2.14) окажется равной

$$B := h \max_j \sum_{l,m=-1}^{2n-1} |\tilde{K}^h(t_{j+1},\ t_{l+1})|\, w_{lm}. \tag{11.2.15}$$

Приняв во внимание значения чисел w_{lm} (формула (11.2.13)), найдем, что

$$\sum_{l=-1}^{2n-1} w_{lm} |\tilde{K}^h(t_{j+1},\ t_{l+1}| \leqslant \max |\tilde{K}^h(t,\ \tau)| = q_2.$$

Отсюда $B \leqslant h \cdot 2nq_2 = q_2$, и при фиксированном индексе j итерационный процесс (11.2.14) сходится. Если ограничиться N итерациями, то погрешность алгоритма для процесса (11.2.14) оценивается величиной

$$q_2^{N+1} \|\tilde{f}^h\|_C/(1-q_2). \tag{11.2.16}$$

§ 3. РАЗЛИЧНЫЕ МЕТОДЫ СВЕДЕНИЯ К АЛГЕБРАИЧЕСКОЙ СИСТЕМЕ

Один метод сведения к алгебраической системе рассмотрен в § 1. В данном параграфе мы рассмотрим еще несколько таких методов.

1°. Рассмотрим уравнение (11.1.1) в следующих предположениях: а) $K \in L_2([0,1]\times[0,1])$; б) единица не есть характеристическое число ядра $K(t,\tau)$; в) $f \in L_2(0,1)$. Пусть построено такое вырожденное ядро

$$K^{(n)}(t,\ \tau) = \sum_{k=1}^{n} \alpha_k(t)\, \beta_k(\tau), \tag{11.3.1}$$

что

$$\int_0^1 \int_0^1 [K(t,\ \tau) - K^{(n)}(t,\ \tau)]^2\, dt d\tau := \gamma_n \underset{n\to\infty}{\to} 0. \tag{11.3.2}$$

В качестве $K^{(n)}(t,\tau)$ можно взять, например, конечноэлементную аппроксимацию ядра $K(t,\tau)$, если только последнее

обладает необходимой гладкостью; отметим, что числовые эксперименты подтверждают целесообразность такой аппроксимации. Другие способы выбора аппроксимирующего ядра указаны в [119]. Будем считать, что свободный член $f(t)$ также заменен другим, вообще говоря, свободным членом $f^{(n)}(t)$, близким к $f(t)$ в норме $L_2(0, 1)$, так что $\|f - f^{(n)}\|_2 \underset{n\to\infty}{\to} 0$. Таким образом, мы заменяем уравнение (11.1.1) новым уравнением

$$x^{(n)}(t) + \int_0^1 K^{(n)}(t, \tau)\, x^{(n)}(\tau)\, d\tau = f^{(n)}(t); \qquad (11.3.3)$$

оценим погрешность аппроксимации такой замены. Пусть K и $K^{(n)}$ означают операторы Фредгольма с ядрами $K(t, \tau)$ и $K^{(n)}(t, \tau)$ соответственно. Положим $R = (I + K)^{-1}$, $R^{(n)} = (I + K^{(n)})^{-1}$. Обозначая соответственно через x_* и $x_*^{(n)}$ точные решения уравнений (11.1.1) и (11.3.3), имеем

$$\|x_* - x_*^{(n)}\| = \|R(f - f^{(n)})\| + \|(R - R^{(n)} f^{(n)})\| \leqslant \|R\|\gamma_n' + \\ + \|R - R^{(n)}\|(\|f\| + \gamma_n'), \quad \gamma_n' = \|f - f^{(n)}\|. \qquad (11.3.4)$$

Далее, положим $K - K^{(n)} = \Delta^{(n)}$. Тогда $R^{(n)} = (I + K - \Delta^{(n)})^{-1} = R(I - \Delta^{(n)}R)^{-1}$; отсюда $R - R^{(n)} = -R\Delta^{(n)}R(I - \Delta^{(n)}R)^{-1}$, и если n достаточно велико, то $\|\Delta^{(n)}R\| \leqslant \gamma_n\|R\| < 1$, и

$$\|R - R^{(n)}\| \leqslant \frac{\gamma_n \|R\|^2}{1 - \gamma_n \|R\|}.$$

Подставив это в (11.3.4), получим оценку погрешности аппроксимации

$$\|x_* - x_*^{(n)}\| \leqslant \|R\|\gamma_n' + \frac{\gamma_n \|R\|^2}{1 - \gamma_n \|R\|}(\|f\| + \gamma_n'). \qquad (11.3.5)$$

Обычным способом решение уравнения (11.3.3) сводится к решению некоторой линейной алгебраической системы порядка n. При этом возникает необходимость в вычислении интегралов вида $\int_0^1 \alpha_j(t)\,\beta_k(t)\,dt$, $\int_0^1 f^{(n)}(t)\,\beta_k(t)\,dt$; эти вычисления, как правило, выполняются по приближенным квадратурным формулам. Кроме того, приходится производить некоторые арифметические действия. Все это приводит к погрешности искажения, которую можно оценить, зная оценки погрешности квадратурных формул. Далее, необходимо еще оценить погрешность, связанную с решением системы; это можно сделать, используя методы глав 5—7.

2°. Уравнение Фредгольма можно свести к линейной алгебраической системе по методу Бубнова — Галеркина. В этом случае можно оценить погрешность аппроксимации по формуле (2.6.3). При этом, если координатная система сильно минимальна в $L_2(0, 1)$, то процесс вычисления коэффициентов устойчив (в смысле второго определения) в последовательности

(R_n, R_n), а процесс вычисления приближенного решения устойчив в последовательности $(H^{(n)}, R_n)$; смысл обозначений очевиден.

Отметим два частных случая, когда можно уточнить оценки метода Бубнова — Галеркина. Пусть выбрана координатная последовательность $\{\omega_j(t)\}$, ортонормированная и полная в $L_2(0, 1)$, и пусть ядро $K(t, \tau)$ симметричное и неотрицательное: в данном случае это означает, что $K(t, \tau) = K(\tau, t)$ и что

$$(Kx, x) = \int_0^1 \int_0^1 K(t, \tau)\, x(t)\, x(\tau)\, dt d\tau \geqslant 0; \qquad (11.3.6)$$

для простоты считаем пространство $L_2(0, 1)$ вещественным. Верхняя грань квадратичной формы (11.3.6) равна $\|K\|$, а ее нижняя грань равна нулю.

По методу Бубнова — Галеркина приближенное решение уравнения (11.1.1) строится в виде

$$x^{(n)}(t) = \sum_{k=1}^n a_k^{(n)} \omega_k(t), \qquad (11.3.7)$$

причем коэффициенты $a_k^{(n)}$ определяются из алгебраической системы

$$a_j^{(n)} + \int_0^1 \int_0^1 K(t, \tau) \sum_{k=1}^n a_k^{(n)} \omega_k(\tau)\, \omega_j(t)\, dt d\tau = \int_0^1 f(t)\, \omega_j(t)\, dt. \qquad (11.3.8)$$

Матрицу этой системы обозначим через M_n. Если γ — произвольный вектор с числовыми составляющими $\gamma_1, \gamma_2, \ldots, \gamma_n$, то

$$(M_n\gamma, \gamma) = \|\gamma\|^2 + \int_0^1 \int_0^1 K(t, \tau) \sum_{j,k=1}^n \gamma_j \gamma_k \omega_j(t)\, \omega_k(\tau)\, dt d\tau.$$

Отсюда следует, что $\|\gamma\|^2 \leqslant (M_n\gamma, \gamma) \leqslant (1 + \|K\|)\|\gamma\|^2$ и, далее,

$$\|M_n\| \leqslant 1 + \|K\|, \quad \|M_n^{-1}\| \leqslant 1, \quad P_{M_n} \leqslant 1 + \|K\|. \qquad (11.3.9)$$

Таким образом, нормы матриц M_n и M_n^{-1} и число обусловленности ограничены известными постоянными, не зависящими от n. Это упрощает и делает более конкретными оценки погрешностей, данные в разделе II.

Второй частный случай таков. По-прежнему считаем координатную систему ортонормированной и полной в $L_2(0, 1)$; о ядре $K(t, \tau)$ допустим, что в разложении $K = K_1 + K_2$, $K_1 = 2^{-1}(K + K^*)$, $K_2 = 2^{-1}(K - K^*)$ ядро $K_1(t, \tau)$ неотрицательное. Напомним, что оператор K_2 кососимметричен, поэтому $(K_2x, x) \equiv 0$.

Оценим норму $\|(I + K_1)^{-1}\|$. Уравнение (11.1.1) запишем в виде $x + K_1x + K_2x = f$ и умножим скалярно на x; при этом член с K_2 исчезает, и мы получим $\|x\|^2 + (K_1x, x) = (f, x)$. Так как по предположению скалярное произведение (K_1x, x) неотрицательно, то $\|x\|^2 \leqslant (f, x) \leqslant \|f\| \cdot \|x\|$. Отсюда $\|x\| \leqslant \|f\|$, или

$\|(I+K)^{-1}\| \leqslant 1$. Несколько проще доказывается, что $\|(I+K_1)^{-1}\| \leqslant 1$.

Приближенное решение по Бубнову — Галеркину по-прежнему определяется формулами (11.3.7), (11.3.8). Как и выше, обозначим через M_n матрицу системы (11.3.8), и пусть $\mu_n^{(1)}$ и $\mu_n^{(2)}$ — матрицы, порожденные операторами K_1 и K_2, так что $M_n = I_n + \mu_n^{(1)} + \mu_n^{(2)}$. Как и выше, найдем, что $\|M_n\| \leqslant 1 + \|K\|$, $\|\mu_n^{(2)}\| \leqslant \|K_2\|$. Оценим еще величину $\|M_n^{-1}\|$. Имеем

$$M_n^{-1} = (I_n + \mu_n^{(1)} + \mu_n^{(2)})^{-1} = (I_n + \mu_n^{(1)})^{-1}\left[I_n + \mu_n^{(2)}(I_n + \mu_n^{(1)})^{-1}\right]^{-1}.$$

Матрица $\mu_n^{(1)}$ неотрицательная, поэтому $\|(I_n + \mu_n^{(1)})^{-1}\| \leqslant 1$. Потребуем еще, чтобы

$$\varkappa := \|K_2\| < 1. \tag{11.3.10}$$

Тогда

$$\|M_n^{-1}\| \leqslant (1-\varkappa)^{-1}, \tag{11.3.11}$$

и в этом случае величины $\|M_n\|$, $\|M_n^{-1}\|$, P_{M_n} также ограничены известными постоянными, не зависящими от n.

3°. Еще один способ сведения к линейной алгебраической системе дает метод наименьших квадратов. Мы не будем останавливаться здесь на сущности этого метода — об этом достаточно полно сказано в [84] — и сделаем только два общих замечания.

1. Пусть метод наименьших квадратов применен к уравнению $Ax = f$, где A — линейный оператор, действующий в некотором гильбертовом пространстве и ограниченно обратимый. Если приближенное решение $x_*^{(n)}$ этого уравнения построено, то апостериорная погрешность этого решения оценивается очень просто:

$$\|x_* - x_*^{(n)}\| \leqslant \|A^{-1}\| \cdot \|f - Ax_*^{(n)}\|. \tag{11.3.12}$$

2 Пусть еще оператор A ограничен. Тогда погрешность аппроксимации метода наименьших квадратов можно оценить по формуле (4.1.5).

Для уравнений Фредгольма оба эти утверждения очевидно справедливы.

4°. Просто решается вопрос об устойчивости относительно искажения В [90] приведены следующие простые факты. Если оператор A замкнут в гильбертовом пространстве H и имеет ограниченный обратный и если $f \in D(A^*)$, то уравнения $Ax = f$ и $A^*Ax = A^*f$ равносильны. Оператор A^*A — положительно определенный, его энергетическое пространство совпадает с множеством значений оператора A, а энергетическая норма $|x| = \|Ax\|$. Если оператор A ограничен, то очевидно, что его энергетическая норма эквивалентна норме пространства H. Из сказанного сразу вытекает следующее утверждение, вер-

ное, если уравнение Фредгольма разрешимо единственным образом:

Процесс наименьших квадратов для уравнения Фредгольма устойчив в последовательности пар пространств (R_n, R_n) *тогда и только тогда, когда координатная система сильно минимальна в* $L_2(0, 1)$.

Для погрешностей алгоритма и округления системы метода наименьших квадратов справедливо все сказанное по этому поводу в разделе II.

Можно еще заметить, что если координатная система метода наименьших квадратов почти ортонормирована в $L_2(0, 1)$, то число обусловленности матрицы этого метода ограничено независимо от n.

Сказанное в пп. 3°, 4° данного параграфа полностью относится и к уравнениям второго рода с вполне непрерывным оператором в сепарабельном гильбертовом пространстве.

5°. К линейной алгебраической системе можно приближенно свести уравнение Фредгольма и по методу коллокации. В [30] доказывается, что оптимальная по порядку оценка погрешности аппроксимации достигается при использовании подходящим образом выбранных сплайнов в качестве координатных функций. Если ядро и свободный член уравнения (11.1.1) принадлежат к классу $C^{(r)}$, а их модули непрерывности мажорируются одной и той же функцией $\omega(\delta)$, то погрешность аппроксимации коллокационного метода имеет оценку $O(n^{-r}\omega(1/n))$. В той же книге [30] приведены и некоторые другие результаты, близкие к только что сформулированному. По поводу устойчивости процесса коллокации можно повторить то, что было сказано в общем случае в главе 3, § 8. В частности, если координатная система $\varphi_k(t)$, $1 \leqslant k < \infty$, такова, что $|\varphi_k'(t)| \leqslant Ck^{\alpha}$, $\alpha < 1$, то процесс коллокации неустойчив.

При фиксированном порядке приближения и при заданных координатных функциях и узлах коллокации мы имеем дело с вполне определенной алгебраической системой. Для нее можно оценить погрешности искажения, алгоритма и округления так, как это описано в разделе II.

§ 4. СИНГУЛЯРНЫЕ ИНТЕГРАЛЬНЫЕ УРАВНЕНИЯ. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

1°. Будем рассматривать одномерное сингулярное интегральное уравнение вида

$$Ax := a(t)\,x(t) + b(t)\,(Sx)(t) + (Tx)(t) = f(t), \qquad (11.4.1)$$

где S — оператор Коши

$$(Sx)(t) = \frac{1}{\pi i}\int_{\Gamma} \frac{x(\tau)}{\tau - t}\,d\tau, \qquad (11.4.2)$$

в следующих предположениях: Γ — замкнутый плоский контур класса $C^{(1,\alpha)}$, $0<\alpha\leqslant 1$, ограничивающий некоторую область, односвязанную или многосвязанную, $a(t)$ и $b(t)$ — функции, скалярные или матричные, непрерывные на Γ, T — оператор, вполне непрерывный в $L_2(\Gamma)$, индекс уравнения (11.4.1) равен нулю и это уравнение имеет не более одного решения, так что оно разрешимо при любой (соответственно скалярной или векторной) функции $f\in L_2(\Gamma)$. При перечисленных условиях операторы A и A^{-1} ограничены в $L_2(\Gamma)$.

2°. Уравнение (11.4.1) можно решать по методу наименьших квадратов, и к этому уравнению в полной мере применимо все сказанное в пп. 3, 4° предшествующего параграфа.

Отметим некоторые координатные системы.

1) Пусть $\Gamma=\bigcup\limits_{j=0}^{a}\Gamma_j$, где Γ_j — замкнутые кривые; предположим, что Γ есть граница некоторой конечной области D, лежащей в плоскости комплексной переменной, и что Γ_0 ограничивает эту область извне. Внутри Γ_j, $1\leqslant j\leqslant n$, выберем точку t_j. Обозначим через $\rho_0, \rho_1, \ldots, \rho_n$ числа, удовлетворяющие неравенствам

$$0<\rho_0<\min_{\tau\in\Gamma_1}|\tau|;\quad \rho_j>\max_{\tau\in\Gamma_j}|\tau-t_j|,\quad 1\leqslant j\leqslant n.$$

Система функций

$$\left(\frac{t}{\rho_0}\right)^{N-1},\ \left(\frac{\bar t}{\rho_0}\right)^{N-1},\ \left(\frac{\rho_j}{t-t_j}\right)^{N},\ \left(\frac{\rho_j}{\bar t-\bar t_j}\right)^{N},\quad N\geqslant 1,\ 1\leqslant j\leqslant n,$$

полна и сильно минимальна в $L_2(\Gamma)$ (см. [90]).

2) Пусть Γ — вещественная ось и пусть вполне непрерывный оператор T — сглаживающий, так что $T: W_2^{(\nu)}(R_1)\to W_2^{(\nu+1)}(R_1)$. Если в качестве координатных взять функции

$$\varphi_k^{(1)}(t)=\sqrt{\frac{2}{1+t^2}}\,T_k\left(\frac{2t}{1+t^2}\right),\quad k=0,\ 1,\ 2,\ \ldots;$$

$$\varphi_k^{(2)}(t)=\frac{\sqrt{2}\,(1-t^2)}{(1+t^2)^{3/2}}\,U_k\left(\frac{2t}{1+t^2}\right),\quad k=1,\ 2,\ \ldots,$$

где T_k и U_k — полиномы Чебышева соответственно первого и второго рода, то общую оценку (4.1.5) можно уточнить: при $f\in W_2^{(s)}(R_1)$ и $a(t)$, $b(t)\in W_2^{(s)}(R_1)$ будет

$$\|x_*-x_*^{(n)}\|_{2,R_1}=0\left(n^{-s}\right)$$

(см. [183]). При этом процесс наименьших квадратов устойчив относительно погрешностей искажения, и число обусловленности матрицы метода наименьших квадратов ограничено.

3°. Для простоты примем, что Γ — окружность $|\tau|=1$ и что уравнение (11.4.1) — скалярное. Допустим, что решение назван-

ного уравнения $x_* \in W_2^{(2s)}(\Gamma)$; для этого достаточно, например, чтобы оператор T был сглаживающим (или чтобы он действовал из $L_2(\Gamma)$ в $W_2^{(2s)}(\Gamma)$), а $f(t)$, $a(t)$, $b(t) \in W_2^{(2s)}(\Gamma)$. В качестве координатных возьмем функции, которые получаются из кусочно-полиномиальных, степени $2s-1$, исходных функций МКЭ (см. главы 9, 10). В этом случае, как нетрудно доказать, погрешность аппроксимации имеет оценку $O(n^{-2s})$. Легко доказать также, что при выборе координатных функций, указанных в данном пункте, справедливо следующее утверждение: процесс наименьших квадратов устойчив в последовательности пар пространств $(\overline{X}_n, \overline{Y}_n)$, где $\overline{X}_n$ и $\overline{Y}_n$ суть n-мерные пространства с нормами

$$\forall a \in R_n, \quad \|a\|_{\overline{X}_n} = h^{1/2}\|a\|_{R_n}, \quad \|a\|_{\overline{Y}_n} = h^{-1/2}\|a\|_{R_n}.$$

Точно так же доказывается, что процесс построения приближенного решения по методу наименьших квадратов устойчив в последовательности пар пространств $(H_n, \overline{Y}_n)$, где H_n есть n-мерное подпространство пространства $L_2(\Gamma)$, натянутое на соответствующие координатные функции. При том же выборе координатных функций число обусловленности матрицы метода наименьших квадратов ограничено независимо от n.

4°. Отметим результаты статьи [52], в которой изучаются одномерные сингулярные интегральные уравнения на разомкнутом контуре. В цитируемой статье ее автор ограничивается случаем уравнений

$$ax(t) + \frac{b}{\pi i}\int_{-1}^{1} \frac{x(\tau)}{\tau - t}\,d\tau + \frac{1}{\pi}\int_{-1}^{1} K(t, \tau)\,x(\tau)\,d\tau = f(t), \quad (11.4.3)$$

где a, b — постоянные, $a^2 - b^2 \neq 0$. Уравнение изучается в пространстве $H = L_2(-1, 1; \rho)$; вес $\rho(t)$ выбирается в зависимости от индекса оператора (11.4.3). Интегральный оператор $\int_{-1}^{1} K(t, \tau)\,x(\tau)\,d\tau$ предполагается вполне непрерывным в H; предполагается также, что уравнение (11.4.3) имеет в H единственное решение. Приближенное решение строится по методу Бубнова — Галеркина; в качестве координатных функций выбираются полиномы Чебышева и Якоби. Доказывается, что алгебраическая система Бубнова — Галеркина достаточно высокого порядка однозначно разрешима и что приближенные решения сходятся к точному в H. Утверждается устойчивость процесса.

5°. Перейдем к многомерным сингулярным уравнениям. Рассмотрим уравнение (скалярное или векторное — в данном случае безразлично)

$$Ax := a(t)\,x(t) + \int_{\Gamma} K(t, \tau)\,x(\tau)\,d\tau + (Tx)(t) = f(t). \quad (11.4.4)$$

Здесь Γ — достаточно гладкое многообразие m измерений без края, $K(t, \tau)$ — сингулярное ядро, T — оператор, вполне непре-

рывный в банаховом пространстве, в котором сингулярный оператор ограничен. Для определенности будем говорить о пространстве $L_2(\Gamma)$. Примем следующие допущения: 1) символ $\Phi(x, \tau)$ уравнения (11.4.4) — достаточно гладкая функция своих аргументов; 2) индекс уравнения равен нулю; 3) уравнение (11.4.4) имеет в $L_2(\Gamma)$ не более одного решения — тогда оно разрешимо при любой правой части из $L_2(\Gamma)$. Операторы A и A^{-1} при сформулированных допущениях ограничены; повторяя многократно использованные выше рассуждения, приходим к следующим результатам.

Уравнение (11.4.4) допускает применение метода наименьших квадратов; погрешность аппроксимации оценивается по формуле (4.1.4). Если координатная система сильно минимальна в $L_2(\Gamma)$, то соответствующий вычислительный процесс устойчив; если координатная система почти ортонормирована в $L_2(\Gamma)$, то число обусловленности матрицы метода наименьших квадратов ограничено независимо от порядка этой матрицы. Таким образом, если систему метода наименьших квадратов решать методом простой итерации (см. главу 4, § 2, п. 1°), то эти итерации сходятся как прогрессия с фиксированным знаменателем, меньшим единицы, и погрешности округления при возрастании порядка системы остаются ограниченными.

В статьях [98] и [112] приведены два примера координатных систем, для которых общая оценка (4.1.4) может быть уточнена.

§ 5. МЕТОДЫ МЕХАНИЧЕСКИХ КВАДРАТУР И КОЛЛОКАЦИИ ДЛЯ ОДНОМЕРНЫХ СИНГУЛЯРНЫХ УРАВНЕНИЙ

1°. Приложениям метода механических квадратур к одномерным интегральным уравнениям посвящен ряд работ Габдулхаева и некоторых других авторов. Довольно полная библиография приведена в [31]; мы ограничимся здесь изложением статьи [29].

Рассмотрим уравнение (11.4.1), в котором вполне непрерывный оператор интегральный:

$$(Tx)(t) = \int_\Gamma K(t, \tau)\, x(\tau)\, d\tau. \qquad (11.5.1)$$

Примем, что Γ — окружность $|\tau| = 1$ и что функции $a(t)$, $b(t)$, $f(t)$ принадлежит к классу $C^{(r, \alpha)}$, где $r \geqslant 0$, $0 < \alpha < 1$. Допустим далее, что индекс рассматриваемого уравнения равен нулю и что это уравнение имеет не более одного решения. Приближенное решение будем искать в виде тригонометрического полинома

$$x^{(n)}(t) = \sum_{k=-n}^{n} a_k t^k. \qquad (11.5.2)$$

Из элементарных свойств оператора Коши вытекает, что

$$(Ax^{(n)})(t) = [a(t) + b(t)] \sum_{k=0}^{n} a_k t^k + [a(t) - b(t)] \sum_{k=-1}^{-n} a_k t^k + \\ + \int_\Gamma K(t, \tau) \sum_{k=-n}^{n} a_k \tau^k d\tau.$$

Заменим здесь интеграл его приближенным значением по формуле прямоугольников, соответствующей разбиению окружности Γ на $2n+1$ равных дуг с точками деления $t_j = \exp[2\pi ij/(2n+1)]$, и потребуем, чтобы полученная таким образом функция от t совпала в $f(t)$ в точках t_j. Это приведет нас к системе линейных алгебраических уравнений

$$[a(t_j) + b(t_j)] \sum_{k=0}^{n} a_k t_j^k + [a(t_j) - b(t_j)] \sum_{k=-1}^{-n} a_k t_j^k + \\ + \sum_{k=-n}^{n} a_k \sum_{p=0}^{2n-1} K(t_j, t_p) t_p^k [t_{p+1} - t_p] = f(t_j), \\ j = 0, 1, \ldots, 2a-1. \qquad (11.5.3)$$

В статье [29] эта система трактуется как система метода механических квадратур для уравнения (11.4.1). Может быть, стоит сказать, что метод названной статьи представляет собой некоторый синтез метода коллокации и метода механических квадратур: он переходит в метод коллокации при $K(t, \tau) \equiv 0$ и в метод механических квадратур при $b(t) \equiv 0$.

Основной результат цитируемой статьи содержится в следующей теореме.

Теорема 11.5.1. *Пусть* β *— любое число из интервала* $0 < \beta < \alpha$. *Существуют такие постоянные* A_1, B_1, A_2, B_2, *что если*

$$(A_1 \ln n + B_1) n^{-r-\alpha+\beta} < 1, \qquad (11.5.4)$$

то система (11.5.3) *имеет единственное решение. Погрешность аппроксимации приближенного решения* (11.5.2) *оценивается по формуле*

$$\| x_* - x_*^{(n)} \|_{H_\beta} \leqslant (A_2 \ln n + B_2) n^{-r+\beta-\alpha}, \qquad (11.5.5)$$

H_β *— пространство функций, заданных на* Γ, *с нормой*

$$\| x \|_{H_\beta} = \max_{t \in \Gamma} |x(t)| + \sup_{t', t'' \in \Gamma} \frac{|x(t') - x(t'')|}{|t' - t''|^\beta}.$$

Доказательство этой теоремы использует общую теорию приближенных методов, построенную Канторовичем [64].

2°. Рассмотрим теперь уравнение (11.4.1) в следующих предположениях: $T \equiv 0$; $a(t)$, $b(t)$ непрерывны на Γ; индекс уравнения равен нулю. Заметим, что результаты статьи [29] здесь неприменимы, потому что $r = \alpha = 0$. Данный случай рассмотрен в статье [193]. В качестве узлов коллокации взяты точки $t_k = t_k^{(n)} = \exp(2\pi ik/n)$, а в качестве координатных функций —

кусочно-линейные функции, широко употребляемые в МКЭ и определяемые формулой

$$\varphi_k^{(n)}(t)=\begin{cases}\dfrac{t-t_{k-1}}{t_k-t_{k-1}}, & \dfrac{2\pi(k-1)}{n}\leqslant \arg t\leqslant \dfrac{2\pi k}{n},\\ \dfrac{t_{k+1}-t}{t_{k+1}-t_k}, & \dfrac{2\pi k}{n}\leqslant \arg t\leqslant \dfrac{2\pi(k+1)}{n}.\end{cases} \tag{11.5.6}$$

Приближенное решение имеет вид

$$x^{(n)}(t)=\sum_{k=0}^{a-1} a_k^{(n)}\varphi_k^{(n)}(t); \tag{11.5.7}$$

коэффициенты $a_k^{(n)}$ определяются из системы

$$a(t_j)a_j^{(n)}+b(t_j)\sum_{k=0}^{n-1}a_k^{(a)}\left(S\varphi_k^{(n)}\right)(t_j)=f(t_j). \tag{11.5.8}$$

Основные результаты статьи [193] содержатся в следующих теоремах.

Теорема 11.5.2. *Пусть коэффициенты $a(t)$, $b(t)$ непрерывны, а решение $x_*\in L_2(\Gamma)$. Пусть выполнено условие*

$$\forall\lambda\in[-1,1],\quad \forall t\in\Gamma;\quad a(t)+\lambda b(t)\neq 0. \tag{11.5.9}$$

Тогда при достаточно большом n система (11.5.8) *разрешима единственным образом и $\|x_*-x_*^{(n)}\|_{2,\Gamma}\underset{n\to\infty}{\to}0$.*

Теорема 11.5.3. *Если $a(t)$, $b(t)\in H_\mu$, $0<\mu\leqslant 1$, и $f\in H_\nu$, $0<\nu\leqslant 1$, то верна оценка*

$$\|x_*-x_*^{(n)}\|_{2,\Gamma}\leqslant Cn^{-\sigma}\ln n,\quad \sigma=\min(\mu,\nu). \tag{11.5.10}$$

Здесь H_μ — пространство функций, удовлетворяющих условию Гельдера с показателем μ.

Теорема 11.5.4. *Пусть $a(t)$, $b(t)\in H_\mu$, $1/2<\mu\leqslant 1$. Если при любой непрерывной функции $f(t)$ система* (11.5.8) *разрешима, когда n достаточно велико, и $\|x_*-x_*^{(n)}\|_{2,\Gamma}\underset{n\to\infty}{\to}0$, то справедливо неравенство* (11.5.9).

Перечисленные здесь результаты распространены в [193] на случай кусочно-непрерывных коэффициентов и ляпуновского контура интегрирования.

3°. При фиксированном n можно получить оценки погрешностей алгоритма и округления так, как об этом сказано в разделе II.

4°. В статье [53] метод коллокации применен к уравнениям вида (11.5.3) при ограничениях, перечисленных в п. 4° § 4 данной главы. Доказывается, что если за узлы коллокации взять корни полиномов Якоби, выбранных подходящим образом, то системы метода коллокации достаточно высокого порядка однозначно разрешимы и приближенные решения сходятся к точному в соответствующей метрике.

Глава 12

РЕЗОЛЬВЕНТНЫЙ МЕТОД И ЕГО ПОГРЕШНОСТИ

Резольвентный метод приближенного решения уравнений Фредгольма разработан автором в статьях [109—113]. Сущность метода заключается в том, что ядро уравнения заменяется его конечноэлементной аппроксимацией (этот прием использован здесь в главе 11, § 2); соответственно резольвента данного ядра заменяется резольвентой аппроксимирующего ядра и затем используется представление решения через резольвенту. Как и в предшествующей главе, мы будем рассматривать скалярное уравнение вида (11.1.1); обобщения во многих случаях очевидны. По поводу менее очевидных случаев мы отсылаем читателя к статьям [111, 112].

§ 1. ПРИБЛИЖЕННОЕ ПРЕДСТАВЛЕНИЕ РЕЗОЛЬВЕНТЫ

1°. Рассмотрим интегральное уравнение Фредгольма

$$x(t) - \lambda \int_0^1 K(t, \tau)\, x(\tau)\, d\tau = f(t) \tag{12.1.1}$$

с ядром $K \in C^{(\tilde{s})}$, $\tilde{s} > 1$; λ — произвольный, вообще говоря, комплексный параметр. Напомним основные факты, относящиеся к резольвенте Фредгольма.

1) Резольвента Фредгольма имеет вид

$$\Gamma(t, \tau; \lambda) = D(t, \tau; \lambda)/D(\lambda), \tag{12.1.2}$$

где *определитель Фредгольма* $D(\lambda)$ *и первый минор Фредгольма* $D(t, \tau; \lambda)$ суть целые функции от λ.

2) Если $D(\lambda) \neq 0$ при данном λ, то уравнение (12.1.1) имеет одно и только одно решение

$$x(t) = f(t) + \lambda \int_0^1 \Gamma(t, \tau; \lambda)\, f(\tau)\, d\tau. \tag{12.1.3}$$

3) Коэффициенты рядов

$$D(\lambda) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} c_k \lambda^k, \tag{12.1.4}$$

$$D(t, \tau; \lambda) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} B_k(t, \tau) \lambda^k \tag{12.1.5}$$

удовлетворяют рекуррентным соотношениям

$$c_0 = 1, \quad c_k = \int_0^1 B_{k-1}(\tau, \tau)\, d\tau, \tag{12.1.6}$$

$$B_0(t, \tau) = K(t, \tau), \quad B_k(t, \tau) = c_k K(t, \tau) - \int_0^1 K(t, \tau')\, B_{k-1}(\tau', \tau)\, d\tau'. \tag{12.1.7}$$

4) Справедлива формула, выражающая коэффициенты $B_k(t, \tau)$ непосредственно через ядро:

$$k > 0,\ B_k(t,\ \tau) =$$

$$= \int_0^1 \dots \int_0^1 \begin{vmatrix} K(t,\ \tau) & K(t,\ \tau_1) & \dots & K(t,\ \tau_k) \\ K(\tau_1,\ \tau) & K(\tau_1,\ \tau_1) & \dots & K(\tau_1,\ \tau_k) \\ \dots & \dots & \dots & \dots \\ K(\tau_k,\ \tau) & K(\tau_k,\ \tau_1) & \dots & K(\tau_k,\ \tau_k) \end{vmatrix} d\tau_1 d\tau_2 \dots d\tau_k. \tag{12.1.8}$$

2°. Обозначим через Q квадрат $\{(t, \tau): 0 \leqslant t,\ \tau \leqslant 1\}$ и допустим, что ядро $K(t, \tau)$ уравнения (12.1.1) определено в некоторой области, для которой Q является внутренней подобластью, и что в этой области $K(t, \tau)$ принадлежит к классу $C^{(2s)}$, где s — натуральное число. Построим (см. главу 9) систему кусочно-полиномиальных исходных функций $\omega_q(t)$, $q = 0, 1, \dots, s-1$, степени $2s-1$, дающих наилучший порядок аппроксимации. В [94, 182] показано, что эту систему можно заменить единственной исходной функцией

$$\breve{\omega}(t) = \sum_{q=0}^{s-1} \sum_{l=-s+1}^{s-1} c_{ql}^{(s)} \omega_q(t + l), \tag{12.1.9}$$

где коэффициенты $c_{ql}^{(s)}$ определяются из s систем уравнений

$$\sum_{l=-s+1}^{s-1} l^r c_{ql}^{(s)} = \delta_{qr} r!,\ r = 0,\ 1,\ \dots,\ 2s-2, \tag{12.1.10}$$

соответствующих значениям $q = 0, 1, \dots, s-1$. Введем обозначения

$$\begin{aligned} &j = (j_1,\ j_2),\ z = (t,\ \tau),\ \tilde{\omega}(z) = \breve{\omega}(t)\,\breve{\omega}(\tau); \\ &b_j^{(0)} = b_{j_1 j_2}^{(0)} = K(j_1 h,\ j_2 h); \\ &K^h(t,\ \tau) = \sum_{j=-s}^{2n+s} b_j^{(0)} \tilde{\omega}(z/h - j). \end{aligned} \tag{12.1.11}$$

Как доказано в [94, 182],

$$|K(t,\ \tau) - K^h(t,\ \tau)| \leqslant Ch^{2s}. \tag{12.1.12}$$

Введем еще обозначения

$$\int_0^1 \breve{\omega}(t/h - j_1)\,\breve{\omega}(t/h - j_2)\,dt = h\breve{w}_{j_1 j_2} = h\breve{w}_j, \tag{12.1.13}$$

$$\beta_k^h(t,\ \tau) = B_k^h(t,\ \tau)/k!,\quad \gamma_k^h = c_k^k/k! \tag{12.1.14}$$

Используя соотношения (12.1.6), (12.1.7), легко найдем, что

$$\begin{aligned} &\beta_k^h(t,\ \tau) = \sum_{j=-s}^{2n+s} b_{jh}^{(k)} \tilde{\omega}(z/h - j), \\ &\gamma_k^h = (h/k) \sum_{j=-s}^{2n+s} b_{jh}^{(k-1)} \breve{w}_j,\ b_{jh}^{(k)} = \text{const}, \end{aligned} \tag{12.1.15}$$

$$b_{jh}^{(k)} = \gamma_k^h b_j^{(0)} - h \sum_{\rho=-s}^{2n+s} b_{j_1\rho}^{(0)} b_{\rho_1 j_2}^{(k-1)} \breve{w}_\rho; \tag{12.1.16}$$

из этих формул определяются коэффициенты $b_{jk}^{(k)}$, $k>0$ и $\gamma_k^{(h)}$, $k>0$, а следовательно, и величины $\beta_k^h(t,\tau)$.

Вырожденное ядро $K^h(t,\tau)$ можно представить в виде

$$K^h(t,\tau)=\sum_{j_1=-s}^{2n+s}\breve{\omega}(t/h-j_1)\sum_{j=-s}^{2n+s}b_j^{(0)}\breve{\omega}(\tau/h-j_2),$$

откуда видно, что оно содержит не более чем $\nu=2n+2s+1$ независимых слагаемых. Соответствующие этому ядру ряды Фредгольма суть полиномы от λ степени не выше ν, и рекуррентную формулу (12.1.16) надо применять не более чем ν раз — последующие коэффициенты рядов Фредгольма равны нулю.

3°. Пусть $D^h(t,\tau;\lambda)$ и $D^h(\lambda)$ суть ряда Фредгольма ядра $K^h(t,\tau)$ и пусть

$$D^h(t,\tau;\lambda)=\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}B_k^h(t,\tau)\lambda^k,\qquad D^h(\lambda)=\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}c_k^h\lambda^k.$$

Оценим разность $B_k(t,\tau)-B_k^h(t,\tau)$. Используя формулу (12.1.8), можно представить эту разность в виде (полагаем $t=t_0$)

$$B_k(t,\tau)-B_k^h(t,\tau)=\sum_{\nu=0}^{k}\int_0^1\dots\int_0^1\Delta_k^{(\nu)}\,dt_1\dots dt_k,$$

$$\Delta_k^{(\nu)}=$$

$$=\begin{vmatrix}K^h(t_0,\tau)\dots K^h(t_0,t_{\nu-1}) & K(t_0,t_\nu)-K^h(t_0,t_\nu) & K(t_0,t_{\nu+1})\dots K(t_0,t_k)\\ K^h(t_1,\tau)\dots K^h(t_1,t_{\nu-1}) & K(t_1,t_\nu)-K^h(t_1,t_\nu) & K(t_1,t_{\nu+1})\dots K(t_1,t_k)\\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\ K^h(t_k,\tau)\dots K^h(t_k,t_{\nu-1}) & K(t_k,t_\nu)-K^h(t_k,t_\nu) & K(t_k,t_{\nu+1})\dots K(t_k,t_k)\end{vmatrix}. \tag{12.1.17}$$

Будем считать, что $|K(t,\tau)|<1$ — этого можно добиться, изменив соответственно, в случае необходимости, параметр λ. Тогда в силу неравенства (12.1.12) при достаточно малом h будет также $|K^h(t,\tau)|<1$. Применяя то же неравенство (12.1.12) и теорему Адамара об определителях, получим $|\Delta_k^{(\nu)}|\leqslant Ch^{2s}(k+1)^{k/2}$. Интегрируя, найдем

$$|B_k(t,\tau)-B_k^h(t,\tau)|\leqslant Ch^{2s}(k+1)^{(k+2)/2}$$

и, как следствие,

$$|D(t,\tau;\lambda)-D^h(t,\tau;\lambda)|\leqslant Ch^{2s}\sum_{k=0}^{\infty}\frac{(k+1)^{(k+2)/2}}{k!}|\lambda|^k. \tag{12.1.18}$$

Остается оценить сумму последнего ряда. Из формулы Стирлинга следует неравенство

$$(k+1)^{k/2+3/4}<\left(1\Big/\sqrt[4]{2\pi}\right)\sqrt{(k+1)!}\,e^{(k+1)/2}$$

и ряд в (12.1.18) мажорируется рядом

$$\sqrt[4]{\frac{e^2}{2\pi}}\sum_{k=0}^{\infty}\frac{(k+1)^{3/4}(e|\lambda|^2)^{k/2}}{\sqrt{k!}}<$$

$$<\sqrt[4]{\frac{e^2}{2\pi}}\left[\sum_{k=0}^{\infty}\frac{1}{(k+1)^{3/2}}\right]^{1/2}\left[\sum_{k=0}^{\infty}\frac{(k+1)^3}{k!}(e|\lambda|^2)^k\right]^{1/2}.$$

Из неравенства $\frac{1}{(k+1)^{3/2}}<\int_k^{k+1}\frac{dy}{y^{3/2}}$ следует, что

$$\sum_{k=0}^{\infty}\frac{1}{(k+1)^{3/2}}<1+\int_1^{\infty}\frac{dy}{y^{3/2}}=3,$$

и для ряда (12.1.18) получается более простая мажоранта

$$\sqrt[4]{\frac{9e^2}{2\pi}}\left[\sum_{k=0}^{\infty}\frac{(k+1)^3}{k!}(e|\lambda|^2)^k\right]^{1/2}.$$

В последнем ряде выделим члены с индексами $k=0,\ 1,\ 2$, тогда мажоранта примет вид

$$\sqrt[4]{\frac{9e^2}{2\pi}}\left[1+8e|\lambda|^2+\frac{27}{2}e^2|\lambda|^4+\sum_{k=0}^{\infty}\frac{(k+4)^3}{(k+3)!}(e|\lambda|^2)^k\right]^{1/2}<$$

$$<\sqrt[4]{\frac{9e^2}{2\pi}}\left[1+8e|\lambda|^2+\frac{27}{2}e^2|\lambda|^4+\frac{32}{3}e^3|\lambda|^6\exp(e|\lambda|^2)\right]^{1/2}.$$

Отсюда следует оценка величины $|D(t,\ \tau;\ \lambda)-D^h(t,\ \tau;\ \lambda)|$:

$$\sqrt[4]{\frac{9e^2}{2\pi}}\left[1+8e|\lambda|^2+\frac{27}{2}e^2|\lambda|^4+\frac{32}{3}e^3|\lambda|^6\exp(e|\lambda|^2)\right]^{1/2}Ch^{2s}. \tag{12.1.19}$$

Аналогичные рассуждения приводят к оценке

$$|D(\lambda)-D^h(\lambda)|\leqslant\sqrt[4]{9/2\pi}\left[(e|\lambda|^2+2e^2|\lambda|^4)\exp(e|\lambda|^2)\right]^{1/2}Ch^{2s}. \tag{12.1.20}$$

В формулах (12.1.19), (12.1.20) C — постоянная неравенства (12.1.12); ее можно оценить, используя результаты § 6 главы 10. Формулы (12.1.19), (12.1.20) дают оценки аппроксимации для рядов Фредгольма.

4°. В п. 3° данного параграфа получены оценки погрешности числителя и знаменателя резольвенты в отдельности. Проще оценивается погрешность самой резольвенты как оператора (разумеется, в предположении, что значение λ — не характеристическое для ядра $K(t,\tau)$). Обозначим соответственно через K и K^h интегральные операторы с ядрами $K(t,\tau)$ и $K^h(t,\tau)$; положим еще $A=I-\lambda K$, $A^h=I-\lambda K^h$. Если h достаточно мало, то λ будет не характеристическим и для ядра $K^h(t,\tau)$; таким образом, будут существовать ограниченные операторы A^{-1} и $(A^h)^{-1}$. Оценим нормы их разности

$$\|A^{-1}-(A^h)^{-1}\|\leqslant\|A^{-1}\|\cdot\|I-A(A^h)^{-1}\|.$$

Далее,

$$A(A^h)^{-1} = (A^h - \lambda(K - K^h))(A^h)^{-1} = I - \lambda(K - K^h)(A^h)^{-1};$$

$$\|I - A(A^h)^{-1}\| \leqslant |\lambda| \cdot \|(A^h)^{-1}\| \cdot \|K - K^h\|$$

и по неравенству (12.1.12)

$$\|A^{-1} - (A^h)^{-1}\| \leqslant |\lambda| \cdot \|A^{-1}\| \cdot \|(A^h)^{-1}\| Ch^{2s}.$$

Если h достаточно мало, то $|\lambda| ch^{2s} \|(A^h)^{-1}\| < 1$, и получается оценка нормы $\|A^{-1}\|$:

$$\|A^{-1}\| \leqslant \frac{\|(A^h)^{-1}\|}{1 - |\lambda| Ch^2 \|(A^h)^{-1}\|}. \qquad (12.1.21)$$

Принимая, что оператор $(A^h)^{-1}$ уже вычислен, можно написать оценку погрешности аппроксимации решения уравнения (12.1.1), вызванную заменой ядра $K(t, \tau)$ на $K^h(t, \tau)$:

$$\|A^{-1}f - (A^h)^{-1} f\| \leqslant \frac{|\lambda| Ch^{2s} \|(A^h)^{-1}\|^2}{1 - |\lambda| Ch^2 \|(A^h)^{-1}\|} \|f\|. \qquad (12.1.22)$$

§ 2. ОЦЕНКА ОСТАТКА РЕЗОЛЬВЕНТЫ

1°. Если число s достаточно велико, то можно ограничиться сравнительно небольшими значениями n и ряды Фредгольма для аппроксимирующего ядра $K^h(t, \tau)$ можно вычислить до конца, используя рекуррентные соотношения (12.1.6), (12.1.7) $2n + 2s$ раз. Если же n велико, то вычисления можно вести до тех пор, пока отбрасываемые остатки рядов Фредгольма не станут достаточно малыми. Ниже в данном параграфе выводятся оценки этих остатков для произвольного ядра.

2°. Пусть в квадрате Q задано произвольное ядро $F(t, \tau)$. Введем обозначения

$$\max_t |F(t, \tau)| = \rho_F(\tau), \quad \max_\tau |F(t, \tau)| = \bar{\rho}_F(t),$$

$$\|\rho_F\|_{L_2(0, 1)} = \sigma_F, \quad \|\bar{\rho}_F\|_{L_2(0, 1)} = \bar{\sigma}_F \qquad (12.2.1)$$

и допустим, что все величины, входящие в формулы (12.2.1), конечны. Вместо σ_K и $\bar{\sigma}_K$ будем писать σ и $\bar{\sigma}$.

Оценим величину $\sigma_{B_k} := \sigma_k$. Пусть Δ — определитель под знаком интеграла в (12.1.6). По неравенству Адамара

$$|B_k(t, \tau)| \leqslant \int_0^1 \ldots \int_0^1 |\Delta| \, d\tau_1 \ldots d\tau_k \leqslant$$

$$\leqslant \left[\int_0^1 \ldots \int_0^1 \Delta^2 \, d\tau_1 \ldots d\tau_k\right]^{1/2} \leqslant (k+1)^{(k+1)/2} \rho_K(\tau) \sigma^k.$$

Отсюда

$$\sigma_k \leqslant (k+1)^{(k+1)/2} \sigma^{k+1}. \qquad (12.2.2)$$

Если в ряде (12.1.5) сохранить только N первых членов и обозначить через V_N сумму отброшенных членов, то

$$\sigma_{V_N} \leqslant \sum_{k=N+1}^{\infty} \frac{(k+1)^{(k+1)/2}}{k!} \mu^k \sigma, \quad \mu = |\lambda| \sigma.$$

Далее, $(k+1)^{(k+1)/2} = k^{k/2}(1+1/k)^{k/2}(k+1)^{1/2} \leqslant \sqrt{2e}\, k^{(k+1)/2}$, поэтому

$$\sigma_{V_N} \leqslant \sqrt{2e} \sum_{k=N+1}^{\infty} \frac{k^{(k+1)/2}}{k!} \mu^k \sigma.$$

Формула Стирлинга дает неравенство $1/k! < (2\pi)^{-1/2} e^k k^{-(k+1)/2}$, из которого следует

$$\sigma_{V_N} < \sqrt{e/\pi}\, \sigma \sum_{k=N+1}^{\infty} \frac{(\mu e)^k}{k^{k/2}}. \tag{12.2.3}$$

Возьмем N столь большим, чтобы члены ряда (12.2.3) убывали. Для этого достаточно, чтобы

$$N+1 \geqslant (\mu e)^2. \tag{12.2.4}$$

К рядам с неотрицательными убывающими членами можно применить интегральную оценку: если $\varphi(t) \geqslant 0$ и $\varphi(t)$ убывает, то $\varphi(k) \leqslant \int_{k-1}^{k} \varphi(t)\, dt$, поэтому

$$\sigma_{V_N} \leqslant \sqrt{\frac{e}{\pi}}\, \sigma \int_N^{\infty} \frac{(\mu e)^t}{t^{t/2}}\, dt \leqslant \sqrt{\frac{e}{\pi}}\, \sigma \int_N^{\infty} e^{-\frac{t}{2}(\ln N - 2\ln(\mu e))}\, dt =$$

$$= 2\sqrt{\frac{e}{\pi}}\, \sigma\, \frac{e^{-\frac{N}{2}(\ln N - 2\ln(\mu e))}}{\ln N - 2\ln(\mu e)} =$$

$$= 2\sqrt{\frac{e}{\pi}}\, \sigma \left[\frac{N}{(\mu e)^2}\right]^{-N/2} \left[\ln \frac{N}{(\mu e)^2}\right]^{-1}; \tag{12.2.5}$$

эта оценка верна при условии (12.2.4).

Рассмотрим численный пример. Пусть ядро $K(t, \tau)$ есть конечноэлементная аппроксимация некоторого ядра при шаге сетки $h = 0{,}001$. В данном случае ядро $K(t, \tau)$ содержит несколько более 1000 независимых слагаемых; чтобы полностью вычислить резольвенту, нужно применить рекуррентную формулу (12.1.16) также несколько более 1000 раз. Допустим, что для нашего ядра $\mu e < 10$; в соответствии с формулой (12.2.4) следует взять $N \geqslant 99$. Приведем оценки для σ_{V_N}

N	120	130	140	150
σ_{V_N}	10^{-3}	10^{-6}	10^{-9}	10^{-12}

При $N = 120$ точность оказывается достаточной для многих приложений. При $N = 150$ получается точность, значительно превосходящая обычные потребности приложений.

Аналогично получается оценка для остатка v_N знаменателя резольвенты Фредгольма

$$|v_N| \leqslant \frac{1}{\sqrt{2\pi N}} \left[\ln\left(\frac{N}{(\tilde{\mu}e)^2}\right)\right]^{-1} \left[\frac{N}{(\tilde{\mu}e)^2}\right]^{-N/2}; \quad \tilde{\mu} = |\lambda| \max(\sigma, \bar{\sigma}). \tag{12.2.6}$$

Оценки (12.2.5), (12.2.6) можно рассматривать как оценки погрешности алгоритма для процесса (12.1.6), (12.1.7).

§ 3. ПОГРЕШНОСТИ ИСКАЖЕНИЯ И ОКРУГЛЕНИЯ

1°. Рекуррентная система (12.1.16) искажается из-за неточного вычисления коэффициентов $b_j^{(0)} = K(j_1 h, j_2 h)$. Обозначим $\vartheta_0 = \max_j |\delta(b_j^{(0)})|$ и найдем оценки величин $\vartheta_k^h = \max_j |\delta(b_{jh}^{(k)})|$ и $\nu_k^h = |\delta(\gamma_k^h)|$.

Выше (п. 3 § 2) было принято, что $|K^h(t, \tau)| < 1$. Из неравенства Адамара легко следует, что

$$|\beta_h^{(k)}(t, \tau)| \leqslant (k+1)^{(k+1)/2}/k!, \tag{12.3.1}$$

$$|\gamma_k^h| \leqslant k^{k/2}/k! \tag{12.3.2}$$

2°. Допустим, что промежуточные вычисления производятся с повышенной точностью. Тогда по неравенству (1.1.9)

$$|\vartheta_k^h| \leqslant \varepsilon_1 \max_j |b_{jh}^{(k)}|, \quad |\nu_k^h| \leqslant \varepsilon_1 |\gamma_k^h|.$$

Далее, $b_{jh}^{(k)} = \beta_k^h(j_1 h, j_2 h)$, и по неравенствам (12.3.1), (12.3.2)

$$|\vartheta_k^h| \leqslant \frac{\varepsilon_1 (k+1)^{(k+1)/2}}{k!}, \quad |\nu_k^h| \leqslant \frac{\varepsilon_1 k^{k/2}}{k!}. \tag{12.3.3}$$

Суммарная погрешность числителя резольвенты (первого минора Фредгольма) оценивается величиной

$$\bar{\vartheta}(\lambda) = \varepsilon_1 \sum_{k=0}^{\infty} \max_j |b_{jh}^{(k)}| \sum_{j=-s}^{2n+s} |\breve{\omega}(z/h - j)| |\lambda|^k; \tag{12.3.4}$$

для простоты мы пренебрегаем погрешностью умножения на вторую сумму. По неравенству (12.3.1)

$$\max_j |b_{jh}^{(k)}| = \max_j |\beta_h^{(k)}(j_1 h, j_2 h)| \leqslant (k+1)^{(k+1)/2}/k!.$$

Кроме того, функция $\breve{\omega}$ ограничена, и ее носитель лежит в промежутке $[-s, s]$, поэтому вторая сумма в (12.3.4) содержит не более чем $2s+1$ слагаемых, отличных от нуля. Отсюда следует, что названная сумма ограничена; пусть она не превосходит некоторой постоянной c. Тогда по аналогии с неравенством

(12.1.19) найдем

$$\bar{\vartheta}(\lambda) \leqslant \varepsilon_1 c \sum_{k=0}^{\infty} \frac{(k+1)^{(k+1)/2}}{k!} |\lambda|^k \leqslant$$

$$\leqslant \varepsilon_1 c \sqrt[4]{\frac{9e^2}{2\pi}} |\lambda| \left[1 + 4e|\lambda|^2 + \frac{9}{2} e^2 |\lambda|^4 \exp\left(e|\lambda|^2\right)\right]^{1/2}. \quad (12.3.5)$$

Аналогично погрешность $\bar{\nu}(\lambda)$ знаменателя резольвенты (определителя Фредгольма) оценивается величиной

$$\bar{\nu}(\lambda) \leqslant \varepsilon_1 \sum_{k=1}^{\infty} \frac{k^{k/2}}{k!} |\lambda|^k \leqslant \varepsilon_1 \sqrt[4]{\frac{9e^2}{2\pi}} |\lambda| \exp\left(e|\lambda|^2/2\right). \quad (12.3.6)$$

Формулы (12.3.5), (12.3.6) дают оценки искажения рядов Фредгольма.

Если $|\lambda|$ не очень велик, то по формуле (1.1.9) можно оценить погрешность искажения резольвенты:

$$|\delta(\Gamma(t, \tau; \lambda))| \leqslant \varepsilon_1 |\Gamma(t, \tau; \lambda)|.$$

Если справа заменить Γ на Γ^h, то это приводит к появлению справа слагаемого $O(\varepsilon_1 h^{2s})$. Пренебрегая им, получаем

$$|\delta(\Gamma(t, \tau; \lambda))| \leqslant \varepsilon_1 |\Gamma^h(t, \tau; \lambda)|; \quad (12.3.7)$$

преимущество нового неравенства состоит в том, что его правая часть известна.

Погрешность искажения решения интегрального уравнения (12.1.1) имеет оценку

$$\varepsilon_1 |\lambda| \int_0^1 |\Gamma^h(t, \tau; \lambda)| \cdot |f(\tau)| d\tau. \quad (12.3.8)$$

3°. Оценим погрешность округления. Обозначим через $\bar{K}^h(t, \tau)$ ядро, полученное вместо $K^h(t, \tau)$ в результате искажения. Очевидно, что

$$\bar{K}^h(t, \tau) = \sum_{j=-\underline{s}}^{2n+\underline{s}} \bar{b}_j^{(0)} \tilde{\omega}(z/h - j), \quad (12.3.9)$$

где

$$\bar{b}_j^{(0)} = b_j^{(0)} + \delta(b_j^{(0)}). \quad (12.3.10)$$

Искаженная система (12.1.16) имеет вид

$$\bar{b}_j^{(k)} = \bar{\gamma}_k - h \sum_{\sigma=-\underline{s}}^{2n+\underline{s}} \bar{b}_{j_1\rho_2}^{(0)} \bar{b}_{\rho_1 j_2}^{(k-1)} \breve{w}_\rho, \quad (12.3.11)$$

где

$$\bar{\gamma}_k = (h/k) \sum_{j=-\underline{s}}^{2n+\underline{s}} \bar{b}_j^{(k-1)} \breve{w}_j. \quad (12.3.12)$$

Оценки коэффициентов $\bar{b}_j^{(k)}$ и $\bar{\gamma}_k$ получены в предшествующем параграфе.

Исключив $\bar{\gamma}_k$ из (12.3.11), получим рекуррентную систему для $\bar{b}_j^{(k)}$:

$$\bar{b}_j^{(k)} = h \sum_{\rho=-\underline{s}}^{2n+\underline{s}} \left[k^{-1} \bar{b}_\rho^{(k-1)} - \bar{b}_{j_1\rho_2}^{(0)} \bar{b}_{\rho_1 j_2}^{(k-1)}\right] \breve{w}_\rho. \quad (12.3.13)$$

Погрешность округления величины $\bar{b}_j^{(k)}$ обозначим через $\bar{\delta}_j^{(k)}$ и положим $\delta_k = \max_j |\delta_j^{(k)}|$. Вычислив выражение в квадратных скобках в (12.3.13), мы получим величину, отличающуюся от этого выражения на $O(\varepsilon_1^2)$. Теперь по формуле (1.1.9) найдем, что с точностью до величины порядка $O(\varepsilon_1^2)$

$$|\delta_j^{(k)}| \leqslant h \sum_{\rho=-\underline{s}}^{2n+\underline{s}} \left| [k^{-1}\bar{b}_\rho^{(k-1)} - \bar{b}_{j_1\rho_2}^{(0)} \bar{b}_{\rho j_1}^{(k-1)}] \right| \varepsilon_1.$$

Отсюда вытекает оценка погрешности округления величины $b_j^{(k)}$:

$$|\delta_j^{(k)}| \leqslant \bar{\delta}_k \leqslant \frac{(2n+2\underline{s}+1)^2}{2n} \varepsilon_1 \left(1 + \frac{1}{k}\right) \max_\rho |\bar{b}_\rho^{(k)}|. \quad (12.3.14)$$

§ 4. ПРИБЛИЖЕННОЕ ПРЕДСТАВЛЕНИЕ РЕЗОЛЬВЕНТЫ С ЗАДАННЫМ ЗНАМЕНАТЕЛЕМ

1°. Пусть $\Gamma(t, \tau; \lambda)$ — резольвента Фредгольма некоторого ядра $K(t, \tau)$, которое может быть скалярным или матричным, и пусть $D_0(\lambda)$ — произвольная аналитическая функция от λ, регулярная в окрестности точки $\lambda = 0$, причем $D_0(0) = 1$. Полагая $D_1(t, \tau; \lambda) = D_0(\lambda)\Gamma(t, \tau; \lambda)$, получаем представление резольвенты в виде частного двух степенных рядов

$$\Gamma(t, \tau; \lambda) = D_0(t, \tau; \lambda)/D_0(\lambda) \quad (12.4.1)$$

с заданным знаменателем; эти ряды сходятся в некоторой окрестности точки $\lambda = 0$.

Можно отметить очевидные случаи. Если $D_0(\lambda) \equiv 1$, то формула (12.4.1) дает разложение резольвенты в ряд по итерированным ядрам. Если $D_0(\lambda)$ есть определитель Фредгольма для ядра $K(t, \tau)$, то получается данное Фредгольмом представление резольвенты в виде частного двух целых функций. Если $D_0(\lambda)$ есть произведение определителя Фредгольма на $e^{g(\lambda)}$, где $g(\lambda)$ — произвольная целая функция, то получается другое представление резольвенты Фредгольма в виде частного целых функций, использованное в работах Пуанкаре (см., например, [43]).

2°. Пусть

$$D_0(\lambda) = \sum_{k=0}^{\infty} (-1)^k \gamma_k \lambda^k, \quad \gamma_0 = 1. \quad (12.4.2)$$

Положим

$$D_0(t, \tau; \lambda) = \sum_{k=0}^{\infty} (-1)^k \beta_k(t, \tau) \lambda^k. \quad (12.4.3)$$

Используя интегральное уравнение резольвенты

$$\Gamma(t, \tau; \lambda) = K(t, \tau) - \lambda \int_0^1 \Gamma(t, t'; \lambda) K(t', \tau)\, dt',$$

найдем, что коэффициенты $\beta_k(t, \tau)$ вычисляются по обычной формуле Фредгольма (12.1.7). Эти коэффициенты просто выражаются через итерированные ядра и коэффициенты γ_ν, $\nu \leqslant k$.

Действительно,

$$\beta_1(t,\ \tau)=\gamma_1 K(t,\ \tau)-\int_0^1 K(t,\ t')\,K(t',\ \tau)\,dt'=$$

$$=\gamma_1 K_1(t,\ \tau)-\gamma_0 K_2(t,\ \tau);$$

$$\beta_2(t,\ \tau)=\gamma_2 K(t,\ \tau)-\int_0^1 [\gamma_1 K(t,\ t')-\gamma_0 K_2(t,\ t')]\times$$

$$\times K(t',\ \tau)\,dt'=\gamma_2 K_1(t,\ \tau)-\gamma_1 K_2(t,\ \tau)+\gamma_0 K_3(t,\ \tau);$$

по индукции находим общую формулу

$$\beta_k(t,\ \tau)=\sum_{\nu=0}^{k}(-1)^{\nu}\gamma_{k-\nu}K_{\nu+1}(t,\ \tau). \qquad (12.4.4)$$

3°. Пусть дано уравнение Фредгольма (12.1.1). Примем, что значение λ — не характеристическое, ядро и свободный член — функции достаточно гладкие. Допустим еще, что нам известны первые, в порядке возрастания модулей, μ характеристических чисел ядра $K(t,\tau)$, а также порядки этих чисел как полюсов резольвенты. Самые характеристические числа обозначим через $\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_\mu$, а их кратности как полюсов резольвенты — через $p_1,\ p_2,\ \ldots,\ p_\mu$. Наконец, допустим, что $|\lambda|<|\lambda_{\mu+1}|$. В качестве $D_0(\lambda)$ возьмем полином

$$D_0(\lambda)=\prod_{\nu=0}^{\mu}(1-\lambda/\lambda_\nu)^{p_\nu}. \qquad (12.4.5)$$

Этот полином можно записать в виде (12.4.2); при этом будет $\gamma_\nu=0,\quad \nu>\nu_0=p_1+p_2+\ \ldots\ +p_\mu$. Соответствующий ряд $D_0(t,\tau;\lambda)$ сходится в круге $|\lambda|<|\lambda_{\mu+1}|$, и решение уравнения (12.1.1) допускает представление

$$x(t)=f(t)+\frac{\lambda}{D_0(\lambda)}\int_0^1\sum_{k=0}^{\infty}(-1)^k\lambda^k\beta_k(t,\ \tau)f(\tau)\,d\tau=$$

$$=f(t)+\frac{\lambda}{D_0(\lambda)}\sum_{k=0}^{\infty}\lambda^k\sum_{\nu=0}^{k}(-1)^{k-\nu}\gamma_{k-\nu}\int_0^1 K_{\nu+1}(t,\ \tau)f(\tau)\,d\tau. \qquad (12.4.6)$$

Менять порядок суммирования в формуле (12.4.6) нельзя — это привело бы к представлению решения в виде ряда

$$x(t)=f(t)+\lambda\sum_{k=0}^{\infty}\lambda^k\int K_{k+1}(t,\ \tau)f(\tau)\,d\tau,$$

который при произвольной $f(t)$ сходится только, если $|\lambda|<|\lambda_1|$.

4°. Выберем произвольное натуральное число N и положим

$$x_N(t)=f(t)+\frac{\lambda}{D_0(\lambda)}\int_0^1\sum_{k=0}^{N}(-1)^k\lambda^k\beta_k(t,\ \tau)f(\tau)\,d\tau=$$

$$=f(t)+\frac{\lambda}{D_0(\lambda)}\sum_{k=0}^{N}\lambda^k\sum_{\nu=0}^{k}(-1)^{k-\nu}\gamma_{k-\nu}\int_0^1 K_{\nu+1}(t,\ \tau)f(\tau)\,d\tau. \qquad (12.4.7)$$

Функцию $x_N(t)$ можно рассматривать как приближенное решение уравнения (12.1.1), потому что $x_N(t)\to x(t)$, если $N\to\infty$.

В конечной сумме (12.4.7) изменим порядок суммирования.

Введя обозначения

$$f_0(t)=f(t),\quad f_{\nu+1}(t)=\int_0^1 K_{\nu+1}(t,\ \tau)f(\tau)\,d\tau, \qquad (12.4.8)$$

получим

$$x_N(t)=f_0(t)+\frac{\lambda}{D_0(\lambda)}\sum_{\nu=0}^{N}\lambda^\nu f_{\nu+1}(t)\sum_{k=0}^{N-\nu}(-1)^k\gamma_k\lambda^k,$$

или, так как $\gamma_k=0$ при $k>\nu_0$, то

$$x_N(t)=f_0(t)+\frac{\lambda}{D_0(\lambda)}\sum_{\nu=0}^{N}\lambda^\nu f_{\nu+1}(t)\sum_{k=0}^{\min(N-\nu,\ \nu_0)}(-1)^k\gamma_k\lambda^k=$$

$$=f_0(t)+\frac{\lambda}{D_0(\lambda)}\Big[\sum_{\nu=0}^{N-\nu_0}\lambda^\nu f_{\nu+1}(t)\sum_{k=0}^{\nu_0}(-1)^k\gamma_k\lambda^k+$$

$$+\sum_{\nu=N-\nu_0+1}^{N}\lambda^\nu f_{\nu+1}(t)\sum_{k=0}^{N-\nu}(-1)^k\gamma_k\lambda^k\Big]=\sum_{\nu=0}^{N-\nu_0+1}\lambda^\nu f_\nu(t)+$$

$$+\frac{1}{D_0(\lambda)}\sum_{\nu=N-\nu_0+2}^{N+1}\lambda^\nu f_\nu(t)\sum_{k=0}^{N-\nu+1}(-1)^k\gamma_k\lambda^k. \qquad (12.4.9)$$

Первая сумма в (12.4.9) получается, если к уравнению (12.1.1) применить метод простой итерации; вторая сумма — поправка, обеспечивающая сходимость при $|\lambda|<|\lambda_{\mu+1}|$.

В частности, пусть $\nu_0=1$ — это возможно, если существует только одно наименьшее по модулю характеристическое число уравнения (12.1.1) и это число есть простой полюс резольвенты Фредгольма. Тогда

$$x_N(t)=\sum_{\nu=0}^{N}\lambda^\nu f_\nu(t)+\frac{\lambda^{N+1}}{1-\lambda/\lambda_1}f_{N+1}(t); \qquad (12.4.10)$$

при этом $x_N(t)\xrightarrow[N\to\infty]{}x(t)$, если $|\lambda|<|\lambda_2|$ и $\lambda\neq\lambda_1$.

Допуская, что $|\lambda|<|\lambda_{\mu+1}|-\varepsilon$, $\varepsilon>0$, легко получить оценку

$$|x(t)-x_N(t)|\leqslant C(\varepsilon)\left[\frac{|\lambda|}{|\lambda_{\mu+1}|-\varepsilon|}\right]^{N+1}. \qquad (12.4.11)$$

Пусть характеристические числа $\lambda_1, \lambda_2, \ldots, \lambda_{\mu+1}$, а с ними и коэффициенты γ_ν, $1\leqslant\nu\leqslant\nu_0$, известны только приближенно. Это приведет к искажению приближенного решения. Пусть γ'_ν и $x'_N(t)$ суть искажения коэффициентов λ_ν и приближенного решения $x_N(t)$ соответственно и пусть $|\gamma_\nu-\gamma'_\nu|\leqslant\delta$, $1\leqslant\nu\leqslant\nu_0$. Нетрудно получить оценку

$$|x_N(t)-x'_N(t)|\leqslant C(N,\ \varepsilon)\,\delta\int_0^1|f(t)|\,dt. \qquad (12.4.12)$$

5°. В условиях настоящего параграфа, когда знаменатель резольвенты задан заранее, резольвентный метод упрощается: как видно из формулы (12.4.10), достаточно вычислить итерации $f_\nu(t)$.

Введем в рассмотрение одномерную исходную функцию $\breve{\omega}(t)$, описанную в § 1, п. 2° данной главы. Определим ядро $K^h(t,\tau)$ формулой (12.1.11). Положим еще

$$f^h(t)=\sum_{j=-s}^{2n+s}f(jh)\,\breve{\omega}(t/h-j). \qquad (12.4.13)$$

Тогда, если $K(t,\tau)$ и $f(t)$ обладают некоторой достаточной гладкостью, то

$$|K(t,\tau)-K^h(t,\tau)|=O(h^{2s}),\quad |f(t)-f^h(t)|=O(h^{2s}). \qquad (12.4.14)$$

Если $x^h(t)$ — решение интегрального уравнения

$$x^h(t)-\lambda\int_0^1 K^h(t,\tau)\,x^h(\tau)\,d\tau=f^h(t), \qquad (*)$$

то $|x^h(t)-x(t)|=O(h^{2s})$. Будем строить приближенное решение уравнения (*) по формуле (12.4.9), причем будем считать, что $\lambda_1, \lambda_2, \ldots, \lambda_\mu$ суть характеристические числа данного ядра $K(t,\tau)$. Нетрудно убедиться, что это приведет к погрешности коэффициентов порядка $O(h^{2s})$. Соответствующая погрешность решения оценивается по формуле (12.4.12). Дело сводится к вычислению итераций $f_\nu^h(t)$. Допустим, что справедлива формула

$$f_{\nu-1}^h(t)=\sum_{l=-s}^{2n+s} f_l^{h,\,\nu-1}\,\breve{\omega}(t/h-l). \qquad (12.4.15)$$

Тогда

$$f_\nu^h(t)=h\sum_{j_1,\,j_2,\,l=-s}^{2n+s} b_{j_1j_2}f_l^{h,\,\nu-1}\breve{\omega}(t/h-j_1)\,\breve{w}_{lj_2}=$$
$$=\sum_{l=-s}^{2n+s} f_l^{h,\,\nu}\breve{\omega}(t/h-l). \qquad (12.4.16)$$

Здесь

$$f_l^{h,\,\nu}=h\sum_{j_1,\,j_2}^{2n+s} b_{lj_2}f_{j_1}^{h,\,\nu-1}\breve{w}_{j_1j_2}, \qquad (12.4.17)$$

а $\breve{w}_{j_1j_2}$ определяется формулой (12.1.13). Таким образом, итерации $f_\nu^h(t)$ определяются рекуррентно по формулам (12.4.15), (12.4.16).

6°. В качестве примера рассмотрим интегральное уравнение, к которому по методу потенциалов сводится решение внутренней задачи Дирихле для уравнения Лапласа на плоскости:

$$\sigma(t)=\frac{1}{2\pi}\int_L \frac{\cos(r,\nu)}{r}\,\sigma(\tau)\,d_\tau L=-\frac{1}{2\pi}\varphi(t). \qquad (12.4.18)$$

Область, ограниченную контуром L, считаем односвязной. Если $L\in C^{(s+1)}$, то ядро уравнения (12.4.18) принадлежит классу $C^{(s)}$; примем, что $\varphi(x)$ принадлежит тому же классу. Известно (см., например, [56]), что наименьшее по модулю характеристическое число ядра уравнения (12.4.18) равно -1, и это есть простой полюс резольвенты. Все остальные характеристические числа по модулю больше единицы.

Чтобы приближенно решить уравнение (12.4.18), можно положить $D_0(\lambda)=1+\lambda$. В этом случае $\nu_0=1$; $\gamma_1=-1$; $\gamma_k=0$, $k>0$; при $|\lambda|<|\lambda_2|$ можно вычислить приближенное решение по формуле (12.4.10). В частности, при $\lambda=1$ получаем приближенное решение в виде

$$\sigma_N(t)=\sum_{\nu=0}^{N} f_\nu(t)+\frac{1}{2}f_{N+1}(t),\quad f_0(t)=-\frac{1}{2\pi}\varphi(t). \qquad (12.4.19)$$

Итерации $f_\nu(t)$ можно вычислить приближенно по способу п. 5°. Формула (12.4.19) была ранее получена Канторовичем из других соображений (см. [65]).

Раздел IV
НЕЛИНЕЙНЫЕ ВЫЧИСЛИТЕЛЬНЫЕ ПРОЦЕССЫ

Глава 13
ОБЩИЕ ВОПРОСЫ

§ 1. О ПОГРЕШНОСТИ АППРОКСИМАЦИИ МЕТОДА РИТЦА

В пп. 1—4° настоящего параграфа мы коротко изложим постановку задачи о минимуме функционала и метод Ритца для ее решения; подробно эти вопросы изложены в [90, глава IX]. В п. 5° будет рассмотрена погрешность аппроксимации процесса Ритца.

1°. Пусть B — рефлексивное банахово пространство и F — функционал, определенный на линейном множестве $D(F)$, плотном в B. Ограничимся случаем, когда функционал F непрерывно дифференцируем на любой конечномерной гиперплоскости, лежащей в $D(F)$. Мы называем конечномерной гиперплоскостью в банаховом пространстве B совокупность элементов вида

$$x_0 + \sum_{k=1}^{n} a_k x_k, \tag{13.1.1}$$

где n — фиксированное натуральное число, называемое размерностью гиперплоскости; a_k, $1 \leqslant k \leqslant n$, — произвольные числа; x_k, $0 \leqslant k \leqslant n$, — фиксированные элементы, лежащие в B. Говоря, что F обладает той или иной гладкостью на гиперплоскости (13.1.1), мы понимаем под этим, что такой гладкостью обладает функция переменных a_1, a_2, ..., a_n, равная $F(x_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n)$.

Если функционал F имеет локальный минимум в точке $x_* \in D(F)$, то

$$(\operatorname{grad} F)(x_*) = 0. \tag{13.1.2}$$

Ниже будем обозначать $\operatorname{grad} F = P$; очевидно, что P действует из B в B^*.

Функционал F называется выпуклым, если

$$\forall x,\ y \in D(F),\ 0 \leqslant \alpha \leqslant 1,\ F((1-\alpha)x + \alpha y) \leqslant (1-\alpha)F(x) + \alpha F(y); \tag{13.1.3}$$

он называется существенно (или строго) выпуклым, если в соотношении (13.1.3) знак равенства имеет место только при $x = y$. Функционал F называется возрастающим, если $F(x) \to +\infty$ тогда и только тогда, когда $\|x\| \to \infty$. Тот же функционал называется полунепрерывным снизу (сверху), если $\varliminf_{x \to x_0} F(x) \geqslant F(x_0)$ (соответственно $\varlimsup_{x \to x_0} F(x) \leqslant F(x_0)$); он называется слабо полунепрерывным снизу (сверху), если упомянутое соотношение выполняется каждый раз, когда $x \rightharpoonup x_0$; символ $\rightharpoonup$ означает слабую сходимость.

Теорема 13.1.1. *Если функционал F возрастающий, существенно выпуклый и слабо полунепрерывный снизу, то он ограничен снизу и его нижняя грань достигается в единственной точке, к которой слабо сходится любая минимизирующая последовательность.*

Эта теорема доказана в [36].

2°. В данном пункте излагаются результаты статьи [8].

Пусть линейный оператор A действует из рефлексивного банахова пространства B в сопряженное пространство B^*; сопряженный оператор также действует из B в B^*. Будем говорить, что оператор A: 1) симметричен, если $D(A) \subset D(A^*)$ и $Ax = A^*x$, $x \in D(A)$; 2) самосопряжен, если он симметричен и $D(A) = D(A^*)$; 3) положителен, если он симметричен и $(Ax, x) > 0$ при $x \neq 0$; 4) положительно определен, если он симметричен и $(Ax, x) \geqslant \gamma^2 \|x\|^2$, где γ^2 — положительная постоянная. Если A — положительный (в частности, положительно определенный) оператор, то с ним можно связать энергетическое пространство B_A так же, как и в случае, когда пространство B само гильбертово [84]. Именно за B_A принимается гильбертово пространство, полученное замыканием множества $D(A)$ в норме, порожденной скалярным произведением $[x, y]_A = [x, y] = (Ax, y)$; здесь (Ax, y) есть значение функционала $Ax \in B^*$ на элементе $y \in B$. Норму в B_A будем обозначать через $|\cdot|$, так что $|x| = \sqrt{[x, x]}$.

Если A — положительно определенный оператор, то B_A ограниченно вкладывается в B; для положительного, но не положительно определенного оператора это утверждение неверно.

3°. Пусть оператор $P = \operatorname{grad} F$ имеет производную Гато P'_u со следующими свойствами: а) эта производная существует при любом $u \in D(P)$; б) $D(P'_u) \supset D(P)$; в) оператор P'_u — положительно определенный при любом $u \in D(P)$, так что

$$\forall x \in D(P'_u),\ (P'_u x,\ x) \geqslant \gamma^2 \|x\|^2, \tag{13.1.4}$$

причем величина γ не зависит от u. Доказывается, что в этих условиях функционал F ограничен снизу, и любая минимизирующая последовательность сходится в метрике B к пределу, который не зависит от минимизирующей последовательности.

Этот предел будем рассматривать как обобщенное решение задачи о минимуме функционала F.

Представляет интерес случай, когда P'_u удовлетворяет не только неравенству (13.1.4), но и более сильному неравенству

$$\forall x \in D(P'_u),\ (P'_u x,\ x) \geqslant \alpha^2 (P'_{u_0} x,\ x),\quad \alpha = \text{const}, \tag{13.1.5}$$

где u_0 — некоторый фиксированный элемент из $D(P)$. Обозначим через B_0 энергетическое пространство оператора P'_{u_0} и через $[\,,\,]_0$, $|\cdot|_0$ — скалярное произведение и норму в B_0. Можно доказать, что в этом случае обобщенное решение принадлежит B_0.

4°. Коротко опишем метод Ритца для задачи о минимуме функционала F. Напомним, что МКЭ является частным случаем метода Ритца. Зададим последовательность конечных последовательностей координатных элементов

$$\{\varphi_{n1},\ \varphi_{n2},\ \ldots,\ \varphi_{nN}\},\quad n = 1,\ 2,\ \ldots;\quad N = N(n), \tag{13.1.6}$$

подчиненных условиям: а) $\varphi_{nk} \in D(F) \cap B_0$; б) при данном n элементы φ_{nk} линейно независимы; в) последовательность (13.1.6) полна в B_0. Обозначим через B_n подпространство пространства B_0 с базисом $\{\varphi_{nk}\}$, $k = 1,\ 2,\ \ldots,\ N$, и будем искать минимум F на B_n. Решение $x_*^{(n)}$ этой новой задачи — приближенное решение данной задачи по Ритцу — существует и единственно, если F удовлетворяет условиям пп. 2, 3°; это приближенное решение имеет вид

$$x_*^{(n)} = \sum_{k=1}^{N} a_k^{(n)} \varphi_{nk}, \tag{13.1.7}$$

$a_k^{(n)}$ — числовые коэффициенты, определяемые из системы (нелинейной) Ритца

$$\frac{\partial}{\partial a_j^{(n)}} F\left(\sum_{k=1}^{N} a_N^{(n)} \varphi_{nk}\right) = 0,\quad 1 \leqslant j \leqslant N. \tag{13.1.8}$$

Будем рассматривать только те решения системы (13.1.8), которые доставляют величине $F\left(\sum_{k=1}^{N} a_k^{(n)} \varphi_{nk}\right)$ абсолютный минимум (эта оговорка не нужна, если функционал F существенно выпуклый — в этом случае система (13.1.8) имеет единственное решение).

Если последовательность пространств B_n расширяющаяся, то [90] последовательность $\{x_*^{(n)}\}$ приближенных решений по Ритцу — минимизирующая для функционала F, если последний функционал возрастающий и полунепрерывный сверху. Нетрудно показать, что это утверждение верно и тогда, когда подпространства B_n не обязательно расширяющиеся (так будет, например, в случае МКЭ), но их последовательность полна в B_0. В [90, § 70] доказывается (без предположения, что пространства B_n расширяющиеся), что из любой последовательности приближенных решений по Ритцу можно выделить мини-

мизирующую подпоследовательность. Допустим, что последовательность всех приближенных по Ритцу решений не минимизирующая. Тогда существуют такое число $\varepsilon_0 > 0$ и такая подпоследовательность $\{x_*^{(n_j)}\}$ приближенных по Ритцу решений, что $F(x_*^{(n_j)}) \geqslant d + \varepsilon_0$, $d = \inf F(x)$. Но из этой подпоследовательности можно выделить минимизирующую подпоследовательность, что невозможно. Полученное противоречие доказывает наше утверждение. Из этого утверждения вытекает, что о сходимости метода Ритца можно повторить все сказанное в п. 3° о сходимости минимизирующей последовательности.

5°. Для некоторого класса функционалов можно получить оценку погрешности аппроксимации процесса Ритца. Допустим, что производная P'_u оператора $P = \operatorname{grad} F$ удовлетворяет кроме неравенств (13.1.4), (13.1.5) еще и неравенству

$$\forall x \in D(P'_u),\ (P'_u x,\ x) \leqslant \beta^2 (P'_{u_0} x,\ x),\quad \beta = \text{const} > 0. \tag{13.1.9}$$

Воспользуемся неравенством [90, § 65]

$$F(x_* + y) - F(x_*) = \int_0^1 \tau^{-1} d\tau \int_0^1 (P'_{x_* + t\tau y} \tau y,\ \tau y)\, dt.$$

Применив неравенство (13.1.9) и положив $x_* + y = x$, получим

$$F(x) - F(x_*) \leqslant \beta^2 \int_0^1 \tau^{-1} d\tau \int_0^1 (P'_u(\tau(x - x_*),\ \tau(x - x_*))\, dt) =$$
$$= 2^{-1}\beta^2 |x - x_*|_0^2. \tag{13.1.10}$$

Пусть $y^{(n)}$ — элемент подпространства B_n, на котором достигается наилучшее приближение элемента x_*:

$$\mathscr{E}_n(x_*) := \inf_{x^{(n)} \in B_n} |x_* - x^{(n)}|_0 = |x_* - y^{(n)}|_0.$$

Положив в (13.1.10) $x = y^{(n)}$, находим

$$F(y^{(n)}) - F(x_*) \leqslant 2^{-1}\beta^2 |x_* - y^{(n)}|_0^2 = 2^{-1}\beta^2 \mathscr{E}_n^2(x_*). \tag{13.1.11}$$

Если $x_*^{(n)}$ — приближенное решение по Ритцу, то $F(x_*^{(n)}) \leqslant F(y^{(n)})$; по неравенствам (13.1.5), (13.1.9) получаем

$$2^{-1}\alpha^2 |x_* - x_*^{(n)}|_0^2 \leqslant F(x_*^{(n)}) - F(x_*) \leqslant$$
$$\leqslant F(y^{(n)}) - F(x_*) \leqslant 2^{-1}\beta^2 \mathscr{E}_n^2(x_*),$$

что дает следующую оценку погрешности аппроксимации:

$$|x_* - x_*^{(n)}|_0 \leqslant (\beta/\alpha)\, \mathscr{E}_n(x_*). \tag{13.1.12}$$

Если пространство B функциональное, то оценка порядка величины $\mathscr{E}_n(x_*)$ как функции от n довольно часто зависит от

гладкости решения x_*. По этому поводу см., например, [74, 186]. Для МКЭ оценка величины $\mathscr{E}_n(x)$ в зависимости от гладкости решения исследована в работах [96, 182, 125, 199, 134].

§ 2. ПОГРЕШНОСТЬ ИСКАЖЕНИЯ И УСТОЙЧИВОСТЬ СВОБОДНОГО ВЫЧИСЛИТЕЛЬНОГО ПРОЦЕССА

1°. Пусть даны две последовательности банаховых пространств X_n и Y_n, $n=1, 2, \ldots$, и пусть при любом n оператор (вообще говоря, нелинейный) A_n действует из X_n в Y_n, определен на всем пространстве X_n (это требование можно ослабить), удовлетворяет равенству $A_n(0)=0$, имеет обратный оператор A_n^{-1}, который, в свою очередь, определен на всем пространстве Y_n, и удовлетворяет следующему условию Липшица: если $\|y_1\|$, $\|y_2\| \leqslant t$, то

$$\| A_n^{-1}(y_1) - A_n^{-1}(y_2)\| \leqslant C_n(t)\| y_1 - y_2\|, \tag{13.2.1}$$

где $C_n(t)$ не зависит от y_1 и y_2.

Здесь и ниже, там, где это не может вызвать недоразумений, мы не ставим обозначения пространства при знаке нормы.

Рассмотрим свободный вычислительный процесс, состоящий в решении последовательности независимых уравнений

$$A_n(x^{(n)}) = f^{(n)}, \quad n=0, 1, 2, \ldots, \tag{13.2.2}$$

и соответствующий искаженный процесс

$$A_n(z^{(n)}) + \Gamma_n(z^{(n)}) = f^{(n)} + \delta^{(n)}. \tag{13.2.3}$$

Допустим, что искажения Γ_n и $\delta^{(n)}$ зависят от параметра $\delta > 0$ так, что

$$\|\delta^{(n)}\| \leqslant \delta, \tag{13.2.4}$$

и существуют числовые функции $\varphi_n(t, \delta)$ и $\psi_n(t, \delta)$ от положительных числовых переменных t и δ со следующими свойствами: $\lim\limits_{\delta\to 0} \varphi_n(t, \delta) = \lim\limits_{\delta\to 0} \psi_n(t, \delta) = 0$; если x, x', $x'' \in X_n$ и нормы этих элементов не превосходят t, то

$$\|\Gamma_n(x)\| \leqslant \varphi_n(t, \delta), \tag{13.2.5}$$

$$\|\Gamma_n(x') - \Gamma_n(x'')\| \leqslant \psi_n(t, \delta)\|x' - x''\|. \tag{13.2.6}$$

Уравнение (13.2.3) равносильно следующему:

$$z^{(n)} = A_n^{-1}\left(-\Gamma_n(z^{(n)}) + f^{(n)} + \delta^{(n)}\right). \tag{13.2.7}$$

Зафиксируем n и обозначим $\|f^{(n)}\| = T_n$. Докажем, что оператор в правой части уравнения (13.2.7) есть оператор сжатия в шаре

$$S_n = \{z^{(n)} \in X_n : \|z^{(n)}\| \leqslant 2T_nC_n(T_n)\},$$

если только δ достаточно мало. Имеем

$$\|-\Gamma_n(z^{(n)} + f^{(n)} + \delta^{(n)}\|) \leqslant \varphi_n(2T_nC_n(T_n), \delta) + T_n + \delta,$$

что не превосходит $2T_n$ при достаточно малом δ. Теперь по неравенству (13.2.1)

$$\| A_n^{-1}(-\Gamma_n(z^{(n)}) + f^{(n)} + \delta^{(n)})\| \leqslant 2T_n C_n(T_n),$$

и оператор (13.2.7) отображает шар S_n в себя. Далее, если $\zeta^{(n)} \in S_n$, то $\|\zeta^{(n)}\| \leqslant 2T_n C_n(T_n)$, и по неравенствам (13.2.1), (13.2.6)

$$\| A_n^{-1}(-\Gamma_n(z^{(n)}) + f^{(n)} + \delta^{(n)}) - A_n^{-1}(-\Gamma_n(\zeta^{(n)}) + f^{(n)} + \delta^{(n)})\| \leqslant$$
$$\leqslant C_n(2T_n C_n(T_n))\,\|\Gamma_n(z^{(n)}) - \Gamma_n(\zeta^{(n)})\| \leqslant$$
$$\leqslant C_n(2T_n C_n(T_n))\,\psi_n(C_n(T_n),\ \delta)\,\|z^{(n)} - \zeta^{(n)}\|,$$

если δ мало, то множитель при норме $\|z^{(n)} - \zeta^{(n)}\|$ меньше единицы, и наше утверждение доказано. В силу принципа сжатых отображений уравнение (13.2.3) имеет в шаре S_n ровно одно решение. Если $z_*^{(n)}$ — это решение, то

$$\|z_*^{(n)} - x_*^{(n)}\| = \| A_n^{-1}(-\Gamma_n(z_*^{(n)}) + f^{(n)} + \delta^{(n)}) - A_n^{-1}(f^{(n)})\| \leqslant$$
$$\leqslant C_n(T_n)\,\|\Gamma_n(z_*^{(n)}) - \delta^{(n)}\| \leqslant C_n(T_n)\,[\varphi_n(2T_n C_n(T_n),\ \delta) + \delta]. \quad (13.2.8)$$

Формула (13.2.8) дает оценку погрешности искажения для процесса (13.2.2) в довольно общих условиях; при фиксированном n эта оценка стремится к нулю вместе с δ. Представляет интерес случай, когда указанное стремление происходит равномерно относительно n. Этим мы займемся в следующем пункте.

2°. В [90] дано определение устойчивости свободного нелинейного вычислительного процесса и получена теорема о достаточных условиях устойчивости процесса Ритца для некоторого класса нелинейных уравнений. Несколько более общее определение устойчивости дал Т. С. Таккер [201]. Б. Дж. Омодеи в своей диссертации [189] доказал некоторые теоремы об устойчивости по Таккеру и применил их для исследования устойчивости МКЭ.

Здесь мы дадим другое определение устойчивости нелинейного процесса, естественным образом обобщающее определения главы 3 (см. также статью [106]).

Допустим, что в некотором шаре $\|x^{(n)}\| \leqslant \bar{l}$ уравнение (13.2.2) однозначно разрешимо при любом n. Относительно искаженного процесса (13.2.3) допустим, что справедливы неравенства (13.2.4), (13.2.5), причем функции φ_n и ψ_n не зависят от n; таким образом, если x, x', $x'' \in X_n$ и нормы этих элементов не превосходят числа l, то

$$\|\Gamma_n(x)\| \leqslant \varphi(l,\ \delta), \quad (13.2.9)$$

$$\|\Gamma_n(x') - \Gamma_n(x'')\| \leqslant \psi(l,\ \delta)\,\|x' - x''\|, \quad (13.2.10)$$

$$\lim_{\delta \to 0} \varphi(l,\ \delta) = \lim_{\delta \to 0} \psi(l,\ \delta) = 0. \quad (13.2.11)$$

Назовем процесс (13.2.2) устойчивым относительно искажения в последовательности пар пространств (X_n, Y_n), если для любого $\varepsilon > 0$ существует такое $\delta > 0$, что при выполнении неравенств (13.2.4), (13.2.9)—(13.2.11) уравнение (13.2 3) однозначно разрешимо в некоторой окрестности точки $x_*^{(n)}$, причем радиус этой окрестности не зависит от n, и справедливо неравенство

$$\| z_*^{(n)} - x_*^{(n)} \| < \varepsilon. \qquad 13.2.12)$$

Теорема 13.2.1. *Пусть* $\|f^{(n)}\| \leqslant T = \text{const}$, $A_n(0) = 0$, *оператор* A_n^{-1} *определен в шаре* $\|f^{(n)}\| \leqslant T$ *и удовлетворяет в этом шаре неравенству Липшица*

$$\| A_n^{-1}(f_1) - A_n^{-1}(f_2) \| \leqslant C \| f_1 - f_2 \| \qquad (13.2.13)$$

с постоянной C, *не зависящей от* n. *Пусть еще искажения* $\delta^{(n)}$ *и* Γ_n *удовлетворяют неравенствам* (13.2.4), (13.2.9)—(13.2.11). *Тогда процесс* (13.2.2) *устойчив в последовательности* (X_n, Y_n).

Для доказательства достаточно заметить, что в условиях теоремы 13.2.1 радиус шара S_n, в котором разрешимо искаженное уравнение (13.2.3), не зависит от n и что правая часть формулы (13.2.8) также не зависит от n и стремится к нулю вместе с δ.

Замечание 13.2.1. Условие $\|f^{(n)}\| \leqslant T$ можно отбросить, если оператор A_n^{-1} определен на всем пространстве Y_n и удовлетворяет условию Липшица (13.2.13), в котором C не зависит от n.

§ 3. ОБ УСТОЙЧИВОСТИ ПРОЦЕССОВ РИТЦА И МКЭ

1°. В [90, § 76] доказано, что классический процесс Ритца для задачи $F(x) = \min$ устойчив, если: 1) оператор P имеет производную Гато P'_u, удовлетворяющую неравенствам (13.1.4), (13.1.5); 2) координатная система сильно минимальна в энергетическом пространстве оператора P'_{u_0}. Доказательство проведено в предположении, что искажение Γ_n удовлетворяет требованиям, несколько более ограничительным, чем требования (13.2.9), (13.2.10). Легко, однако, провести доказательство, предполагая выполненными только эти последние требования. Мы не будем останавливаться на этом подробнее; ниже будет доказана более общая теорема, аналогичная теореме 3.3.3 для линейных задач.

2°. Пусть функционал F, определенный в банаховом пространстве B, обладает следующими свойствами: оператор $P = \operatorname{grad} F$ имеет в любой точке $u \in B$ производную Гато P'_u, удовлетворяющую неравенствам (13.1.4), (13.1.5); $P(0) = 0$. Поставим задачу о минимуме функционала $F(x) - (f, x)$, где f — заданный элемент сопряженного пространства B^*. Пусть

координатная система имеет вид (13.1.6). Составим матрицу чисел $[\varphi_{nj}, \varphi_{nk}]_{j,\,k=1}^{N}$, где квадратные скобки означают скалярное произведение в энергетическом пространстве B_0 оператора P'_{u_0} (формула (13.1.5)); эту матрицу обозначим через M_n. Пусть $\lambda_n^{(1)}$ — наименьшее собственное число матрицы M_n и пусть $\gamma(n)$ — положительная функция натурального числа n, удовлетворяющая неравенству

$$\lambda_1^{(n)} \geqslant C\gamma^2(n), \quad C = \text{const}, \tag{13.3.1}$$

Введем N-мерные гильбертовы пространства $\bar{X}_n$, $\bar{Y}_n$ с нормами

$$\forall b \in R_N, \quad \|b\|_{\bar{X}_N} = \gamma(n)\|b\|_{R_N}, \quad \|b\|_{\bar{Y}_N} = \frac{\|b\|_{R_N}}{\gamma(n)}. \tag{13.3.2}$$

Примем, что координатная система удовлетворяет обычным условиям: 1) элементы $\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nN}$ образуют базис натянутого на них подпространства $B_0^{(n)}$ пространства B_0; 2) последовательность $\{B_0^{(n)}\}$ полна в B_0. Добавим к этому еще одно условие, более ограничительное, чем в линейном случае: 3) $\varphi_{nk} \in D(P)$. Как всегда в методе Ритца, будем искать приближенное решение задачи о минимуме как элемент пространства $B_0^{(n)}$:

$$x^{(n)} = \sum_{k=1}^{N} a_k^{(n)} \varphi_{nk}.$$

Условие 3) позволяет записать уравнения Ритца в форме Бубнова — Галеркина:

$$\left(P\left(\sum_{k=1}^{N} a_k^{(n)}\varphi_{nk}\right), \varphi_{nj}\right) = (f, \varphi_{nj}), \quad 1 \leqslant j \leqslant N. \tag{13.3.3}$$

Обозначим $a^{(n)} = (a_1^{(n)}, a_2^{(n)}, \ldots, a_N^{(n)})$, $f^{(n)} = ((f, \varphi_{n1}), (f, \varphi_{n2}), \ldots$ $\ldots, (f, \varphi_{nN}))$. Положим еще

$$P\left(\sum_{k=1}^{N} a_k^{(n)}\varphi_{nk}, \varphi_{nj}\right) = P_j^{(n)}(a^{(n)})$$

и

$$P_n(a^{(n)}) = (P_1^{(n)}(a^{(n)}), P_2^{(n)}(a^{(n)}), \ldots, P_N^{(n)}(a^{(n)})).$$

Тогда систему Ритца можно записать в виде

$$P_n(a^{(n)}) = f^{(n)}; \tag{13.3.4}$$

будем при этом рассматривать $a^{(n)}$ и $f^{(n)}$ как элементы пространств $\bar{X}_N$ и $\bar{Y}_N$ соответственно.

Составление оператора P_n и вектора $f^{(n)}$ сопряжено с погрешностями вычислений. Пусть в результате вычислений мы вместо P_n и $f^{(n)}$ получили искаженные объекты $P_n + \Gamma_n$ и $f^{(n)} + \delta^{(n)}$. Тогда вместо системы (13.3.4) мы на самом деле получим и будем решать систему

$$(P_n + \Gamma_n)(c^{(n)}) = f^{(n)} + \delta^{(n)} \tag{13.3.5}$$

3° **Теорема 13.3.1.** *В условиях п. 2° данного параграфа процесс* (13.3.3) *устойчив в последовательности пар пространств* $(\bar{X}_N, \bar{Y}_N)$.

Прежде всего покажем, что нормы $\|a^{(n)}\|_{\bar{X}_N}$ ограничены независимо от n. Так как

$$P(x_*^{(n)}) = P(x_*^{(n)}) - P(0) = \int_0^1 P'_{tx_*^{(n)}} x_*^{(n)}\, dt,$$

то уравнениям Ритца можно придать вид

$$\int_0^1 \left(\left(P'_{tx_*^{(n)}} x_*^{(n)},\ x_*^{(n)} \right),\ \varphi_{nj} \right) dt = (f,\ \varphi_{nj}), \quad 1 \leqslant j \leqslant N.$$

Умножая это на $a_j^{(n)}$ и суммируя, находим

$$\int_0^1 \left(\left(P'_{tx_*^{(n)}},\ x_*^{(n)} \right),\ x_*^{(n)} \right) dt = (f,\ x_*^{(n)}).$$

Слева заменим подынтегральную функцию по неравенству (13.1.4) меньшей величиной $\alpha^2 (P'_{u_0} x_*^{(n)},\ x_*^{(n)})$, а справа заменим $(f,\ x_*^{(n)})$ большей величиной $\|f\| \cdot \|x_*^{(n)}\|$:

$$\alpha^2 |x_*^{(n)}|_0^2 \leqslant \|f\| \cdot \|x_*^{(n)}\|;$$

$|\cdot|_0$ — норма в B_0. По неравенству (13.1.4) $\|x_*^{(n)}\| \leqslant \gamma^{-1} |x_*^{(n)}|_0$, поэтому

$$|x_*^{(n)}|_0 \leqslant \|f\| / \alpha^2 \gamma. \tag{13.3.6}$$

Далее,

$$|x_*^{(n)}|_0^2 = (M_n a^{(n)},\ a^{(n)})_{R_N} \geqslant \lambda_1^{(n)} \|a^{(n)}\|_{R_N}^2 \geqslant$$

$$\geqslant C\gamma^2(n) \|a^{(n)}\|_{R_N}^2 = C \|a^{(n)}\|_{\bar{X}_N}^2,$$

и мы приходим к нужной оценке

$$\|a^{(n)}\|_{\bar{X}_N} \leqslant C^{-1/2} |x_*^{(n)}|_0 \leqslant \|f\| / \alpha^2 \gamma \sqrt{C}. \tag{13.3.7}$$

4°. Дальнейшее опирается на результаты статьи [176], которые мы здесь опишем с небольшими изменениями. В этой статье рассматриваются операторы, действующие в гильбертовом пространстве, но рассуждения и результаты этой работы без труда обобщаются на операторы, действующие из рефлексивного банахова пространства B в сопряженное пространство B^*. Пусть P — оператор, вообще говоря, нелинейный, определенный на линейном множестве $D(P)$, плотном в B, и действующий из B в B^*. Пусть $P(0) = 0$. Предположим, что производная Гато P'_u определена при любом $u \in D(P)$ и $D(P'_u) \supset D(P)$. Допустим далее, что эта производная удовлетво-

ряет неравенствам (13.1.4), (13.1.5). Оператор P совпадает на $D(P)$ с градиентом функционала

$$F(x)=\int_0^1 (P(tx),\ x)\,dt,$$

поэтому решение (если оно существует) уравнения

$$Px=f,\ f\in B^*, \tag{13.3.8}$$

реализует абсолютный минимум функционала

$$\Phi_f(x)=F(x)-(f,\ x). \tag{13.3.9}$$

Задача о минимуме функционала (13.3.9) имеет одно и только одно решение (вообще говоря, обобщенное) при любом $f\in B^*$; это решение принадлежит энергетическому пространству B_0 (см. § 1 данной главы). Расширим функционал F на все множество обобщенных решений задачи о минимуме функционала (13.3.9), полученных при всевозможных $f\in B^*$: если $\bar{x}$ — такое решение и оно соответствует элементу $\bar{f}\in B^*$, то положим

$$F(\bar{x})=\inf_{x\in D(P)} \Phi_{\bar{f}}(\bar{x})+(\bar{f},\ x).$$

После такого расширения задача о минимуме функционала (13.3.9) имеет при любом $f\in B^*$ одно и только одно решение, и на этом решении $\operatorname{grad}\Phi_f$ существует и равен нулю. Решение задачи о минимуме функционала (13.3.9) обозначим через $\tilde{P}^{-1}(f)$.

Различным f соответствуют различные решения. Допустим противное, и пусть $x_*=\tilde{P}^{-1}(f_1)=\tilde{P}^{-1}(f_2)$. Тогда при любом $\xi\in B_0$ будет

$$(\operatorname{grad}_{x_*}\Phi_{f_1})(\xi)=(\operatorname{grad}_{x_*}\Phi_{f_2})(\xi)=0.$$

Отсюда $(f_1-f_2,\ \xi)=0$; так как B_0 плотно вкладывается в B, то $f_1=f_2$, и наше утверждение доказано. Из него следует, что существует оператор $\tilde{P}$: $=(\tilde{P}^{-1})^{-1}$ и $\tilde{P}(x)=f$. При этом, если $x\in D(P)$, то $P(x)=f$; это значит, что $\tilde{P}$ есть расширение оператора P.

Докажем, что $\tilde{P}^{-1}$ удовлетворяет условию Липшица

$$\left|\tilde{P}^{-1}f_1-\tilde{P}^{-1}f_2\right|_0\leqslant\alpha^{-2}\gamma^{-1}\|f_1-f_2\|, \tag{13.3.10}$$

где α и γ — постоянные неравенств (13.1.4), (13.1.5). Обозначим $\tilde{P}^{-1}f_k=x_k$, $k=1,\ 2$. Рассмотрим величину

$$(f_1-f_2,\ x_1-x_2)=(\tilde{P}(x_1)-\tilde{P}(x_2),\ x_1-x_2).$$

Для любого оператора P, дифференцируемого по Гато, известно соотношение (см., например, [90, с. 302])

$$P(x)-P(y)=\int_0^1 P'_{y+t(x-y)}(x-y)\,dt.$$

Отсюда

$$(f_1 - f_2,\ x_1 - x_2) = \int_0^1 \left(\tilde{P}'_{x_2 + t(x_1 - x_2)}(x_1 - x_2),\ x_1 - x_2\right) dt$$

и далее, по формуле (13.1.5)

$$|x_1 - x_2|_0^2 = (\tilde{P}'_{u_0}(x_1 - x_2),\ x_1 - x_2) \leqslant$$

$$\leqslant \alpha^{-2} \int_0^1 \left(\tilde{P}'_{x_2 + t(x_1 - x_2)}(x_1 - x_2),\ x_1 - x_2\right) dt =$$

$$= \alpha^{-2} (f_1 - f_2,\ x_1 - x_2) \leqslant \alpha^{-2} \|f_1 - f_2\|_{B^*} \cdot \|x_1 - x_2\|_B.$$

В силу формулы (13.1.4) $\|x_1 - x_2\|_B \leqslant \gamma^{-1} |x_1 - x_2|_0$; из двух последних неравенств вытекает условие Липшица (13.3.10).

Рассмотрим теперь уравнение

$$\tilde{P}(z) + \Gamma(z) = f, \tag{13.3.11}$$

в котором Γ удовлетворяет условию

$$\|\Gamma(z_1) - \Gamma(z_2)\| \leqslant \varkappa |z_1 - z_2|, \quad \varkappa = \text{const}. \tag{13.3.12}$$

Доказывается, что при $\varkappa \leqslant \alpha^2 \gamma$ уравнение (13.3.11) имеет одно и только одно решение. Если x_* и z_* — решения уравнений (13.3.8), (13.3.11) соответственно, то $|z_* - x_*|_0 \xrightarrow[\varkappa \to 0]{} 0$.

5°. Вернемся к теореме 13.3.1. Воспользуемся результатами, изложенными в п. 4°. С этой целью отождествим $\tilde{P}$ с P_n, Γ — с Γ_n, B — с $\tilde{X}_N$. Тогда следует отождествить B^* с Y_N и B_0 — с R_N. Далее, отождествим x с $a^{(n)}$, f — с $f^{(n)}$ и z — с $c^{(n)}$. Оценим значения величин γ, α, $\varkappa$. Прежде всего

$$P'_{n,\, a^{(n)}b^{(n)}} = \left(P^{(n)'}_{1,\, a^{(n)}b^{(n)}},\ P^{(n)'}_{2,\, a^{(n)}b^{(n)}},\ \ldots,\ P^{(n)'}_{N,\, a^{(n)}b^{(n)}}\right).$$

Имеем

$$P^{(n)'}_{l,\, a^{(n)}b^{(n)}} = \frac{d}{dt}\left(P\left(x^{(n)} + t y^{(n)}\right),\ \varphi_{nl}\right)\Big|_{t=0},$$

здесь

$$x^{(n)} = \sum_{k=1}^N a_k^{(n)} \varphi_{nk}, \quad y^{(n)} = \sum_{k=1}^N b_k^{(n)} \varphi_{nk}.$$

Далее,

$$\frac{d}{dt}\left(P\left(x^{(n)} + t y^{(n)}\right)\varphi_{nl}\right)\Big|_{t=0} = \left(\frac{d}{dt} P\left(x^{(n)} + t y^{(n)}\right)\Big|_{t=0},\ \varphi_{nl}\right) =$$

$$= \left(P'_{x^{(n)}} y^{(n)},\ \varphi_{nl}\right).$$

Умножая это на $b_l^{(n)}$ и суммируя, получаем

$$\left(P'_{n,\, a^{(n)}b^{(n)}},\ b^{(n)}\right)_{R_N} = \left(P'_{x^{(n)}} y^{(n)},\ y^{(n)}\right). \tag{13.3.13}$$

По неравенству (13.1.5)

$$\left(P'_{n,\, a^{(n)}b^{(n)}},\ b^{(n)}\right)_{R_N} \geqslant \alpha^2 \left(P'_{u_0} y^{(n)},\ y^{(n)}\right) = \alpha^2 |y^{(n)}|^2. \tag{13.3.14}$$

Воспользуемся теперь неравенством (13.1.7):

$$\left(P'_{n,\,a^{(n)}}b^{(n)},\ b^{(n)}\right)_{R_N}\geqslant C\alpha^2\left\|b^{(n)}\right\|^2_{\overline{X}_N}. \tag{13.3.15}$$

Неравенство (13.3.14) является аналогом неравенства (13.1.4); отсюда видно, что в утверждениях п. 4° можно применительно к нашему случаю заменить γ на $\alpha\sqrt{C}$; можно отметить, что последние величины не зависят от n. Далее, по формулам (13.3.13), (13.3.14)

$$\left(P'_{n,\,a^{(n)}}b^{(n)},\ b^{(n)}\right)_{R_N}=\left(P'_{x^{(n)}}y^{(n)}y^{(n)}\right)\geqslant$$

$$\geqslant\alpha^2\left(P'_0y^{(n)},\ y^{(n)}\right)=\alpha^2\left(P'_{n,\,0}b^{(n)},\ b^{(n)}\right), \tag{13.3.16}$$

мы приняли здесь $u_0=0$, что, очевидно, не уменьшает общности. Неравенство (13.1.16) есть аналог неравенства (13.1.5) с той же постоянной α.

Из результатов п. 4° теперь следует, что операторы $P'_{n,\,a^{(n)}}$ удовлетворяют условию Липшица с постоянной C, которая не зависит от n и $f^{(n)}$.

Рассмотрим теперь уравнение Ритца (13.3.4), искаженное уравнение Ритца (13.3.5) и вспомогательное уравнение

$$P_ng^{(n)}=f^{(n)}+\delta^{(n)}. \tag{13.3.17}$$

Так как P_n^{-1} удовлетворяет условию Липшица с постоянной C, то

$$\left\|g^{(n)}-a^{(n)}\right\|_{\overline{X}_N}\leqslant C\left\|\delta^{(n)}\right\|_{\overline{Y}_N}\xrightarrow[\delta\to0]{}0. \tag{13.3.18}$$

Далее, Γ_n удовлетворяет условию Липшица (13.3.12), в котором следует положить $\varkappa=\psi(t,\delta)$. Как было доказано в п. 3°, нормы $\|a^{(n)}\|_{\overline{X}_N}$ ограничены, поэтому можно считать число t фиксированным; в силу сказанного в конце п. 4° $\|c^{(n)}-g^{(n)}\|_{\overline{X}_N}\xrightarrow[\delta\to0]{}0$; Вместе с неравенством (13.3.18) это дает

$$\left\|c^{(n)}-a^{(n)}\right\|_{\overline{X}_N}\xrightarrow[\delta\to0]{}0; \tag{13.3.19}$$

последнее соотношение означает, что вычислительный процесс (13.3.3) устойчив в последовательности $(\overline{X}_N,\overline{Y}_N)$.

Замечание 13.3.1. Если координатные функции образуют систему, сильно минимальную в энергетической норме оператора P'_0, то можно принять $\gamma(n)=1$. Отсюда следует, что в рассматриваемом случае процесс (13.3.3) устойчив в последовательности (R_N,R_N).

6°. **Теорема 13.3.2.** *Если координатная система почти ортонормирована в B_0, то процесс вычисления приближенного решения по Ритцу в условиях настоящего параграфа устойчив в последовательности $(B_0^{(n)},R_N)$.*

Доказательство очень просто. Если $z_*^{(n)}$ — искаженное приближенное решение по Ритцу, то

$$\left|z_*^{(n)} - x_*^{(n)}\right|_0^2 = \left(M_n\left(c^{(n)} - a^{(n)}\right),\ c^{(n)} - a^{(n)}\right)_{R_N} \leqslant \Lambda_0 \left\|c^{(n)} - a^{(n)}\right\|_{R_N}^2,$$

где Λ_0 — верхняя граница собственных чисел матрицы M_n. Почти ортонормированная система сильно минимальна, поэтому процесс вычисления коэффициентов Ритца устойчив в (R_N, R_N), и при достаточно малом δ будет $\|c^{(n)} - a^{(n)}\| \leqslant \varepsilon$. Но тогда $\left|z_*^{(n)} - x_*^{(n)}\right|_0 \leqslant \sqrt{\Lambda_0}\,\varepsilon$, и теорема доказана.

7°. Обратимся к МКЭ. Для определенности остановимся на варианте этого метода, развитом в [182]. В этом случае система координатных функций не сильно минимальна в энергетической норме. В частности, в случае первой краевой задачи, а также в случае кубической области и более произвольных краевых условий наименьшее собственное число матрицы МКЭ имеет порядок $O(h^m)$, где h — шаг сетки и m — размерность евклидова пространства задачи.

Справедлива теорема, которая является частным случаем теоремы 13.3.1.

Теорема 13.3.3. *Пусть достаточно гладкий функционал $F(x)$ определен на функциях пространства $B = \overset{\circ}{W}{}_p^{(s)}(\Omega)$, где Ω — ограниченная область в R_m. Пусть операторы $P = \operatorname{grad} F$ и P_u' определены на множестве, плотном в B, причем P_u' удовлетворяет неравенствам* (13.1.4), (13.1.5). *Пусть $h = 1/(2n)$ — шаг сетки и $N = N(n)$ — число координатных функций, входящих в выражение приближенного решения. Вычислительный процесс МКЭ устойчив в последовательности пар N-мерных гильбертовых пространств $(\bar{X}_N, \bar{Y}_N)$ с нормами*

$$\forall a \in R_N;\ \|a\|_{\bar{X}_N} = h^{m/2}\|a\|_{R_N},\ \|a\|_{\bar{Y}_N} = h^{-m/2}\|a\|_{R_N}.$$

Аналогичная теорема верна и для задачи с естественными краевыми условиями, если Ω — куб в R_m.

§ 4. ПОГРЕШНОСТИ ИСКАЖЕНИЯ И ОКРУГЛЕНИЯ НЕЛИНЕЙНОГО РЕКУРРЕНТНОГО ПРОЦЕССА

1°. Рассмотрим нелинейный рекуррентный процесс

$$A_n\left(x^{(0)}, x^{(1)}, \ldots, x^{(n)}\right) = f^{(n)},\quad n = 0, 1, 2, \ldots, \qquad (13.4.1)$$

где $x^{(k)} \in X_k$, $f^{(n)} \in Y_n$, а X_N, Y_N — банаховы пространства. Искаженный процесс запишем в виде

$$A_n\left(z^{(0)}, z^{(1)}, \ldots, z^{(n)}\right) + \Gamma_n\left(z^{(0)}, z^{(1)}, \ldots, z^{(n)}\right) = f^{(n)} + \delta^{(n)},$$
$$n = 0, 1, 2, \ldots \qquad (13.4.2)$$

Примем следующие допущения: 1) если элементы $x^{(0)}$, $x^{(1)}$, ...

..., $x^{(n-1)}$ фиксированы и их нормы не превосходят некоторого числа l, то n-е уравнение (13.4.1) имеет не более одного решения в шаре $\|x^{(n)}\| \leqslant l$; 2) существует такое число $T = \text{const}$, что если $\|f^{(n)}\| \leqslant T$ при любом n, то бесконечная система (13.4.1) разрешима и $\|x^{(n)}\| \leqslant l$.

Нетрудно построить свободный вычислительный процесс, эквивалентный рекуррентному процессу (13.4.1). Построим две новые последовательности пространств $\bar{X}_n$ и $\bar{Y}_n$, $n = 0, 1, 2, \ldots$ Элементами пространства $\bar{X}_n$ являются конечные последовательности $\bar{x}^{(n)} = (x^{(0)}, x^{(1)}, \ldots, x^{(n)})$; норму в $\bar{X}_n$ определим формулой

$$\|\bar{x}^{(n)}\|_{\bar{X}_N} = \max_{k \leqslant n} \|x^{(k)}\|_{X_k} \tag{13.4.3}$$

(допустимы и другие определения нормы); аналогично определяются и пространства $\bar{Y}_n$. Определим еще оператор $\bar{A}_n$, действующий из $\bar{X}_n$ в $\bar{Y}_n$, по формуле

$$\bar{A}_n(\bar{x}^{(n)}) = \{A_k(x^{(0)}, \ldots, x^{(k)})\}_{k=0}^{n}; \tag{13.4.4}$$

аналогично определим оператор $\bar{\Gamma}_n$. Уравнения (13.4.1) и (13.4.2) соответственно равносильны уравнениям свободного вычислительного процесса, точного и искаженного:

$$\forall n \geqslant 0, \quad \bar{A}_n(\bar{x}^{(n)}) = \bar{f}^{(n)}; \tag{13.4.5}$$

$$\forall n \geqslant 0, \quad \bar{A}_n(\bar{z}^{(n)}) + \bar{\Gamma}_n(\bar{z}^{(n)}) = \bar{f}^{(n)} + \bar{\delta}^{(n)}. \tag{13.4.6}$$

Мы дадим здесь определение устойчивости рекуррентного вычислительного процесса, отличное от аналогичного определения для линейного процесса, данного в главе 3. Рекуррентный процесс (13.4.1) назовем устойчивым по отношению к погрешностям искажения, если устойчив соответствующий свободный процесс (13.4.5).

2°. Условия устойчивости рекуррентного процесса теперь получаются из результатов § 2 данной главы. Эти условия таковы: 1) $\bar{A}_n(0) = 0$; 2) оператор $\bar{A}_n^{-1}$ определен в шаре $\|f^{(n)}\| \leqslant T$ и удовлетворяет условию Липшица

$$\|\bar{A}_n^{-1}(\bar{f}_1^{(n)}) - \bar{A}_n^{-1}(\bar{f}_2^{(n)})\| \leqslant C\|\bar{f}_1^{(n)} - \bar{f}_2^{(n)}\|$$

с постоянной, не зависящей от n; 3) $\|\bar{\delta}^{(n)}\| \leqslant \delta$; 4) если $\bar{x}$, $\bar{x}'$, $\bar{x}'' \in \bar{X}_n$ и нормы этих элементов не превосходят t, то

$$\|\bar{\Gamma}_n(\bar{x})\| \leqslant \varphi(t, \delta), \quad \|\bar{\Gamma}_n(\bar{x}') - \bar{\Gamma}_n(\bar{x}'')\| \leqslant \psi(t, \delta)\|\bar{x}' - \bar{x}''\|.$$

Переформулируем условия 1)—4) в терминах операторов A_n и Γ_n. Условие 1) означает, что

$$\forall n \geqslant 0, \quad A_n(0, 0, \ldots, 0) = 0. \tag{13.4.7}$$

Далее, $\bar{A}_n^{-1}$ есть оператор, определяющий решение конечной системы $A_k(x^{(0)}, x^{(1)}, \ldots, x^{(k)}) = f^{(k)}$, $k = 0, 1, \ldots, n$.

Это решение имеет вид

$$x^{(k)} = B_k(f^{(0)}, f^{(1)}, \ldots, f^{(k)}), \quad k = 0, 1, \ldots, n.$$

Отсюда следует, что $\overline{A}_n^{-1} = \{B_0, B_1, \ldots, B_n\}$, и условие 2) принимает вид

$$\forall n \geqslant 0, \quad \|B_n(f_1^{(0)}, f_1^{(1)}, \ldots, f_1^{(n)}) - B_n(f_2^{(0)}, f_2^{(1)}, \ldots, f_2^{(n)})\| \leqslant C \max_k \|f_1^{(k)} - f_2^{(k)}\|. \tag{13.4.8}$$

Условие 3) означает, что

$$\|\delta^{(n)}\| \leqslant \delta. \tag{13.4.9}$$

Наконец, условия 4) сводятся к следующим:

$$\|\Gamma_n(x^{(0)}, x^{(1)}, \ldots, x^{(n)})\| \leqslant \varphi(t, \delta); \tag{13.4.10а}$$

$$\|\Gamma_n(x_1^{(0)}, x_1^{(1)}, \ldots, x_1^{(n)}) - \Gamma_n(x_2^{(0)}, x_2^{(1)}, \ldots, x_2^{(n)})\| \leqslant \leqslant \psi(t, \delta) \max_{0 \leqslant k \leqslant n} \|x_1^{(k)} - x_2^{(k)}\|. \tag{13.4.10б}$$

На основании теоремы 13.2.1 можно утверждать, что нелинейный рекуррентный процесс (13.4.1) устойчив в смысле определения настоящего параграфа, если $\|f^{(n)}\| \leqslant T = \text{const}$ и выполнены условия (13.4.7)—(13.4.10).

Если процесс (13.4.1) устойчив, то при малых δ погрешность искажения ограничена независимо от n.

3°. Отметим случай, когда погрешность округления рекуррентного процесса также остается ограниченной. Обозначим $A_n + \Gamma_n = \tilde{A}_n$, $f^{(n)} + \delta^{(n)} = \tilde{f}^{(n)}$, и пусть точное решение искаженной системы (13.4.2) имеет вид

$$\forall n \geqslant 0, \quad z^{(n)} = \tilde{B}_n(z^{(0)}, z^{(1)}, \ldots, z^{(n-1)}, f^{(n)}). \tag{13.4.11}$$

Допустим, что операторы B_n удовлетворяют условию Липшица

$$\|\tilde{B}_n(z_1^{(0)}, z_1^{(1)}, \ldots, z_1^{(n-1)}, f^{(n)}) - \tilde{B}_n(z_2^{(0)}, z_2^{(1)}, \ldots, z_2^{(n-1)}, f^{(n)})\| \leqslant \leqslant q \max_{0 \leqslant k \leqslant n} \|z_1^{(k)} - z_2^{(k)}\|, \quad q = \text{const} < 1. \tag{13.4.12}$$

Пусть $v^{(0)}, v^{(1)}, \ldots, v^{(n-1)}$ — точно известные элементы и $\tilde{z}^{(n)}$ есть точное значение выражения $\tilde{B}_n(v^{(0)}, v^{(1)}, \ldots, v^{(n-1)}, f^{(n)})$. Допустим, что в результате погрешностей вычисления мы вместо $\tilde{z}^{(n)}$ получим элемент $v^{(n)} = \tilde{z}^{(n)} + \alpha^{(n)}$; будем считать, что относительная погрешность округления ограничена числом $\varepsilon > 0$, так что $\|\alpha^{(n)}\| \leqslant \varepsilon \|v^{(n)}\|$. По неравенству (13.4.12) имеем

$$\forall n \geqslant 0, \quad \|v^{(n)} - \tilde{z}^{(n)}\| \leqslant q \max_{0 \leqslant k \leqslant n-1} \|v^{(k)} - \tilde{z}^{(k)}\| + \varepsilon \|v^{(n)}\|.$$

Допуская, что $\|z^{(n)}\| \leqslant C = \text{const}$, получаем из последнего неравенства

$$\hat{\xi}_n := \|v^{(n)} - \tilde{z}^{(n)}\| \leqslant q' \max_{0 \leqslant k \leqslant n-1} \hat{\xi}_k + C\varepsilon, \quad q' = q/(1 - \varepsilon). \tag{13.4.13}$$

Пусть $\max\limits_{0\leqslant k\leqslant n-1}\hat{\xi}_k=\hat{\xi}_l$, $0\leqslant l\leqslant n-1$. Неравенство (13.4.13) дает при $n=l$ неравенство $\hat{\xi}_l\leqslant q'\hat{\xi}_l+C\varepsilon$, или, если $q'<1$, $\hat{\xi}_l\leqslant C\varepsilon/(1-q')$. Отсюда следует искомая оценка погрешности округления

$$\hat{\xi}_n\leqslant C\varepsilon/(1-q'). \tag{13.4.14}$$

§ 5. МЕТОД НЬЮТОНА — КАНТОРОВИЧА

1°. В настоящем параграфе мы исследуем погрешности искажения и округления двух методов Ньютона — Канторовича: модифицированного и основного. Эти методы приводят к рекуррентным процессам более специального вида, нежели рассмотренные в § 4. Устойчивость таких процессов можно исследовать проще и в более общих предложениях С рассмотрения процессов такого рода мы и начнем.

Пусть дан рекуррентный процесс вида

$$A_nx^{(n)}+Q_n\left(x^{(n-1)}\right)=f^{(n)}, \tag{13.5.1}$$

где A_n — линейные, а Q_n — нелинейные операторы. Для простоты допустим, что все эти операторы действуют в одном и том же банаховом пространстве X. Искаженный процесс запишем в виде

$$A_nz^{(n)}+\Gamma_{1n}z^{(n)}+Q_n\left(z^{(n-1)}\right)+\Gamma_{2n}\left(z^{(n-1)}\right)=f^{(n)}+\delta^{(n)}. \tag{13.5.2}$$

Примем следующие допущения: 1) операторы A_n^{-1} равномерно ограничены, так что $\|A_n^{-1}\|\leqslant a=\text{const}$; 2) при любом n оператор $A_n^{-1}Q_n$ есть оператор сжатия

$$\|A_n^{-1}Q_n(x)-A_n^{-1}Q_n(y)\|\leqslant q\|x-y\|,\quad q=\text{const}<1; \tag{13.5.3}$$

3) справедливо неравенство

$$\|\delta^{(n)}\|\leqslant\delta, \tag{13.5.4}$$

где δ — достаточно малая постоянная; 4) если $\|x\|$, $\|x'\|$, $\|x''\|\leqslant t$, то $(k=1, 2)$

$$\|A_n^{-1}\Gamma_{kn}\|\leqslant\varphi(t,\delta); \tag{13.5.5}$$

$$\|A_n^{-1}\Gamma_{kn}(x')-A_n^{-1}\Gamma_{kn}(x'')\|\leqslant\psi(t,\delta)\|x'-x''\|. \tag{13.5.6}$$

Функции φ и ψ такие же, как в предшествующих параграфах.

Имеем

$$x^{(n)}=A_n^{-1}f^{(n)}-A_n^{-1}Q_n\left(x^{(n-1)}\right),$$

$$z^{(n)}=A_n^{-1}f^{(n)}+A_n^{-1}\delta^{(n)}-A_n^{-1}Q_n\left(z^{(n-1)}\right)-A_n^{-1}\Gamma_{1n}\left(z^{(n)}\right)-$$
$$-A_n^{-1}\Gamma_{2n}\left(z^{(n-1)}\right).$$

Отсюда

$$z^{(n)} - x^{(n)} = A_n^{-1}\delta^{(n)} - \left[A_n^{-1}Q_n\left(z^{(n-1)}\right) - A_n^{-1}Q_n\left(x^{(n-1)}\right)\right] - \\ - A_n^{-1}\Gamma_{1n}\left(z^{(n-1)}\right) - A_n^{-1}\Gamma_{2n}\left(z^{(n-1)}\right).$$

Как и в предшествующих параграфах, доказывается, что если нормы $\|x^{(n)}\|$ ограничены в совокупности, $\|x^{(n)}\| \leqslant t$, то уравнение (13.5.2) имеет в шаре $\|z\| \leqslant 2t$ одно и только одно решение. Оценим разность $z^{(n)} - x^{(n)}$, где $x^{(n)}$ и $z^{(n)}$ — решения уравнений (13.5.1) и (13.5.2) соответственно. Имеем

$$\|z^{(n)} - z^{(n)}\| \leqslant q\|z^{(n-1)} - x^{(n-1)}\| + a\delta + 2\varphi(2t, \delta),$$

или, обозначая $\|z^{(n)} - x^{(n)}\| = \hat{\xi}_n$, $a\delta + 2\varphi(2t, \delta) = g$,

$$\hat{\xi}_n \leqslant q\,\hat{\xi}_{n-1} + g.$$

Ясно, что $\hat{\xi}_n \leqslant \tau_n$, где τ_n — решение системы $\tau_0 = \hat{\xi}_0$, $\tau_n = q\tau_{n-1} + g$, $n = 1, 2, \ldots$. Это решение есть

$$\tau_n = q^n\hat{\xi}_0 + g(1 - q^n)(1 - q)^{-1} < q^n\hat{\xi}_0 + g(1 - q)^{-1}$$

и, следовательно,

$$\|z^{(n)} - x^{(n)}\| \leqslant q^n\|z^{(0)} - x^{(0)}\| + (1 - q)^{-1}[a\delta + 2\varphi(2t, \delta)]. \qquad (13.5.7)$$

Неравенство (13.5.7) дает оценку погрешности искажения для процесса (13.5.1); из этого же неравенства вытекает и устойчивость упомянутого процесса при условиях (13.5.3)—(13.5.6).

Замечание 13.5.1. В статьях [49—51] рассмотрен частный случай процесса (13.5.1), который получается применением классического метода Ритца к уравнению $Ax + Qx = f$, где A — самосопряженный положительно определенный оператор в сепарабельном гильбертовом пространстве H, а Q — нелинейный непрерывный потенциальный оператор, имеющий положительно определенный дифференциал Фреше. Предполагается, что Q определен на всем пространстве H и $Q(0) = 0$. Наконец, $f \in H$. Принимается, что рассматриваемое уравнение имеет одно и только одно решение.

В качестве координатных выбираются собственные элементы оператора B, сходного с A [90]. Доказываются следующие утверждения. Невязка приближенного по Ритцу решения стремится к нулю; в метрике H погрешность аппроксимации имеет оценку (ω_{n+1} есть $(n+1)$-е собственное число оператора B)

$$\|x_* - x_*^{(n)}\|_H = O\left(\|\delta^{(n)}\|/\omega_{n+1}\right), \quad \delta^{(n)} = Ax_*^{(n)} + Q\left(x_*^{(n)}\right) - f,$$

а в энергетической норме оператора A — оценку

$$\left|x_* - x_*^{(n)}\right| = O\left(\|\delta^{(n)}\|/\sqrt{\omega_{n+1}}\right).$$

Доказана устойчивость процесса в том же предположении, что координатные элементы суть собственные элементы оператора, сходного с A.

2°. Переходим к ошибкам округления. Обозначим $A_n + \Gamma_{1n} = \tilde{A}_n$, $Q_n + \Gamma_{2n} = \tilde{Q}_n$ — оба эти оператора известны точно, и оператор $\tilde{A}_n^{-1}$ существует. Допустим, что оператор $\tilde{A}_n^{-1}\tilde{Q}_n$ есть оператор сжатия в некотором шаре, содержащем z^n, с коэффициентом сжатия $\tilde{q}$, не зависящим от n. Уравнению (13 5.2) можно придать вид

$$z^{(n)} = a_n(z^{(n-1)}) + b^{(n)},$$

$$a_n = -\tilde{A}_n^{-1}\tilde{Q}_n,\ b^{(n)} = \tilde{A}_n^{-1}(f^{(n)} + \delta^{(n)}). \qquad (13.5.8)$$

Уравнение (13.5.8) разрешимо итерациями, которые можно записать так:

$$y^{(m)} = a_n(y^{(m-1)}) + b^{(n)},\quad m = 0,\ 1,\ 2,\ \ldots, \qquad (13.5.9)$$

где $y^{(m)}$ есть m-е к решению $z^{(n)}$. Допустим, что, вычисляя правую часть формулы (13.5.9), мы совершаем ошибку округления, относительная величина которой не превосходит числа $\varepsilon > 0$. Пусть, далее, известное нам округленное значение $(m-1)$-го приближения $y^{(m-1)}$ есть $v^{(m-1)}$. Тогда в результате вычислений мы вместо элемента $y^{(m)}$ получим некоторый элемент $v^{(m)}$, удовлетворяющий соотношению

$$v^{(m)} = a_n(v^{(m-1)}) + b^{(n)} + \alpha^{(m)}. \qquad (13.5.10)$$

Вычитая отсюда равенство (13.5.9), получаем

$$v^{(m)} - y^{(m)} = a_n(v^{(m-1)}) - a_n(y^{(m-1)}) + \alpha^{(m)}$$

и

$$\gamma_m := \|v^{(m)} - y^{(m)}\| \leqslant \tilde{q}\gamma_{m-1} + \varepsilon\gamma_m + \varepsilon\|y^{(m-1)}\|.$$

Допустим, что нормы $\|b^{(n)}\|$ ограничены в совокупности: $\|b^{(n)}\| \leqslant C$. Тогда нормы $\|y^{(m)}\|$ также ограничены; пусть $\|y^{(m)}\| \leqslant C_0$. В этом случае

$$\gamma_m \leqslant \tilde{q}\gamma_{m-1}/(1-\varepsilon) + C_0\varepsilon/(1-\varepsilon). \qquad (13.5.11)$$

Обозначим $\tilde{q}/(1-\varepsilon) = r$, $C_0\varepsilon/(1-\varepsilon) = \lambda$, так что $\gamma_m \leqslant r\gamma_{m-1} + \lambda$. Заметим, что $v^{(0)} = y^{(0)}$; так как $y^{(0)}$ — начальное значение — не вычисляется, а задается, то $\gamma_0 = 0$. Теперь из (13.5.11) следует

$$\gamma_m \leqslant \frac{1-r^m}{1-r}\,\frac{C_0\varepsilon}{1-\varepsilon}. \qquad (13.5.12)$$

Если ε столь мало, что $r < 1$, то оценка (13.5.12) дает далее

$$\gamma_m < \frac{C_0\varepsilon}{(1-r)(1-\varepsilon)}; \qquad (13.5.13)$$

ошибка округления ограничена независимо от m и n и равномерно стремится по этим индексам к нулю, если $\varepsilon \to 0$. Если ε

не очень мало, так что $r > 1$, то оценка (13.5.12) стремится к бесконечности вместе с m.

3°. Модифицированный метод Ньютона — Канторовича применяется к решению нелинейных задач вида $P(x) = 0$. Достаточно полное описание этого метода, условия и скорость его сходимости даны в [64]; здесь мы отметим только, что самый метод состоит в применении рекуррентного процесса

$$x^{(n)} = x^{(n-1)} - A^{-1}P(x^{(n-1)}), \quad A = P'_{x^{(0)}}, \qquad (13.5.14)$$

$x^{(0)}$ — начальное приближение к искомому решению. Нам будет удобнее записать этот процесс в виде

$$Ax^{(n)} + Q(x^{(n-1)}) = 0, \quad Q = A - P. \qquad (13.5.15)$$

Заметим, что оператор A не задан, а вычисляется, естественно, с некоторой погрешностью.

Проверим выполнение условий п. 1°. Условие $\|A_n^{-1}\| \leqslant \text{const}$ вытекает из того, что $A_n = A = P'_{x^{(0)}}$ не зависит от n и что ограниченность оператора $\left[P'_{x^{(0)}}\right]^{-1}$ принадлежит к числу условий сходимости модифицированного метода Ньютона — Канторовича. Условие (13.5.3) будет выполнено, если P имеет непрерывную первую производную; в теореме о сходимости метода требуется существование второй производной. Мы предполагаем также, что выполнены условия (13.5.4)—(13.5.6), характеризующие искажение процесса. По доказанному в пп. 1—2° вычислительный процесс модифицированного метода Ньютона — Канторовича устойчив; если уравнения этого процесса решать итерациями с достаточно малыми относительными погрешностями, то погрешность округления остается ограниченной.

4°. В [64] доказано следующее утверждение. Пусть $x^{(0)}$ — начальное приближение, а $x^{(1)}, x^{(2)}, \ldots, x^{(n)}, \ldots$ — приближения, построенные по основному методу Ньютона — Канторовича. Если в точке $x^{(0)}$ выполнены условия сходимости модифицированного метода, то они выполнены в точках $x^{(1)}, x^{(2)}, \ldots, x^{(n)}, \ldots$ Отсюда нетрудно установить, что для вычислительного процесса, связанного с основным методом Ньютона — Канторовича, верны заключения, сделанные в конце п. 3° для модифицированного метода.

Глава 14

НЕКОТОРЫЕ ОДНОСТОРОННИЕ ВАРИАЦИОННЫЕ ЗАДАЧИ

В обширной литературе по задачам, связанным с вариационными неравенствами, видное место занимают так называемые односторонние вариационные задачи. Их обычная поста-

новка такова: на рефлексивном банаховом пространстве B заданы полунепрерывный снизу возрастающий функционал $J(x)$ и непустое замкнутое выпуклое множество M. Ставится задача: найти такой элемент $x_* \in M$, чтобы

$$\forall x \in M,\ J(x_*) \leqslant J(x).$$

Условие $x \in M$ называется *односторонним ограничением*, а M — *множеством ограничений.*

В [76] доказано, что такая задача имеет решение; оно единственно, если функционал J — существенно выпуклый. В [55] рассмотрен ряд задач физики и механики, которые приводятся к вариационным неравенствам и, в частности, к односторонним вариационным задачам. В книге [38] излагается ряд методов приближенного решения таких задач и даны приложения этих методов к большому числу более конкретных задач. Для многих приближенных методов даны доказательства сходимости. Однако, по мнению авторов этой книги, получение оценок погрешности приближенных методов — задача более деликатная.

В настоящей главе будут изложены результаты работ автора [99—103] по оценке погрешностей аппроксимации и искажения для односторонних вариационных задач; в основном речь пойдет о некотором классе таких задач, подробное описание которых дано ниже, в § 1. Будут изложены также (в § 4), но без доказательств, некоторые результаты работ [46] и [136]. В § 7 данной главы будет дан обзор некоторых работ последних лет по оценкам погрешности аппроксимации для вариационных неравенств; этот обзор написан В. Б. Тюхтиным.

§ 1. ПОСТАНОВКА ЗАДАЧИ И ЕЕ ПРИБЛИЖЕННОЕ РЕШЕНИЕ

1°. Пусть требуется решить задачу о минимуме функционала

$$F(x) - 2lx \tag{14.1.1}$$

при одностороннем ограничении

$$\|\!| x - g |\!\| \leqslant \alpha, \quad \alpha = \text{const}. \tag{14.1.2}$$

Здесь F — квадратичная форма, определенная и положительная на некотором линейном множестве. Прежде чем переходить к описанию остальных данных задачи, отметим, что на упомянутом множестве можно построить энергетическое пространство H со скалярным произведением $[u, v] = F(u, v)$ и нормой $|u| = \sqrt{F(u)}$; через $F(u, v)$ обозначена билинейная форма, соответствующая квадратичной форме $F(u)$. Примем, что l — линейный функционал, ограниченный в энергетической норме, и что $\|\!| \cdot |\!\|$ — полунорма, определенная (но не обязательно конечная) на всем пространстве H

Введем еще гильбертово пространство $\hat{H}$, по отношению к которому H является подпространством (может быть, несобственным). Скалярное произведение и норму в $\hat{H}$ будем обозначать теми же символами, что и в H. Будем считать, что g — заданный элемент пространства $\hat{H}$ и что полунорма $|||\cdot|||$ определена (хотя, разумеется, она необязательно всюду конечна) на всех элементах этого пространства. Необходимо потребовать, чтобы множество (14.1.2) не было пустым — для этого необходимо и достаточно, чтобы $|||g||| < \infty$ и чтобы расстояние $\rho(g, H)$, вычисленное в метрике $|||\cdot|||$, не превосходило числа α.

2°. Последний вопрос рассмотрим подробнее. Вычислим величину $\varkappa = \rho(g, H)$. Если $\varkappa > \alpha$, то, как только что было отмечено, множество

$$M = \{x \in H : |||x - g||| \leqslant \alpha\} \tag{14.1.3}$$

пусто, и задача (14.1.1), (14.1.2) не имеет решения. Пусть $\varkappa < \alpha$, тогда существует такой элемент $x' \in H$, что $|||x' - g||| < \alpha$, и множество M не пусто. Положим $g - x' = h$, $x - x' = y$, тогда множество ограничений (14.1.2) определяется соотношением $|||y - h||| \leqslant \alpha$, в котором $|||h||| < \alpha$.

Пусть теперь $\varkappa = \alpha$. Может случиться, что $\inf\limits_{x \in H} |||x - g|||$ не достигается, тогда $\forall x \in H$, $|||x - g||| > \alpha$, и множество M пусто. Если же инфимум достигается, то множество M не пусто, и задача (14.1.1), (14.1.2) имеет решение (см. ниже, п. 3°). Однако это решение неустойчиво относительно малых изменений числа α. Действительно, пусть α заменено на $\alpha' < \alpha$, тогда $\varkappa > \alpha'$, новое множество $M' = \{x \in H : |||x - g||| \leqslant \alpha'\}$ пусто и возмущенная задача не имеет решения.

Ниже, рассматривая множество вида (14.1.3), мы всегда будем считать, что $\beta := |||g||| < \alpha$.

3°. По теореме Ф. Риса существует единственный элемент $x_0 \in H$, удовлетворяющий тождеству $\forall x \in H$, $lx = [x, x_0]$; этот элемент, решающий задачу о минимуме функционала (14.1.1) при отсутствии ограничений, будем считать известным.

Задача (14.1.1), (14.1.2) равносильна следующей:

$$|x - x_0| = \min, \quad |||x - g||| \leqslant \alpha. \tag{14.1.4}$$

Множество M ограничений, определяемое неравенством (14.1.2) (или (14.1.3)), — выпуклое; будем считать полунорму $|||\cdot|||$ такой, что это множество замкнуто в метрике H. В силу результатов, изложенных в [76, глава 1], задача (14.1.4) имеет одно и только одно решение, которое мы обозначим через x_*. Если $|||x_0 - g||| < \alpha$, то очевидно, что $x_* = x_0$; если же $|||x_0 - g||| \geqslant \alpha$, то x_* решает задачу

$$|x_* - x_0| = \min, \quad |||x_* - g||| = \alpha. \tag{14.1.5}$$

Докажем это. Рассмотрим прямолинейный отрезок Λ, соединяющий точки x_0 и x_*: $\Lambda = \{x \in H: x = \lambda x_* + (1 - \lambda) x_0;\ 0 \leqslant$

$\leqslant \lambda \leqslant 1\}$. Установим на Λ направление от x_* к x_0 и в соответствии с этим назовем x_* левым, а x_0 — правым концом отрезка Λ. Левый конец отрезка Λ принадлежит множеству M, правый — множеству $H \setminus M$. Введем на Λ расстояние, порожденное нормой $|\cdot|$, и разделим этот отрезок пополам. Один из двух отрезков деления таков, что его левый конец принадлежит M, а правый — $H \setminus M$. Этот отрезок разделим пополам, и т. д. В результате получим последовательность вложенных отрезков с длинами, стремящимися к нулю; эти отрезки таковы, что их левые концы принадлежат M, а правые — $H \setminus M$. Отрезки этой последовательности имеют ровно одну общую точку; обозначим ее через y. Она — предельная для множества M; так как M замкнуто, то $y \in M$ и, следовательно, $\|\!| y - g |\!\| \leqslant \leqslant \alpha$. Одновременно точка y — предельная для $H \setminus M$: если $\{y_n\}$ — последовательность правых концов рассматриваемых вложенных отрезков, то $|y - y_n| \xrightarrow[n \to \infty]{} 0$. На одномерном отрезке полунорма $\|\!|\cdot|\!\|$ непрерывна, поэтому $\|\!| y - y_n |\!\| \to 0$, но $\|\!| g - y_n |\!\| > \alpha$ и потому $\|\!| g - y |\!\| \geqslant \alpha$. Окончательно $\|\!| g - y |\!\| = \alpha$.

Докажем теперь, что $x_* = y$. Допустим противное. Тогда на отрезке Λ точка y — промежуточная между x_* и x_0, поэтому $|x_0 - y| \leqslant |x_0 - x_*| = \min\limits_{x \in M} |x_0 - x|$, что невозможно. Отсюда следует, что $x_* = y$, $\|\!| x_* - g |\!\| = \alpha$, и наше утверждение доказано.

Множество $\{x \in H : \|\!| x - g |\!\| = \alpha\}$ назовем *границей множества* M и обозначим через ∂M. Можно, таким образом, считать, что при $\|\!| x_0 - g |\!\| > \alpha$ элемент x_* решает задачу

$$|x_0 - x| = \min, \quad x \in \partial M. \tag{14.1.6}$$

Ниже всюду принимается, что $\|\!| x_0 - g |\!\| > \alpha$.

4°. Задачу (14.1.6) будем решать приближенно, сводя ее к конечномерной задаче. Выберем натуральное число n и N-мерное ($N = N(n)$) подпространство H_n пространства H. Потребуем, чтобы на элементах подпространств H_n полунорма $\|\!|\cdot|\!\|$ была конечной. Поставим одностороннюю вариационную задачу

$$|x_0 - x^{(n)}| = \min, \quad x^{(n)} \in H_n \cap M. \tag{14.1.7}$$

Эту задачу можно несколько упростить. Пусть x_{0n} — ортогональная проекция элемента x_0 на H_n, тогда

$$|x_0 - x^{(n)}|^2 = |x_{0n} - x^{(n)}|^2 + |x_0 - x_{0n}^2|^2.$$

Второе слагаемое справа — постоянное, поэтому задача (14.1.7) равносильна следующей:

$$|x_{0n} - x^{(n)}| = \min, \quad x^{(n)} \in H_n \cap M. \tag{14.1.8}$$

Преимуществом этой новой формы нашей задачи является то, что как данный элемент x_{0n}, так и искомый элемент $x^{(n)}$ принадлежат одному и тому же подпространству H_n.

Пересечение подпространства H_n и замкнутого выпуклого множества M есть замкнутое выпуклое множество. Потребуем еще, чтобы оно было непустым; тогда задача (14.1.8) будет иметь одно и только одно решение, которое будем рассматривать как приближенное решение задачи (14.1.6). Допустим, что последовательность подпространств H_n, соответствующих всевозможным натуральным n, полна в H. Тогда, если $x_0 \notin M$, то при n достаточно большом будет $x_{0n} \notin M$ и по доказанному в п. 3° $x_*^{(n)} \in \partial M$.

5°. Известны многие способы решения задач типа (14.1.8), а также более общих задач, дающие приближенные решения тех или иных вариационных неравенств. Мы здесь ограничимся ссылкой на книгу [38, глава II]. Укажем еще один прием, который может оказаться небесполезным [99].

Пусть $\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nN}$ — базис в H_n. Допустим, что существует такая монотонная функция $\omega : R_1^+ \to R_1^+$, что $\omega(\|x^{(n)} - g\|)$, где

$$x^{(n)} = \sum_{k=1}^{N} \alpha_k \varphi_{nk}$$

есть достаточно гладкая функция коэффициентов α_k. Положим

$$x_*^{(n)} = \sum_{k=1}^{n} a_k^{(n)} \varphi_{nk},$$

обозначим через $a^{(n)}$ вектор $(a_1^{(n)}, a_2^{(n)}, \ldots, a_N^{(n)})$, через M_n — матрицу чисел $[\varphi_{nk}, \varphi_{nj}]$, $j, k = 1, 2, \ldots, N$. Положим еще $c_j^{(n)} = [u_{0n}, \varphi_{jn}]$, $j = 1, 2, \ldots, N$, и $\Phi_n(a^{(n)}) = \omega(\|x_*^{(n)} - g\|)$. По способу множителей Лагранжа задача (14.1.8) сводится к системе уравнений относительно коэффициентов $a_j^{(n)}$.

$$\frac{\partial}{\partial a_j^{(n)}}(M_n a^{(n)}, a^{(n)}) - 2\lambda \frac{\partial \Phi_n(a^{(n)})}{\partial a_j^{(n)}} = 2c_j^{(n)}, \quad 1 \leqslant j \leqslant N,$$

или, короче,

$$M_n a^{(n)} - \lambda \operatorname{grad} \Phi_n(a^{(n)}) = c^{(n)}. \tag{14.1.9}$$

Будем рассматривать λ как независимую переменную и $a_j^{(n)}$ как функции от λ. Дифференцируя (14.1.9) по λ, получим систему обыкновенных дифференциальных уравнений

$$(M_n - \lambda \Psi_n) \frac{da^{(n)}}{d\lambda} - \operatorname{grad} \Phi_n(a^{(n)}) = 0, \tag{14.1.10}$$

которую надо решать при начальном условии

$$a^{(n)}\big|_{\lambda=0} = M_n^{-1} c^{(n)}; \tag{14.1.11}$$

в уравнении (14.1.10) Ψ_n означает матрицу вторых производных функции Φ_n.

Задачу (14.1.1), (14.1.2) будем решать каким-нибудь разностным методом. Пусть h — шаг соответствующей сетки. Положим $\lambda_k = \pm kh$, $k = 1, 2, \ldots$ Определим вектор $a^{(n)}(k)$, соответствующий некоторому k, и вычислим величину $\left\| \sum_{j=1}^{N} a_j^{(n)}(k) \varphi_{jn} - g \right\|$. При малых k она близка к $\|u_0 - g\|$ и, следовательно, больше, чем α. Процесс решения задачи (14.1.10), (14.1.11) следует приостановить при том значении k, положительном или отрицательном, при котором в пределах принятой точности будет выполняться равенство

$$\left\| \sum_{j=1}^{N} a_j^{(n)}(k) \varphi_{jn} - g \right\| = \alpha. \tag{14.1.12}$$

Вычисление значений $a_j^{(n)}$ упрощается, если полунорма $\|\cdot\|$ определяется равенством $\|u\|^2 = b(u, u)$, где $b(u, u)$ — неотрицательная квадратичная форма, которой соответствует билинейная форма $b(u, v)$, определенная на всем пространстве H. Если в этом случае положить $\omega(t) = t^2$, то система (14.1.9) примет вид

$$(M_n - \lambda F_n) a^{(n)} = c^{(n)} - \lambda f^{(n)}; \tag{14.1.13}$$

здесь F_n — матрица чисел $b(\varphi_{jn}, \varphi_{kn})$, $j, k = 1, 2, \ldots, N$, а $f^{(n)}$ — вектор с составляющими $f_k^{(n)} = b(g, \varphi_{kn})$. Матрица M_n — положительно определенная и потому невырожденная; систему (14.1.13) можно решить, обратив матрицу $M_n - \lambda F_n$ по методу Фаддеева ([139]; см. также [14]). Определив коэффициенты $a_j^{(n)}$, мы затем найдем λ из уравнения $b(x^{(n)} - g, x^{(n)} - g) = \alpha^2$.

§ 2. ПОГРЕШНОСТЬ АППРОКСИМАЦИИ

1° Пусть B — банахово пространство с нормой $\|\cdot\|$, обладающее следующими свойствами:

1) B ограниченно вкладывается в H; при этом

$$\forall x \in B, \quad |x| \leqslant \|x\|, \quad \|x\| \leqslant \|x\|; \tag{14.2.1}$$

2) элементы x_0 и x_* содержатся в B;

3) при любом n подпространство $H_n \subset B$, а последовательность $\{H_n\}$ полна в B.

Если величина $\|x_0\|$ конечна, что мы ниже всегда предполагаем, то условиям 1), 2) можно удовлетворить, положив, например,

$$\|x\|^q = |x|^q + \|x\|^q, \quad 1 \leqslant q \leqslant \infty. \tag{14.2.2}$$

Условие 3) будет выполнено, если на элементах подпространств H_n полунорма $\|\cdot\|$ конечна и последовательность этих подпространств полна в той же полунорме. В качестве примера можно

указать такой случай. Пусть $\tilde{H} \sim W_2^{(k)}(\Omega)$, где k — натуральное число и Ω — конечная область евклидова пространства, удовлетворяющая условию конуса, и пусть для некоторых $s \geqslant k$ и $p \geqslant 1$ справедливо неравенство $\|\cdot\| \leqslant C\|\cdot\|_{W_p^{(s)}(\Omega)}$, $C = \text{const}$.

В качестве H_n возьмем подпространства, натянутые на координатные функции МКЭ (см. главу 9), которые получены преобразованием независимых переменных из соответствующих исходных функций; эти функции можно выбрать так, чтобы последовательность $\{H_n\}$ была полной в $W_p^{(s)}(\Omega)$. Пусть еще p таково, что $W_p^{(s)}(\Omega)$ ограниченно вкладывается в $W_2^{(k)}(\Omega)$. Тогда $H_n \subset B$ при любом n и последовательность $\{H_n\}$ полна в B.

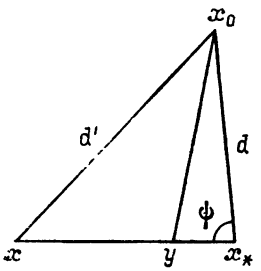

Рис. 1.

Ниже будем пользоваться равенством (14.2.2) при $q = 1$.

2°. **Лемма 14.2.1.** *Пусть M — выпуклое и замкнутое в норме $|\cdot|$ непустое множество. Далее, пусть x_* — решение односторонней вариационной задачи*

$$|x_0 - x_*| = \min, \quad x_0 \in H \setminus M, \quad x_* \in M, \tag{14.2.3}$$

а x — произвольный элемент множества M. Тогда

$$|x_* - x|^2 \leqslant d'^2 - d^2, \tag{14.2.4}$$

где обозначено

$$d = |x_0 - x_*|, \quad d' = |x_0 - x|. \tag{14.2.5}$$

В пространстве H проведем двумерную плоскость P через точки x_0, x_*, x. Гильбертова норма пространства H индуцирует на P евклидову метрику. Построим треугольник с вершинами x_0, x_*, x и обозначим через ψ угол при вершине x_*. Докажем, что $\psi \geqslant \pi/2$. Допустим противное. Через точку x_0 проведем в P прямую, наклоненную к прямой $\overline{x_0 x_*}$ под углом, меньшим, чем $\pi/2 - \psi$, и пересекающую отрезок $\overline{x_* x}$; пусть y — точка пересечения (рис. 1). В треугольнике с вершинами x_0, x_*, y угол при вершине y тупой, поэтому $|x_0 - y| < d$. Множество M выпуклое, а x, $x_* \in M$. Отсюда следует, что $y \in M$. Теперь неравенство $|x_0 - y| < d$ означает, что x_* не есть решение задачи (14.2.3), вопреки предположению. Неравенство $\psi \geqslant \pi/2$ доказано. Далее, по теореме косинусов $d'^2 = |x_* - x|^2 + d^2 - 2d|x_* - x|\cos\psi$, но $\cos\psi \leqslant 0$, и $d'^2 \geqslant |x_* - x|^2 + d^2$. Лемма доказана.

3°. **Лемма 14.2.2.** *Пусть выполнены условия* 1) — 3) *п.* 1° *настоящего параграфа, а множество ограничений M определено формулой* (14.1.2), *причем $\beta = \|g\| < \alpha$. Пусть x — некоторая точка на ∂M, так что $\|x - g\| = \alpha$. Допустим, что существует*

точка $x^{(n)} \in H_n$, *удовлетворяющая неравенству* $\|x - x^{(n)}\| \leqslant \gamma_n$, *где* $\gamma_n < (\alpha - \beta)/2$, *и* $\|\cdot\| = |\cdot| + |||\cdot|||$. *Тогда существует элемент* $y^{(n)} \subset M \cap H_n$, *такой, что* $\|x - y^{(n)}\| \leqslant C_x \gamma_n$, *где* C_x *может зависеть от* x, *но не зависит от* n.

Положим $y^{(n)} = ax^{(n)}$, $0 < a < 1$; очевидно, что $y^{(n)} \in H_n$. Выясним, при каких a справедливо включение $y^{(n)} \in M$. Имеем

$$|||y^{(n)} - g||| \leqslant a\,|||x^{(n)} - g|| + (1 - a)\,\beta,$$

$$|||x^{(n)} - g||| \leqslant |||x^{(n)} - x||| + |||x - g||| \leqslant \gamma_n + \alpha.$$

Отсюда $|||y^{(n)} - g||| \leqslant a(\gamma_n + \alpha) + (1 - a)\,\beta$. Если $a(\gamma_n + \alpha) + (1 - a)\,\beta = \alpha$, или

$$a = (\alpha - \beta)/(\alpha - \beta + \gamma_n), \tag{14.2.6}$$

то $y^{(n)} \in M$. При этом

$$\|x - y^{(n)}\| \leqslant \|x - x^{(n)}\| + (1 - a)\,\|x^{(n)}\| \leqslant \gamma_n + \gamma_n \|x^{(n)}\|/(\alpha - \beta + \gamma_n).$$

Далее, $\|x^{(n)}\| \leqslant \|x\| + \|x - x^{(n)}\| \leqslant \|x\| + \gamma_n$ и потому

$$\|x - y^{(n)}\| \leqslant \gamma_n [1 + (\|x\| + \gamma_n)/(\alpha - \beta + \gamma_n)]. \tag{14.2.7}$$

Отсюда вытекает более простая оценка

$$\|x - y^{(n)}\| \leqslant C_x \gamma_n, \quad C_x = 3/2 + \|x\|/(\alpha - \beta). \tag{14.2.8}$$

Лемма доказана. Аналогично можно получить оценку

$$|x - y^{(n)}| \leqslant C'_x \gamma_n, \quad |||x - y^{(n)}||| \leqslant C''_x \gamma_n, \tag{14.2.9}$$

где

$$C'_x = 3/2 + |x|/(\alpha - \beta), \quad C''_x = 3/2 + |||x|||/(\alpha - \beta). \tag{14.2.10}$$

Теорема 14.2.1. *Если пространство* B *удовлетворяет условиям* 1)—3) п. 1°, *а* x_0, x_*, $x_*^{(n)}$ *имеют тот же смысл, что и выше, то*

$$|x_* - x_*^{(n)}| = O(\sqrt{e_n(x_*)}), \tag{14.2.11}$$

где $e_n(x)$ *есть наилучшее приближение элемента* $x \in B$ *элементами подпространства* H_n *в норме* $\|\cdot\|$.

Заметим, что по предположению последовательность $\{H_n\}$ полна в B, поэтому $e_n(x) \xrightarrow[n\to\infty]{} 0$. В частности, $e_n(x_*) \xrightarrow[n\to\infty]{} 0$.

Как было доказано в п. 4° § 1, $x_*^{(n)} \in \partial M$, если n достаточно велико. По определению наилучшего приближения при любом заданном $\varepsilon > 0$ существует такой элемент $x^{(n)} \in H_n$, что $\|x_* - x^{(n)}\| \leqslant (1 + \varepsilon)\,e_n(x_*)$. По лемме 14.2.2 существует элемент $y^{(n)} \in M_n := M \cap H_n$, такой, что $|x_* - y^{(n)}| \leqslant C_* e_n(x_*)$, $C_* = (1 + \varepsilon)\,C'_{x_*}$. По лемме 14.2.1 $|x_* - x_*^{(n)}|^2 \leqslant d_n^2 - d^2$, $d_n = |x_0 - x_*^{(n)}|$. Но $x_*^{(n)}$ есть решение задачи (14.1.7), поэтому $d_n \leqslant d'_n = |x_0 - y^{(n)}|$ и

$$|x_* - x_*^{(n)}|^2 \leqslant d_n'^2 - d^2 = (d'_n + d)(d'_n - d).$$

Кроме того,

$$d_n' = |x_0 - y^{(n)}| \leqslant |x_0 - x_*| + |x_* - y^{(n)}| \leqslant d + C_* e_n(x_*).$$

Отсюда

$$|x_* - x_*^{(n)}|^2 \leqslant [2d + C_* e_n(x_*)]\, C_* e_n(x_*). \qquad (14.2.12)$$

Теорема доказана.

§ 3. СЛУЧАИ, КОГДА ОЦЕНКА ПОГРЕШНОСТИ УЛУЧШАЕТСЯ

Существуют случаи, когда оценка (14.2.11) может быть существенно улучшена. В данном параграфе приведены два таких случая, описанных в [100, 101].

1°. На рис. 1 заменим точку x на $x_*^{(n)}$. По теореме косинусов

$$|x_* - x_*^{(n)}|^2 = (|x_0 - x_*^{(n)}| - d)^2 + 4d\,|x_0 - x_*^{(n)}| \sin^2(\theta_n/2);$$

через θ_n здесь обозначен угол при вершине x_0 в треугольнике $(x_0, x_*, x_*^{(n)})$. Величина $|x_0 - x_*^{(n)}| \xrightarrow[n\to\infty]{} d$, поэтому при n достаточно большом $|x_0 - x_*^{(n)}| \leqslant 2d$. Обозначим еще через $x^{(n)}$ элемент подпространства H_n, на котором достигается равенство $|x_* - x^{(n)}| = e_n(x_*)$. Тогда

$$d \leqslant |x_0 - x_*^{(n)}| \leqslant |x_0 - x^{(n)}| \leqslant |x_0 - x_*| + |x_* - x^{(n)}| = d + e_n(x_*).$$

Отсюда $0 \leqslant |x_0 - x_*^{(n)}| - d \leqslant e_n(x_*)$ и, следовательно,

$$|x_* - x_*^{(n)}|^2 \leqslant e_n^2(x_*) + 2d^2\theta_n^2; \qquad (14.2.1)$$

дело сводится к оценке угла θ_n. Допустим, что многообразие ∂M достаточно гладкое, тогда вектор $x_0 - x_*$ направлен по нормали к ∂M. Приведем доказательство этого простого утверждения. Выберем на ∂M произвольную точку y и проведем через точки x_0, x_*, y двумерную плоскость Q (рис. 2), которая пересечет ∂M по достаточно гладкой кривой L. Введем в Q декартову систему координат (ξ_1, ξ_2); за ось ξ_1 примем прямую, проходящую через точки x_0 и x_* и направленную от x_0 к x_*; эту же ось примем за ось полярных координат, центр которых поместим в x_0. Уравнение касательной к кривой L в точке x_* можно написать в виде (ξ — абсцисса переменной точки x)

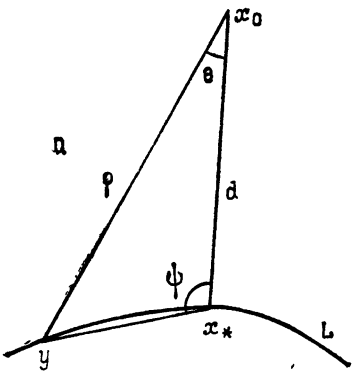

Рис. 2.

$$(\xi_1 - \xi_{*1})/d\xi_{*1} = \xi_2/d\xi_2.$$

Пусть $\rho = \rho(\theta)$ есть уравнение кривой L в полярных координатах вблизи точки x_*, тогда $d\xi_{*1} = d(\rho\cos\theta) = \cos\theta d\rho - \rho\sin\theta d\theta$. Но $d\rho = 0$ в точке x_*, потому что в этой точке ρ достигает минимума; кроме того, в этой же точке $\sin\theta = 0$. Таким образом, $d\xi_{*1} = 0$, уравнение касательной к L в точке x_* есть $\xi_1 = \xi_{*1}$ и прямая x_*x_0 направлена по нормали к ∂M.

Как было только что отмечено, в введенных нами полярных координатах можно записать уравнение кривой L вблизи точки x_* в виде $\rho = \rho(\theta)$, где $\rho(\theta)$ — достаточно гладкая функция от θ, причем $\rho(0) = d$, $\rho'(0) = 0$. Наложим на многообразие ∂M следующее *требование* T: существуют такие положительные числа k_0, k_1, $\bar{k}$, что

$$k_0 \leqslant 2^{-1}\rho''(0) \leqslant k_1, \quad 6^{-1}|\rho'''(\theta)| \leqslant \bar{k}. \tag{14.3.2}$$

Эти числа могут зависеть от x_0, но не зависят от двумерного сечения, проходящего через точки x_0 и x_*; второе соотношение (14.3.2) должно выполняться при достаточно малых θ.

Обозначим $\rho(\theta) - d = l(\theta)$. Разлагая $\rho(\theta)$ по формуле Тейлора, находим $l(\theta) = k\theta^2 + \rho'''(\tau)\theta^3/6$; $k = \rho''(0)/2$. Отсюда $\theta^2 = l(\theta)/k - \rho'''(\tau)\theta^3/6k$ и, следовательно,

$$\frac{1}{k_1}l(\theta) - \frac{\bar{k}}{k_0}|\theta|^3 \leqslant \theta^2 \leqslant \frac{1}{k_0}l(\theta) + \frac{\bar{k}}{k_0}|\theta|^3.$$

Пусть $|\theta| \leqslant k_0/2\bar{k}$. Из последнего неравенства находим

$$\frac{2}{3k_1}l(\theta) \leqslant \theta^2 \leqslant \frac{2}{k_0}l(\theta). \tag{14.3.3}$$

По теореме косинусов

$$|x_* - y|^2 = d^2 + \rho^2(0) - 2d\rho(\theta)\cos\theta = l^2(\theta) + 4d(d + l(\theta))\sin^2(\theta/2);$$

отсюда и из неравенства (14.3.3) вытекает неравенство

$$\beta_1\theta \leqslant |x_* - y| \leqslant \beta_2\theta \tag{14.3.4}$$

с постоянными β_1 и β_2, которые не зависят от выбора описанного выше двумерного сечения.

Теорема 14.3.1. *Пусть выполнено требование* T *и предположения* 1)—3) *п.* 1° § 2. *Тогда*

$$|x_* - x_*^{(n)}| \leqslant Ce_n(x_*), \quad C = \text{const}. \tag{14.3.5}$$

Существует такой элемент $x^{(n)} \in H_n$, что $|x_* - x^{(n)}| \leqslant (1 + \varepsilon)e_n(x_*)$. По лемме 14.2.2 найдется элемент $y^{(n)} \in M \cap H_n$, удовлетворяющий неравенству $|x_* - y^{(n)}| \leqslant C_*e_n(x_*)$. Воспользуемся обозначениями рис. 3; все расстояния и углы измеряются в метрике H. По построению приближенного решения $\rho_n < \rho_n'$. Обозначим $l_n = \rho_n - d$ и $l_n' = \rho_n' - d$, тогда $l_n < l_n'$.

В силу неравенства (14.3.3)

$$\theta_n^2 \leqslant \frac{2}{k_0} l_n < \frac{2}{k_0} l'_n \leqslant \frac{3k_1}{k_0} \theta_n'^2$$

или $|\theta_n| < |\theta'_n| \sqrt{3k_1/k_0}$. Теперь по неравенству (14.3.4)

$$|x_* - x_*^{(n)}| \leqslant \beta_2 \theta_n < \beta_2 \theta'_n \sqrt{3k_1/k_0} \leqslant (\beta_2/\beta_1) \sqrt{3k_1/k_0} |x_* - y^{(n)}| \leqslant$$
$$\leqslant (\beta_2/\beta_1) \sqrt{3k_1/k_0} C_* e_n(x_*).$$

Таким образом, если выполнено требование Т и предположения 1)—3) п. 1° § 2, то погрешность аппроксимации приближенного решения есть величина порядка $O(e_n(x_*))$.

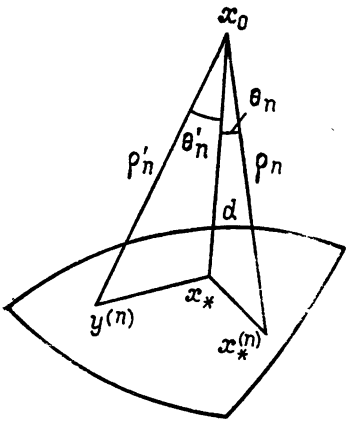

Рис. 3.

2°. В настоящем пункте мы укажем некоторый простой класс полунорм, удовлетворяющих требованию Т. Именно пусть $|||x|||^2 = b(x, x)$, где $b(x, x)$ — квадратичная форма, такая же, как в конце § 1; дополнительно предположим существование такой постоянной $\bar{c} > 0$, что $b(x, x) \leqslant \bar{c}|x|^2$, $\forall x \in H$. Для упрощения выкладок будем считать в данном пункте, что $g = 0$, $\bar{H} = H$.

Произвольную двумерную плоскость, проходящую через x_0 и x_*, можно построить так. В пространстве H возьмем точку y такую, что $|x_0 - y| = d$ и $[x_0 - x_*, x_0 - y] = 0$, и положим $e_1 = (x^* - x_0)/d$, $e_2 = (y - x_0)/d$. Двумерное множество, определяемое уравнением $x - x_0 = \xi e_1 + \eta e_2$, в котором ξ и η — произвольные вещественные числа, есть плоскость, проходящая через точки x_0, x_*, y; элементы вида $x - x_0$, соответствующие числам $\xi = 0$ и $\eta = 0$, ортогональны, поэтому ξ и η можно рассматривать как декартовы координаты на нашей плоскости; началу координат соответствует точка x_0.

Введем полярные координаты по обычным формулам $\xi = \rho \cos\theta$, $\eta = \rho \sin\theta$. Отметим полярные координаты точек x_0, x_*, y, именно x_0: $\rho = 0$; x_*: $\rho = d$, $\theta = 0$; y: $\rho = d$, $\theta = \pi/2$. В полярных координатах $x = x_0 + \rho v_\theta/d$, где $v_\theta = (x_* - x_0)\cos\theta + (y - x_0)\sin\theta$.

Если $\omega(x, y)$ — билинейная форма, то

$$\omega(x) := \omega(x, x) = d^{-2}\rho^2 \omega(v_\theta) + 2d^{-1}\rho\omega(x_0, v_\theta) + \omega(x_0).$$

В частности, многообразие ∂M определяется уравнением

$$d^{-2}\rho^2 b(v_\theta) + 2d^{-1}\rho b(x_0, v_\theta) + b(x_0) - \alpha^2 = 0. \qquad (14.3.6)$$

Нетрудно видеть, что $b(v_0) > 0$. Действительно, если $b(v_0) = 0$, то $|||x_0 - x_*||| = 0$ и $|||x_0||| = |||x_*||| = \alpha$; это означает, что x_0 есть решение односторонней вариационной задачи, вопреки предположению.

Итак, $b(v_0) > 0$. Но тогда при достаточно малых θ будет $b(v_\theta) > 0$. Уравнение (14.3.6) — квадратное относительно ρ; из его корней один равен $\rho(\theta)$, а другой больше, чем $\rho(\theta)$. Отсюда вытекают неравенства

$$-b(x_0, v_0) = b(x_0, x_0 - x_*) > 0, \qquad (14.3.7)$$

$$b(x_0) - a^2 > 0. \qquad (14.3.8)$$

Положив в (14.3.6) $\theta = 0$, $\rho = d$, получим

$$b(x_0) - \alpha^2 = -[b(v_0) + 2b(x_0, v_0)] = b(x_0 + x_*, x_0 - x_*),$$

и уравнение (14.3.6) можно записать в виде

$$d^{-2}\rho^2 b(v_\theta) - 2d^{-1}\rho b(x_0, -v_0) + b(x_0 + x_*, x_0 - x_*) = 0, \qquad (14.3.9)$$

а неравенство (14.3.8) — в виде

$$b(x_0 + x_*, x_0 - x_*) > 0. \qquad (14.3.10)$$

Вычислим теперь величину $\rho''(0)$. Продифференцируем (14.3.6) два раза по θ. Положив $\theta = 0$, получим

$$\rho''(0) = d\,\frac{b(x_*, x_0 - x_*) + b(x_0 - y)}{b(x_*, x_0 - x_*)}. \qquad (14.3.11)$$

Докажем, что знаменатель последней дроби положителен. При $\theta = 0$ уравнение (14.3.9) принимает вид

$$d^{-2}\rho^2 b(x_0 - x_*) - 2d^{-1}\rho b(x_0, x_0 - x_*) + b(x_0 + x_*, x_0 - x_*) = 0, \qquad (14.3.12)$$

его корни суть

$$\frac{d}{b(x_0 - x_*)}\{b(x_0, x_0 - x_*) \pm \pm [b^2(x_0, x_0 - x_*) - b(x_0 - x_*)\, b(x_0 + x_*, x_0 - x_*)]^{1/2}\} =$$

$$= \frac{d}{b(x_0 - x_*)}[b(x_0, x_0 - x_*) \pm b(x_*, x_0 - x_*)]. \qquad (14.3.13)$$

Если в (14.3.13) выбрать знак минус, то получится меньший корень уравнения (14.3.12) $\rho = d$; отсюда следует, что $b(x_*, x_0 - x_*) > 0$

Из формулы (14.3.11) вытекает теперь, что $\rho''(0) \geqslant d$. Далее, по сделанному выше предположению $b(x_0 - y) \leqslant \bar{c}\,|x_0 -$

$-y|^2 = \bar{c}d^2$, и мы приходим к оценке

$$k_0 = \frac{d}{2} \leqslant \frac{1}{2}\rho''(0) \leqslant \frac{d}{2} + \frac{\bar{c}d^3}{2b(x_*, x_0 - x_*)} := k_1, \qquad (14.3.14)$$

в которой k_0 и k_1 не зависят от выбора элемента y.

Докажем теперь, что при малых θ третья производная $\rho'''(\theta)$ ограничена по модулю независимо от θ и y. Меньший корень уравнения (14.3.10) равен

$$\frac{d}{b(v_\theta)}\{b(x_0, -v_\theta) - [b^2(x_0, -v_\theta) - b(x_0 + x_*, x_0 - x_*)\, b(v_\theta)]^{1/2}\}. \qquad (14.3.15)$$

Как мы видели, $b(v_\theta) > 0$ при малых θ. Из формулы (14.3.15) видно, что любая производная от $\rho(\theta)$ будет ограничена, если ограничено снизу подкоренное выражение. Выше было установлено, что при $\theta = 0$ оно равно $b^2(x_*, x_0 - x_*) > 0$. Далее,

$$-v_\theta = x_0 - x_* - 2(x_0 - x_*)\sin^2(\theta/2) + (x_0 - y)\sin\theta := x_0 - x_* + \eta;$$

заметим, что $|\eta| \leqslant d(|\theta| + \theta^2/2)$, и потому при малых θ будет $|\eta| \leqslant \beta|\theta|$, где β не зависит от θ и от y. Подкоренное выражение в (14.3.15) равно

$$b^2(x_0, x_0 - x_*) + 2b(x_0, \eta)\, b(x_0, x_0 - x_0) +$$
$$+ b^2(x_0, \eta) - [2b(x_0 - x_*, \eta) + b(\eta)]\, b(x_0 + x_*, x_0 - x_*) =$$
$$= b^2(x_0, x_0 - x_*) + O(|\theta|),$$

и при достаточно малых θ оно ограничено снизу положительным числом, которое не зависит от θ и y.

Заметим еще, что требование T выполнено, если для любого элемента y линия L — прямая или ее часть. В этом случае $\rho(\theta) = d/\cos\theta$ и можно положить $k_0 = k_1 = d/2$ и, при $|\theta| \leqslant \pi/6$, $\bar{k} = 7d/9$.

3°. Рассмотрим еще один случай. Как и выше, возьмем произвольную точку $y \in \partial M$, проведем в H двумерную плоскость Q через точки x_0, x_*, y (см. рис. 2) и введем на этой плоскости полярные координаты ρ и θ. Отметим, что полунорма $\|\cdot\|$ конечна на всех элементах плоскости Q; кроме того, как было показано в § 2, $\psi \geqslant \pi/2$. Пусть по-прежнему $L = \partial M \cap Q$, и уравнение $\rho = \rho(\theta)$ есть уравнение кривой L.

Функция $\rho(\theta)$ достигает минимума при $\theta = 0$, поэтому $\rho(\theta)$ возрастает при возрастании $|\theta|$. Из выпуклости кривой L следует, что с возрастанием $|\theta|$ угол ψ также возрастает.

Сделаем следующее допущение: существуют такие положительные постоянные θ_0, k и $\delta < \pi/2$, что при $|\theta| \leqslant \theta_0$ справедливы неравенства $|\rho'(\theta)| \leqslant k$ и $\delta + \pi/2 \leqslant \psi$; эти постоянные могут зависеть от x_0, но не от выбора двумерной плоскости Q, проходящей через точки x_0 и x_*.

Неравенство $\psi \geqslant \delta + \pi/2$ означает, что x_* — коническая точка множества ∂M: вектор $x_0 - x_*$ образует угол, не меньший, чем $\delta + \pi/2$, с любой секущей множества ∂M, проходящей через точку x_*.

Докажем, что при сформулированном здесь допущении верна оценка (14.3.5) для погрешности аппроксимации приближенного решения.

Пусть $|\theta| \leqslant \theta_0$. Обозначим $l(\theta) = \rho(\theta) - d$. Имеем

$$l(\theta) = \rho(\theta) - \rho(0) = \int_0^\theta \rho'(\theta)\, d\theta \leqslant k|\theta|, \qquad (14.3.16)$$

поэтому, если x — точка с полярными координатами $(\rho(\theta), \theta)$, то

$$|x_* - x|^2 = \rho^2(\theta) + d^2 - 2d\rho(\theta)\cos\theta =$$
$$= l^2(\theta) + 4d\rho(\theta)\sin^2\theta/2 \leqslant c_1^2\theta^2, \quad c_1 = \text{const}. \qquad (14.3.17)$$

Обозначая $\lambda = |x_* - x|$, находим

$$\sin|\theta| = \frac{\lambda \sin\psi}{\rho(\theta)}, \quad |\theta| \leqslant \frac{\theta_0}{\sin\theta_0}\frac{\lambda}{d} := c_2|x_* - x|. \qquad (14.3.18)$$

Кроме того,

$$\lambda = \frac{\rho(\theta)\sin|\theta|}{\sin\psi} \geqslant \frac{\rho(\theta)\sin|\theta|}{\cos\delta}$$

и, следовательно,

$$l^2(\theta) + 4\rho(\theta)\, d\sin^2\frac{\theta}{2} \geqslant \frac{\rho^2(\theta)\sin^2\theta}{\cos^2\delta}. \qquad (14.3.19)$$

Докажем, что при достаточно малом θ_0 справедливо неравенство

$$4\rho(\theta)\, d\sin^2\frac{\theta}{2} \leqslant \sigma\rho^2(\theta)\frac{\sin^2\theta}{\cos^2\delta}, \qquad (14.3.20)$$

в котором σ — постоянная, $0 < \sigma < 1$. Неравенство (14.3.20) равносильно следующему: $d \leqslant \sigma\rho(\theta)\cos^2\left(\frac{\theta}{2}\right)/\cos^2\delta$, и достаточно, чтобы $d \leqslant \sigma\rho(\theta)\cos^2\frac{\theta_0}{2}/\cos^2\delta$. Если взять $\theta_0 < 2\delta$, то последнее неравенство будет выполнено при $\delta = \cos^2\delta/\cos^2\frac{\theta_0}{2}$. Теперь

$$l^2(\theta) \geqslant (1-\sigma)\frac{\rho^2(\theta)\sin^2\theta}{\cos^2\delta} \geqslant (1-\sigma)\, d^2\frac{\sin^2\theta}{\cos^2\delta}$$

и окончательно

$$l(\theta) \geqslant c_3|\theta|, \quad c_3 = \frac{d\sqrt{1-\sigma}\sin\theta_0}{\theta_0\cos\delta}. \qquad (14.3.21)$$

Пусть $\bar{x}^{(n)}$ — элемент пространства H_n, на котором достигается наилучшее приближение элемента x_*, так что $e_n(x_*) = \|x_* - \bar{x}^{(n)}\|$. По лемме 14.2.2 существует элемент $y^{(n)} \in M \cap H_n$ такой, что $\|x_* - y^{(n)}\| \leqslant Ce_n(x_*)$. Воспользуемся обозначениями рис. 3, где на этот раз будем считать $x^{(n)} = y^{(n)}$; все углы и

расстояния измеряются в метрике H. По построению приближенного решения $\rho'_n > \rho_n$. Пусть $l_n = \rho_n - d$, $l'_n = \rho'_n - d$, так что $l_n < l'_n$. По формулам (14.3.17), (14.3.21), (14.4.16), (14.3.18) получаем

$$|x_* - x_*^{(n)}| \leqslant c_1 |\theta_n| \leqslant \frac{c_1}{c_3} l_n < \frac{c_1}{c_3} l'_n \leqslant \frac{c_1 k}{c_3} |\theta'_n| \leqslant$$

$$\leqslant \frac{c_1 c_2 k}{c_3} |x_* - y^{(n)}| \leqslant C e_n(x_*), \quad C = c_1 c_2 c_* c_3^{-1} k,$$

и наше утверждение доказано.

§ 4. ОЦЕНКА ПОГРЕШНОСТИ АППРОКСИМАЦИИ ДЛЯ БОЛЕЕ ОБЩЕЙ ЗАДАЧИ

1°. В статье [136] исследована погрешность аппроксимации для задачи, сформулированной в начале данной главы. Напомним постановку этой задачи. В рефлексивном банаховом пространстве B с нормой $\|\cdot\|$ заданы непустое выпуклое замкнутое множество и выпуклый слабонепрерывный снизу функционал $J(x)$; если множество M неограничено, то дополнительно требуется, чтобы $J(x) \xrightarrow[\|x\| \to \infty]{} +\infty$. Задача состоит в нахождении элемента $x_* \in M$, реализующего минимум функционала $J(x)$ на множестве M. Решение этой задачи существует; оно единственно, если J — строго выпуклый функционал.

2°. Приближенное решение x_n этой задачи будем строить в существенном так же, как в § 2; выберем полную в B последовательность конечномерных подпространств $\{B_n\}$ и примем за x_n решение задачи о минимуме функционала $J(x)$ на множестве $M \cap B_n$. Полнота последовательности $\{B_n\}$ понимается в [136] в следующем смысле. Пусть $P_M(x)$ — функционал Минковского для множества M. Этот функционал определяется следующим образом. Рассмотрим произвольное линейное множество L и в нем произвольное выпуклое множество M. Точка $\xi \in M$ принадлежит ядру множества M (терминология книги [71]; в книге [5] ξ называется алгебраически внутренней точкой множества M), если для любого $\eta \in L$ существует такое число $a > 0$, что при любом t, $|t| < a$, имеет место соотношение $(\xi + t\eta) \in M$. Функционал Минковского определяется формулой (см. [71, 5])

$$\forall x \in L, \quad P_M(x) = \inf_{\lambda > 0} \left\{\lambda : \frac{x - \xi_0}{\lambda} \in M\right\}, \qquad (14.4.1)$$

здесь ξ_0 — произвольно фиксированная точка ядра множества M. Ниже будем считать, что $\xi_0 = 0$ — это, очевидно, не ограничивает общности.

Положим теперь

$$\forall x \in B, \quad \mathscr{E}_n(x) = \inf_{x^{(n)} \in B_n} [\|x - x^{(n)}\| + |P_M(x) - P_M(x^{(n)})|]. \tag{14.4.2}$$

Будем считать, что на элементах подпространства B_n функционал Минковского конечен. По определению, принятому в [136], последовательность $\{B_n\}$ полна в точке $x \in B$, если $\lim\limits_{n\to\infty} \mathscr{E}_n(x) = 0$. Очевидно, для такой полноты необходимо, чтобы $P_M(x) < \infty$.

3°. Введем в рассмотрение функцию

$$g(\alpha) = \inf_{y \in M_\alpha} [J(y) - J(x_*)], \qquad M_\alpha = \{y \in M, \ . \ \|y - x_*\| = \alpha\}, \tag{14.4.3}$$

которую назовем *минорантой функционала J на множестве M.*

Оценка погрешности аппроксимации дается следующей теоремой:

Теорема 14.4.1. *Пусть выполнены следующие условия: а) на элементах подпространств B_n функционал Минковского конечен; б) в некоторой окрестности (в смысле нормы пространства B) точного решения x_* функционал $J(x)$ конечен и удовлетворяет условию Липшица; в) при малых значениях числа $\alpha \geqslant 0$ миноранта $g(\alpha)$ строго монотонна; г) последовательность $\{B_n\}$ полна в B в указанном выше смысле. Тогда имеет место оценка погрешности аппроксимации*

$$\|x_* - x_*^{(n)}\| \leqslant g^{-1}(\delta \mathscr{E}_n(x_*)), \qquad \delta = \beta \max(1, \|x_*\|), \tag{14.4.4}$$

β — постоянная Липшица для функционала J в окрестности точки x_*.

Замечание 14.4.1. В [136] доказано, что J удовлетворяет условию Липшица, если M имеет внутреннюю точку.

Замечание 14.4.2. При доказательстве теоремы 14.4.1 играет важную роль лемма, которая является обобщением леммы 14.2.2.

Лемма 14.4.1. *Пусть $x \in M$ и $x^{(n)} \in B_n$. Существует точка $y^{(n)} \in B_n \cap M$, такая, что*

$$\|x - y^{(n)}\| \leqslant \|x - x^{(n)}\| + \|x\| \cdot |P_M(x) - P_M(x^{(n)})|. \tag{14.4.5}$$

Отметим важный частный случай, рассмотренный в [136]. Пусть B — гильбертово пространство и $J(x) = \|x - x_0\|^2$, $x_0 \notin M$; принимаем, что M — произвольное непустое выпуклое замкнутое множество в B. Тогда верно следующее утверждение.

Теорема 14.4.2. *Если функционал Минковского множества M конечен на элементах подпространств B_n и последовательность $\{B_n\}$ полна в точке x_* в упомянутом выше смысле, то*

$$\|x_* - x_*^{(n)}\| = O\left(\sqrt{\mathscr{E}_n(x_*)}\right). \tag{14.4.6}$$

4°. В данном пункте мы изложим, также без доказательств, некоторые результаты работы [46]. В ней рассмотрена задача п. 3° § 1 о минимуме расстояния от заданной точки $x_0 \in H$, где H — гильбертово пространство с нормой $|\cdot|$, до непустого замкнутого выпуклого множества $M \in H$. Пусть H_n — подпространство H, $x_*^{(n)}$ и $x^{(n)}$ имеют тот же смысл, что и во всем предшествующем. Обозначим

$$\forall x \in H, \quad E_n(x) = \inf_{x^{(n)} \in H_n} |x - x^{(n)}|, \tag{14.4.7}$$

пусть еще p_n — ортопроектор из H на H_n. Доказывается следующая теорема.

Теорема 14.4,3. *Пусть* $M_n = M \cap H_n$ *и* $p_n x_* \in M_n$. *Тогда верна оценка*

$$|x_* - x_*^{(n)}| \leqslant E_n(x_*) + [E_n^2(x_*) + E_n(x_*) E_n(x_0)]^{1/2}. \tag{14.4.8}$$

Заметим, что если $E_n(x_*)$ и $E_n(x_0)$ суть величины одного порядка, то из (14.4.8) следует, что

$$|x_* - x_*^{(n)}| = O(E_n(x_*)); \tag{14.4.9}$$

к сожалению, условие $p_n x_* \in M_n$ трудно проверить

§ 5. ОБ УСТОЙЧИВОСТИ ТОЧНОГО РЕШЕНИЯ ОДНОСТОРОННЕЙ ВАРИАЦИОННОЙ ЗАДАЧИ

1°. Будем рассматривать задачу о минимуме расстояния в гильбертовом пространстве H от фиксированной точки $x_0 \in H$ до непустого выпуклого замкнутого множества $M \in H$:

$$|x - x_0| = \min, \quad x \in M. \tag{14.5.1}$$

Поставим вопрос [102] об устойчивости точного решения этой задачи относительно малых изменений ее данных: элемента x_0, нормы $|\cdot|$ пространства H, а также множества M. Этот вопрос, который кажется нам интересным и сам по себе, поможет нам потом (см. § 6) разобраться в погрешности искажения приближенного решения задачи (14.5.1). Последнюю задачу будем называть простейшей односторонней вариационной задачей.

2°. Пусть норма $|\cdot|$ и множество M остаются невозмущенными, а элемент x_0 заменен элементом $y_0 \in H$, причем $|x_0 - y_0| \leqslant \delta$; через δ здесь и ниже в данном параграфе обозначается малое положительное число. Решение возмущенной таким образом задачи (14.5.1) обозначим через y_*. По обычному предположению считаем, что $x_0 \notin M$, поэтому при достаточно малом δ будет также $y_0 \notin M$, а тогда $x_* \neq x_0$, $y_* \neq y_0$; $x_*, y_* \in \partial M$.

Рассмотрим четырехугольник (рис. 4). Его четыре вершины определяют трехмерное (в частном случае — двумерное) пространство E, которое является подпространством для H. Из доказательства леммы 14.2.1 вытекает, что углы четырехугольника в вершинах x_* и y_* — тупые или прямые, поэтому, если

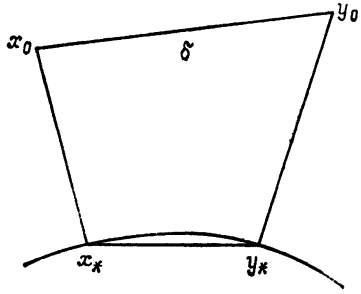

Рис. 4.

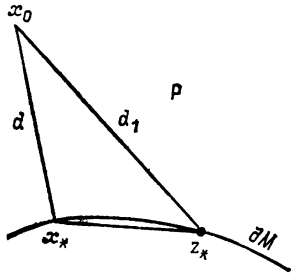

Рис. 5.

через точки x_* и y_* провести в E плоскости (прямые, если пространство E двумерное), нормальные к отрезку $\overline{x_*y_*}$, то названные точки будут лежать вне или на границе слоя, ограниченного проведенными плоскостями. Отсюда следует, что

$$|x_* - y_*| \leqslant |x_0 - y_0| \leqslant \delta, \tag{14.5.2}$$

и решение простейшей односторонней вариационной задачи устойчиво относительно малых возмущений элемента x_0.

3°. В пространстве H заменим норму $|\cdot|$ другой гильбертовой нормой $|\cdot|_1$ так, чтобы

$$\forall x \in H, \quad (1-\delta)|x| \leqslant |x|_1 \leqslant (1+\delta)|x|. \tag{14.5.3}$$

Пространство H, снабженное новой нормой, обозначим через H_1, а решение возмущенной задачи $|x - x_0|_1 = \min$, $x \in M$, — через z_*. Оценим величину $|x_* - z_*|$. Через точки x_0, x_*, z_* проведем в H двумерную плоскость P и введем на ней евклидову метрику, порожденную невозмущенной нормой $|\cdot|$. В обозначениях рис. 5 имеем $|x_0 - z_*|^2 = d^2 + |z_* - x_*|^2 - 2d|z_* - x_*|\cos\psi$. Очевидно, что $\psi \geqslant \pi/2$ — в противном случае x_* не была бы точкой минимума, поэтому

$$|z_* - x_*|^2 \leqslant |x_0 - z_*|^2 - d^2. \tag{14.5.4}$$

Аналогично, если на плоскости P ввести евклидову метрику, индуцированную возмущенной нормой $|\cdot|_1$, и обозначить $|x_0 - z_*|_1 = d_1$, то получим

$$|x_* - z_*|_1^2 \leqslant |x_0 - x_*|_1^2 - d_1^2. \tag{14.5.5}$$

Если $d_1 \leqslant d$, то из (14.5.3), (14.5.4) следует

$$|x_* - z_*|^2 \leqslant \frac{|x_0 - z_*|_1^2}{(1-\delta)^2} - d^2 = \frac{d_1^2}{(1-\delta)^2} - d^2 \leqslant \frac{d^2(2-\delta)}{(1-\delta)^2}\,\delta,$$

или

$$|x_* - z_*| \leqslant d\sqrt{2-\delta}\,\sqrt{\delta}\,(1-\delta)^{-1}; \qquad (14.5.6)$$

если же $d_1 > d$, то по неравенству (14.5.5)

$$|x_* - z_*|_1^2 \leqslant (1+\delta)^2 d^2 - d_1^2 \leqslant (1+\delta)^2 d^2 - d^2 = (2+\delta)\, d^2\delta,$$

откуда

$$|x_* - z_*| \leqslant \frac{|x_* - z_*|_1}{1-\delta} \leqslant \frac{d\sqrt{2+\delta}}{1-\delta}\sqrt{\delta}. \qquad (14.5.7)$$

Из неравенств (14.5.6), (14.5.7) следует, что

$$|x_* - z_*| \leqslant C_\delta\sqrt{\delta}, \quad C_\delta = d\sqrt{2+\delta}\,(1-\delta)^{-1}, \qquad (14.5.8)$$

и, значит, решение простейшей задачи устойчиво относительно малых возмущений нормы пространства H.

4°. Допустим теперь, что элемент x_0 и норма пространства H остались невозмущенными, а множество M заменено близким к нему множеством M', которое мы также предполагаем непустым, выпуклым и замкнутым в норме H. Близость множеств M и M' мы здесь определяем следующим требованием: существует такое биективное преобразование S: $M' \to M$, при котором

$$\forall x \in M, \quad |x - S^{-1}x| \leqslant \delta(|x|+1). \qquad (14.5.9)$$

Отметим одно следствие из неравенства (14.5.9). Пусть $x' \in M'$, тогда $Sx' \in M$ и по неравенству (14.5.9)

$$|x' - Sx'| = |Sx' - S^{-1}Sx'| \leqslant \delta(|Sx'|+1).$$

Полагая в (14.5.9) $x = Sx'$, получаем $|Sx'| \leqslant (1-\delta)^{-1}(|x'|+1)$ и, следовательно,

$$\forall x' \in M', \quad |x' - Sx'| \leqslant \delta(1-\delta)^{-1}(|x'|+1). \qquad (14.5.10)$$

Наряду с задачей (14.5.1) рассмотрим возмущенную задачу

$$|x' - x_0| = \min, \quad x' \in M'; \qquad (14.5.11)$$

ее решение обозначим через x'_*. Пусть, как и выше, $x_0 \notin M$. Докажем, что при достаточно малом δ будет $x_0 \notin M'$. Действительно, если $x_0 \in M'$, то $Sx_0 \in M$ и $|Sx_0 - x_0| \geqslant d$; как и выше, здесь $d = |x_* - x_0|$. С другой стороны, по неравенству (14.5.10) $|x_0 - Sx_0| \leqslant \delta(1-\delta)^{-1}(|x_0|+1)$, что противоречит предшествующему неравенству, если δ достаточно мало.

Обозначим $d' = \min\limits_{x' \in M'} |x' - x_0| = |x'_* - x_0|$. По неравенству (14.5.9)

$$d' \leqslant |x_0 - S^{-1}x_*| \leqslant |x_0 - x_*| + |x_* - S^{-1}x_*| \leqslant d + (|x_*|+1)\,\delta.$$

Аналогично, так как $x'_* \in M'$, то $Sx'_* \in M$, и по неравенству (14.5.10)

$$d \leqslant |x_0 - Sx'_*| \leqslant |x_0 - x'_*| + |x'_* - Sx'_*| \leqslant$$
$$\leqslant d' + \delta(1-\delta)^{-1}(|x'_*| + 1).$$

Таким образом,

$$-\delta(|x_*| + 1) \leqslant d - d' \leqslant \delta(1-\delta)^{-1}(|x'_*| + 1). \quad (14.5.12)$$

Теперь займемся оценкой разности $x_* - x'_*$. Имеем

$$|x_* - x'_*| \leqslant |x_* - S^{-1}x_*| + |S^{-1}x_* - x'_*| \leqslant$$
$$\leqslant \delta(|x_*| + 1) + |S^{-1}x_* - x'_*|. \quad (14.5.13)$$

Точка $S^{-1}x_* \in M'$. По лемме 14.2.1

$$|S^{-1}x_* - x'_*|^2 \leqslant |x_0 - S^{-1}x_*|^2 - d'^2 =$$
$$= [|x_0 - S^{-1}x_*| + d'][|x_0 - S^{-1}x_*| - d']$$

и по неравенству (14.5.12)

$$|S^{-1}x_* - x'_*|^2 \leqslant |\delta(1-\delta)^{-1}(|x_*| + 1) + \delta(|x_*| + 1)] \times$$
$$\times [2d + \delta(1-\delta)^{-1}(|x'_*| + 1) + \delta(|x_*| + 1)]. \quad (14.5.14)$$

Обозначим $|S^{-1}x_* - x'_*| = \varepsilon$. Из неравенства (14.5.13) получаем $|x'_*| \leqslant (1+\delta)(|x_*| + 1) + \varepsilon$. Подставив это в (14.5.14), приходим к квадратному неравенству для ε вида

$$(1 - a\delta^2)\varepsilon^2 - 2b\,\delta\varepsilon - c\delta \leqslant 0, \quad (14.5.15)$$

здесь a, b, c суть функции от δ и от $|x_*|$, ограниченные и положительные при фиксированной $|x_*|$ и при достаточно малых δ. Из (14.5.15) следует, что $\varepsilon \leqslant C\sqrt{\delta}$, $C = \text{const}$ и в силу неравенства (14.5.13)

$$|x'_* - x_*| \leqslant C'\sqrt{\delta}, \quad C' = \text{const}. \quad (14.5.16)$$

Соотношение (14.5.16) показывает, что решение задачи (14.5.1) устойчиво относительно возмущений множества M, малых в смысле неравенства (14.5.9)

5°. Обратимся к задаче (14.1.1), (14.1.2). Она является частным случаем задачи (14.5.1), поэтому для нее верны утверждения пп. 2—4°, и достаточно рассмотреть случаи, когда множество M возмущено так, что возмущения эти естественно считать малыми, хотя они могут не описываться соотношениями вида (14.5.9). В настоящем пункте мы рассмотрим случай, когда M' — возмущенное множество — определяется формулой

$$M' = \{x \in H : |||x - g||| \leqslant \alpha'\}, \quad (14.5.17)$$

в которой α' — число, близкое к α. Как было выяснено в § 1, можно считать, что $\|g\| := \beta < \alpha'$

Заметим, что если $g = 0$ (к этому, очевидно, сводится более общий случай $g \in H$), то мы оказываемся в условиях п. 4°: за S можно принять преобразование $Sx' = \alpha x'/\alpha'$; поэтому ниже предполагается, что $g \in \bar{H}$.

Пусть сначала $\alpha' = \alpha - \delta$. Положим

$$M_- = \{x \in H : \|x - g\| \leqslant \alpha - \delta\}. \tag{14.5.18}$$

Положим еще

$$\forall x \in M, \quad Qx = (1 - \sigma)x = x_-, \quad \sigma > 0, \tag{14.5.19}$$

и покажем, что при σ достаточно малом Q есть инъекция $M \to M_-$. Действительно, имеем

$$\|x_- - g\| = \|(1 - \sigma)(x - g) - \sigma g\| \leqslant (1 - \sigma)\alpha + \sigma\beta = \\ = \alpha - \sigma(\alpha - \beta)$$

и достаточно принять $\sigma = \delta(\alpha - \beta)^{-1}$; отметим еще, что $\sigma = c\delta$, $c = (\alpha - \beta)^{-1} = \text{const}$, и что справедливо равенство

$$|x - Qx| = c\delta|x|. \tag{14.5.20}$$

Наряду с задачей (14.1.1), (14.1.2) рассмотрим возмущенную задачу $|x - x_0| = \min$, $x \in M_-$, где M_- определено формулой (14.5.18). Решения этих задач пусть будут соответственно x_* и y_*. Пусть еще $d = |x_0 - x_*|$ и $d_- = |x_0 - y_*|$. Так как $M_- \subset M$, то $d \leqslant d_-$; кроме того, $x_* \in M$ и поэтому $Qx_* \in M_-$. Отсюда в силу неравенства (14.5.20)

$$d \leqslant d_- \leqslant |x_0 - Qx_*| \leqslant |x_0 - x_*| + |x_* - Qx_*| = d + c\delta|x_*|. \tag{14.5.21}$$

Оценим разность $x_* - y_*$. Имеем

$$|x_* - y_*| \leqslant |x_* - Qx_*| + |Qx_* - y_*| = c\delta|x_*| + |Qx_* - y_*|. \tag{14.5.22}$$

Точка $Qx_* \in M_-$; по лемме 14.2.1 $|Qx_* - y_*|^2 \leqslant |x_0 - Qx_*|^2 - d_-^2$. Теперь по неравенству (14.5.21) $|Qx_* - y_*|^2 \leqslant C\delta|x_*|(2d + C\delta|x_*|)$. Подставив это г (14.5.22), найдем

$$|x_* - y_*| \leqslant C|x_*|\delta + [C|x_*|(2d + C\delta|x_*|)]^{1/2}\sqrt{\delta}. \tag{14.5.23}$$

Если элемент x_* фиксирован, то отсюда получается более простое неравенство

$$|x_* - y_*| \leqslant C\sqrt{\delta}, \quad C = \text{const}, \tag{14.5.24}$$

которое доказывает устойчивость решения задачи (14.1.1), (14.1.2) по отношению к малому уменьшению числа α.

Пусть теперь α заменено на $\alpha + \delta$. Обозначим

$$M_+ = \{x \in H : \|x - g\| \leqslant \alpha + \delta\}. \tag{14.5.25}$$

Если $d_+ = |x_0 - z_*|$, где z_* — решение задачи

$$|x - x_0| = \min, \quad x \in M_+, \tag{14.5.26}$$

то справедливы соотношения, аналогичные соотношениям (14.5.21)—(14.5.23):

$$d_+ \leqslant d \leqslant d_+ + C\delta |z_*|, \tag{14.5.27}$$

$$|x_* - z_*| \leqslant C|z_*|\delta + [C|z_*|(2d + C\delta|z_*|)]^{1/2}\sqrt{\delta}. \tag{14.5.28}$$

В неравенстве (14.5.28) заменим левую часть меньшей величиной $||z_*| - |x_*||$. Теперь легко получить для $|z_*|$ квадратное неравенство

$$|z_*|^2 - 2|z_*|\frac{|x_*|(1 - C\delta) + 2dC\delta}{1 - 2C\delta - C^2\delta^2} + \frac{|x_*|^2}{1 - 2C\delta - C^2d^2} \leqslant 0,$$

из которого следует, что

$$|z_*| \leqslant (1 - 2C\delta - C^2\delta^2)^{-1}\{1 - C\delta|x_*| + 2Cd\delta + \\ + \sqrt{2C^2\delta^2|x_*|^2 + 2Cd\delta|x_*| + 4C^2\delta^2d^2}\}.$$

При малых δ последний радикал есть величина порядка $O(|x_*|\delta + \sqrt{|x_*|\delta} + \delta)$ и, следовательно, $|z_*| \leqslant |x_*| + O(\sqrt{|x_*|\delta})$. Подставив это в правую часть неравенства (14.5.28), убедимся, что при фиксированном x_* будет

$$|x_* - z_*| = O(\sqrt{\delta}); \tag{14.5.29}$$

этим доказана устойчивость решения задачи (14.1.1), (14.1.2) относительно малого увеличения числа α.

6°. Принимая по-прежнему, что $|||g||| < \alpha$, допустим, что g заменено элементом g', таким, что $|||g - g'||| \leqslant \delta$; остальные данные не изменены. Получаем новую задачу

$$|x - x_0| = \min, \quad x \in M', \\ M' = \{x \in H : |||x - g'||| \leqslant \alpha\}. \tag{14.5.30}$$

При этом

$$|||x - g||| - |||g - g'||| \leqslant |||x - g'||| \leqslant |||x - g||| + |||g - g'|||,$$

или

$$|||x - g||| - \delta \leqslant |||x - g'||| \leqslant |||x - g||| + \delta. \tag{14.5.31}$$

Обозначим, как и выше, $M_\pm = \{x \in H, |||x - g||| \leqslant \alpha \pm \delta\}$.

Пусть $x \in M_-$. Тогда $|||x - g'||| \leqslant |||x - g||| + \delta \leqslant \alpha$, т. е. $x \in M'$. Таким образом, $M_- \subset M'$. Точно так же найдем, что $M' \subset M_+$; неравенство (14.5.31) означает, следовательно, что

$$M_- \subset M' \subset M_+. \tag{14.5.32}$$

При расширении множества минимум не увеличивается, поэтому (смысл обозначений очевиден)

$$d_+ \leqslant d' \leqslant d_-, \quad d_+ \leqslant d \leqslant d_-. \tag{14.5.33}$$

Из формул (14.5.21), (14.5.27) следует, что $d_- - d_+ = O(\delta)$, а тогда тот же порядок имеют и разности $d_- - d'$, $d_- - d$, $d' - d_+$, $d - d_+$. Далее, пусть x'_* и z_* — решения, соответствующие множествам M' и M_+. Так как $M' \subset M_+$, то $x'_* \in M_+$; по лемме 14.2.1 $|x'_* - x_*|^2 \leqslant d'^2 - d_+^2$ и величина $|x'_* - z_*| = O(\sqrt{\delta})$. Точно так же $x_* \in M \subset M_+$ и $|x_* - z_*| = O(\delta)$. Теперь

$$|x_* - x'_*| \leqslant |x_* - z_*| + |z_* - x'_*| = O(\sqrt{\delta}) \qquad (14.5.34)$$

и решение задачи (14.1.1), (14.1.2) оказывается устойчивым по отношению к малым (в метрике $|||\cdot|||$) возмущениям элемента g.

7°. Пусть одновременно мало возмущены все параметры задачи (14.1.1), (14.1.2): элемент x_0 заменен на x_1, норма $|\cdot|$ — на $|\cdot|_1$, число α заменено на $\alpha' = \alpha \pm \delta$ и, наконец, элемент g заменен близким, в метрике $|||\cdot|||$, элементом g'. По-прежнему пусть x_* означает решение невозмущенной задачи (14.1.1), (14.1.2); через x_{1*}, x_{2*}, x_{3*}, x_{4*} обозначим решения следующих возмущенных задач: x_{1*} соответствует элементу x_1 при остальных невозмущенных параметрах; x_{2*} соответствует элементу x_1 и норме $|\cdot|_1$ при невозмущенных значениях α и g; x_{3*} — решение, полученное при невозмущенном элементе g и остальных возмущенных параметрах; наконец, x_{4*} соответствует возмущенным значениям всех четырех параметров. По-прежнему положим $d = |x_0 - x_*|$; пусть, кроме того, $d_i = |x_i - x_{i*}|$, $i = 1, 2, 3, 4$. Оценим разность $x_* - x_{4*}$. Имеем

$$|x_* - x_{4*}| \leqslant \sum_{i=1}^{4} |x_{i-1,*} - x_{i*}|, \qquad x_{0*} = x_*.$$

В силу неравенств (14.5.2), (14.5.3)

$$|x_* - x_{4*}| \leqslant \delta + \sum_{i=2}^{4} |x_{i-1,*} - x_{i*}|. \qquad (14.5.35)$$

Как видно из формул (14.5.6), (14.5.29), (14.5.34), оценки норм в (14.5.35) определяются через оценки величин d_i в зависимости от d. Найдем эти последние оценки. Очевидно,

$$d_1 \leqslant |x_1 - x_0| + |x_0 - x_*| + |x_* - x_{1*}| \leqslant d + 2\delta; \qquad (14.5.36)$$

по формулам (14.5.3), (14.5.8), (14.5.36)

$$d_2 = |x_1 - x_{2*}| \leqslant (1-\delta)^{-1} |x_1 - x_{2*}|_1 \leqslant (1-\delta)^{-1} \times$$
$$\times (|x_1 - x_{1*}|_1 + |x_{1*} - x_{2*}|_1) \leqslant (1+\delta)(1-\delta)^{-1} \times$$
$$\times (d + 2\delta + C_\delta (d + 2\delta)\sqrt{\delta}) = d + O(\sqrt{\delta}). \qquad (14.5.37)$$

Оценим величину d_3. По формуле (14.5.27) в случае $\alpha' = \alpha + \delta$ будет $d_3 = d_2 + O(\sqrt{\delta}) = d + O(\sqrt{\delta})$; то же соотношение верно и для $\alpha' = \alpha - \delta$ — это следует из формулы (14.5.21). Теперь из (14.5.33) вытекает, что $d_4 = d_3 + O(\sqrt{\delta}) = d + O(\sqrt{\delta})$.

Легко проверить, что все слагаемые в (14.5.35) суть величины порядка $O(\sqrt{\delta})$; это означает, что решение x_* устойчиво относительно малых изменений всех параметров, перечисленных в начале данного пункта.

Аналогичные рассуждения доказывают устойчивость решения задачи (14.5.1) относительно одновременных малых возмущений элемента x_0, нормы $|\cdot|$ и множества M, если последнее возмущение соответствует условиям п. 4°.

§ 6. ОЦЕНКА ПОГРЕШНОСТИ ИСКАЖЕНИЯ

1°. Пусть $a(x, y)$ — симметричная билинейная форма, определенная на некотором линейном множестве L, и соответствующая ей квадратичная форма $a(x) = a(x, x)$ положительна — это значит, что $a(x) \geqslant 0$ и $a(x) = 0 \Rightarrow x = 0$. Введя скалярное произведение $[x, y] = a(x, y)$ и норму $|x| = \sqrt{a(x)}$, мы превратим множество L в предгильбертово пространство; замкнув его указанной норме, получим гильбертово пространство, которое обозначим через H. Примем, что H сепарабельно; для упрощения последующих записей примем еще, что оно вещественно.

2°. Пусть $f(x)$ — линейный ограниченный в H функционал, определенный на всем пространстве, и $M \subset H$ — непустое выпуклое множество, замкнутое в H. Задача

$$a(x) - 2f(x) = |x|^2 - 2f(x) = \min, \quad x \in M, \tag{14.6.1}$$

имеет одно и только одно решение; она равносильна задаче

$$|x - x_0| = \min, \quad x \in M, \tag{14.6.2}$$

где x_0 — элемент пространства H, определяемый тождеством

$$\forall x \in H, \quad f(x) = [x, x_0], \tag{14.6.3}$$

существование и единственность элемента x_0 вытекает из известной теоремы Ф. Риса.

3°. Приближенное решение задачи (14.6.1) можно строить, например, по методу Ритца (в [38] вместо термина «метод Ритца» употребляется термин «внутренние методы»). Зададим полную в H последовательность конечномерных подпространств $\{H_n\}$, и пусть $(\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nN})$, $N = N(n)$ — базис подпространства H_n. Задачу (14.1.6) заменим следующей:

$$|x^{(n)}|^2 - 2f(x^{(n)}) = \min, \quad x^{(n)} \in H_n \cap M, \tag{14.6.4}$$

или, что то же,

$$|x^{(n)} - x_0| = \min, \quad x^{(n)} \in H_n \cap M. \tag{14.6.5}$$

Если пересечение $H_n \cap M$ не пусто, то задача (14.6.4) имеет одно и только одно решение.

4°. Пусть

$$x^{(n)} = \sum_{k=1}^{N} a_k^{(a)} \varphi_{nk}, \tag{14.6.6}$$

тогда

$$|x^{(n)}|^2 = \sum_{j,\,k=1}^{N} [\varphi_{nk},\ \varphi_{nj}]\, a_k^{(n)} a_j^{(n)}. \tag{14.6.7}$$

Скалярные произведения $[\varphi_{nk},\ \varphi_{nj}]$ вычисляются с некоторыми погрешностями $\gamma_{jk}^{(n)}$. Положим $[\varphi_{nk},\ \varphi_{nj}] + \gamma_{jk}^{(n)} = \alpha_{jk}^{(n)}$; числа $\alpha_{jk}^{(n)}$ нам известны. Обозначим через M_n и Γ_n матрицы элементов $[\varphi_{nk},\ \varphi_{nj}]$ и $\gamma_{jk}^{(n)}$; $j,\ k = 1,\ 2,\ \ldots,\ N$, и через $\|M_n\|$, $\|\Gamma_n\|$ — нормы этих матриц как операторов в R_N. Матрица M_n симметрична, и мы вправе считать, что матрица Γ_n также симметрична. Можно также считать, что матрица $M_n + \Gamma_n$ — положительно определенная — это будет всегда, когда Γ_n достаточно мала по норме.

Если $x^{(n)}$ — произвольный элемент из H_n, т. е. произвольный элемент вида (14.6.6), то по формуле (14.6.7) $|x^{(n)}|^2 = (M_n a^{(n)},\ a^{(n)})$, где $a^{(n)} = (a_1^{(n)},\ a_2^{(n)},\ \ldots,\ a_N^{(n)})$. Искаженная матрица $\tilde{M}_n := M_n + \Gamma_n$ порождает в H_n новую норму

$$|x^{(n)}|_1^2 := (\tilde{M}_n a^{(n)},\ a^{(n)}) = |x^{(n)}|^2 + (\Gamma_n a^{(n)},\ a^{(n)}).$$

Допустим, что по сравнению с M_n матрица Γ_n мала в следующем смысле:

$$\forall a^{(n)} \in R_N, \qquad \frac{|(\Gamma_n a^{(n)},\ a^{(n)})|}{(M_n a^{(n)},\ a^{(n)})} \leqslant \delta_1, \tag{14.6.8}$$

где δ_1 — малое число. Заметим, что условие (14.6.8) выполнено, если

$$\|\Gamma_n\| \cdot \|M_n^{-1}\| \leqslant \delta_1. \tag{14.6.9}$$

Действительно, обозначим $M_n^{1/2} a^{(n)} = b^{(n)}$, тогда отношение (14.6.8) принимает вид

$$\frac{|(\Gamma_n M_n^{-1/2} b^{(n)},\ M_n^{-1/2} b^{(n)})|}{\|b^{(n)}\|^2} \leqslant \|\Gamma_n\| \cdot \|M_n^{-1/2}\|^2 = \|\Gamma_n\| \cdot \|M_n^{-1}\|.$$

Если условие (14.6.8) соблюдено, то

$$(1 - \delta)\,|x^{(n)}| \leqslant |x^{(n)}|_1 \leqslant (1 + \delta)\,|x^{(n)}|, \tag{14.6.10}$$

$$\delta = \max\left(\sqrt{1 + \delta_1} - 1,\ 1 - \sqrt{1 - \delta_1}\right) = \delta_1/(1 + \sqrt{1 - \delta_1}).$$

5°. Положим $x_0 = x_{0n} + x_{1n}$, где $x_{0n} \in H_n$ и $x_{1n} \perp H_n$. Задача (14.6.5) равносильна следующей:

$$|x^{(n)} - x_{0n}| = \min, \qquad x^{(n)} \in H_n \cap M; \tag{14.6.11}$$

элемент x_{0n} можно построить как проекцию x_0 на H_n, т. е. как решение экстремальной задачи

$$\left|x_0 - \sum_{k=1}^{N} c_k^{(n)} \varphi_{nk}\right|^2 = \min, \qquad x_{0n} = \sum_{k=1}^{N} c_k^{(n)} \varphi_{nk},$$

что приводит к системе уравнений

$$\sum_{k=1}^{N} c_k^{(n)} [\varphi_{nk}, \varphi_{nj}] = [x_0, \varphi_{nj}] = f(\varphi_{nj}), \quad 1 \leqslant j \leqslant N. \quad (14.6.12)$$

Эту систему можно записать в виде

$$[x_{0n}, \varphi_{nj}] = f(\varphi_{nj}), \quad 1 \leqslant j \leqslant N;$$

таким образом, x_{0n} есть элемент, даваемый теоремой Ф. Риса, если функционал f сузить на H_n.

При составлении системы (14.6.12) будут допущены погрешности: вместо $[\varphi_{nk}, \varphi_{nj}]$ появятся коэффициенты $[\varphi_{nk}, \varphi_{nj}] + \gamma_{jk}^{(n)} = \alpha_{jk}^{(n)}$; кроме того, числа $f(\varphi_{nj})$ также будут вычислены с погрешностями; пусть последние равны $\delta_j^{(n)}$. Вместо системы (14.6.12) мы на самом деле построим искаженную систему

$$\tilde{M}_n c^{(n)} = f^{(n)} + \delta^{(n)}; \quad (14.6.13)$$

здесь $\tilde{c}^{(n)} = (\tilde{c}_1^{(n)}, \tilde{c}_2^{(n)}, \ldots, \tilde{c}_N^{(n)})$ — искаженный вектор коэффициентов и $\delta^{(n)} = (\delta_1^{(n)}, \delta_2^{(n)}, \ldots, \delta_N^{(n)})$.

Допустим, что линейный вычислительный процесс определения элемента x_0 по методу Ритца (процесс (14.6.2)) устойчив в последовательности пар пространств (R_N, R_N); напомним, что для этого необходимо и достаточно, чтобы координатная система была сильно минимальна в H. Тогда существуют такие постоянные p, q, $r > 0$, что если $\|\Gamma_n\| \leqslant r$, то $|y_{0n} - x_{0n}| \leqslant p \times \times \|\Gamma\| + q\|\delta^{(n)}\|$; здесь y_{0n} — искаженное значение элемента x_{0n}. Пусть $\|\Gamma_n\|$ и $\|\delta^{(n)}\|$ достаточно малы. Тогда $\|y_{0n} - x_{0n}\| \leqslant \delta$, где δ — малое число, не зависящее от n. Одновременно в силу необходимого условия устойчивости нормы $\|M_n^{-1}\|$ ограничены равномерно по n, и неравенства (14.6.9), (14.6.10) также выполнены. Таким образом, искажение сводится к тому, что при каждом n возмущаются элемент x_{0n} и норма в подпространстве H_n, и эти возмущения равномерно малы относительно n. По доказанному в § 5 искажение элемента x_{0n} приводит к искажению на $O(\|\Gamma_n\| + \|\delta^{(n)}\|)$, а искажение нормы — к искажению решения на $O\left(\sqrt{\|\Gamma_n\| \cdot \|M_n^{-1}\|}\right) = O\left(\sqrt{\|\Gamma_n\|}\right)$. Окончательно, если $y_*^{(n)}$ — искаженное решение n-й приближенной задачи, то

$$\|y_*^{(n)} - x_*^{(n)}\| = O\left(\sqrt{\|\Gamma_n\|} + \|\delta^{(n)}\|\right). \quad (14.6.14)$$

Как показывает неравенство (14.6.14), с целью избежать больших искажений следует вычислять матрицу Ритца точнее, чем свободные члены системы Ритца, — так, чтобы $\|\Gamma_n\| = O(\|\delta^{(n)}\|^2)$.

Можно исследовать устойчивость процесса вычисления коэффициентов $c_k^{(n)}$ не в последовательности (R_N, R_N), а в последовательности $(\bar{X}_N, \bar{Y}_N)$ (см. формулу (3.3.4)). Сформулируем ре-

зультат: упомянутый процесс устойчив и при достаточно малых нормах $\|\Gamma_n\|_{\overline{X}_N \to \overline{Y}_N}$ верна оценка

$$|y_*^{(n)} - x_*^{(n)}| = O\left(\sqrt{\|\Gamma_n\|_{\overline{X}_N \to \overline{Y}_N}} + \|\delta^{(n)}\|_{\overline{Y}_N}\right). \qquad (14.6.15)$$

Эта оценка пригодна, в частности, для метода конечных элементов.

§ 7. ДОБАВЛЕНИЕ. О ПОГРЕШНОСТИ АППРОКСИМАЦИИ ДЛЯ НЕКОТОРЫХ ВАРИАЦИОННЫХ НЕРАВЕНСТВ

В настоящем добавлении, написанном В. Б. Тюхтиным, излагаются результаты некоторых работ, опубликованных в последние годы и посвященных оценкам погрешности аппроксимации для решений односторонних вариационных задач, отличных, вообще говоря, от простейшей задачи, изученной в данной главе. Приближенные решения в этих работах обычно строятся по МКЭ.

В данном добавлении приняты несколько иные термины и обозначения, чем в основном тексте настоящей главы. Соболевские пространства $W_2^{(k)}$ и $\overset{\circ}{W}{}_2^{(k)}$ обозначаются соответственно через H^k и H_0^k. Множество ограничений обозначается через K. Точка в пространстве R_m чаще всего обозначается через x или y. Если ограничительное условие имеет вид $u(x) \leqslant \chi(x)$ или $u(x) \geqslant \chi(x)$, то множество ограничений называется *препятствием*; различаются препятствия в области или на границе, в зависимости от того, где выполняется упомянутое неравенство. Задачи, как правило, формируются в терминах вариационных неравенств.

1°. В статье Р. С. Фалка [165] рассматривается двумерная задача с препятствием в области. Пусть $\Omega \in R_2$ — выпуклая область с границей класса $C^{(2)}$ и (по повторяющимся индексам производится суммирование от 1 до 2)

$$a(u, v) = \int\limits_\Omega \int \left(a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + cuv \right) dx_1\, dx_2.$$

Принимаются следующие допущения:

$$a_{ij}(x) \in C^{(1)}(\overline{\Omega}), \quad a_{ij} = a_{ji}, \quad c(x) \in L_\infty(\Omega),$$
$$a_{ij}\xi_i\xi_j \geqslant \alpha |\xi|^2, \quad c(x) \geqslant \lambda, \quad \alpha,\ \lambda = \text{const} > 0.$$

Для $f \in L_2(\Omega)$ рассматривается задача об отыскании такой функции $u \in K$, что

$$a(u, v-u) \geqslant \int\limits_\Omega \int f(v-u)\, dx_1\, dx_2, \quad \forall v \in K, \qquad (14.7.1)$$

$$K = \{v \in H_0^1(\Omega):\ v(x) \geqslant \chi(x) \text{ почти всюду в } \Omega\}.$$

Здесь $\chi \in H^2(\Omega)$ — заданная функция, $\chi|_{\partial\Omega} = 0$. При перечисленных допущениях решение задачи принадлежит к классу $H^2(\Omega)$ [158].

Для приближенного решения задачи (14.7.1) используется МКЭ с кусочно-линейными координатными функциями. Пусть Ω_h — вписанный в Ω многоугольник, длины сторон которого не превосходят h. Этот многоугольник разбивается на треугольники со сторонами, не превышающими h. Через V_h обозначим пространство функций, заданных на Ω, непрерывных, линейных в каждом треугольнике разбиения и равных нулю на $\Omega \setminus \Omega_h$. Для всякой функции $v \in C(\Omega)$, равной нулю на $\partial\Omega$, обозначим через v_I ее интерполянт, т. е. функцию из V_h, совпадающую с v в вершинах триангуляции. Приближенное решение задачи (14.7.1) определяется как решение задачи

$$\forall v_h \in K_h, \quad u_h \in K_h, \quad a(u_h, v_h - u_h) \geqslant \int\int_{\Omega} (v_h - u_h) f \, dx_1 \, dx_2, \tag{14.7.2}$$

где $K_h = \{v_h \in V_h: v_h \geqslant \chi_I$ в узлах триангуляции$\}$. При некоторых «условиях регулярности» описанного здесь варианта МКЭ устанавливается оценка

$$\| u - u_h \|_{H^1(\Omega)} \leqslant Ch \| u \|_{H^2(\Omega)}. \tag{14.7.3}$$

В [166] Р. С. Фалк обобщил результаты своей предшествующей работы [165] на случай невыпуклой области $\Omega \in R_2$. Пусть кривая $\partial\Omega \in C^{(2)}$ и имеет лишь конечное число точек, в которых кривизна меняет знак. Тогда оценка (14.7.3) по-прежнему справедлива.

2° В [197] рассмотрена двумерная задача с препятствием на границе. Пусть $\Omega \in R_2$ — выпуклая конечная область с границей класса $C^{(2)}$, $a(u, v)$ — билинейная форма, такая же, как в п. 1°, $f \in L_2(\Omega)$, $\psi \in H^2(\Omega)$. Рассматривается задача (14.7.1), но с другим препятствием, именно

$$K = \{v \in H^1(\Omega): v \geqslant \psi \text{ на } \partial\Omega\}. \tag{14.7.4}$$

Известно [157], что при указанных условиях решение задачи (14.7.1), (14.7.4) $u \in H^2(\Omega)$.

Для приближенного решения задачи применяется метод, близкий к методу п. 1°. Впишем в Ω многоугольник Ω_h, длины сторон которого не превосходят h. Разобьем Ω_h на треугольники, длины сторон которых также не превосходят h. Обозначим через $I_0 = \{1, 2, \ldots, N_0\}$ индексы внутренних узлов, через $I_1 = \{N_0 + 1, \ldots, N\}$ — индексы граничных узлов; самые узлы обозначим через x_i. Положим еще $I = I_0 \cup I_1$. Далее, пусть Γ_i — часть $\partial\Omega$ между узлами x_i и x_{i+1}, Σ_i — часть Ω, ограниченная отрезком $[x_i, x_{i+1}]$ и кривой Γ_i.

Если T_i — треугольник со стороной $[x_i, x_{i+1}]$, $i \in I_1$, то объединение $T_i \cup \Sigma_i$ рассматривается как криволинейный элемент.

Через $\varphi_i^h(x)$, $x \in \Omega$, $i \in I$, обозначена непрерывная в Ω функция, линейная на каждом треугольнике или криволинейном элементе, равная единице в узле x_i и нулю во всех остальных узлах.

Пусть

$$H_h^1(\Omega) = \{v_h:\ v_h = \sum_{i \in I} v_i^h \varphi_i^h(x);\ v_i^h \in R_1,\ \forall i \in I\}. \quad (14.7.5)$$

Приближенная задача имеет вид (14.7.2), но препятствие определяется формулой

$$K_h = \{v_h \in H_h^1:\ v_h(x_i) \geqslant \psi(x_i), \quad \forall i \in I\}. \quad (14.7.6)$$

При условии регулярности использованного МКЭ верна оценка

$$\|u - u_h\|_{H^1(\Omega)} \leqslant Ch^{3/4} \|u\|_{H^2(\Omega)}. \quad (14.7.7)$$

3°. В статье [159] рассматривается конечноэлементная аппроксимация задачи с препятствием в области или на границе для некоторых билинейных форм.

Задача с препятствием в области ставится в виде

$$u \in K,$$

$$\int_\Omega [\nabla u \cdot \nabla(v-u) + u(v-u)]\, dx \geqslant \int_\Omega f(v-u)\, dx, \quad \forall v \in K, \quad (14.7.8)$$

$$K = \{v \in H^1(\Omega):\ v \geqslant \psi \text{ в } \Omega,\ v|_{\partial\Omega} = g\}.$$

Здесь Ω — выпуклая область класса $C^{(1,1)}$; g, $\psi \in H^2(\Omega)$, $\psi|_{\partial\Omega} \leqslant g$; $f \in L_2(\Omega)$.

Пусть Ω_h — вписанный в Ω многоугольник, длины сторон которого не превосходят h, и пусть этот многоугольник регулярным образом триангулирован. Для каждого треугольника триангуляции T_i, прилегающего к $\partial\Omega_h$, построим треугольник T_i' — зеркальное отражение треугольника T_i в его стороне, принадлежащей $\partial\Omega_h$. Обозначим $R_h = \bigcup T_i'$. Будем считать h столь малым, что $(\Omega_h \cup R_h) \supset \Omega$. Обозначим еще через $\hat{\Omega}$ некоторую область, содержащую $\Omega_h \cup R_h$ при любом достаточно малом h, и через V_h — множество непрерывных кусочно-линейных функций в $\Omega_h \cup R_h$. Продолжим решение $u \in H^2(\Omega)$ задачи (14.7.8) до функции $u \in H^2(\hat{\Omega})$. Пусть $\hat{u}_I$, ψ_I — интерполянты функций $\hat{u}$ и ψ соответственно на $\Omega_h \cup R_h$. Теперь можно определить множество ограничений K_h для приближенной задачи:

$$K_h = \{v_h \in V_h:\ v_h \geqslant \psi_I \text{ в узлах на } \Omega_h,\ v_h = \hat{u}_I \text{ на } \Omega \setminus \Omega_h\}.$$

Приближенное решение u_h определяется согласно (14.7.8) с заменой K на K_h. Устанавливается оптимальная по порядку оценка погрешности аппроксимации

$$\|u - u_h\|_{H^1(\Omega)} = O(h) \quad (14.7.9)$$

Наряду с кусочно-линейными координатными функциями для приближенного решения задачи (14.7.8) применяются также кусочно-квадратичные функции. В этом случае в предположении, что точное решение $u \in H^{5/2-\varepsilon}(\Omega)$, устанавливается оценка

$$\|u - u_h\|_{H^1(\Omega)} = O(h^{3/2-\varepsilon}). \tag{14.7.10}$$

Задача с препятствием на границе также ставится в форме (14.7.8), но K определяется формулой $K = \{v \in H^1(\Omega)\colon v|_{\partial\Omega} \geqslant g\}$. Пространство V_h определяется следующим образом. Если $v_h \in V_h$, то v_h непрерывна и кусочно-линейна на Ω_h; при этом, если точка $Q \in \Omega \setminus \Omega_h$, то $v_h(Q) = v_h(P)$, где P — проекция Q на $\partial\Omega_h$. Полагаем в данном случае $K_h = \{v_h \in V_h\colon v_h \geqslant q$ в узлах на $\partial\Omega\}$. Приближенное решение строится как решение задачи (14.7.8), в которой K заменено на K_h. В предположении, что $u \in H^2(\Omega)$, а u, $\tilde{g} \in W_\infty^1$ вблизи $\partial\Omega$ и свободная граница состоит из конечного числа точек, устанавливается оптимальная по порядку оценка погрешности аппроксимации:

$$\|u - u_h\|_{H^1(\Omega)} = O(h). \tag{14.7.11}$$

Таким образом, при более сильных ограничениях на решение получается более высокая по порядку оценка погрешности сравнительно с работой [197].

4°. В статье [188] рассматривается задача о минимуме функционала

$$J(v) = \int_\Omega |\nabla v|^2\, dx - 2\int_\Omega fv\, dx \tag{14.7.12}$$

на множестве $K = \{v\colon v \in H_0^1(\Omega),\ v \leqslant z$ почти всюду в $\Omega\}$. Здесь $\Omega \in R_2$ — область с достаточно гладкой границей, z и f — заданные функции, $z \in W_\infty^{(2)}(\Omega)$, $f \in L_2(\Omega)$.

Производится регулярная триангуляция с параметром h области Ω и строится пространство S_h непрерывных кусочно-линейных функций, равных нулю на $\partial\Omega$. Пусть $K_h = K \cap S_h$. Для задачи (14.7.12) приближенная задача состоит в минимизации функционала $J(v)$ на множестве K_h. Пусть u и u_h — соответственно точное и приближенное решение. Если $u \in W_\infty^{(2)}(\Omega)$, то устанавливается следующая оценка погрешности аппроксимации:

$$\|u - u_h\|_{L_\infty(\Omega)} \leqslant Ch^2 |\ln h| \cdot \|u\|_{W_\infty^{(2)}(\Omega)}, \tag{14.7.13}$$

где $C > 0$ не зависит от h.

5°. В [185] рассматривается задача с препятствием в области Ω для некоторого неквадратичного функционала. Именно пусть $\Omega \in R_n$, $K = \{v \in W_2^{(1)}(\Omega)\colon v(x) \geqslant \psi(x)$ в Ω, $v|_{\partial\Omega} = f\}$, где ψ и f — заданные функции, причем $\psi|_{\partial\Omega} = f$. Рассматривается

задача о минимуме функционала

$$\int_\Omega F(\nabla u)\,dx \tag{14.7.14}$$

на множестве K. В интеграле (14.7.14) $F \in C^{(2)}(R_n)$. Обозначим $\alpha(p) = \nabla F(p) = (\alpha_1(p), \alpha_2(p), \ldots, \alpha_n(p))$, $p \in R_n$, и допустим, что $(\alpha(p) - \alpha(q))(p-q) \geqslant C \max(|p|, |q|) \cdot |p-q|$; $|\cdot|$ — евклидова норма в R_n. Задача (14.7.14) равносильна задаче

$$a(u, v-u) := \int_\Omega \alpha_i(\nabla u) \frac{\partial}{\partial x_i}(v-u)\,dx \geqslant 0, \qquad \forall v \in K;$$

эта последняя аппроксимируется по МКЭ с кусочно-линейными координатными функциями. Пусть $\partial\Omega \in C^{(2)}$. Обозначим через $T_1, T_2, \ldots, T_m$ симплексы разбиения и через Ω_h — объединение $\bigcup_{i=1}^{m} T_i$. Через h обозначен параметр, которого не превосходят длины ребер симплексов T_i. Предполагается, что внутренности симплексов T_i не пересекаются и что каждая грань любого симплекса либо есть грань другого симплекса, либо ее вершины лежат на $\partial\Omega$. Через V_h обозначено пространство функций, непрерывных на Ω_h и линейных в каждом симплексе T_i. Обозначим еще через $N = N_h$ множество узлов в Ω_h. Пусть

$$K_h = \{v_h \in V_h: v_h(x) \geqslant \psi(x),\ x \in \Omega \cap N;\ v_h(x) = f(x),\ x \in \partial\Omega \cap N\}.$$

Приближенное решение определяется как решение задачи

$$u_h = K_h,$$
$$\int_{\Omega_h} \alpha_i(\nabla u_h) \frac{\partial}{\partial x_i}(v_h - u_h)\,dx \geqslant 0, \qquad \forall v_h \in K_h;$$

в предположении, что $u \in H^2(\Omega) \cap W_\infty^1(\Omega)$, устанавливается оценка погрешности

$$\|u - u_h\|_{H^1(\Omega_h)} = O(h). \tag{14.7.15}$$

6°. В [171] рассмотрена следующая задача. Пусть Ω — выпуклая область в R_n и $p > 1$. Требуется найти функцию $u \in \overset{\circ}{W}{}_p^{(1)}(\Omega)$, такую, что

$$J(u) = \inf J(v), \qquad v \in \overset{\circ}{W}{}_p^{(1)}(\Omega),$$
$$J(v) = p^{-1} \int_\Omega |\nabla u|^p\,dx - \int_\Omega fv\,dx, \qquad f \in W_{p'}^{(-1)}(\Omega), \tag{14.7.16}$$
$$1/p + 1/p' = 1.$$

Пусть $\{V_n\}$ — семейство конечномерных подпространств в $\overset{\circ}{W}{}_p^{(1)}(\Omega)$. Приближенное решение задачи (14.7.16) определяется

как решение задачи

$$J(u_h) = \inf_{v_h \in V_h} J(v_h). \tag{14.7.17}$$

Обе задачи (14.7.16), (14.7.17) имеют единственные решения u и u_h соответственно (см., например, [77]). Пусть $E_h(u) = \inf_{v_h \in V_h} \|v_h - u\|_{\overset{\circ}{W}{}_p^{(1)}(\Omega)}$. Устанавливаются следующие оценки погрешности аппроксимации:

$$\begin{aligned} &\|u_h - u\|_{\overset{\circ}{W}{}_p^{(1)}(\Omega)} \leqslant C\,[E_h(u)]^{1/(p-1)}, \quad p \geqslant 2; \\ &\|u_h - u\|_{\overset{\circ}{W}{}_p^{(1)}(\Omega)} \leqslant C\,[E_h(u)]^{1/(3-p)}, \quad 1 < p < 2, \end{aligned} \tag{14.7.18}$$

где C не зависит от h. Оценки (14.7.18) улучшены в [137]:

$$\|u_h - u\|_{\overset{\circ}{W}{}_p^{(1)}(\Omega)} \leqslant C\,[E_h(u)]^{\varkappa}; \quad \varkappa = \begin{cases} 2/p, & p \geqslant 2, \\ p/2, & 1 < p \leqslant 2. \end{cases} \tag{14.7.19}$$

7°. Пусть Ω — единичный квадрат на двумерной плоскости: $\Omega = [0, 1] \times [0, 1]$; на Ω заданы функции u_0, v_0, причем $u_0 \leqslant v_0$ почти всюду в Ω, $u_0|_{\partial\Omega} \leqslant 0 \leqslant v_0|_{\partial\Omega}$, $p > 2$ и

$$J(x) = \frac{1}{p} \int_\Omega \sum_{i=1}^{2} \left| \frac{\partial u}{\partial x_i} \right|^p dx - \int_\Omega f u\, dx. \tag{14.7.20}$$

В работе [196] рассматривается, в частности, проблема оценки приближенного решения задачи минимизации функционала (14.7.20) на множестве

$$K = \{v \in \overset{\circ}{W}{}_p^{(1)}(\Omega):\ u_0(x) \leqslant v(x) \leqslant v_0(x) \text{ почти всюду в } \Omega\}.$$

Пусть u — точное решение. Если $f \in L_{p'}(\Omega)$; $u_0,\ v_0 \in W_p^{(1+\sigma)}(\Omega)$; $\sigma = 1/(p-1)$, то $u \in W_p^{(1+\sigma)}(\Omega)$ (см. [163, 187]). Пусть еще u_h — приближенное решение, полученное по МКЭ при кусочно-линейных координатных функциях. Доказывается, что

$$\|u_h - u\|_{\overset{\circ}{W}{}_p^{(1)}(\Omega)} \leqslant C h^{\sigma'}, \tag{14.7.21}$$

где C не зависит от h.

8°. Статья [164] посвящена следующей задаче. Пусть $\Omega \subset R_2$ — выпуклый многоугольник. Требуется найти такую функцию $u \in H_0^1(\Omega)$, что

$$J(u) := \alpha \int_\Omega u^2\, dx + \beta \int_\Omega |\nabla u|^2\, dx + j(u) - \int_\Omega f u\, dx = \min, \tag{14.7.22}$$

или, что то же,

$$\begin{aligned} &\alpha(u, v - u) + \beta a(u, v - u) + j(v) - j(u) \geqslant (f, v - u); \\ &a(u, v) = \int_\Omega \nabla u \cdot \nabla v\, dx, \quad (u, v) = \int_\Omega uv\, dx, \quad f \in L_2(\Omega). \end{aligned} \tag{14.7.23}$$

Здесь α и β — положительные постоянные,

$$j(v)=\int_\Omega \Phi(v(x))\,dx,$$

а функция Φ определяется формулой

$$\Phi(\lambda)=2^{-1}\alpha_2(\lambda-u_0)_+^2+2^{-1}\alpha_1(\lambda-u_0)_-^2+L(\lambda-u_0)_+,$$

в которой u_0, α_1, α_2, L — положительные постоянные, $\lambda_+ = \max(\lambda, 0)$, $\lambda_- = \min(\lambda, 0)$, $\lambda \in R_1$. Существование и единственность решения задачи (14.7.22) следует из общих теорем [76].

Задача (14.7.22) аппроксимируется по МКЭ с кусочно-линейными координатными функциями. Область Ω покрывается семейством τ^h треугольников, h — наибольшая из их сторон. Пусть V_h — пространство непрерывных кусочно-линейных функций, равных нулю на $\partial\Omega$. Через $v_I \in V_h$ обозначается интерполянт функции v по узлам триангуляции. Обозначим еще

$$j_h(v)=\int_\Omega \{\Phi(v)\}_I\,dx=6^{-1}\sum_{T\in\tau^h} A(T)\,\Phi(v_i^T),$$

где $A(T)$ есть площадь треугольника T, а v_i^T есть значение функции v в i-й вершине треугольника T. Определяется дискретное скалярное произведение

$$(w, v)_h=\sum_{T\in\tau^h} 3^{-1}A(T)\sum_{i=1}^3 w_i^T v_i^T.$$

Приближенная задача ставится так: найти функцию $u_h \in V_h$, удовлетворяющую дискретному вариационному неравенству

$$\alpha(u_h, v_h-u_h)+\beta a(u_h, v_h-u_h)+j_h(v_h)-j_h(u_h)\geqslant \geqslant (f, v_h-u_h)_h, \quad \forall v_h\in V_h. \tag{14.7.24}$$

Решение этой задачи существует и единственно. При некоторых предположениях о свойствах точного решения устанавливается следующая оценка погрешности аппроксимации:

$$\|u-u_h\|_{H^1(\Omega)}\leqslant Ch, \quad C=\text{const}. \tag{14.7.25}$$

9°. В статье [190] рассмотрено вариационное неравенство

$$a(u, v-u)+\Phi_\nu(u, v)-\Phi_\nu(u, u)\geqslant (f, v-u), \quad \forall v\in K, \tag{14.7.26}$$

где a — билинейная форма, ограниченная и коэрцитивная, заданная на гильбертовом пространстве V; $K\in V$ — непустое замкнутое в V выпуклое множество; Φ_ν: $V\times V\to R$ — функционал, зависящий от положительного параметра ν и удовлетворяющий следующим условиям: 1) $\Phi_\nu(u, v)\geqslant 0$, $\forall u, v\in V$; 2) Φ_ν линеен по первому аргументу; 3) существует такое

$k_0(\nu) > 0$, что

$$\Phi_\nu(u, v) \leqslant k_0(\nu) \| u \|_V \cdot \| v \|_V;$$

4) при фиксированном $u \in V$ $j(v) := \Phi_\nu(u, v)$ является собственным выпуклым полунепрерывным снизу функционалом на V; 5) для любых u, v, $w \in V$ справедливы неравенства $|\Phi_\nu(u, v) - \Phi_\nu(u, w)| \leqslant \Phi_\nu(u, v - w)$ и $\Phi_\nu(u, v \pm w) \leqslant \Phi_\nu(u, v) + \Phi_\nu(u, w)$.

Неравенство (14.7.26) характеризует задачу Синьорини с нелокальным трением, коэффициент которого обозначен через ν. В более ранней работе тех же авторов доказывается, что упомянутое вариационное неравенство допускает единственное решение при достаточно малом ν.

Пусть V_h — конечномерное подпространство пространства V и K_h — замкнутое выпуклое множество, в некотором смысле близкое к K. Приближенным решением u_h задачи (14.7.26) называется решение вариационного неравенства

$$a(u_h, v_h - u_h) + \Phi_\nu(u_h, v_h) - \Phi_\nu(u_h, u_h) \geqslant (f, v_h - u_h), \quad \forall v_h \in V_h. \tag{14.7.27}$$

Это неравенство также допускает единственное решение при достаточно малом ν. Пусть U — гильбертово пространство, плотно и непрерывно вложенное в V, и оператор A определен по формуле $(Au, v) = a(u, v)$; круглые скобки означают скалярное произведение в V. Основным результатом статьи [190] является следующая теорема.

Пусть $(Au - f) \in U$. Тогда при перечисленных выше условиях и при достаточно малом ν существует такая постоянная $C > 0$, что

$$\| u - u_h \|_V \leqslant C \{ \| u - v_h \|_V + [\| Au - f \|_U (\| u - v_h \|_V + \| u_h - v \|_V)]^{1/2} + [\| u \|_V (\| u - v_h \|_V + \| u_h - v \|_V)]^{1/2}; \quad \forall v_h \in K_h, \quad \forall v \in K. \tag{14.7.28}$$

Глава 15

УПРУГОПЛАСТИЧЕСКОЕ СОСТОЯНИЕ ПО СЕН-ВЕНАНУ — МИЗЕСУ И ХААРУ — КАРМАНУ

Как хорошо известно, впервые математическая теория пластичности была предложена Б. де Сен-Венаном в 1870 г. Эта теория давала описание пластического состояния в условиях плоской деформации; на трехмерный случай ее распространил М. Леви в том же 1870 г. В 1913 г. Р. Мизес предложил описывать трехмерное пластическое состояние условием, близким к условию Сен-Венана и в то же время значительно более про-

стым. Как выяснилось потом, условие Мизеса имеет простую механическую трактовку. Для плоской деформации уравнения Сен-Венана и Мизеса совпадают.

В 1909 г. А. Хаар и Т. Карман опубликовали вариационный принцип, позволяющий описывать напряженное состояние в упругопластической среде, т. е. в среде, часть которой находится в упругом состоянии, а остальная часть — в пластическом. Принцип Хаара — Кармана опирается на условие пластичности Сен-Венана — Леви, но аналогичный принцип очень легко получить, исходя из условия Мизеса.

Впоследствии выяснилось, что явление пластического состояния значительно сложнее и многообразнее и что уравнения Сен-Венана — Леви и Мизеса описывают только так называемое пластическое течение. В этом плане следует прежде всего отметить работы Г. Хенки, Л. Прандтля, В. Прагера, Ф. Ходжа, Р. Хилла, А. А. Ильюшина, Л. М. Качанова и ряда других авторов.

В настоящей главе мы рассмотрим некоторые из задач, к которым приводит принцип Хаара — Кармана. Как мы увидим ниже, эти задачи подходят под схему (14.1.1), (14.1.2). В § 1 мы для облегчения ссылок формулируем уравнения Сен-Венана, Леви и Мизеса и вариационный принцип Хаара — Кармана. В § 2 рассматривается задача упругопластического кручения цилиндрических стержней, в § 3 исследуется задача о плоской деформации и в § 4 — трехмерная задача упругопластического состояния. Мы будем заниматься в этих параграфах исследованием устойчивости точных решений перечисленных задач относительно малых возмущений данных и оценкой погрешностей аппроксимации и искажения.

Ниже мы будем ссылаться на статьи Сен-Венана, Леви, Мизеса, Хаара — Кармана. Отметим, что русские переводы почти всех этих статей даны в сборнике [128].

§ 1. УРАВНЕНИЯ СЕН-ВЕНАНА — МИЗЕСА И ВАРИАЦИОННЫЙ ПРИНЦИП ХААРА — КАРМАНА

1°. Тензор напряжений будем обозначать так:

$$T = \begin{pmatrix} \tau_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{xy} & \tau_{yy} & \tau_{yz} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} \end{pmatrix}. \tag{15.1.1}$$

Если $F = (X, Y, Z)$ есть вектор объемных сил, то в состоянии равновесия тензор напряжений удовлетворяет уравнению равновесия

$$\operatorname{div} T + F = 0, \tag{15.1.2}$$

или в более подробной записи

$$\begin{aligned} \tau_{xx/x} + \tau_{xy/y} + \tau_{xz/z} + X &= 0, \\ \tau_{xy/x} + \tau_{yy/y} + \tau_{yz/z} + Y &= 0, \\ \tau_{xz/x} + \tau_{yz/y} + \tau_{zz/z} + Z &= 0, \end{aligned} \tag{15.1.3}$$

индекс за косой черточкой означает дифференцирование по соответствующему направлению.

Вектор смещений будем обозначать через $u = (u_x, u_y, u_z)$, вектор скоростей смещений — через v, так что $v = \dot{u} = (u_{x/t}, u_{y/t}, u_{z/t})$. Будем считать деформации и их скорости малыми, тогда тензор скоростей деформации можно записать так:

$$\frac{1}{2}\begin{pmatrix} 2v_{x/x} & v_{x/y} + v_{y/x} & v_{x/z} + v_{z/x} \\ v_{x/y} + v_{y/x} & 2v_{y/y} & v_{y/z} + v_{z/y} \\ v_{x/z} + v_{z/x} & v_{y/z} + v_{z/y} & 2v_{z/z} \end{pmatrix}. \tag{15.1.4}$$

Главные нормальные напряжения обозначим через $\tau_1 \geqslant \geqslant \tau_2 \geqslant \tau_3$; они суть корни уравнения

$$\begin{vmatrix} \tau_{xx} - \tau & \tau_{xy} & \tau_{xz} \\ \tau_{xy} & \tau_{yy} - \tau & \tau_{yz} \\ \tau_{xz} & \tau_{yz} & \tau_{zz} - \tau \end{vmatrix} = 0. \tag{15.1.5}$$

Наибольшее касательное напряжение равно

$$\tau_m = (\tau_1 - \tau_3)/2. \tag{15.1.6}$$

2°. Сен-Венан [195] сформулировал условие пластичности и написал уравнение пластического состояния для случая плоской деформации; в этом случае напряжения и смещения не зависят от одной из координат, скажем, от z, и $u_z = \tau_{xz} = = \tau_{yz} = 0$. Одно из главных нормальных напряжений равно τ_{zz}, а остальные два равны

$$2^{-1}\left[\tau_{xx} + \tau_{yy} \pm \sqrt{(\tau_{xx} - \tau_{yy})^2 + 4\tau_{xy}^2}\right]. \tag{15.1.7}$$

Опираясь на опыты Треска, Сен-Венан сформулировал следующие гипотезы о пластическом состоянии.

I. Наибольшее касательное напряжение в пластическом состоянии есть величина постоянная, характерная для данной среды. Сен-Венан обозначил эту постоянную буквой K. Мы будем называть данную гипотезу условием пластичности Сен-Венана.

II. Направления наибольшей скорости скольжения и наибольшего касательного напряжения совпадают.

III. Вещество в пластическом состоянии несжимаемо.

Кроме этих гипотез, явно сформулированных, Сен-Венан пользуется еще двумя, явно не сформулированными гипотезами.

IV. τ_{zz} лежит между главными напряжениями (15.1.7).

V. Производные по координатам от смещений и скоростей смещений малы.

Перечисленные гипотезы приводят к следующим уравнениям пластического состояния в условиях плоской деформации:

$$\tau_{xx/x} + \tau_{xy/y} + X = 0, \quad \tau_{xy/x} + \tau_{yy/y} + Y = 0; \tag{15.1.8}$$

$$(\tau_{xx} - \tau_{yy})^2 + 4\tau_{xy}^2 = 4K^2; \tag{15.1.9}$$

$$(v_{y/y} - v_{x/x}) : (v_{y/x} + v_{x/y}) = (\tau_{yy} - \tau_{xx}) : 2\tau_{xy}; \tag{15.1.10}$$

$$v_{x/x} + v_{y/y} = 0. \tag{15.1.11}$$

Уравнения (15.1.8) суть общие уравнения равновесия механики сплошной среды, уравнение (15.1.9) есть условие пластичности Сен-Венана, уравнения (15.1.10) и (15.1.11) соответствуют гипотезам II и III.

3°. Уравнения теории пластичности в трехмерном случае вывел М. Леви [179] на основании условия пластичности Сен-Венана. Гипотеза IV становится ненужной, гипотеза II заменяется следующей, более общей.

IIa. Девиатор напряжений и тензор скоростей смещений пропорциональны.

Гипотезы I, III, V Сен-Венана сохраняются.

Гипотезы М. Леви приводят к следующим уравнениям пластического состояния в трехмерной среде:

$$(v_{x/x} - v_{y/y}) : (\tau_{xx} - \tau_{yy}) = (v_{y/y} - v_{z/z}) : (\tau_{yy} - \tau_{zz}) =$$

$$= (v_{x/y} + v_{y/x}) : 2\tau_{xy} = (v_{y/z} + v_{z/y}) : 2\tau_{yz} = (v_{z/x} + v_{x/z}) : 2\tau_{xz};$$

$$v_{x/x} + v_{y/y} + v_{z/z} = 0. \tag{15.1.13}$$

Условие пластичности можно записать в виде

$$\tau_1 - \tau_3 = 2K, \tag{15.1.14}$$

где τ_1 и τ_3 суть наибольший и наименьший из корней уравнения (15.1.5). К этим уравнениям следует добавить уравнения равновесия (15.1.3).

Заметим, что в случае плоской деформации из уравнений М. Леви вытекает гипотеза IV Сен-Венана. В самом деле, в этом случае $v_z = 0$, $\tau_{xz} = \tau_{yz} = 0$, уравнение (15.1.13) дает $v_{x/x} = -v_{y/y}$, и из первого уравнения (15.1.12) следует $\tau_{zz} = (\tau_{xx} + \tau_{yy})/2$. Отсюда и из (15.1.7) вытекает гипотеза IV.

4°. Коротко изложим теорию Р. Мизеса [184]. Существенным ее отличием является иная формулировка условия пластичности. Пусть τ_1, τ_2, τ_3 — главные нормальные напряжения. Положим

$$\zeta_1 = (\tau_2 - \tau_3)/2, \quad \zeta_2 = (\tau_3 - \tau_1)/2, \quad \zeta_3 = (\tau_1 - \tau_2)/2, \tag{15.1.15}$$

так что

$$\zeta_1 + \zeta_2 + \zeta_3 = 0. \tag{15.1.16}$$

Мизес принимает следующую гипотезу, более общую, чем гипотеза I Сен-Венана: в пластическом состоянии напряжения остаются на пределе упругости, причем в пространстве координат ζ_1, ζ_2, ζ_3 предел упругости представляется в виде замкнутой кривой, лежащей в плоскости (15.1.16) и обходящей начало координат.

По Сен-Венану, $|\zeta_i| \leqslant K$, $i = 1, 2, 3$. Эти неравенства определяют в пространстве $(\zeta_1, \zeta_2, \zeta_3)$ куб, который пересекается с плоскостью (15.1.16) по правильному шестиугольнику. Мизес предложил заменить куб Сен-Венана сферой

$$\zeta_1^2 + \zeta_2^2 + \zeta_3^2 = 2K'^2, \qquad K' = \text{const}. \tag{15.1.17}$$

Уравнение (15.1.17) легко преобразуется к виду

$$\tau_i^2 := \frac{1}{6}\left[(\tau_{xx} - \tau_{yy})^2 + (\tau_{yy} - \tau_{zz})^2 + (\tau_{zz} - \tau_{xx})^2 + \right.$$
$$\left. + 6\,(\tau_{xy}^2 + \tau_{yz}^2 + \tau_{xz}^2)\right] = \frac{4}{3}K'^2. \tag{15.1.18}$$

Величина τ_i называется интенсивностью касательных напряжений.

В условиях плоской деформации $\tau_i^2 = 4^{-1}(\tau_{xx} - \tau_{yy})^2 + \tau_{xy}^2$; условие пластичности Сен-Венана будет получаться как частный случай условия (15.1.18) Мизеса, если положить

$$K' = \sqrt{3}\,K/2, \tag{15.1.19}$$

при этом условие Мизеса несколько упрощается:

$$\tau_i = K. \tag{15.1.20}$$

Остальные уравнения пластического состояния по Мизесу совпадают с соответствующими уравнениями М. Леви.

5°. Вариационный принцип Хаара и Кармана был опубликован в статье [173], вышедшей в 1909 г., за несколько лет до появления статьи Мизеса [184]; естественно, что этот принцип был основан на теории пластичности Сен-Венана — Леви. Однако принцип Хаара — Кармана легко формулируется в терминах теории Мизеса, и в этом случае названный принцип утверждает следующее.

Пусть среда, заполняющая область Ω трехмерного (или двумерного) пространства, находится в упругопластическом состоянии. Для простоты допустим, что объемные силы отсутствуют, так что вектор $F = 0$, и тензор напряжений T удовлетворяет однородному уравнению

$$\operatorname{div} T = 0. \tag{15.1.21}$$

Пусть $W(T)$ — плотность энергии упругих деформаций, выраженная в терминах тензора напряжений T. Тогда в упругопластическом состоянии этот тензор реализует минимум интеграла

$$J(T) = \int_\Omega W(T)\,dx \tag{15.1.22}$$

на множестве симметричных тензоров T, удовлетворяющих следующим условиям: а) интеграл (15 1.22) конечен; б) T удовлетворяет уравнению (15.1.21) и краевым условиям Коши, означающим, что известны внешние силы, приложенные к границе области Ω; в) T удовлетворяет одностороннему ограничению

$$\tau_i \leqslant K, \tag{15.1.23}$$

где τ_i — интенсивность касательных напряжений (формула (15.1.18)).

В статье [173] доказывается некоторая общая теорема, из которой, в частности, вытекает: если T решает сформулированную выше одностороннюю вариационную задачу, то в каждой точке области Ω удовлетворяется либо уравнение Эйлера для функционала (15.1.22), либо условие пластичности (15.1.20).

Следует, однако, иметь в виду, что в пластической зоне должны выполняться также уравнения, содержащие скорости деформаций; так, в трехмерной задаче должны быть удовлетворены уравнения (15.1.12), (15.1.13). Поэтому из существования решения вариационной задачи Хаара — Кармана еще не вытекает автоматически существование решения задачи об упругопластическом состоянии.

Принцип Хаара — Кармана тесно связан с известным принципом Кастильяно в теории упругости. Напомним этот последний принцип: если тензор напряжений реализует минимум функционала (15 1.22) при условиях а) и б), то этот тензор решает задачу теории упругости при краевых условиях, упомянутых в б). Таким образом, можно формально получить принцип Хаара — Кармана из принципа Кастильяно, если потребовать, чтобы функционал Кастильяно достигал минимума на множестве тензоров напряжений, удовлетворяющих не только условиям а), б), но и неравенству (15.1.23).

§ 2. КРУЧЕНИЕ СТЕРЖНЯ

1°. Пусть нормальное сечение G скручиваемого стержня есть конечная и в общем случае многосвязная область плоскости координат (x, y). Если стержень весь находится в упругом состоянии, то отличные от тождественного нуля напряжения $\tau_x := \tau_{xz}, \tau_y := \tau_{yz}$ удовлетворяют уравнениям

$$\tau_{x/x} + \tau_{y/y} = 0, \quad (x, y) \in G; \tag{15.2.1}$$

$$\tau_x \cos(\nu, x) + \tau_y \cos(\nu, y) = 0, \quad (x, y) \in \partial G; \tag{15.2.2}$$

$$\iint_G (x\tau_y - y\tau_x)\, dx\, dy = Q; \tag{15.2.3}$$

$$\Delta\tau_x = \Delta\tau_y = 0, \tag{15.2.4}$$

Здесь ν — нормаль к ∂G, Q — момент внешних сил, приложенных к основанию стержня. Если стержень находится в упругопластическом состоянии, то уравнения (15.2.1)—(15.2.3) сохраняют силу; уравнения (15.2.4) остаются справедливыми в упругой зоне. Приняв, что напряжения в пластической зоне удовлетворяют условию Сен-Венана (15.1.14) или Мизеса (15.1.20), мы должны заменить уравнения (15.2.4) одним уравнением

$$\tau_x^2 + \tau_y^2 = K^2. \tag{15.2.5}$$

Введем предгильбертово пространство, элементы которого суть двумерные векторы $\tau = (\tau_x, \tau_y)$, удовлетворяющие уравнению (15.2.1) и краевому условию (15.2.2); норму в этом пространстве зададим формулой

$$|\tau|^2 = \iint_G (\tau_x^2 + \tau_y^2)\, dx\, dy. \tag{15.2.6}$$

Пополнив это пространство в метрике (15.2.6), получим гильбертово пространство, которое обозначим через $\tilde{H}$. Об элементах этого пространства будем говорить, что они удовлетворяют (в обобщенном смысле) уравнениям (15.2.1), (15.2.2).

Если стержень находится в упругом состоянии, то в силу принципа Кастильяно задача кручения равносильна задаче о минимуме функционала (15.2.6) на множестве векторов, удовлетворяющих уравнениям (15.2.1)—(15.2.3), — иначе говоря, на множестве элементов пространства $\tilde{H}$, удовлетворяющих уравнению (15.2.3). Пусть $-g = -(g_x, g_y)$ — какой-нибудь вектор из $\tilde{H}$, удовлетворяющий уравнению (15.2.3). Положим $g + \tau = \sigma = (\sigma_x, \sigma_y)$. Тогда σ удовлетворяет однородному уравнению

$$\iint_G (x\sigma_y - y\sigma_x)\, dx\, dy = 0. \tag{15.2.7}$$

Пусть H — подпространство тех элементов из $\tilde{H}$, которые удовлетворяют уравнению (15.2.7). В силу принципа Хаара — Кармана решение задачи упругопластического кручения заменяется решением следующей односторонней вариационной задачи:

$$|\sigma - g| = \min, \quad \sigma \in M; \quad M = \{\sigma \in H : |||\sigma - g||| \leqslant K\};$$
$$|||\tau||| = \sup_{(x,\,y)\,\in\, G} \operatorname{ess} \sqrt{\tau_x^2 + \tau_y^2}. \tag{15.2.8}$$

2°. Исследуем множество M, определенное в (15.2.8). При больших по модулю значениях Q это множество пусто. Действительно,

$$|Q| \leqslant \iint_G \rho \sqrt{\tau_x^2 + \tau_y^2}\, dx\, dx, \quad \rho^2 = x^2 + y^2;$$

в упругопластическом состоянии $\tau_x^2 + \tau_y^2 \leqslant K^2$, поэтому

$$|Q| \leqslant K \iint_G \rho \, dx \, dy. \tag{15.2.9}$$

Правая часть неравенства (15.2.9) есть некоторая постоянная; если $|Q|$ больше этой постоянной, то M пусто.

Назовем момент Q *допустимым*, если ему соответствует непустое множество M. Изменив в случае необходимости ориентацию осей координат, можно считать, что $Q \geqslant 0$. Обозначим через $\bar{Q}$ точную верхнюю границу допустимых моментов. Докажем, что множество допустимых моментов есть либо отрезок $0 \leqslant Q \leqslant \bar{Q}$, либо полуинтервал $0 \leqslant Q < \bar{Q}$. Достаточно убедиться, что последний полуинтервал содержится в множестве допустимых моментов.

В любой левой полуокрестности точки $\bar{Q}$ содержится хотя бы один допустимый момент Q'. Ему соответствует по крайней мере один вектор $\tau' = (\tau'_x, \tau'_y)$, такой, что

$$\iint_G (x\tau'_y - y\tau'_x) \, dx \, dy = Q', \qquad \|\tau'\| \leqslant K.$$

Пусть $0 \leqslant Q'' < Q'$. Положим $\tau'' = Q''\tau'/Q'$. Тогда $\tau'' \in \tilde{H}$ и

$$\iint_G (x\tau''_y - y\tau''_x) \, dx \, dy = Q'', \qquad \|\tau''\| = \frac{Q''}{Q'} \|\tau'\| < K.$$

Последнее соотношение означает, что множество M'', которое соответствует моменту Q'', содержит элемент τ'' и потому не пусто. Этим доказано, что полуинтервал $0 \leqslant Q < \bar{Q}$ содержится в множестве допустимых моментов.

Докажем теперь, что множество M, определенное в (15.2.8), замкнуто в норме $\|\cdot\|$. Пусть $\{\tau_n\}$, $\tau_n = (\tau_{nx}, \tau_{ny})$ — сходящаяся в $\tilde{H}$ последовательность векторов из M и $\lim \tau_n = \tau_0$. Из $\{\tau_n\}$ можно выделить последовательность $\{\tau_{n_k}\}$, сходящуюся к τ_0 почти всюду на G. Так как $\|\tau_{n_k}\| = \sup \sqrt{\tau_{n_k x}^2 + \tau_{n_k y}^2}$, то и $\sup \sqrt{\tau_{0x}^2 + \tau_{0y}^2} \leqslant K$ почти всюду в G. Это и означает, что $\tau_0 \in M$ и, следовательно, множество M замкнуто.

Из общих теорем, упомянутых в § 1 главы 14, следует, что задача (15.2.8) имеет одно и только одно решение.

3°. Задача (15.2.8) есть частный случай задачи (14.1.1), (14.1.2), в которой, в частности, следует положить $\alpha = K$. Если $0 \leqslant Q < \bar{Q}$, то, как мы видели в п. 2°, множество M содержит элемент с полунормой, меньшей K. Примем этот элемент за g. На основании результатов § 5 главы 14 можно утверждать, что решение задачи (15.2.8) устойчиво относительно малых возмущений вектора g и постоянной пластичности K, если крутящий момент лежит в полуинтервале $0 \leqslant Q < \bar{Q}$. Пусть момент Q

заменен моментом Q', мало отличающимся от Q и по-прежнему лежащим в полуинтервале $[0, \bar{Q})$. Если моменту Q соответствует вектор g, $\|g\| < K$, то моменту Q' можно привести в соответствие вектор $g' = Q'g/Q$; если Q и Q' достаточно близки, то $\|g'\| < K$ и величина $\|g - g'\|$ достаточно мала. Из сказанного следует устойчивость решения задачи (15.2.8) относительно малых возмущений крутящего момента, не выводящих из полуинтервала $[0, \bar{Q})$.

Ниже мы покажем, что при том же условии $0 \leqslant Q < \bar{Q}$ решение задачи (15.2.8) устойчиво относительно малых возмущений области G.

4°. Для облегчения последующих ссылок приведем некоторые известные факты из теории кручения упругих стержней.

Пусть область G — $(n+1)$-связная. Тогда ∂G состоит из $n+1$ связных компонент, которые мы обозначим через ∂G_j, $j = 0, 1, \ldots, n$; через ∂G_0 обозначим ту компоненту, которая ограничивает область G извне. Обозначим еще через G_j, $j = 1, 2, \ldots, n$, ту область, которая ограничена извне контуром ∂G_j. Пусть $G_0 = G \cup G_1 \cup \ldots \cup G_n$; G_0 есть конечная односвязная область, ограниченная контуром ∂G_0. Можно определить G_0 как наименьшую односвязную область, содержащую область G.

Уравнение (15.2.1) позволяет ввести так называемую функцию Прандтля $u(x, y)$:

$$\tau_x = \frac{\partial u}{\partial y}, \qquad \tau_y = -\frac{\partial u}{\partial x}. \tag{15.2.10}$$

Из краевого условия (15.2.2) следует, что на границе области G функция Прандтля удовлетворяет условию

$$u|_{\partial G_j} = A_j = \text{const}. \tag{15.2.11}$$

При изменении функции Прандтля на постоянное слагаемое напряжения не меняются, поэтому можно положить $A_0 = 0$.

Как видно из уравнений (15.2.4), в упругой области функция Прандтля удовлетворяет уравнению Пуассона

$$-\Delta u = c_0, \qquad c_0 = \text{const}. \tag{15.2.12}$$

Известна формула, связывающая крутящий момент Q с функцией Прандтля $u(x, y)$, именно

$$Q = 2\iint_G u\,dx\,dy + 2\sum_{j=1}^{n} A_j |G_j|. \tag{15.2.13}$$

Продолжим функцию u на всю область G_0, положив $u(x, y) = A_j$, $(x, y) \in G_j$. Тогда формула (15.2.13) принимает более простой вид:

$$Q = 2\iint_{G_0} u\,dx\,dy. \tag{15.2.14}$$

5°. Задачу упругопластического кручения можно свести к вариационной задаче, отличной по форме от задачи (15.2.8). Если скручиваемый стержень находится в упругом состоянии, то задачу (15.2.8) можно записать в терминах функции Прандтля следующим образом:

$$\iint_G |\nabla u|^2\,dx\,dy = \min,$$

$$u = \begin{cases} 0 & \text{на } \partial G_0, \\ A_j = \text{const} & \text{на } \partial G_j, \quad 1 \leqslant j \leqslant n, \end{cases} \tag{15.2.15}$$

$$2\iint_G u\,dx\,dy + 2\sum_{j=1}^{n} A_j |G_j| = Q.$$

Это задача на условный минимум; по теореме Эйлера существует такая постоянная c_0, что

$$\delta\left\{\iint_G [|\nabla u|^2 - 2c_0 u]\,dx\,dy - 2c_0 \sum_{j=1}^{n} A_j |G_j|\right\} = 0; \tag{15.2.16}$$

δ — знак вариации; функция u должна удовлетворять краевому условию из (15.2.15). Если эту функцию продолжить на G_0, как об этом сказано в п. 4°, то уравнение (15.2.16) примет вид

$$\delta \iint_{G_i} [|\nabla u|^2 - 2c_0 u]\,dx\,dy = 0. \tag{15.2.17}$$

Обозначим через $\bar{H}_0$ подпространство пространства $W_2^{(1)}(G_0)$, образованное функциями, которые остаются постоянными в каждой из внутренних областей G_j и равны нулю на ∂G_0. Норму в $\bar{H}_0$ зададим формулой

$$|u|^2 = \iint_{G_0} |\nabla u|^2\,dx\,dy;$$

соответственно скалярное произведение есть интеграл

$$[u, v] = \iint_{G_0} \nabla u \cdot \nabla v\,dx\,dy.$$

Вариационная задача (15.2.15) приводится к виду

$$|u - u_0| = \min, \qquad u \in \bar{H}_0,$$

где u_0 в соответствии с теоремой Ф. Риса определяется тождеством

$$\forall u \in \bar{H}_0, \qquad [u, u_0] = c_0 \iint_{G_0} u\,dx\,dy.$$

Применяя принцип Хаара — Кармана, мы сведем задачу упругопластического кручения к следующей односторонней вариа-

ционной задаче:

$$|u-u_0|=\min, \quad u\in\bar{M};$$
$$\bar{M}=\{u\in\bar{H}_0:|||u|||\leqslant K\}; \tag{15.2.18}$$
$$|||u|||=\sup_G \operatorname{ess}|\nabla u|.$$

Задача (15.2.18) сформулирована в статье Эллен Ланшон [178]. При любом конечном значении c_0 эта задача имеет одно и только одно решение, которое мы обозначим через $u_*=u_*(x, y, c_0)$. В [178] доказывается, что это решение удовлетворяет всем дифференциальным уравнениям и краевым условиям задачи упругопластического кручения.

Если задан крутящий момент, то c_0 следует определить из уравнения

$$2\iint_G u_*(x, y, c_0)\,dx\,dy=Q. \tag{15.2.19}$$

Каждому конечному значению c_0 соответствует по формуле (15.2.19) одно и только одно допустимое значение Q. Обратно, как видно из результатов пп. 2—4°, каждому допустимому значению Q из полуинтервала $[0, \bar{Q})$ соответствует одно и только одно конечное значение c_0, и при любом значении $Q\in[0, \bar{Q})$ уравнение (15.2.19) имеет одно и только одно решение.

6°. Займемся вопросом об устойчивости решения задачи (15.2.18) относительно малых возмущений области G. Наряду с задачей (15.2.18) рассмотрим такую же задачу для стержня из того же материала и с нормальным сечением G', близким к G. Близость областей G и G' определим следующим образом. Обозначим через ξ, η и x, y декартовы координаты в областях G' и G соответственно и допустим, что существует преобразование координат S, которое взаимно-однозначно переводит G' в G и G'_0 в G (здесь G'_0 наименьшая односвязаная область, содержащая область G'), причем якобиева матрица преобразования S имеет вид $I+O(\varepsilon)$, где I — единичная матрица, а ε — малое положительное число. Тогда, если J — якобиан преобразования S, то $J=1+O(\varepsilon)$.

Задачу кручения для области G' можно записать так:

$$\iint_{G'_0}\left[\left(\frac{\partial u'}{\partial\xi}\right)^2+\left(\frac{\partial u'}{\partial\eta}\right)^2-2c'_0u'\right]d\xi\,d\eta=\min, \quad u'\in\bar{H}'_0;$$
$$\sup_{G'}\operatorname{ess}\left[\left(\frac{\partial u'}{\partial\xi}\right)^2+\left(\frac{\partial u'}{\partial\eta}\right)^2\right]\leqslant K^2; \tag{15.2.20}$$

здесь $\bar{H}'_0$ — подпространство пространства $\overset{\circ}{W}{}_2^{(1)}(G'_0)$, образованное функциями, постоянными в каждой из связных компонент множества $G'_0\setminus G'$.

В (15.2.20) выполним преобразование S; заметим, что оно переводит $\bar{H}_0'$ в $\bar{H}_0$. Обозначим $\nabla' = (\partial/\partial\xi,\ \partial/\partial\eta)$, тогда

$$\iint\limits_{G_0} \{(u_{/x})^2 |\nabla' x|^2 + 2u_{/x}u_{/y}(\nabla' x,\ \nabla' y) + \\ + (u_{/y})^2 |\nabla' y|^2 - 2c_0' u\}\, dx\, dy = \min,\quad u(x,\ y) = u'(\xi,\ \eta),\quad u \in \bar{H}; \tag{15.2.21}$$

$$\sup_G \operatorname{ess} \{(u_{/x})^2 |\nabla' x|^2 + 2u_{/x}u_{/y}(\nabla' x,\ \nabla' y) + (u_{/y})^2 |\nabla' y|^2\} \leqslant K^2.$$

В пространстве $\bar{H}_0$ введем новую норму $|\cdot|_1$ и соответственно новое скалярное произведение $[\ ,\ ]_1$, положив

$$|u|_1^2 = \iint\limits_{G_0} \{(u_{/x})^2 |\nabla' x|^2 + 2u_{/x}u_{/y}(\nabla' x,\ \nabla' y) + (u_{/y})^2 |\nabla' y|^2\}\, J\, dx\, dy. \tag{15.2.22}$$

Функционал

$$l_1 u = \int_{G_0} c_0' J u\, dx\, dy$$

ограничен в норме (15.2.22), поэтому $l_1 u = [u, u_1]_1$, где u_1 — некоторый элемент из $\bar{H}_0$. Наконец, обозначим через M' множество функций из $\bar{H}_0$, удовлетворяющих последнему из соотношений (15.2.21), и через $\|\!|u|\!\|_1$ — левую часть этого соотношения. Задача (15.2.22) принимает вид

$$|u - u_1|_1 = \min,\ u \in \overline{M}';\ \overline{M}' = \{u \in \bar{H}_0 : \|\!|u|\!\|_1 \leqslant K\}. \tag{15.2.23}$$

7°. При переходе от задачи (15.2.18) к задаче (15.2.23) меняются все параметры, определяющие задачу: элемент u_0, норма $|\cdot|$ и множество $\bar{M}$. Докажем, что все эти изменения малы в смысле определений § 5 главы 14.

Прежде всего докажем, что нормы $|\cdot|$ и $|\cdot|_1$ связаны неравенством вида (14.5.3). Выражение под знаком интеграла в (15.2.22) есть квадратичная форма относительно $u_{/x}$ и $u_{/y}$; нетрудно видеть, что ее собственные числа суть $\lambda_k = 1 + O(\varepsilon)$, $k = 1,\ 2$ В п. 6° было отмечено, что $J = 1 + O(\varepsilon)$. Из сказанного следует существование такой постоянной δ_1, зависящей только от ε, что $\lim\limits_{\varepsilon \to 0} \delta_1 = 0$ и $(1 - \delta_1)^2 \leqslant J\lambda_k \leqslant (1 + \delta_1)^2$, $k = 1,\ 2$. А тогда

$$(1 - \delta_1)|u| \leqslant |u_1| \leqslant (1 + \delta_1)|u|.$$

Теперь оценим величину $|u_0 - u_1|$. Рассмотрим разность

$$[u,\ u_1] - [u,\ u_1]_1 = \\ = \iint_{G_0} \{u_{/x}u_{1/x}(1 - J|\Delta' x|^2) + u_{/y}u_{1/y}(1 - J|\nabla' y|^2) - \\ - (u_{/x}u_{1/y} + u_{/y}u_{1/x}) J (\nabla' x,\ \nabla' y)\}\, dx\, dy.$$

Из последнего тождества находим $[u, u_1]_1 - [u, u_1]_1 = |u| \cdot |u_1| \times$ $\times O(\varepsilon)$ и, следовательно,

$$[u, u_0 - u_1] = [u, u_0] - [u, u_1]_1 + |u| \cdot |u_1| \cdot O(\varepsilon).$$

Далее, принимая, что $c_0 - c_0' = O(\varepsilon)$, получаем

$$[u, u_0] - [u, u_1]_1 = c_0 \iint_{G_0} u\, dx\, dy - c_0' \iint_{G_1} uJ\, dx\, dy =$$

$$= (c_0 - c_0') \iint_{G_0} u\, dx\, dy + c_0' \iint_{G_0} u(1 - J)\, dx\, dy = |u| \cdot O(\varepsilon)$$

и $[u, u_0 - u_1] = |u|(1 + |u_1|) O(\varepsilon)$. Полагая здесь $u = u_0 - u_1$, находим

$$|u_0 - u_1| \leqslant C\varepsilon(1 + |u_1|), \quad C = \text{const}. \tag{15.2.24}$$

Отсюда $|u_1| - |u_0| \leqslant C\varepsilon(1 + |u_1|)$; $|u_1| \leqslant |u_0| + O(\varepsilon)$. Теперь то же неравенство (15.2.24) дает $|u_0 - u_1| = O(\varepsilon)$; существует, следовательно, такая величина δ_2, зависящая только от ε, что

$$\lim_{\varepsilon \to 0} \delta_2 = 0, \quad |u_0 - u_1| \leqslant \delta_2.$$

8°. Исследуя влияние замены множества $\overline{M}$ на $\overline{M}'$, примем, что $G' \subset G$. Ниже это ограничение будет снято. Введенные выше пространства $\overline{H}_0$ и $\overline{H}_0'$ будем теперь обозначать через $H_0(G)$ и $H_0(G')$. Функции класса $H_0(G')$ доопределим в $G_0 \setminus G_0'$, положив их там равными нулю; после этого $H_0(G')$ станет подпространством пространства $H_0(G)$.

Выше мы определили S как преобразование координат; перенесем это определение также на функции, заданные в соответствующих областях: если функция $u(\xi, \eta)$ определена в G_0', то положим $(Su)(x, y) = u(\xi(x, y), \eta(x, y))$, где $(x, y) = S(\xi, \eta)$. При таком определении S переводит $\overline{M}'$ в $\overline{M}$; докажем, что при этом выполняется неравенство (14.5.9). Пусть $u \in \overline{M}$. При этом

$$S^{-1}u = \begin{cases} u(\xi(x, y), \eta(x, y)); & (\xi(x, y), \eta(x, y)) \in G_0', \\ 0, & (\xi(x, y), \eta(x, y)) \in G_0 \setminus G_0' \end{cases}$$

и

$$|u - S^{-1}u| = \iint_{G_0 \setminus G_0'} |\nabla u|^2\, dx\, dy +$$

$$+ \iint_{G_0'} [\nabla'(u(\xi, \eta) - (S^{-1}u)(\xi, \eta))]^2\, d\xi\, d\eta.$$

Первый интеграл справа в силу условия пластичности Сен-Венана (или Мизеса) имеет порядок $O(\varepsilon)$. Второй интеграл ра-

вен

$$\iint_{G_0'} \{(u_{/\xi})^2[(1-\xi_{/x})^2+(\xi_{/y})^2]+2u_{/\xi}u_{/\eta}[(1-\xi_{/x})\eta_{/y}+$$

$$+\xi_{/y}(1-\eta_{/x})]+(u_{/y})^2[(\eta_{/x})^2+(1-\eta_{/y})^2]\}\,dx\,dy.$$

Якобиева матрица преобразования S имеет вид $I+O(\varepsilon)$; отсюда следует, что последний интеграл оценивается величиной $|u|^2 O(\varepsilon^2)$. Окончательно $|u-S^{-1}u|^2=O(\varepsilon)+|u|^2 O(\varepsilon^2)$ и $|u-S^{-1}u|=O(\sqrt{\varepsilon})+|u|O(\varepsilon)=(1+|u|)O(\sqrt{\varepsilon})$. Это означает существование такой величины δ_3, зависящей только от ε, что $\lim_{\varepsilon\to 0}\delta_3=0$ и $|u-S^{-1}u|\leqslant\delta_3(1+|u|)$.

Из доказанного в настоящем пункте следует, что задача упругопластического кручения стержня устойчива относительно малых возмущений области нормального сечения стержня, если возмущенная область лежит внутри первоначальной.

9°. Пусть область G' получена малым возмущением области G, но G' не вкладывается в G. Построим область $\tilde{G}$, которая может быть получена малым возмущением каждой из областей G и G' и притом так, что как G, так и G' вкладываются в $\tilde{G}$. По доказанному в п. 8° функции Прандтля u и u', соответствующие областям G и G', мало отличаются от $\tilde{u}$ — функции Прандтля для области $\tilde{G}$. Но тогда u и u' мало различаются между собой.

10°. На решение задачи упругопластического кручения естественным образом распространяются результаты § 2 и 6 главы 14. Таким образом, если приближенное решение $u_*^{(n)}$ определено по методу Ритца (в частности, по МКЭ) как элемент соответствующего подпространства H_n, а u_* есть точное значение функции Прандтля, то

$$|u_*-u_*^{(n)}|=O(\sqrt{e_n(u_*)});$$
$$e_n(u)=\inf_{u^{(n)}\in H_{(n)}}[|u-u^{(n)}|+|||u-u^{(n)}|||]; \qquad (15.2.25)$$

норма $|\cdot|$ и полунорма $|||\cdot|||$ определяются формулами

$$|u|^2=\iint_G |\nabla u|^2\,dx\,dy;\quad |||u|||=\operatorname*{sup\,ess}_G|\nabla u|.$$

Погрешность искажения приближенного решения имеет оценку $O(\sqrt{\|\Gamma_n\|}+\|\delta^{(n)}\|)$, где Γ_n — искажение матрицы Ритца, а $\delta^{(n)}$ — искажение столбца свободных членов системы Ритца.

11°. В литературе много внимания уделено приближенному решению задачи (15.2.18). В случае односвязной области этот вопрос обстоятельно изложен в [38]; некоторые результаты содержатся в более ранней книге [146]. Случаю многосвязной

области посвящена работа [170]. Оценки погрешностей в этих работах отсутствуют.

Упругопластическое кручение стержней с односвязным сечением изучалось как в точной, так и в приближенной постановке в ряде работ и независимо от вариационного принципа. Первой в этом ряду была знаменитая работа А. Надаи [121]. Интересные результаты в названном направлении были получены в [33, 34]. Отметим еще работу [123].

§ 3. ПЛОСКАЯ ЗАДАЧА

1°. Введем следующие обозначения: σ — постоянная Пуассона упругой среды,

$$T = \begin{pmatrix} \tau_{xx} & \tau_{xy} \\ \tau_{xy} & \tau_{yy} \end{pmatrix} \tag{15.3.1}$$

— тензор напряжений,

$$\tau_m = \left[4^{-1}(\tau_{xx} - \tau_{yy})^2 + \tau_{xy}^2\right]^{1/2} \tag{15.3.2}$$

— максимальное касательное напряжение, X и Y — составляющие объемных сил, X_ν и Y_ν — составляющие поверхностных усилий. Далее, G — конечная область в плоскости (x, y), $\partial G = \bigcup_{j=0}^{n} \partial G_j$ — граница области G, ∂G_j — ее связные компоненты. Тензор напряжений удовлетворяет внутри G уравнениям равновесия

$$\tau_{xx/x} + \tau_{xy/y} + X = 0, \quad \tau_{xy/x} + \tau_{yy/y} + Y = 0, \tag{15.3.3}$$

а на ∂G — краевым условиям

$$\begin{aligned} \tau_{xx}\cos(\nu, x) + \tau_{xy}\cos(\nu, y) &= X_\nu, \\ \tau_{xy}\cos(\nu, x) + \tau_{yy}\cos(\nu, y) &= Y_\nu. \end{aligned} \tag{15.3.4}$$

Если среда находится в упругом состоянии, то, по принципу Кастильяно, тензор упругих напряжений, который мы в этом случае обозначим через $\hat{g}$, сообщает минимальное значение интегралу

$$\iint_G \left[(1 - 2\sigma)(\tau_{xx} + \tau_{yy})^2 + 4\tau_m^2\right] dx\, dy \tag{15.3.5}$$

на множестве тензоров, удовлетворяющих условиям (15.3.3), (15.3.4). По условию пластичности Сен-Венана, в пластическом состоянии и в условиях плоской деформации выполняется равенство

$$\tau_m^2 = 4^{-1}(\tau_{xx} - \tau_{yy})^2 + \tau_{xy}^2 = K^2, \tag{15.3.6}$$

а тогда, по принципу Хаара — Кармана, тензор T упругопластических напряжений решает одностороннюю вариационную

задачу о минимуме интеграла (15.3.5) на множестве тензоров, удовлетворяющих уравнениям (15.3.3), (15.3.4) и, кроме того, неравенству

$$(\tau_{xx} - \tau_{yy})^2 + 4\tau_{xy}^2 \leqslant 4K^2. \qquad (15.3.7)$$

2°. Сформулированную здесь одностороннюю вариационную задачу сведем к задаче вида (14.1.1), (14.1.2). Введем гильбертово пространство $\bar{H}$ тензоров вида (15.3.1), составляющие которых квадратично суммируемы в G, с нормой

$$|T|^2 = \left\{\iint_G [(1-2\sigma)(\tau_{xx} - \tau_{yy})^2 + 4\tau_m^2]\, dx\, dy\right\}^{1/2} \qquad (15.3.8)$$

и соответствующим скалярным произведением, а также подпространство H этого пространства, образованное тензорами, удовлетворяющими однородным уравнениям равновесия

$$\tau_{xx|x} + \tau_{xy|y} = 0, \quad \tau_{xy|x} + \tau_{yy|y} = 0 \qquad (15.3.9)$$

и однородными краевым условиям

$$\tau_{xx}\cos(\nu, x) + \tau_{xy}\cos(\nu, y) = 0, \quad \tau_{xy}\cos(\nu, x) + \tau_{yy}\cos(\nu, y) = 0. \qquad (15.3.10)$$

Аналогично тому, как это сделано в [84] для трехмерной задачи теории упругости, можно доказать, что подпространство $\bar{H} \ominus H$ состоит из тензоров, которые удовлетворяют соотношениям (15.3.3), (15.3.4) при подходящих X, Y, X_ν, Y_ν и уравнениям плоской задачи теории упругости; в частности, $\hat{g} \in \bar{H} \ominus H$.

Обозначим через $\|\cdot\|$ полунорму

$$\|T\| = \sup_{(x,\, y) \in G} \operatorname{ess} \tau_m. \qquad (15.3.11)$$

Положим $\bar{g} = \hat{g} + V$, где V — произвольный тензор из H. Если $\inf_{V \in H} \|\bar{g}\| > K$, то односторонняя вариационная задача п. 1° неразрешима (см. главу 14, § 1, п. 2°). Будем считать, что $\inf_{V \in H} \|\bar{g}\| < K$. Тогда найдется тензор $V_0 \in H$, такой, что $\|g\| < K$, где $g = \hat{g} + V_0$. Обозначив $T = V - g$, $V \in H$, мы приведем задачу п. 1° к виду

$$|V - g| = \min, \quad V \in M; \quad M = \{V \in H : \|V - g\| \leqslant K\}. \qquad (15.3.12)$$

Последнюю задачу можно слегка упростить. Имеем $V - g = (V - V_0) - \hat{g}$; слагаемые справа ортогональны в метрике (15.3.8) и $|V - g|^2 = |V - V_0|^2 + |\hat{g}|^2$ Постоянная $|\hat{g}|^2$ не влияет на минимизирующий элемент V, и ее можно отбросить. В результате (15.3.12) заменяется следующей:

$$|V - V_0| = \min, \quad V \in M; \quad M = \{V \in H : \|V - g\| \leqslant K\}. \qquad (15.3.13)$$

Аналогично тому, как это сделано в п. 2° § 2, можно доказать, что множество M последней задачи замкнуто в норме $|\cdot|$.

Если граница области достаточно гладкая, то основные задачи линейной теории упругости можно свести к интегральным (Фредгольмовым или сингулярным) уравнениям; подробнее об этом см. [120, 81, 87, 74]. Несложный анализ этих уравнений показывает, что при подходящем выборе необходимых метрик тензор напряжений, решающий вторую краевую задачу теории упругости, устойчив в метрике C (или L_2) по отношению к малым изменениям объемных и поверхностных сил, а также области, занятой упругой средой.

Опираясь на этот факт, сделаем следующее допущение: при подходящем выборе метрик для объемных и поверхностных сил и при подходящем определении малого возмущения области G тензор напряжений $\hat{g}$ устойчив в метрике $C(\bar{G})$ относительно малых возмущений перечисленных параметров; очевидно, при этих же условиях тензор $\hat{g}$ устойчив в метрике (15.3.8).

При малых возмущениях тензора $\hat{g}$ и невозмущенном тензоре V_0 тензор $g = \hat{g} + V_0$ оказывается также мало возмущенным; неравенство $\|g\| < K$ при этом сохраняется. Отсюда следует, что тензор g устойчив как в метрике $C(\bar{G})$, так и в метрике (15.3.8) относительно малых возмущений внешних сил и области.

Если малое возмущение испытывают только внешние усилия, а область G остается неизменной, то элемент V_0 и метрика (15.3.8) остаются неизменными, меняется только множество M. Тогда из результатов пп 5—7° § 5 главы 14 вытекает, что тензор напряжений, решающий плоскую задачу упругопластического состояния, устойчив относительно малых возмущений вешних усилий, а также относительно малых возмущений постоянной пластичности K.

3°. Рассмотрим вопрос об устойчивости решения плоской задачи упругопластического состояния при малых возмущениях области G. Будем предполагать, что эти возмущения удовлетворяют требованиям, сформулированным в § 2. Временно предположим еще, что возмущенная область $G' \subset G$.

Объекты, связанные с областью G', будем обозначать теми же символами, что и аналогичные объекты, связанные с областью G, но со штрихом сверху. Возмущенная задача имеет вид

$$|V' - V_0'| = \min, \quad V' \in M'; \quad M' = \{V_0' \in H_0' : \|V' - g'\| \leqslant K\}. \tag{15.3.14}$$

Введем в рассмотрение преобразование S, описанное в § 2. Объекты, полученные в результате этого преобразования, будем отмечать индексом «1» внизу; в частности,

$$|T|_1^2 = \iint_G [(1 - 2\sigma)(\tau_{xx} + \tau_{yy})^2 + 4\tau_m^2] J\, dx\, dy, \tag{15.3.15}$$

где J — якобиан преобразования S. Из свойств этого преобразования и из формул (15.3.15) легко вытекает, что норма $|\cdot|_1$ удовлетворяет неравенствам вида (14.5.3), (14.5.9). Тензор $\hat{g}$ по предположению устойчив относительно малых возмущений области G; отсюда легко заключить, что такой же устойчивостью обладают тензоры g и V_0. Из результатов § 5 главы 14 следует, что тензор напряжений, решающий плоскую задачу упругопластического состояния, устойчив относительно малых возмущений области, если возмущенная область $G' \subset G$. Последнее требование несущественно — от него можно освободиться так же, как в п. 9° § 2.

4°. Для плоской задачи упругопластического состояния верны оценки, аналогичные оценкам п. 10° § 2, для погрешностей аппроксимации и искажения. Мы не будем останавливаться на этом подробнее.

5°. Поставим вопрос: при каких обстоятельствах можно утверждать, что решение плоской задачи упругопластического состояния, полученное из принципа Хаара — Кармана, является на самом деле решением этой задачи? По самой постановке задачи тензор T упругих напряжений, полученный на основе этого принципа, удовлетворяет уравнениям равновесия (15.3.9) — мы принимаем для простоты, что объемные силы равны нулю. Далее, как показано в статье А. Хаара и Т. Кармана [173], в упругой зоне этот тензор удовлетворяет уравнению Эйлера, которое в данном случае сводится к дифференциальному уравнению

$$\Delta(\tau_{xx} + \tau_{yy}) = 0 \qquad (15.3.16)$$

и к краевому условию (если границы области G и упругой зоны имеют общую часть) $T \cdot \nu = F_\nu$, где ν — внешняя нормаль к ∂G и F_ν — вектор напряжений, действующих на элементарную площадку границы ∂G; это условие равносильно системе краевых условий (15.3.4). В силу результатов той же статьи [173] в пластической зоне тензор T удовлетворяет условию пластичности Сен-Венана.

Таким образом, в упругой зоне тензор напряжений удовлетворяет системе трех дифференциальных уравнений (15.3.9) и (15.3.16). Как известно, если T удовлетворяет этой системе, то можно найти такой вектор смещений, что соответствующий тензор деформаций вместе с тензором T удовлетворяет уравнениям закона Гука.

В пластической зоне тензор T удовлетворяет системе трех уравнений (15.3.9) и (15.1.9). Этот тензор будет решать плоскую задачу упругопластического состояния, если в пластической зоне существует вектор смещений, который удовлетворяет уравнениям (15.1.10), (15.1.11) и вместе со своей производной по времени непрерывно переходит в вектор упругих смещений при переходе через общую границу упругой и пластической зон

[198]. Можно указать простое достаточное условие, при котором последнее требование будет выполнено, так что принцип Хаара — Кармана приводит к решению плоской задачи упругопластического состояния. Это будет, если данные задачи не зависят от времени. В этом случае тензор напряжений и вектор смещений в упругой зоне не зависят от времени; на границе раздела зон упругости и пластичности вектор скоростей смещений равен нулю. Но тогда можно удовлетворить уравнениям (15.1.10), (15.1.11), положив вектор скоростей смещений тождественно равным нулю в пластической зоне.

Что касается вектора смещений в пластической зоне, то он не определяется однозначно из уравнений Сен-Венана и его можно подчинить только двум требованиям: он не должен зависеть от времени и он должен непрерывно переходить в вектор упругих смещений на общей границе обеих зон.

Заметим, что высказанные в данном пункте соображения полностью относятся и к трехмерной задаче упругопластического состояния (см. ниже, § 4).

6°. В заключение данного параграфа заметим следующее. Полунорма (15.3.11) не непрерывно дифференцируема, поэтому вычисление приближенного решения наталкивается на некоторые трудности. Можно упростить задачу, заменив полунорму (15.3.11) следующей:

$$\|\!|T|\!\|_p = \frac{1}{|G|^{1/2p}} \|\tau_m\|_{2p} = \frac{1}{|G|^{1/2p}} \left\{ \iint_G \left[\frac{1}{4} (\tau_{xx} - \tau_{yy})^2 + \right. \right.$$

$$\left. \left. + \tau_{xy}^2 \right]^p dx\, dy \right\}^{1/2p}, \qquad (15.3.17)$$

а неравенство $\|\!|V - g|\!\| \leqslant K$ — неравенством

$$\|\!|V - g|\!\|_p \leqslant K; \qquad (15.3.18)$$

здесь p достаточно большое натуральное число; очевидно, полунорма (15.3.18) непрерывно дифференцируема, и при $p \to \infty$ она стремится к полунорме (15.3.11). Обозначим множество тензоров V, удовлетворяющих неравенству (15.3.18), через M_p. Так как $\|\!|\cdot|\!\|_p \leqslant \|\!|\cdot|\!\|$, то множество M_p не пусто, если не пусто множество M. Очевидно также, что M_p — выпуклое множество. Докажем, что оно замкнуто в норме $|\cdot|$. Пусть $\{V_n\} \subset M_p$ — последовательность, сходящаяся в упомянутой норме (формула (15.3.8)) к некоторому тензору V_0. Тогда $f_n \to f_0$ и $\varphi_n \to \varphi_0$ в $L_2(G)$, где

$$f_n = V_{nxx} - V_{nyy}, \quad \varphi_n = V_{nxy};$$

$$f_0 = V_{0xx} - V_{0yy}, \quad \varphi_0 = V_{0xy}.$$

Положим еще $g_{xx} - g_{yy} = \tilde{f}$, $g_{xy} = \tilde{\varphi}$, тогда в $L_2(G)$

$$f_n - \tilde{f} \to f_0 - \tilde{f}, \quad \varphi_n - \tilde{\varphi} \to \varphi_0 - \tilde{\varphi}. \qquad (15.3.19)$$

Отсюда вытекает существование такой последовательности $\{V_{n_k}\}$, что те же соотношения (15.3.19) выполняются для этой последовательности почти всюду в G. По известной теореме Фату (см., например, [33])

$$\|\!|V_0 - g|\!\|_p = \frac{1}{|G|^{1/2p}} \left\{\iint_G \left[\frac{1}{4}(f_0 - \tilde{f})^2 + (\varphi_0 - \tilde{\varphi})^2\right]^p dx\,dy\right\}^{1/2p} \leqslant$$

$$\leqslant \frac{1}{|G|^{1/2p}} \sup_k \left\{\iint_G \left[\frac{1}{4}(f_{n_k} - \tilde{f})^2 + (\varphi_{n_k} - \tilde{\varphi})^2\right]^p dx\,dy\right\}^{1/2p} =$$

$$= \sup_k \|\!|V_{n_k} - g|\!\| \leqslant K.$$

Это означает, что $V_0 \in M_p$ и, следовательно, множество M_p замкнуто.

Аналогично можно упростить задачи § 2 и 4 (см. ниже).

§ 4. ТРЕХМЕРНАЯ ЗАДАЧА

1°. Пусть известны векторы F — объемных сил и F_ν — поверхностных сил, действующих на упругопластическую среду, которая занимает конечную область Ω трехмерного пространства. Тензор напряжений T удовлетворяет в Ω дифференциальному уравнению

$$\operatorname{div} T + F = 0, \tag{15.4.1}$$

а на границе этой области — краевому условию

$$T \cdot \nu = F_\nu; \tag{15.4.2}$$

ν — внешняя нормаль к $\partial\Omega$.

Если вся среда находится в упругом состоянии, то, по принципу Кастильяно, тензор T сообщает минимальное значение функционалу

$$\iiint_\Omega \left[\frac{1-2\sigma}{6(1+\sigma)}(\tau_{xx} + \tau_{yy} + \tau_{zz})^2 + \tau_i^2\right] dx\,dy\,dz \tag{15.4.3}$$

на множестве тензоров, удовлетворяющих уравнениям (15.4.1), (15.4.2); через τ_i обозначена интенсивность касательных напряжений. Применяя принцип Хаара — Кармана, будем считать, что тензор напряжений в упругопластическом состоянии сообщает минимальное значение функционалу (15.4.3) на множестве векторов, которые удовлетворяют кроме уравнений (15.4.1), (15.4.2) еще и неравенству (15.1.23).

2°. Только что сформулированную одностороннюю вариационную задачу преобразуем к обычному виду. Примем, что тензор напряжений квадратично суммируем в Ω. На множестве таких тензоров введем норму

$$|T|^2 = \iiint_\Omega \left[\frac{1-2\sigma}{6(1+\sigma)}(\tau_{xx} + \tau_{yy} + \tau_{zz})^2 + \tau_i^2\right] dx\,dy\,dz \tag{15.4.4}$$

и обозначим полученное таким образом гильбертово пространство через $\bar{H}$. Выделим множество H элементов пространства $\bar{H}$, удовлетворяющих уравнениям

$$\operatorname{div} T = 0, \quad T \cdot \nu|_{\partial\Omega} = 0. \tag{15.4.5}$$

Как доказано в [84], H есть подпространство в $\bar{H}$. Ортогональное к H подпространство состоит из тензоров упругих напряжений: это значит, что если $T \in \bar{H} \ominus H$, то T удовлетворяет уравнениям (15.4.1), (15.4.2) при подходящем выборе векторов F и F_ν, и существует такой вектор смещений $u = (u(x), u(y), u(z))$, что

$$\tau_{xx} = \lambda \operatorname{div} u + 2\mu u_{x|x}, \quad \tau_{xy} = \mu (u_{x|y} + u_{y|x})$$

и т. п.; λ и μ — постоянные Ляме рассматриваемой упругой среды.

Пусть $\hat{g}$ — тензор напряжений, решающий для области Ω задачу теории упругости при заданных F и F_ν; составляющие этого тензора пусть будут $\hat{g}_{xx}, \hat{g}_{xy}, \ldots, \hat{g}_{zz}$. Ясно, что $\hat{g} \in \bar{H} - H$. Положив $T = V - \hat{g}$, мы сведем вариационную задачу п. 1° к следующей:

$$|V - \hat{g}| = \min, \quad V \in M = \{V \in H : |||V - \hat{g}||| \leqslant K\}, \quad |||T||| = \tau_l. \tag{15.4.6}$$

Как обычно, задача (15.4.6) неразрешима, если $\varkappa = \inf\limits_{V \in H} |||V - \hat{g}||| > K$; она разрешима единственным образом, если $\varkappa < K$; неразрешима или имеет неустойчивое решение, если $\varkappa = K$. Будем считать, что $\varkappa < K$. Тогда существует такой тензор $V_0 \in H$, что $|||g||| < K$, $g = V_0 + \hat{g}$. Обозначим $V = V' - V_0$. Тогда $V - \hat{g} = V' - g$. Далее, $|V - \hat{g}|^2 = |V|^2 + |\hat{g}|^2 = |V' - V_0|^2 + |\hat{g}|^2$; заменив обозначение V' на V, мы приведем задачу (15.4.6) к виду

$$|V - V_0| = \min, \quad V \in M; \quad M = \{V \in H : |||V - g||| \leqslant K\}. \tag{15.4.7}$$

Примем допущение, аналогичное тому, которое было принято при анализе плоской задачи: тензор упругих напряжений $\hat{g}$ устойчив в метриках C и (15.4.4) относительно малых возмущений векторов F и F_ν и области G. Тогда (ср. п. 3° § 2) относительно тех же возмущений устойчив и тензор g. Повторив рассуждения п. 3° § 2, мы убедимся, что решение трехмерной задачи упругопластического состояния устойчиво относительно малых возмущений постоянной пластичности K.

3°. Легко убедиться, что общая оценка погрешности аппроксимации (14.2.11) верна для задачи настоящего параграфа: если T_* — точное решение этой задачи, а $T_*^{(n)}$ — ее приближен-

ное решение, полученное заменой пространства H его конечномерным подпространством H_n, то

$$\|T_* - T_*^{(n)}\| = O(\sqrt{e_n(T_*)}), \quad e_n(T_*) = \inf_{T_n \in H_n} \|T_* - T_n\|; \qquad (15.4.8)$$

$$\|T\| = |T| + |||T|||.$$

Точно так же верна и общая оценка погрешности искажения (14.6.15).

4°. В заключение отметим некоторые работы, в которых исследуются вариационные неравенства, относящиеся к задачам упругопластического состояния. В [167] рассмотрены две такие задачи: о плоском напряженном состоянии и о кручении цилиндрического стержня; обе задачи исследованы в вариационной постановке для односвязной области. Каждая из названных задач сведена к уравнению с оператором, удовлетворяющим условию теоремы Брауэра о неподвижной точке. Аналогичный результат получен в [177] для трехмерной задачи. В работе [168] даны оценки устойчивости для решений некоторых вариационных неравенств; в качестве примера рассмотрено прямолинейное течение вязкопластической жидкости в трубе. В работе [169] принцип Хаара — Кармана использован в рамках теории пластичности Прандтля — Рейсса.

Глава 16

ЗАДАЧА УПРОЧНЕНИЯ В ТЕОРИИ ПЛАСТИЧНОСТИ; АПОСТЕРИОРНАЯ ОЦЕНКА ПОГРЕШНОСТИ

§ 1. ПОСТАНОВКА ЗАДАЧИ

В данном параграфе для облегчения ссылок коротко изложены некоторые хорошо известные факты из теории пластичности.

1°. Пусть изотропное тело, занимающее некоторую конечную область трехмерного пространства координат (x, y, z), находится в состоянии пластического упрочнения под действием вектора объемных сил $F = (X, Y, Z)$. Будем пользоваться обозначениями главы 15; кроме того, составляющие тензора деформации будем обозначать через ε_{xx}, ε_{xy} и т. п. Примем, что деформации малы, так что

$$\varepsilon_{xx} = 2u_{x|x}, \quad \varepsilon_{xy} = u_{x|y} + u_{y|x}, \ldots \qquad (16.1.1)$$

Согласно деформационной теории пластичности в состоянии пластического упрочнения напряжения и деформации связаны соотношениями

$$\tau_{xx} = \left(\frac{1}{3k} - \frac{2}{3}g\right) \operatorname{div} u + 2g\varepsilon_{xx}, \quad \tau_{xy} = g\varepsilon_{xy} \qquad (16.1.2)$$

и другими, получаемыми круговой перестановкой индексов x, y, z. В формулах (16.1.2) k — коэффициент упругого изменения объема, связанный с постоянной Пуассона σ и модулем Юнга E равенством

$$k = (1 - 2\sigma)/E. \tag{16.1.3}$$

Далее, $g = g(\Gamma^2)$ — функция от квадрата интенсивности деформаций сдвига, который определяется формулой

$$\Gamma^2(u) = \frac{2}{3}\Big\{(u_{x/x} - u_{y/y})^2 + (u_{y/y} - u_{z/z})^2 + (u_{z/z} - u_{x/x})^2 + \\ + \frac{3}{2}\left[(u_{x/y} + u_{y/x})^2 + (u_{y/z} + u_{z/y})^2 + (u_{z/x} + u_{x/z})^2\right]\Big\}; \tag{16.1.4}$$

функция g является характеристикой данного материала в состоянии пластического упрочнения. В общем случае эта функция подчинена неравенствам

$$0 < g(\xi^2) \leqslant G, \quad g'(\xi^2) < 0, \quad g(\xi^2) + 2g'(\xi^2)\,\xi^2 \geqslant 0, \tag{16.1.5}$$

где G — модуль сдвига материала в упругом состоянии. Мы примем, что функция g удовлетворяет более строгому неравенству

$$g_0 \leqslant g(\xi) \leqslant G, \tag{16.1.6}$$

g_0 — положительная постоянная, $\xi > 0$.

2°. Систему уравнений равновесия среды с упрочнением можно записать в виде векторного уравнения в смещениях

$$Au = f, \tag{16.1.7}$$

где

$$(Au)_x = -\frac{\partial}{\partial x}\left[\frac{1}{3k}\operatorname{div} u + 2g\left(u_{x/x} - \frac{1}{3}\operatorname{div} u\right)\right] - \\ - \frac{\partial}{\partial y}\left[g\,(u_{x/y} + u_{y/x})\right] - \frac{\partial}{\partial z}\left[g\,(u_{x/z} + u_{z/x})\right], \tag{16.1.8}$$

а $(Au)_y$ и $(Au)_z$ получаются круговой перестановкой индексов x, y, z. Для уравнений (16.1.7) обычно ставится одна из двух краевых задач. Первая краевая задача определяется условием

$$u|_{\partial\Omega} = 0, \tag{16.1.9}$$

вторая — условием

$$t(u)|_{\partial\Omega} = 0; \tag{16.1.10}$$

здесь $t(u)$ — вектор напряжений, действующих на элементарную площадку границы $\partial\Omega$.

В последующих параграфах настоящей главы мы излагаем с некоторыми изменениями содержание диссертации И. В. Алексиевой; см. статью [2].

§ 2. НЕКОТОРАЯ ОБЩАЯ СХЕМА ПОСТРОЕНИЯ РЕШЕНИЯ

1°. Пусть H — гильбертово пространство. Рассмотрим функционал $\Phi(u)$, определенный на всем пространстве H и обладающий следующими свойствами:

а) $P = \operatorname{grad}\Phi$ существует в каждой точке $u \in H$;

б) производная Гато P'_u существует в каждой точке $u \in H$ и область ее определения совпадает с H;

в) P'_u есть равномерно положительно определенный оператор в H; это значит, что P'_u симметричен и что справедливо неравенство

$$(P'_u h, h) \geqslant c_1 \|h\|^2, \tag{16.2.1}$$

где $c_1 > 0$ не зависит ни от u, ни от h;

г) $\Phi(0) = 0$, $P(0) = 0$.

Отметим, что из свойств б) и в) следует ограниченность оператора P'_u, так как симметричный оператор, определенный на всем пространстве, ограничен. Таким образом,

$$\|P'_u h\| \leqslant c_2 \|h\|, \quad \forall h \in H, \quad c_2 = \text{const} > 0. \tag{16.2.2}$$

Примем, что постоянная c_2 не зависит от u. Из неравенств (16.2.1), (16.2.2) следует двойное неравенство

$$c_1 \|h\|^2 \leqslant (P'_u h, h) \leqslant c_2 \|h\|^2. \tag{16.2.3}$$

2°. Докажем, что функционал $\Phi(u)$ непрерывен на H. Пусть $h \in H$ и $\|h\| \to 0$. Воспользуемся известными соотношениями, связывающими функционал с его градиентом и оператор с его производной Гато:

$$\Phi(u+h) - \Phi(u) = \int_0^1 (P(u+th), h)\, dt, \quad P = \operatorname{grad}\Phi, \tag{16.2.4}$$

$$P(u+h) - P(u) = \int_0^1 P'_{u+th} h\, dt. \tag{16.2.5}$$

Из этих формул следует

$$\Phi(u+h) - \Phi(u) = \int_0^1 \frac{d\tau}{\tau} \int_0^1 (P'_{u+t\tau h} \tau h, \tau h)\, dt + (P(u), h).$$

Отсюда по неравенству (16.2.3)

$$|\Phi(u+h) - \Phi(u)| \leqslant 2^{-1} c_2 \|h\|^2 + \|P(u)\| \cdot \|h\| \xrightarrow[h \to 0]{} 0.$$

Очевидно, что непрерывен и функционал

$$F(u) = \Phi(u) - (f, h), \tag{16.2.6}$$

где f — фиксированный элемент пространства H.

Так как функционал $\Phi(u)$ непрерывен на H, а производная Гато его градиента равномерно положительно ограничена снизу, то этот функционал существенно выпуклый (см., например,

[90, § 66]). Существенно выпуклым будет и функционал (16.2.6)

Докажем, что функционал (16.2.6) слабо полунепрерывен снизу в H. Достаточно доказать, что этим свойством обладает функционал $\Phi(u)$. Воспользовавшись неравенством (16.2.1), получим

$$\Phi(u+h)-\Phi(u)=\int_0^1\frac{d\tau}{\tau}\int_0^1\left(P'_{u+t\tau h}\tau h,\ \tau h\right)+(P(u),\ h)\geqslant$$

$$\geqslant\frac{c_1}{2}\|h\|^2+(P(u),\ h).$$

Пусть h слабо сходится к нулю. Тогда $(P(u), h)\to 0$, и из последнего неравенства вытекает, что

$$\varliminf_{h\to 0}\Phi(u+h)\geqslant\Phi(u). \tag{16.2.7}$$

3°. Для функционала $\Phi(u)$ справедливо двустороннее неравенство

$$2^{-1}c_3\|u\|^2\leqslant\Phi(u)\leqslant 2^{-1}c_4\|u\|^2;\quad c_3,\ c_4=\text{const}>0. \tag{16.2.8}$$

Действительно, используя соотношение между функционалом и его градиентом, справедливое, если $\Phi(0)=0$:

$$\Phi(u)=\int_0^1(P(tu),\ u)\,dt,$$

получаем

$$\Phi(u)=\int_0^1(P(tu),\ u)\,dt=\int_0^1(P(tu)-P(0),\ u)\,dt=$$

$$=\int_0^1 dt\int_0^1\left(P'_{\tau tu}tu,\ u\right)d\tau=\int_0^1 t\,dt\int_0^1\left(P'_{\tau tu}u,\ u\right)d\tau,$$

и из неравенства (16.2.1), (16.2.3) вытекает неравенство (16.2.8).

4°. Покажем, что функционал $F(u)$ — возрастающий в H. В силу правого неравенства (16.2.8) имеем $F(u)\leqslant 2^{-1}c_4\|u\|^2+\|f\|\cdot\|u\|$, откуда следует, что функционал F конечен, если конечна норма $\|u\|$. Применяя далее левое неравенство (16.2.8) и неравенство Коши, находим $F(u)\geqslant 2^{-1}c_3\|u\|^2-\|f\|\cdot\|u\|$ и, следовательно,

$$F(u)/\|u\|\xrightarrow[\|u\|\to\infty]{}\infty. \tag{16.2.9}$$

5°. Из теорем, доказанных в [90, § 66], и из свойств функционала F, доказанных выше, вытекает следующее утверждение.

Теорема 16.2.1. *Функционал $F(u)$ имеет в пространстве H абсолютный минимум, который достигается в единственной точке. К этой точке слабо сходится любая минимизирующая F последовательность.*

Пусть u_* есть решение задачи о минимуме функционала $F(u)$. Тогда (см. [90]) элемент u_* обращает в нуль оператор $\operatorname{grad} F$ и, следовательно, удовлетворяет уравнению

$$Pu = f. \tag{16.2.10}$$

Теорема 16.2.2. *Приближенные по Ритцу решения задачи о минимуме функционала F можно построить при любом n; для функционала F последовательность этих решений — минимизирующая.*

Функционал F непрерывен и полунепрерывен снизу в H; кроме того, он возрастающий в H. Тогда утверждение теоремы 16.2 2 вытекает из теоремы, доказанной в [88] (см. также [90, § 70]).

6°. Пусть функционал $\Psi(v)$ — встречный (см. § 7 главы 1) для F, так что $\Psi_{\min} = -F_{\min}$. Возьмем последовательности $\{u_n\}$, $\{v_m\}$, минимизирующие соответственно $F(u)$ и $\Psi(v)$. Воспользуемся тождеством (16.2.4), из которого следует, что

$$F(u_n) - F(u_*) = \int_0^1 (P(u_* + \tau h),\ h)\, d\tau - (f,\ h), \tag{16.2.11}$$

где u_* — точное решение уравнения (16.2.10) и $h = u_n - u_*$. Используя далее тождество (16.2.5) и неравенство (16.2.1), получаем

$$F(u_n) - F(u_*) \geqslant c \| u_n - u_* \|^2, \quad c = \text{const} > 0. \tag{16.2.12}$$

Отсюда и из цепочки соотношений

$$F(u_n) + \Psi(v_m) = F(u_n) - F(u_*) + F(u_*) + \Psi(v_m) \geqslant F(u_n) - F(u_*)$$

следует

$$\| u_n - u_* \| \leqslant c \sqrt{F(u_n) + \Psi(v_m)}, \tag{16.2.13}$$

что дает апостериорную оценку приближенных по Ритцу решений уравнения (16.2.10).

7°. Пусть H_n — подпространство пространства H, в котором строится приближенное решение по Ритцу u_n, и пусть $\mathscr{E}_n(u)$ — наилучшее приближение элемента u элементами подпространства H_n. По неравенству (16.2.12)

$$\forall u^{(n)} \in H_n, \quad \| u_n - u_* \|^2 \leqslant c^{-1} [F(u_n) - F(u_*)] \leqslant$$
$$\leqslant c^{-1} [F(u^{(n)}) - F(u_*)]. \tag{16.2.14}$$

Далее, справедливо тождество

$$F(u_* + h) - F(u_*) = \int_0^1 \frac{d\tau}{\tau} \int_0^1 (P'_{u_* + t\tau h} \tau h,\ \tau h)\, d\tau - (f,\ h). \tag{16.2.15}$$

Действительно,

$$F(u_* + h) - F(u_*) = \int_0^1 (P(u_* + th),\ h)\, d\tau - (f,\ h) =$$

$$= \int_0^1 (P(u_* + \tau h) - (P(u_*),\ h)\, d\tau = \int_0^1 \frac{d\tau}{\tau} \int_0^1 (P'_{u_* + t\tau h} \tau h,\ \tau h)\, dt.$$

Положив здесь $h = u^{(n)} - u_*$ и применив неравенство (16.2.3), найдем

$$\forall u^{(n)} \in H_n, \quad F(u^{(n)}) - F(u_*) \leqslant (c_2/2)\|u^{(n)} - u_*\|^2.$$

Подставим это в (16.2.14) и возьмем за $u^{(n)}$ тот элемент пространства H_n, на котором достигается равенство $\|u^{(n)} - u_*\| = \mathscr{E}_n(u_*)$. В результате получится оценка погрешности аппроксимации метода Ритца для задачи настоящего параграфа:

$$\|u_n - u_*\| \leqslant \sqrt{c_2/2c}\,\mathscr{E}_n(u_*). \tag{16.2.16}$$

§ 3. ПЕРВАЯ КРАЕВАЯ ЗАДАЧА

1°. Нетрудно убедиться, что первая краевая задача, определяемая уравнениями (16.1.7)—(16.1.9), равносильна задаче о минимуме функционала

$$F_1(u) = \Phi_1(u) - (u, f) = \int_0^1 (A(tu), u)\,dt - \int_\Omega uf\,d\Omega \tag{16.3.1}$$

на функциях, удовлетворяющих краевому условию (16.1.9) на $\partial\Omega$. Простые выкладки показывают, что

$$\Phi_1(u) = \int_0^1 (A(tu), u)\,dt = \frac{1}{6k}\int_\Omega (\operatorname{div} u)^2\,d\Omega + \\ + \int_\Omega N(u)\,d\Omega \int_0^1 g(t^2 N(u))\,t\,dt; \tag{16.3.2}$$

здесь и ниже $d\Omega = dxdydz$; $N(u) = \Gamma^2(u)$ — квадрат интенсивности деформации сдвига — определяется формулой (16.1.4).

Введем в рассмотрение гильбертово пространство H_1 с нормой

$$|u|^2 = \int_\Omega [u_{x/x}^2 + u_{y/y}^2 + u_{z/z}^2 + (u_{x/y} + u_{y/x})^2 + \\ + (u_{y/z} + u_{z/y})^2 + (u_{z/x} + u_{x/z})^2]\,d\Omega, \tag{16.3.3}$$

соответствующее скалярное произведение обозначим через $[u, v]$. Пространство H_1 — энергетическое пространство линейного оператора первой краевой задачи теории упругости при значениях постоянных Ляме $\lambda = 0$ и $\mu = 1/2$.

2°. Наша ближайшая цель — показать, что функционал $\Phi_1(u)$ удовлетворяет в H условиям а) — г) п. 1° § 2. Прежде всего отметим, что для функции g, удовлетворяющей неравенству (16.1.6), функционал Φ_1 удовлетворяет, в свою очередь, неравенству

$$\Phi_1(u) \leqslant \frac{1}{6k}\int_\Omega (\operatorname{div} u)^2\,d\Omega + \frac{G}{2}\int_\Omega [2(u_{x/x} + u_{y/y} + u_{z/z})^2 + \\ + (u_{x/y} + u_{y/x})^2 + (u_{y/z} + u_{z/y})^2 + (u_{z/x} + u_{x/z})^2]\,d\Omega \leqslant \\ \leqslant c_1 |u|^2, \quad c_1 = \text{const}.$$

В то же время

$$\Phi_1(u) \geqslant \frac{1}{6k}\int_\Omega (\operatorname{div} u)^2\, d\Omega + \frac{g_0}{2}\int_\Omega N(u)\, d\Omega \geqslant c_2 |u|^2, \qquad c_2 = \text{const}.$$

Таким образом,

$$c_2 |u|^2 \leqslant \Phi_1(u) \leqslant c_1 |u|^2. \tag{16.3.4}$$

Отметим еще, что $\Phi_1(0) = 0$.

3°. Докажем, что в любой точке $u \in H_1$ существует $P_1 = \operatorname{grad} \Phi_1$. Пусть u и h — произвольные элементы пространства H_1. Простые, хотя и несколько громоздкие выкладки показывают, что

$$\frac{d\Phi_1(u+\tau h)}{d\tau}\Big|_{\tau=0} = \frac{1}{3k}\int_\Omega \operatorname{div} u \operatorname{div} h\, d\Omega + \int_\Omega N(u, h)\, g(\Gamma^2(u))\, d\Omega, \tag{16.3.5}$$

$N(u, h)$ — билинейная форма, соответствующая квадратичной форме $N(u) = \Gamma^2(u)$.

Формула (16.3.5) показывает, что $d\Phi_1(u+\tau h)/d\tau|_{\tau=0}$ есть линейный функционал над $h \in H_1$, каков бы ни был элемент $u \in H_1$, а это означает, что $P_1 = \operatorname{grad} \Phi_1$ существует при любом $u \in H_1$. Нетрудно оценить норму функционала $P_1(u)$:

$$|[P_1(u), h]| \leqslant (3k)^{-1}\int_\Omega |\operatorname{div} u| \cdot |\operatorname{div} h|\, d\Omega + G\int_\Omega |N(u, h)|\, d\Omega \leqslant$$

$$\leqslant 3k^{-1}\Big[\int_\Omega (u_{x/x}^2 + u_{y/y}^2 + u_{z/z}^2)\, d\Omega \int_\Omega (h_{x/x}^2 + h_{y/y}^2 + h_{z/z}^2)\, d\Omega\Big]^{1/2} + G\Big[\int_\Omega N(u)\, d\Omega \int_\Omega N(h)\, d\Omega\Big]^{1/2},$$

отсюда следует, что $|P_1(u)| \leqslant c|u|$ и, в частности, $P(0) = 0$.

4°. Докажем теперь, что P_{1u}' существует в любой точке $u \in H_1$ и представляет собой равномерно положительно определенный оператор, заданный на всем пространстве H_1. Для этого рассмотрим выражение

$$\frac{d}{dt}(P_1(u + tv), h)|_{t=0}.$$

По формуле (16.3.5) находим

$$\frac{d}{dt}(P(u+tv), h)|_{t=0} = \int_\Omega \Big[\frac{1}{3k}\operatorname{div} v \operatorname{div} h + N(v, h)\, g(N(u)) + N(u, h) N(u, v)\, g'(N(u))\Big] d\Omega.$$

Правая часть последней формулы есть линейный функционал над h, ограниченный при любых фиксированных u, $v \in H_1$. Отсюда следует, что эта формула определяет линейный оператор над v, который и является производной Гато оператора P_1.

Обозначая его, как обычно, через $P'_{1u}v$, имеем

$$[P'_{1u}v,\ h] = \int_\Omega [(3k)^{-1} \operatorname{div} v \operatorname{div} h + N(v,\ h)\, g(N(u)) + \\ + N(u,\ h)\, N(u,\ v)\, g'(N(u))]\, d\Omega. \qquad (16.3.6)$$

Положим $v = h$:

$$[P'_{1u}h,\ h] = \int_\Omega [(3k)^{-1} (\operatorname{div} h)^2 + N(h)\, g(N(u)) + \\ + N^2(u,\ h)\, g'(N(u))]\, d\Omega. \qquad (16.3.7)$$

Как показывают неравенства (16.1.5) и (16.1.6),

$$g(N(u)) \geqslant g_0,\quad 2^{-1} g(N(u)) + N(u)\, g'(N(u)) \geqslant 0;$$

из соотношения (16.3.7) и из неравенства Корна теперь следует неравенство

$$[P'_{1u}h,\ h] \geqslant \int_\Omega [(3k)^{-1} (\operatorname{div} h)^2 + 2^{-1} g_0 N(h)]\, d\Omega \geqslant \\ \geqslant c\, |h|^2; \qquad c = \text{const} > 0. \qquad (16.3.8)$$

Утверждение настоящего пункта доказано.

5°. Результаты пп. 2° — 4° показывают, что в рассматриваемой задаче выполнены все предположения а) — г) § 2, и потому для этой задачи верны соответствующие выводы, сделанные в § 2. В частности, справедливы теоремы 16.2.1, 16.2.2, оценки погрешности аппроксимации метода Ритца и построение апостериорной оценки с помощью встречных функционалов. Последний вопрос мы подробнее рассмотрим в следующем параграфе.

§ 4. ВСТРЕЧНЫЙ ФУНКЦИОНАЛ

1°. Укажем функционал, встречный для функционала $F(u)$ (формула (16.3.1)). Это хорошо известный в теории пластичности функционал дополнительной работы. Напомним его определение.

Рассмотрим множество симметричных тензоров T, определенных и квадратично суммируемых в Ω; составляющие этих тензоров пусть будут $\tau_{xx}, \tau_{xy}, \ldots, \tau_{zz}$. С каждым тензором этого множества свяжем симметричный тензор $\varepsilon = \varepsilon(T)$ с составляющими $\varepsilon_{xx}, \varepsilon_{xy}, \ldots, \varepsilon_{zz}$, где

$$\varepsilon_{xx} = k\tau + 2^{-1} \bar{g}(\tau_i^2)(\tau_{xx} - \tau), \qquad \varepsilon_{xy} = \bar{g}(\tau_i^2)\, \tau_{xy}, \qquad (16.4.1)$$

а остальные составляющие получаются круговой перестановкой индексов x, y, z. Далее, $\tau = (\tau_{xx} + \tau_{yy} + \tau_{zz})/3$, а τ_i — интенсивность касательных напряжений — определяется формулой (15.1.18). Наконец, функция $\bar{g}$ определяется формулой

$$\bar{g}(\tau_i^2) = [g(\Gamma^2)]^{-1}, \qquad (16.4.2)$$

где Γ^2 следующим образом определяется по тензору ε:

$$\Gamma^2 = 6^{-1}[(\varepsilon_{xx}-\varepsilon_{yy})^2+(\varepsilon_{yk}-\varepsilon_{zz})^2+(\varepsilon_{zz}-\varepsilon_{xx})^2]+ \\ +\varepsilon_{xy}^2+\varepsilon_{yz}^2+\varepsilon_{zx}^2. \qquad (16.4.3)$$

Если ε есть тензор малых деформаций для некоторого вектора смещений u, то формула (16.4.3) совпадает с формулой (16.1.4) для интенсивности деформаций сдвига. Очевидно,

$$\bar{g}'(\tau_i^2)>0, \quad G^{-1}\leqslant \bar{g}(\tau_i^2)\leqslant g_\wedge^{-1}. \qquad (16.4.4)$$

Рассмотрим подмножество M_1 тензоров T, для которых определены почти всюду и суммируемы с квадратом в Ω выражения

$$\frac{\partial\tau_{xx}}{\partial x}+\frac{\partial\tau_{xy}}{\partial y}+\frac{\partial\tau_{xz}}{\partial z}, \quad \frac{\partial\tau_{xy}}{\partial x}+\frac{\partial\tau_{yy}}{\partial y}+\frac{\partial\tau_{yz}}{\partial z}, \\ \frac{\partial\tau_{xz}}{\partial x}+\frac{\partial\tau_{yz}}{\partial y}+\frac{\partial\tau_{zz}}{\partial z}. \qquad (16.4.5)$$

Множество M_1 превратим в предгильбертово пространство, введя в нем норму

$$\|T\|^2=\int_\Omega[\tau_{xx}^2+\tau_{yy}^2+\tau_{zz}^2+2(\tau_{xy}^2+\tau_{yz}^2+\tau_{xz}^2)]\,d\Omega \qquad (16.4.6)$$

и соответствующее скалярное произведение. Пополнив это пространство, получим гильбертово пространство, которое обозначим через H_1. В этом пространстве определим функционал

$$\Psi(T)=\int_\Omega\Big[\frac{3k}{2}\tau^2+\int_0^{\tau_i}\Gamma(\varepsilon)\,d\tau_i\Big]d\Omega. \qquad (16.4.7)$$

Согласно принципу дополнительной работы [69] функционал $\Psi(T)$ достигает минимума на тензоре напряжений T, который соответствует точному решению задачи (16.1.7)—(16.1.9).

2°. Докажем, что

$$\min\Psi(T)=-\min F(u), \qquad (16.4.8)$$

где F — функционал (16.3.1). Для этого достаточно вычислить значение $\Psi(T)$ на тензоре T, соответствующем упомянутому точному решению.

Из соотношений (16.1.2) вытекает, что

$$\operatorname{div}u=3k\tau, \quad \tau_i=g(\Gamma^2)\Gamma, \quad \Gamma=\Gamma(u_*).$$

Дифференцируя второе равенство, получаем

$$\Gamma\,d\tau_i=2^{-1}[g(\Gamma^2)+2\Gamma^2g'(\Gamma^2)]\,d\Gamma^2.$$

Отсюда

$$\Psi_{\min}=\int_\Omega\Big\{\frac{1}{6k}(\operatorname{div}u_*)^2+\frac{1}{2}\int_0^{N(u_*)}[g(\xi)+2\xi g'(\xi)]\,d\xi\Big\}d\Omega.$$

Последнюю формулу преобразуем. Выделим справа и возьмем по частям интеграл

$$\int_\Omega d\Omega \int_0^{\Gamma^2(u_*)} \xi g'(\xi)\, d\xi = \int_\Omega d\Omega \int_0^{N(u_*)} \xi\, dg(\xi) =$$
$$= \int_\Omega \xi g(\xi) \Big|_0^{N(u_*)} d\Omega - \int_\Omega d\Omega \int_0^{N(u_*)} g(\xi)\, d\xi.$$

Теперь

$$\Psi_{\min} = \frac{1}{6k} \int_\Omega (\operatorname{div} u_*)^2\, d\Omega - \frac{1}{2} \int_\Omega d\Omega \int_0^{N(u_*)} g(\xi)\, d\xi + \int_\Omega N(u_*)\, g(N(u_*))\, d\Omega. \tag{16.4.9}$$

Если в последнем интеграле справа в (16.3.2) сделать замену $\xi = t^2 N(u_*)$, то из (16.3.1) найдем

$$F_{1\min} = F_1(u_*) = \frac{1}{6k} \int_\Omega (\operatorname{div} u_*)^2\, d\Omega + \frac{1}{2} \int_\Omega d\Omega \int_0^{N(u_*)} g(\xi)\, d\xi - (u_*, f). \tag{16.4.10}$$

Далее, $(u_*, f) = (A(u_*), u_*)$, где A — оператор (16.1.8), или

$$(u_*, f) = \int_\Omega A(u_*)\, u_*\, d\Omega.$$

Интегрируя по частям и принимая во внимание краевое условие $u_*|_{\partial\Omega} = 0$, получаем

$$\int_\Omega A(u_*)\, u_*\, d\Omega = \frac{1}{3k} \int_\Omega (\operatorname{div} u_*)^2\, d\Omega + \int_\Omega N(u_*)\, g(\Gamma^2(u_*))\, d\Omega.$$

Подставим это в (16.4.10):

$$F_{1\min} = -\frac{1}{6k} \int_\Omega (\operatorname{div} u_*)^2\, d\Omega + \frac{1}{2} \int_\Omega d\Omega \int_0^{N(u_*)} g(\xi)\, d\xi - \int_\Omega N(u_*)\, g(N(u_*))\, d\Omega.$$

Сравнив последнюю формулу с (16.4.9), получим равенство (16.4.8). Этим доказано, что функционалы F_1 и Ψ — встречные. Как показано в § 2, отсюда вытекает апостериорная оценка погрешности для приближенного по Ритцу решения первой краевой задачи теории упрочнения.

Используя общую схему § 2, можно доказать, что метод Ритца, примененный к функционалу Ψ дополнительной работы, дает для этого функционала минимизирующую последовательность. Это создает возможность за счет выбора достаточно далекого элемента этой последовательности сделать апостериорную оценку сколь угодно точной.

§ 5. ВТОРАЯ КРАЕВАЯ ЗАДАЧА

В краевом условии (16.1.10) внешние силы, действующие на поверхность тела, равны нулю, и условия статики сводятся к следующим:

$$\int_\Omega f\, d\Omega = 0, \quad \int_\Omega r \times f\, d\Omega = 0. \tag{16.5.1}$$

Здесь r — радиус-вектор точки (x, y, z), а знак «$\times$» означает векторное произведение. Вектор смещений можно подчинить аналогичным условиям:

$$\int_\Omega u\, d\Omega = 0, \quad \int_\Omega r \times u\, d\Omega = 0. \tag{16.5.2}$$

Примем, что область Ω удовлетворяет известным условиям, описанным, в частности, в [84, глава IV]. Введем в рассмотрение гильбертово пространство H_2, элементы которого суть векторы смещений, удовлетворяющие условиям (16.5.2); норму в H_2 зададим той же формулой (16.3.3), что и для пространства H_1. Очевидно, что H_2 есть энергетическое пространство второй краевой задачи для линейного оператора теории упругости при постоянных Ляме $\lambda = 0$, $\mu = 1/2$. В силу условий (16.5.2) и условий, которым подчинена область Ω, для векторов пространства H_2 справедливо известное неравенство Корна, из которого вытекает вложение $H_2 \subset W_2^{(1)}(\Omega)$.

Скалярное произведение и норму в H_2 обозначим прежними символами $[\ ,\ ]$ и $|\cdot|$.

Зададим функционалы $\Phi_2(u)$ и $F_2(u)$, определенные на элементах пространства H_2 формулами (16.3.1), (16.3.2). Нетрудно установить неравенства

$$\varkappa_2 |u|^2 \leqslant \Phi_2(u) \leqslant \varkappa_1 |u|^2; \quad \varkappa_1,\ \varkappa_2 = \text{const} > 0. \tag{16.5.3}$$

Легко доказывается, что обобщенное (в H_2) решение второй задачи теории упрочнения сообщает минимальное значение функционалу F_2. Нетрудно также доказать, что задача о минимуме функционала F_2 подходит под общую схему § 2, и для этой задачи верны все выводы названного параграфа.

Введем в рассмотрение пространство $\bar{H}_2$, аналогичное пространству $\bar{H}_1$ (§ 4), но отличающееся от него тем, что тензоры из H_2 создают на $\partial\Omega$ усилия, равные нулю. Можно доказать, что функционал дополнительной работы, заданный на пространстве $\bar{H}_2$, является встречным для функционала F_2 и, следовательно, позволяет построить апостериорную оценку погрешности для второй краевой задачи теории упрочнения.

ДОПОЛНЕНИЕ

ИЛЛЮСТРАТИВНЫЕ ПРИМЕРЫ

Цель данного дополнения — на примерах, по возможности простых, показать, как можно фактически вычислять оценки погрешностей, исходя из соотношений основного текста настоящей монографии. Предполагается, что вычисление приближенных решений производится с повышенной точностью — при этом условии выведена бóльшая часть упомянутых оценок, а сами формулы для погрешностей оказываются достаточно простыми.

Автор стремился подобрать примеры так, чтобы оценки погрешностей можно было вычислить, не прибегая к большим ЭВМ и, в частности, к специальному программированию. Это, разумеется, возможно не всегда; автор хотел показать, что случаи, когда такое простое вычисление возможно, не так уже редки, а ведь оценки погрешности содержат важнейшую информацию о качестве полученного (обычно с затратой большого труда) приближенного решения. К этому можно добавить, что, по-видимому, в идеале любые машинные вычисления должны сопровождаться оценкой их погрешности. Заметим, что в ряде случаев использование больших ЭВМ и специального программирования может оказаться необходимым для оценки погрешностей. Это может произойти, например, при решении линейных алгебраических систем методом Гаусса (см. главу 5); если порядок системы достаточно велик, то оценка нормы $\|L\|$, от которой весьма существенно зависит погрешность метода Гаусса, может оказаться трудоемкой и невыполнимой вручную.

Дополнение содержит пять параграфов. В § 1 оценивается погрешность метода Гаусса для алгебраической системы, матрица которой заимствована из книги [26, ч. I]. Для этой задачи оценка нормы матрицы L по формулам главы 5 оказывается грубой, но использование специальной структуры ис-

ходной матрицы позволяет оценить $\|L\|$ гораздо точнее и без затраты больших усилий.

Погрешности решения той же системы оцениваются в § 2 для методов Холецкого, сопряженных направлений и наискорейшего спуска. В § 3 оцениваются погрешности классического метода Ритца для задачи Дирихле (уравнение Пуассона) в части кругового кольца, в § 4 рассмотрена первая краевая задача для обыкновенного дифференциального уравнения 2-го порядка, решаемая по МКЭ. Наконец, в § 5 оценивается погрешность решения некоторого уравнения Фредгольма с симметричным неотрицательным интегральным оператором.

§ 1. МЕТОД ГАУССА

1°. Рассмотрим уравнение

$$Ax = f, \tag{1.1}$$

где f — произвольный вектор евклидова пространства R_m, а матрица A порядка m имеет вид [26, задача № 50]

$$A = \begin{bmatrix} a+b & a & \dots & a \\ a & a+b & \dots & a \\ \dots & \dots & \dots & \dots \\ a & a & \dots & a+b \end{bmatrix}; \tag{1.2}$$

будем считать, что $a > 0$, $b > 0$. Обратная матрица A^{-1} получается из данной заменой a и b соответственно на

$$a' = -a/b\,(b + am), \quad b' = 1/b. \tag{1.3}$$

Оценим нормы $\|A\|$ и $\|A^{-1}\|$. Это можно сделать двумя способами. Первый способ был использован в главе 5: разобьем A на сумму матриц, каждая из которых имеет ненулевые элементы, равные соответственно $a + b$ или a, только на одной диагонали. Нормы этих матриц равны соответственно $a + b$ и a и ясно, что

$$\|A\| \leqslant a + b + 2(m-1)\,a = (2m-1)\,a + b. \tag{1.4}$$

Аналогично

$$\|A^{-1}\| \leqslant \frac{1}{b}\left(1 - \frac{a}{b+am}\right) + \frac{(2m-1)\,a}{b\,(b+am)} = \frac{b + (3m-2)\,a}{b\,(b+am)}. \tag{1.5}$$

Другой способ вытекает из неравенств $\|A\| \leqslant \|A\|_E$, $\|A^{-1}\| \leqslant \|A^{-1}\|_E$:

$$\|A\|^2 \leqslant m\,(a+b)^2 + (m^2-m)\,a^2; \tag{1.6}$$

$$\|A^{-1}\|^2 \leqslant m\,(a'+b')^2 + (m^2-m)\,a'^2 = \frac{m}{b^2}\left(1 - \frac{a}{b+am}\right)^2 + \frac{(m^2-m)\,a^2}{(b+am)^2}. \tag{1.7}$$

2°. Для определенности положим далее $a=b=1$, так что $a'=-1/(m+1)$, $b'=1$:

$$A=\begin{bmatrix} 2 & 1 & \dots & 1 \\ 1 & 2 & \dots & 1 \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & \dots & 2 \end{bmatrix},$$

$$A^{-1}=\begin{bmatrix} \frac{m}{m+1} & -\frac{1}{m+1} & \cdots & -\frac{1}{m+1} \\ -\frac{1}{m+1} & \frac{m}{m+1} & \cdots & -\frac{1}{m+1} \\ \cdot & \cdot & \cdot & \cdot \\ -\frac{1}{m+1} & -\frac{1}{m+1} & \cdots & \frac{m}{m+1} \end{bmatrix}. \qquad (1.8)$$

Формулы (1.4) — (1.7) дают следующие оценки:

$$\|A\|\leqslant 2m, \quad \|A\|\leqslant\sqrt{m^2+3m};$$
$$\|A^{-1}\|\leqslant\frac{3m-1}{m+1}, \quad \|A^{-1}\|\leqslant\frac{m^3+m^2-m}{(m+1)^2}.$$

Для $\|A\|$ выгоднее вторая оценка, для $\|A^{-1}\|$ — первая. Несколько увеличив правые части этих оценок, получим

$$\|A\|<m+2, \quad \|A^{-1}\|<3, \quad P_A<3m+6. \qquad (1.9)$$

Небезынтересно выяснить, насколько точны формулы (1.9). Матрица A — симметричная; в [26, ч. III] приведены значения ее наибольшего и наименьшего собственных чисел: $\lambda_{\max}=ma+b$, $\lambda_{\min}=b$. При $a, b>0$ они положительны, значит, матрица A — положительно определенная; отсюда $\|A\|=\lambda_{\max}=ma+b$. $\|A^{-1}\|=\lambda_{\min}^{-1}=1/b$. При выбранных здесь значениях $a=1$ и $b=1$ будет $\|A\|=m+1$, $\|A^{-1}\|=1$. Ниже мы будем пользоваться значениями (1.9), потому что они получены достаточно простыми средствами и довольно близки к точным значениям.

3°. Займемся прямым ходом по Гауссу. При $a=b=1$ матрица A дана в (1.8). Наибольший ее элемент расположен в левом верхнем углу, поэтому перенумерация строк и столбцов не нужна, и мы получаем $\alpha_{2j}^{(0)}=-1/2$, $j=2, 3, \dots, m$. После исключения первой неизвестной получится новая матрица

$$\left[\begin{array}{c|ccccc} 2 & 1 & 1 & 1 & \dots & 1 \\ \hline 0 & 3/2 & 1/2 & 1/2 & \dots & 1/2 \\ 0 & 1/2 & 3/2 & 1/2 & \dots & 1/2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 1/2 & 1/2 & 1/2 & \dots & 3/2 \end{array}\right].$$

Дальнейшей обработке подлежит матрица, отгороженная пунктирными линиями. Ее наибольший элемент опять расположен в левом верхнем углу, и мы находим $\alpha_{3j}^{(0)} = -1/3$, $j = 3, \ldots, m$. Нетрудно убедиться в справедливости общего соотношения

$$\alpha_{ij}^{(0)} = -1/i, \quad j = i, \ i+1, \ \ldots, \ m. \tag{1.10}$$

Формула (5.3.14) позволяет сразу записать матрицу L^{-1}:

$$L^{-1} = \begin{bmatrix} 1 & & & & \\ 1/2 & 1 & & & \\ 1/3 & 1/2 & 1 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1/m & 1/(m-1) & \ldots & 1/2 & 1 \end{bmatrix}. \tag{1.11}$$

Отсюда

$$\|L^{-1}\| \leqslant 1 + 1/2 + \ldots + 1/m < 1 + \ln m. \tag{1.12}$$

4°. Матрица L определяется формулой (5.2.5), а ее элементы, расположенные ниже первой боковой диагонали, — рекуррентной формулой (5.2.4), в которой надо положить $i = j + \varkappa + 1$ и, следовательно, $i \geqslant \varkappa + 2$. Пусть $\varkappa = 1$, тогда $i \geqslant 3$, $|\alpha_{ij}^{(0)}| \leqslant 1/3$; по формуле (5.2.4) $|\alpha_{ij}^{(1)}| \leqslant 1/3(1 + 1/3) < 1$. Допустим теперь, что при некотором индексе $\varkappa - 1 \geqslant 1$ будет $|\alpha_{ij}^{(\varkappa-1)}| \leqslant 1/(\varkappa - 1)$. Оценим $|\alpha_{ij}^{(\varkappa)}|$. Так как $i \geqslant \varkappa + 2$, то по формуле (1.10) $|\alpha_{ij}^{(0)}| \leqslant 1/(\varkappa + 2)$ и далее по формуле (5.2.4)

$$|\alpha_{ij}^{(\varkappa)}| \leqslant \frac{1}{\varkappa + 2}\left(\frac{1}{\varkappa - 1} - 1\right) = \frac{\varkappa}{\varkappa^2 + \varkappa - 2}.$$

Но $\varkappa \geqslant 2$, и в результате

$$\alpha_{ij}^{(\varkappa)}| \leqslant 1/\varkappa. \tag{1.13}$$

Теперь можно оценить $\|L\|$; это будет оценкой и для $\|L\|$ (определение см. в главе 1, § 2, п. 1°). Воспользуемся для этого приемом п. 3° § 2 главы 5: представим L в виде $L = \sum_{k=0}^{m-1} M_k$, где $M_0 = I$, а M_k, $k > 0$, определяется формулой (5.2.6). Очевидно,

$$\|M_k\| \leqslant \max_{1 \leqslant \nu \leqslant m-k} |\alpha_{k+\nu,\nu}^{(k-1)}| \leqslant \begin{cases} 1/2, & k = 1, \\ 1\,(k-1), & k \geqslant 2; \end{cases}$$

отсюда

$$\|L\| \leqslant \sum_{k=0}^{m-1} M_k \leqslant 1 + \frac{1}{2} + \sum_{k=2}^{m-1} \frac{1}{k-1} = \frac{5}{2} + \sum_{k=3}^{m-1} \frac{1}{k-1} \leqslant$$

$$\leqslant \frac{5}{2} + \ln(m-2). \tag{1.14}$$

Последняя оценка существенно лучше общей оценки (5.2.11).

5°. Приступим к вычислению оценок погрешностей. В целях большей конкретности результатов положим далее $m = 100$. По

формулам (5.3.3), (5.3.4) с учетом результатов формул (1.8), (1.13)

$$\|\Gamma\| \leqslant \varepsilon_1 (1+\sqrt{m}) \|\hat{L}\| \cdot \|A\|; \quad \|\delta\| \leqslant 2\varepsilon_1 \|\hat{L}\| \cdot \|f\|,$$

где Γ и δ суть погрешности соответственно матрицы и вектора свободных членов системы, полученной в результате прямого хода по Гауссу (системы (5.1.7)). Если, например, вычисления проводились на ЭВМ БЭСМ-6, то $\varepsilon_1 < 2 \cdot 10^{-12}$ и $\|\Gamma\| \leqslant 2 \cdot 10^{-8} \cdot \|\delta\| \leqslant 3 \cdot 10^{-11}$. Далее, из формул (5.3.7), (5.3.8) следует, что если $\|A^{-1}\| \cdot \|L^{-1}\| \cdot \|\Gamma\| = \beta < 1$, то погрешность искажения (погрешность решения, возникающая из погрешностей прямого хода по Гауссу) оценивается так:

$$\xi \leqslant \frac{\|A^{-1}\| \cdot \|L^{-1}\|}{1-\beta} [\|\Gamma\| \cdot \|A^{-1}\| \cdot \|f\| + \|\delta\|].$$

В данном случае, как показывают совсем простые подсчеты, $\beta < 2 \cdot 10^{-5}$. Пренебрегая в формуле для ξ достаточно малой величиной β по сравнению с единицей, получаем, что $\xi < 1{,}08 \cdot 10^{-6} \|f\|$; мы пренебрегли здесь также вторым слагаемым в квадратной скобке, так как оно существенно меньше первого.

6°. Погрешность округления (погрешность обратного хода) оценивается формулой (5.4.4)

$$\|\gamma\| \leqslant (1-\beta)^{-1} \|A^{-1}\| \cdot \|L^{-1}\| \cdot \|\alpha\|,$$

причем $\|\alpha\| \leqslant \varepsilon_1 b_0 \|A^{-1}\| \cdot \|f\|$, где $b_0 = \max |b_{kk}|$ и числа b_{kk} суть диагональные элементы матрицы $U = LA$. Имеем (a_{ik} — элементы матрицы (1.8))

$$b_{kk} = \sum_{j=1}^{m} l_{kj} a_{jk},$$

где l_{kj} суть элементы матрицы L. Из формулы (1.13) видно, что элементы матрицы L, лежащие на диагонали с номером $\varkappa > 2$, не превосходят единицы по модулю; тем же свойством обладают и элементы первых двух диагоналей. Таким образом, $|l_{kj}| \leqslant 1$ и $b_{kk} \leqslant m+1$. В нашем примере $b_0 \leqslant 101$. Теперь $\|\alpha\| \leqslant 2{,}02 \cdot 10^{-10} \|f\|$. Пренебрегая опять малой величиной β по сравнению с единицей, получаем

$$\|\gamma\| \leqslant 36{,}36 \cdot 10^{-10} \|f\|; \qquad (1.15)$$

здесь использована формула (1.12), по которой при $m = 100$ будет $\|L^{-1}\| < 6$.

Общая погрешность δ решения системы по методу Гаусса не превосходит величины $\xi + \|\gamma\|$. В нашем примере $\|\gamma\|$ имеет оценку, существенно меньшую, чем ξ, поэтому с точностью до малых высшего порядка общая погрешность δ решения нашего примера по методу Гаусса имеет оценку

$$\delta \leqslant 10^{-6} \|f\|. \qquad (1.16)$$

§ 2. ДРУГИЕ ТОЧНЫЕ МЕТОДЫ РЕШЕНИЯ ЛИНЕЙНОЙ АЛГЕБРАИЧЕСКОЙ СИСТЕМЫ

1°. Будем рассматривать тот же пример, что и в § 1. Матрица (1.2) — положительно определенная, если $a > 0$, $b > 0$, и нашу систему можно решать по методу Холецкого. Формула (6.2.15) дает оценку полной погрешности названного метода. Напомним, что в формуле (6.2.15) $\beta = \varepsilon_1 \sqrt{P_A}$, $c_0 = \max s_{kk}$ и что в рассматриваемом примере $m = 100$. Как было выяснено в § 1, $\|A^{-1}\| < 3$, $\|A\| < 102$ и, следовательно, $P_A < 306$. Чтобы воспользоваться формулой (6.2.15), достаточно оценить c_0. Из формулы (6.2.3) следует; что $s_{kk} \leqslant \sqrt{2}$, и можно принять $c_0 = 1{,}42$. Далее, $\sqrt{P_A} < 18$ и $\beta < 36 \cdot 10^{-12}$. Число β — достаточно малое, и при вычислениях по формуле (6.2.15) мы им пренебрежем. Теперь простые вычисления дают оценку

$$\|\delta(x)\| < 2 \cdot 10^{-9}. \tag{2.1}$$

2° Применим к нашей системе метод сопряженных направлений, погрешности которого исследованы в § 4 главы 6. Матрица (1.8), как уже было отмечено, — симметричная и положительно определенная, поэтому, применяя метод сопряженных направлений, можно принять $B = C = I$ и $K = A$. Если ω_k, $k = 1, 2, \ldots, m$, — система «A-ортонормированных» векторов, то решение имеет вид

$$x = \sum_{k=1}^{m} a_k (f, \omega_k). \tag{2.2}$$

Как и выше, полагаем $a = b = 1$, $m = 100$. Так же, как в § 4 главы 6, систему векторов $\{\omega_k\}$ будем строить через ортогонализацию системы $\{\varphi_k\}$, $\varphi_k = e_k / |e_k|$, где e_k — координатные векторы, и будем считать, что нормы $|e_k|$ и векторы φ_k вычислены столь точно, что их погрешностями можно пренебречь. Как и выше, будем считать, что вычисления производятся на ЭВМ БЭСМ-6 и что все промежуточные вычисления производятся с повышенной точностью, так что верна формула (1.1.9). Было выяснено, что формула (6.4.18), дающая оценку погрешности метода сопряженных направлений, верна с точностью до величин высшего порядка малости, если $P_A^3 \varepsilon_1 \ll 1$. В нашем примере $P_A < 306$ и $P_A^3 \varepsilon_1 < 5{,}3 \cdot 10^{-5}$. Последняя величина достаточно мала, и формула (5.4.18) применима. По этой формуле легко находим

$$\|\delta\|(x) < 5{,}6 \cdot 10^{-9} \|f\|. \tag{2.3}$$

3°. В данном пункте мы применим метод наискорейшего спуска к уравнению с матрицей (1.8), но произвольного порядка m. Как об этом сказано в § 1, матрица A имеет только два различных собственных числа, поэтому (см. главы 7, § 3,

п. 1°) по методу наискорейшего спуска достаточно сделать только две итерации.

Формула (3.8.7) дает оценку погрешности искажения n-й итерации. Нас в данном случае интересуют только два значения $n = 1, 2$; примем еще во внимание, что $\xi_0 = 0$, потому что начальное приближение не вычисляется, а задается. Теперь

$$\xi_1 \leqslant |\vartheta_1|(c_0 M + \|f\|),$$
$$\xi_2 \leqslant \left(\frac{M-\mu}{\mu} + |\vartheta_2| M\right)\xi_1 + |\vartheta_2|(c_0 M + \|f\|); \tag{2.4}$$

обозначения пояснены в главе 3, § 8.

В нашем примере $M = \lambda_{\max} = m + 1$, $\mu = \lambda_{\min} = 1$. Примем, что начальное приближение $x_0 = 0$. Из формулы (3.8.3) легко получить верхнюю границу для $\|x^{(n)}\|$, которую можно принять за c_0. Именно по формуле (3.8.3), принимая во внимание значение $\mu = 1$, находим

$$\|x^{(n)}\| \leqslant \|x_*\| + \|y^{(1)}\| \leqslant \|A^{-1}\| \cdot \|f\| + \|y^{(1)}\| = \|f\| + \|y^{(1)}\|.$$

Вычислим $x^{(1)}$. Имеем

$$x^{(1)} = Ax^{(0)} - \sigma_0 y^{(0)} = -\sigma_0 y^{(0)} = -y^{(0)}\|y^{(0)}\|^2/(Ay^{(0)},\ y^{(0)}).$$

Но $y^{(1)} = Ax^{(0)} - f = -f$, поэтому $x^{(1)} = \|f\|^2 f/(Af,\ f)$. Теперь

$$y^{(1)} = Ax^{(1)} - f = \frac{\|f\|^2}{(Af,\ f)} Af - f = \left(\frac{\|f\|^2}{(Af,\ f)} A - I\right) f$$

и

$$\|y^{(1)}\| \leqslant \left\|\frac{\|f\|^2}{(Af,\ f)} - I\right\| \cdot \|f\|.$$

Верхняя грань оператора справа не превосходит $\lambda_{\max} - 1 = m$. Отсюда $\|y^{(1)}\| \leqslant m\|f\|$, $\|x^{(n)}\| \leqslant (m+1)\|f\|$, и можно положить $c_0 = (m+1)\|f\|$.

Допустим теперь, что σ_0 и σ_1 вычислены с относительной погрешностью ε_1, так что $|\vartheta_k| \leqslant \sigma_{k-1}\varepsilon_1$, $k = 1, 2$. Далее, $\sigma_{n-1} = \|y^{(n-1)}\|^2/(Ay^{(n-1)},\ y^{(n-1)}) \leqslant 1/\lambda_{\min} = 1$. Отсюда следует в силу формул (2.4)

$$\xi_1 \leqslant \varepsilon_1 [(m+1)^2 + 1] \cdot \|f\|,$$
$$\xi_2 \leqslant (m + \varepsilon_1 (m+1))[(m+1)^2 + 1]\varepsilon\|f\| + \varepsilon_1 [(m+1)^2 + 1]\|f\|.$$

Отбросив члены высшего порядка малости, приходим к следующей оценке погрешности искажения в нашем примере:

$$\xi_2 \leqslant \varepsilon_1 (m+1)[(m+1)^2 + 1] \cdot \|f\|. \tag{2.5}$$

В частности, если $m = 100$, то $\xi_2 < 1{,}9 \cdot 10^{-6}\|f\|$.

Обратимся к погрешности округления. Она определяется формулой (4.4.4), в которой следует положить $n = 2$; кроме того, можно считать, что справедливо приведенное выше неравенство $|\vartheta_k| \leqslant \varepsilon_1 \sigma_{k-1}$, а в формуле (4.4.2) можно пренебречь

членом с ε'. В результате получаем $\tau = (M - \mu)/\mu$ и окончательно по формуле (4.4.4)

$$|\gamma^{(2)}| \leqslant (\tau + 1)\varepsilon_1. \tag{2.6}$$

Если $m = 100$, то $\|\gamma^{(2)}\| < 1{,}8 \cdot 10^{-10}$; в этом случае погрешность округления заметно меньше погрешности искажения, и допустимо считать, что общая погрешность метода наискорейшего спуска оценивается правой частью формулы (2.5).

§ 3. КЛАССИЧЕСКИЙ МЕТОД РИТЦА

1°. В плоскости декартовых координат t_1, t_2 рассмотрим краевую задачу

$$t = (t_1, t_2) \in \Omega, \quad \frac{\partial^2 x}{\partial t_1^2} + \frac{\partial^2 x}{\partial t_2^2} = f(t); \quad t \in \partial\Omega, \ x = 0; \tag{3.1}$$

область Ω есть четверть кругового кольца, изображенная на рис. 6. В этой задаче кажется целесообразным перейти к полярным координатам: $t_1 = r\cos\vartheta$, $t_2 = r\sin\vartheta$. В свою очередь будем рассматривать r и ϑ как декартовы координаты некоторой другой плоскости. Задача (3.1) переходит в следующую:

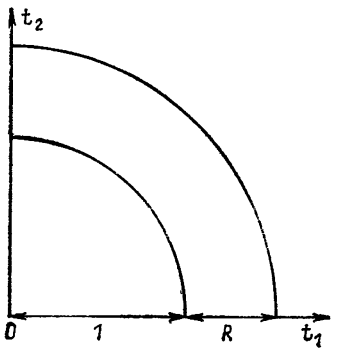

Рис. 6.

$$(r, \vartheta) \in \Pi, \quad \frac{\partial}{\partial r}\left(r\frac{\partial x}{\partial r}\right) + \frac{1}{r}\frac{\partial^2 x}{\partial \vartheta^2} = rf(t) =: \bar{f}(r, \vartheta);$$
$$(r, \vartheta) \in \partial\Pi, \quad x = 0; \tag{3.2}$$

Π — прямоугольник $1 < r < 1 + R$, $0 < \vartheta < \frac{\pi}{2}$.

Обозначим через A оператор задачи (3.2), а через B — оператор, определенный соотношениями

$$(r, \vartheta) \in \Pi, \quad Bx = \frac{\partial^2 x}{\partial r^2} + \frac{\partial^2 x}{\partial \vartheta^2}; \quad (r, \vartheta) \in \partial\Pi, \quad x = 0. \tag{3.3}$$

Операторы A и B — сходные [90], и их энергетические нормы эквивалентны:

$$|x|_A^2 = \int_1^{1+R}\int_0^{\pi/2}\left[r\left(\frac{\partial x}{\partial r}\right)^2 + \frac{1}{r}\left(\frac{\partial x}{\partial \vartheta}\right)^2\right] dr\, d\vartheta \leqslant$$
$$\leqslant (1+R)|x|_B^2 = (1+R)\int_1^{1+R}\int_0^{\pi/2}\left[\left(\frac{\partial x}{\partial r}\right)^2 + \left(\frac{\partial x}{\partial \vartheta}\right)^2\right] dr\, d\vartheta \leqslant$$
$$\leqslant (1+R)^2|x|_A^2. \tag{3.4}$$

Система собственных функций оператора B, ортонормированная и полная в энергетической норме $|\cdot|_B$, есть

$$\varphi_{mn}(r,\ \vartheta) = c_{mn} \sin \frac{(r-1)\,m\pi}{R} \sin 2n\vartheta, \qquad m,\ n = 1,\ 2,\ \ldots, \tag{3.5}$$

где

$$c_{mn} = 2\sqrt{2R}/\sqrt{m^2\pi^3 + 4R^4 n^2 \pi}. \tag{3.6}$$

2°. В силу теоремы 4.3 из [90] система функций (3.5), (3.6), ортонормированная в энергетической метрике оператора B, почти ортонормирована в энергетической метрике оператора A. Пусть A_{mn} — матрица чисел $[\varphi_{ij}, \varphi_{kl}]_A$, $1 \leqslant i,\ k \leqslant m$, $1 \leqslant j,\ l \leqslant n$; она же — матрица Ритца задачи (3.2), и пусть $\mu_1^{(m,\ n)} \leqslant \leqslant \mu_2^{(m,\ n)} \leqslant \ldots \leqslant \mu_{mn}^{(m,\ n)}$ — ее собственные числа. Пользуясь тем, что система (3.5), (3.6) ортонормирована в H_B, и неравенствами (3.4), найдем, что $\mu_1^{(m,\ n)} \geqslant (1+R)^{-1}$, $\mu_{mn}^{(m,\ n)} \leqslant 1+R$; отсюда

$$\|A_{mn}\| \leqslant 1+R, \quad \|A_{mn}^{-1}\| \leqslant 1+R, \quad P_{A_{mn}} \leqslant (1+R)^2. \tag{3.7}$$

3°. Оценим погрешность аппроксимации решения по Ритцу. Для простоты положим $m = n$; обозначим $n^2 = N$. Погрешность будем оценивать в метрике H_A. Пусть x_* — точное решение задачи (3.2) и $x_*^{(N)}$ — ее приближенное решение по Ритцу. Пусть функция $f(t)$ такова, что $x_* \in W_2^{(3)}(\Pi)$. Разложим x_* в ряд Фурье

$$x_*(r,\ \vartheta) = \sum_{i,\ j=1}^{\infty} a_{ij} \sin \frac{i\pi\,(r-1)}{R} \sin 2j\vartheta. \tag{3.8}$$

Оценим коэффициенты a_{ij}. Имеем

$$a_{ij} = \frac{8}{\pi R} \int_1^{1+R} \int_0^{\pi/2} x_*(r,\ \vartheta) \sin \frac{i\pi\,(r-1)}{R} \sin 2j\vartheta\, dr\, d\vartheta. \tag{3.9}$$

Интеграл (3.9) возьмем по частям два раза по ϑ. Приняв во внимание краевое условие задачи, получим

$$a_{ij} = -\frac{2}{\pi R j^2} \iint_{\Pi} \frac{\partial^2 x_*}{\partial \vartheta^2} \sin \frac{i\pi\,(r-1)}{R} \sin 2j\vartheta\, dr\, d\vartheta.$$

Интегрируя по частям еще один раз по r, найдем

$$a_{ij} = -\frac{2}{\pi^2 j^2 i} \iint_{\Pi} \frac{\partial^3 x_*}{\partial r\, \partial \vartheta^2} \cos \frac{i\pi\,(r-1)}{R} \sin 2j\vartheta\, dr\, d\vartheta +$$

$$+ \frac{1}{\pi^2 R j^2 i} \int_0^{\pi/2} \left\{ \left[\frac{\partial^2 x_*}{\partial \vartheta^2} \right]_{r=1+R}^{(-1)^i} - \left[\frac{\partial^2 x_*}{\partial \vartheta^2} \right]_{r=1} \right\} \sin 2j\vartheta\, d\vartheta.$$

Второй интеграл исчезает в силу краевых условий: на контуре прямоугольника Π, в частности, на его сторонах $r = 1$ и $r = 1 + R$, будет $x_* = 0$, и потому на тех же сторонах $\partial^2 x_*/\partial \vartheta^2$

тоже обращается в нуль Теперь

$$a_{ij} = -\frac{2}{\pi^2 ij^2}\iint\limits_{\Pi} \frac{\partial^3 x_*}{\partial r\,\partial\vartheta^2}\cos\frac{i\pi(r-1)}{R}\sin 2j\vartheta\,dr\,d\vartheta; \qquad (3.10)$$

аналогично получается еще одно представление:

$$a_{ij} = -\frac{4R}{\pi^3 i^2 j}\iint\limits_{\Pi} \frac{\partial^3 x^*}{\partial r^2\,\partial\vartheta}\sin\frac{i\pi(r-1)}{R}\cos 2j\vartheta\,dr\,d\vartheta. \qquad (3.11)$$

Формулы (3.10), (3.11) можно представить в виде

$$a_{ij} = -\frac{R}{4\pi ij^2},\ b_{ij} = -\frac{R^2}{\pi^2 i^2 j}\,c_{ij}, \qquad (3.12)$$

где b_{ij} и c_{ij} суть коэффициенты Фурье производных $\partial^3 x_*/\partial r\,\partial\vartheta^2$ и $\partial^3 x_*/\partial r^2\,\partial\vartheta$ соответственно по произведениям $\cos\frac{i\pi(r-1)}{R}\times$ $\times\sin 2j\vartheta$ и $\sin\frac{i\pi(r-1)}{R}\cos 2j\vartheta$; $i,\ j=1,\ 2,\ \ldots$. Заметим, что по неравенству Бесселя

$$\sum_{i,\ j=1}^{\infty} b_{ij}^2 \leqslant \frac{16}{\pi R}\left\|\frac{\partial^3 x_*}{\partial r\,\partial\vartheta^2}\right\|^2_{L_2(\Pi)},\quad \sum_{i,\ j=1}^{\infty} c_{ij}^2 \leqslant \frac{16}{\pi R}\left\|\frac{\partial^3 x_*}{\partial r^2\,\partial\vartheta}\right\|^2_{L_2(\Pi)}. \qquad (3.13)$$

Продифференцировав ряд (3.8), получим ряды Фурье для первых производных:

$$\frac{\partial x_*}{\partial r} = \sum_{i,\ j=1}^{\infty}\beta_{ij}\cos\frac{i\pi(r-1)}{R}\sin 2j\vartheta;\quad \beta_{ij} = -\frac{R}{\pi ij}\,c_{ij},$$
$$\frac{\partial x_*}{\partial\vartheta} = \sum_{i,\ j=1}^{\infty}\gamma_{ij}\sin\frac{i\pi(r-1)}{R}\cos 2j\vartheta;\quad \gamma_{ij} = -\frac{4}{\pi^2 ij}\,b_{ij}. \qquad (3.14)$$

Пусть $x_*^{(N)}$ — приближенное решение задачи (3.2) по Ритцу, а $x^{(N)}$ — сумма

$$x^{(N)} = \sum_{i,\ j=1}^{n} a_{ij}\sin\frac{i\pi(r-1)}{R}\sin 2j\vartheta,\quad n=\sqrt{N}.$$

Очевидно, что $|x_* - x_*^{(N)}|_A \leqslant |x_* - x^{(N)}|_A$. По неравенству (3.4) ($\|\cdot\|$ — норма в $L_2(\Pi)$)

$$|x_* - x^{(N)}|^2 \leqslant (1+R)\left\|\frac{\partial x_*}{\partial r} - \frac{\partial x^{(N)}}{\partial r}\right\|^2 + \left\|\frac{\partial x_*}{\partial\vartheta} - \frac{\partial x^{(N)}}{\partial\vartheta}\right\|^2.$$

Далее, по неравенству Бесселя и формулам (3.12) — (3.14)

$$\left\|\frac{\partial x_*}{\partial r} - \frac{\partial x^{(N)}}{\partial r}\right\|^2 \leqslant \frac{R^2}{\pi^2}\left\{\sum_{i=1}^{n}\sum_{j=n+1}^{\infty}\frac{1}{i^2 j^2}\,c_{ij}^2 + \right.$$
$$\left. + \sum_{i=n+1}^{\infty}\sum_{j=1}^{\infty}\frac{1}{i^2 j^2}\,c_{ij}^2\right\} \leqslant \frac{R^2}{\pi^2 (n+1)^2}\sum_{i,\ j=1}^{\infty} c_{ij}^2 \leqslant \frac{16R}{\pi^3 (n+1)^2}\left\|\frac{\partial^3 x_*}{\partial r^2\,\partial\vartheta}\right\|^2.$$

Аналогично

$$\left\| \frac{\partial x_*}{\partial \vartheta} - \frac{\partial x^{(N)}}{\partial \vartheta} \right\|^2 \leqslant \frac{16}{\pi^4 (n+1)^2} \sum_{i,\, j=1}^{\infty} b_{ij}^2 \leqslant \frac{256}{\pi^5 R (n+1)^2} \left\| \frac{\partial^3 x_*}{\partial r\, \partial \vartheta^2} \right\|^2.$$

Отсюда следует оценка погрешности аппроксимации:

$$|x_* - x_*^{(N)}|_A \leqslant \frac{4\sqrt{R(1+R)}}{\sqrt{\pi^5}\,(n+1)} \left\| \frac{\partial^3 x}{\partial r^2\, \partial \vartheta} \right\| + \frac{16}{\sqrt{\pi^3 R}} \left\| \frac{\partial^3 x}{\partial r\, \partial \vartheta^2} \right\|^2. \quad (3.15)$$

4°. Оценим погрешность искажения. По-прежнему считаем $m = n$. Если вычисления производятся с повышенной точностью, то

$$|\delta([\varphi_{ij}, \varphi_{kl}])| \leqslant \varepsilon_1 |[\varphi_{ij}, \varphi_{kl}]_A|, \quad |\delta(\bar{f}, \varphi_{kl})| \leqslant \varepsilon_1 |(\bar{f}, \varphi_{kl})|.$$

Пусть Γ_N и $\delta^{(N)}$ — погрешности матрицы $A_N = A_{nn}$ и вектора $f^{(N)} = ((\bar{f}, \varphi_{11}), \ldots, (\bar{f}, \varphi_{nn}))$ соответственно. Матрица A_N имеет порядок $N = n^2$, и из неравенства для $\delta([\varphi_{ij}, \varphi_{kl}]_A)$ следует, что $\|\Gamma_N\| \leqslant \varepsilon_1 \|\widehat{A}_N\| \leqslant \varepsilon_1 n \|A_N\|$; нормы матриц берутся в метрике R_N. Норму $\|A_N\|$ нетрудно оценить:

$$\|A_N\| = \max_{g \in {}^N R \setminus \{0\}} (A_N g, g)/\|g\|^2.$$

Далее,

$$(A_N g, g) = \sum_{i,\, j,\, k,\, l=1}^{n} [\varphi_{ij}, \varphi_{kl}]_A g_{ij} g_{kl} = \left| \sum_{i,\, j=1}^{n} \varphi_{ij} g_{ij} \right|^2$$

и по неравенству (3.4)

$$(A_N g, g) \leqslant (1+R) \left| \sum_{i,\, j=1}^{n} \varphi_{ij} g_{ij} \right|_B^2.$$

Функции φ_{ij} ортонормированы в метрике H_B, поэтому

$$(A_N g, g) \leqslant (1+R) \sum_{i,\, j=1}^{n} g_{ij}^2 = (1+R)\|g\|^2.$$

Отсюда следует, что $\|A_N\| \leqslant (1+R)$; это дает оценку для Γ_N:

$$\|\Gamma_N\| \leqslant \varepsilon_1 (1+R) n. \quad (3.16)$$

Оценим теперь норму $\|\delta^{(N)}\|$. Из соотношений (3.5), (3.6) находим

$$\|\varphi_{ij}\|_{L_3(\Pi)} = R^2/(i^2\pi^2 + 4j^2R^2);$$

теперь

$$\|\delta^{(N)}\|^2 \leqslant \varepsilon_1^2 \sum_{i,\, j=1}^{n} (\bar{f}, \varphi_{ij})^2 \leqslant \varepsilon_1^2 \|\bar{f}\|_{L_2(\Pi)}^2 \sum_{i,\, j=1}^{n} \|\varphi_{ij}\|_{L_2(\Pi)}^2 \leqslant$$

$$\leqslant \varepsilon_1^2 R^2 \|\bar{f}\|_{L_2(\Pi)}^2 \sum_{i,\, j=1}^{n} 1/(i^2\pi^2 + 4j^2R^2).$$

Обозначим $\sigma = \min(\pi^2, 4R^2)$, тогда

$$\|\delta^{(N)}\|^2 \leqslant \varepsilon_1^2 R^2 \|\bar{f}\|_{L_2(\Pi)}^2 \sigma^{-1} \sum_{i,\, j=1}^{n} 1/(i^2 + j^2).$$

Оценим последнюю сумму

$$\sum_{i,\, j=1}^{n} 1/(i^2 + j^2) = 2\sum_{i=1}^{n} 1/(i^2+1) - 1 + \sum_{i,\, j=2}^{n} 1/(i^2 + j^2).$$

Но

$$\frac{1}{i^2+1} < \int_{i-1}^{i} \frac{dv}{v^2+1}, \quad \sum_{i=1}^{n} \frac{1}{i^2+1} < \int_0^n \frac{dv}{v^2+1} < \frac{\pi}{2}$$

и

$$\sum_{i,\,j=1}^{n} 1/(i^2+j^2) < \pi - 1 + \sum_{i,\,j=2}^{n} 1/(i^2+j^2).$$

В сумме справа $\frac{1}{i^2+j^2} < \int_{i-1}^{i}\int_{j-1}^{j} \frac{dy\,dz}{y^2+z^2}$; отсюда

$$\sum_{i,\,j=2}^{n} \frac{1}{i^2+j^2} < \int_1^n \int_1^n \frac{dy\,dz}{y^2+z^2} = \int_1^n dy \int_1^n \frac{dz}{y^2+z^2}.$$

Во внутреннем интеграле положим $z = y\zeta$:

$$\int_1^n \frac{dz}{y^2+z^2} = \frac{1}{y}\int_{1/y}^{n/y} \frac{d\zeta}{1+\zeta^2} < \frac{\pi}{2y}$$

и

$$\sum_{i,\,j=2}^{n} 1/(i^2+j^2) < \frac{\pi}{2}\ln n.$$

Таким образом,

$$\|\delta^{(N)}\| \leqslant \varepsilon_1 \|\tilde{f}\| \sqrt{\pi - 1 + \frac{\pi}{2}\ln n}\,. \tag{3.17}$$

5°. Из формул (3.1.7) и (3.16), (3.17) следует: если $\varepsilon_1(1+R)^2 n := \beta < 1$, то погрешность искажения

$$\hat{\xi}_N \leqslant \frac{\varepsilon_1 \|\tilde{f}\|(1+R)}{1-\beta}\left[(1+R)^2 n + \sqrt{\pi - 1 + \frac{\pi}{2}\ln n}\right]. \tag{3.18}$$

Обычно величина $\varepsilon_1 n$ мала по сравнению с $1/n$; сравнение формул (3.15) и (3.18) показывает, что в задаче настоящего параграфа погрешность искажения имеет существенно меньшую оценку, чем погрешность аппроксимации. По-видимому, отсюда можно сделать вывод, что при составлении системы Ритца для задач типа рассмотренных в данном параграфе нет необходимости вычислять матрицу и свободные члены системы с повышенной точностью.

6°. Мы не будем здесь заниматься погрешностями алгоритма и округления — они зависят от выбора метода решения системы Ритца.

§ 4. МЕТОД КОНЕЧНЫХ ЭЛЕМЕНТОВ

1°. Рассмотрим самосопряженную положительно определенную задачу

$$0 < t < 1, \quad -\frac{d}{dt}\left[(2+t)\frac{dx}{dt}\right] + tx = f(t); \tag{4.1}$$

$$x(0) = x(1) = 0. \tag{4.2}$$

Для ее решения воспользуемся МКЭ с кусочно-линейной аппроксимацией: введем в рассмотрение исходную функцию

$$\omega(t)=\begin{cases} t+1, & -1\leqslant t\leqslant 0,\\ 1-t, & 0<t\leqslant 1,\\ 0, & t\notin[-1,\ +1], \end{cases} \tag{4.3}$$

выберем натуральное число n, положим $h=1/(2n)$ и будем искать приближенное решение в виде

$$x_{*h}(t)=\sum_{j=1}^{2n-1} a_j^{(n)}\omega(t/h-j):=\sum_{j=1}^{2n-1} a_j^{(n)}\varphi_{jh}(t), \tag{4.4}$$

где коэффициенты a_{jn} определяются из алгебраической системы МКЭ (9.1.13), в которой надо положить $s=1$.

2°. Погрешность аппроксимации будем оценивать в энергетической норме. В работе [96] получена оценка (формула (4) § 3 главы III)

$$\|x'-x^{h'}\|_{L_p}\leqslant h\|x''\|_{L_p},\quad 1<p<\infty, \tag{4.5}$$

где x^h определяется формулой (9.1.5) при $s=1$. Ниже мы воспользуемся оценкой (4.5) при $p=2$. Норму в $L_2(0,1)$ в данном параграфе будем обозначать через $\|\cdot\|$.

Оценим энергетическую норму $|x-x^h|$, где x — произвольная функция из энергетического пространства задачи (4.1), (4.2). Имеем

$$|x|^2=\int_0^1\left[(2+t)\left(\frac{dx}{dt}\right)^2+tx^2(t)\right]dt\leqslant 3\int_0^1\left(\frac{dx}{dt}\right)^2dt+\int_0^1 x^2(t)\,dt.$$

Далее, из тождества $x(t)=\int_0^t x'(\tau)\,d\tau$ получаем по неравенству Буняковского

$$\int_0^1 x^2(t)\,dt\leqslant\int_0^1[x'(t)]^2\,dt$$

и, следовательно, $|x|\leqslant 2\|x'\|$. Заметим, что множитель 2 можно уменьшить. Заменяя x на $x-x^h$ и используя соотношение (4.5), найдем, что $|x-x^h|\leqslant 2\|dx/dt-dx^h/dt\|\leqslant 2h\|x''\|$. Отсюда следует неравенство

$$|x_*-x_{*h}|\leqslant 2h\|x_*''\|; \tag{4.6}$$

здесь x_* — точное решение задачи (4.1), (4.2). Дело сводится к оценке величины $\|x_*''\|$.

3°. Из уравнения (4.1) находим

$$|x_*''|=\left|-\frac{1}{2+t}(f+tx_*-x_*')\right|\leqslant\frac{1}{2}[\|f\|+\|x_*\|+\|x_*'\|]$$

и, следовательно,

$$\|x_*''\|\leqslant 2^{-1}[\|f\|+\|x_*\|+\|x_*'\|]. \tag{4.7}$$

Для дальнейших оценок воспользуемся представлением решения через функцию Грина:

$$x_*(t) = \int_0^1 G(t, \tau) f(\tau)\, d\tau. \tag{4.8}$$

Функция Грина вычисляется просто; она равна

$$G(t, \tau) = -\frac{1}{\ln(3/2)} \begin{cases} \ln\dfrac{2+t}{2} \ln\dfrac{2+\tau}{3}, & t \leqslant \tau; \\ \ln\dfrac{2+t}{3} \ln\dfrac{2+\tau}{2}, & \tau < t. \end{cases}$$

Легко видеть, что $0 \leqslant G \leqslant \ln(3/2)$, и из формулы (4.8) следует неравенство $\|x_*\| \leqslant \|f\| \ln(3/2) \leqslant 2^{-1}\|f\|$. Столь же просто находим, что $\|x_*'\| \leqslant 2^{-1}\|f\|$. Теперь по формуле (4.7)

$$\|x_*''\| \leqslant 2^{-1}(1 + \ln(3/2) + 2^{-1})\|f\| \leqslant \|f\|,$$

и из неравенства (4.6) вытекает оценка погрешности аппроксимации

$$|x_* - x_{*h}| \leqslant 2h\|f\|. \tag{4.9}$$

4°. Переходим к погрешности искажения. Пусть $N = 2n - 1$ — размерность матрицы МКЭ M_n и Γ_n — ее погрешность. По формуле (9.1.17)

$$\|\Gamma_n\|_{R_N} \leqslant 4\varepsilon_1 h^{-1} \max p(t) = 12\varepsilon_1 h^{-1} \tag{4.10}$$

Пусть, далее, $\delta^{(n)}$ — погрешность столбца свободных членов системы алгебраических уравнений МКЭ. Тогда, по формуле (9.7.6)

$$\|\delta^{(n)}\|_{R_N} \leqslant (2/3) h^{1/2} \varepsilon_1 \|f\|. \tag{4.11}$$

Воспользуемся общей формулой (3.1.7), дающей оценку погрешности искажения

$$\hat{\xi}_n \leqslant (1-\beta)^{-1}\left[\|M_n^{-1}\Gamma_n\| \cdot \|x_{*h}\| + \|M_n^{-1}\delta^{(n)}\|\right];$$

здесь можно положить $\beta = \|M_n^{-1}\| \cdot \|\Gamma_n\|$. Несколько увеличив правую часть, получим оценку искажения:

$$\hat{\xi}_n \leqslant (1-\beta)^{-1}\left[\|M_n^{-1}\| \cdot f^{(n)}\| \cdot 12\varepsilon_1 h^{-1} + \|M_n^{-1}\| \cdot \|f^{(n)}\| \varepsilon_1 h^{1/2}\|\right].$$

Далее, по результатам § 6 главы 9

$$\|f^{(n)}\|^2 = \sum_{k=1}^{2n-1} (f, \varphi_{nk})^2 \leqslant (4/3) h \|f\|^2$$

и

$$\hat{\xi}_n \leqslant \frac{2\varepsilon_1}{(1-\beta)\sqrt{3}} \|M_n^{-1}\| \cdot \|f\| \left[12\|M_n^{-1}\| h^{-1/2} + h^{1/2}\right],$$

и остается оценить $\|M_n^{-1}\| = 1/\mu_1^{(n)}$, где $\mu_1^{(n)}$ есть наименьшее собственное число матрицы M_n. По формуле (9.4.4) $\mu_1^{(n)} \geqslant c'h$,

где c' — положительная постоянная, и $\|M_n^{-1}\| \leqslant ch^{-1}$, $c = 1/c'$. Теперь

$$\hat{\xi}_n \leqslant \frac{2\varepsilon_1 ch^{-1}\|f\|}{\sqrt{3}\,(1 - 12c\varepsilon_1 h^{-2})}\left[12ch^{-3/2} + h^{1/2}\right].$$

При обычных значениях ε_1 и h произведение $\varepsilon_1 h^{-2}$ есть величина достаточно малая; это позволяет заменить скобку в знаменателе единицей. Кроме того, $h^{1/2} \ll h^{-3/2}$, и можно пренебречь вторым слагаемым в квадратной скобке. В результате получаем оценку, верную с точностью до членов высшего порядка малости:

$$\hat{\xi}_n \leqslant 8\sqrt{3}\,c^2\varepsilon_1 h^{-5/2}\|f\|. \tag{4.12}$$

5°. Займемся оценкой числа c. Для этого нам понадобится вкратце повторить вывод неравенства (9.3.3), имея в виду интересующий нас случай, а также сделав результаты этого вывода более определенными. В нашем примере $s = 1$, и формула (9.3.2) упрощается:

$$\int_{lh}^{(l+1)h} [x^h(t)]^2\,dt = h\int_0^1 [a_l\omega(\tau) + a_{l+1}\omega(\tau - 1)]^2\,d\tau.$$

Под интегралом справа $\omega(\tau) = 1 - \tau$, $\omega(\tau - 1) = \tau$; вычислив этот интеграл, найдем

$$\int_{lh}^{(l+1)h} [x^h(t)]^2\,dt = \frac{h}{3}(a_l^2 + a_l a_{l+1} + a_{l+1}^2);$$

заменив здесь $a_l a_{l+1}$ на $\pm 2^{-1}\left(a_l^2 + a_{l+1}^2\right)$, получим двойное неравенство

$$\frac{h}{6}(a_l^2 + a_{l+1}^2) \leqslant \int_{lh}^{(l+1)h} [x^h(t)]^2\,dt \leqslant \frac{h}{2}(a_l^2 + a_{l+1}^2).$$

Просуммируем последнее неравенство по l в пределах $0 \leqslant l \leqslant 2n - 1$; при этом, очевидно, надо считать $a_{2n} = 0$:

$$\sqrt{h/6}\,\|a\|_{R_N} \leqslant \|x^h\|_{L_2} \leqslant \sqrt{h}\,\|a\|_{R_N}. \tag{4.13}$$

Неравенство (4.12) есть частный случай неравенства (9.3.3) со значениями постоянных $c_1 = 1/\sqrt{6}$, $c_2 = 1$. Перепишем (4.13) в виде $\|x^h\|_{L_2}^2 \leqslant h\|a\|_{R_N}^2 \leqslant 6\|x^h\|_{L_2}^2$; отсюда видно, что константы неравенства (9.4.2) в нашем примере равны $c_1' = 1$, $c = 6$, и первая из формул (9.4.3) принимает вид

$$\mu_1^{(n)} \geqslant \frac{h}{6}\frac{|w_1|^2}{\|w_1\|^2} \geqslant \frac{h}{6}\min_{x^h \in H_A^{(N)} \setminus \{0\}} \frac{|x^h|^2}{\|x^h\|^2} = \frac{h}{6}\lambda_1^{(n)};$$

здесь $\lambda_1^{(n)}$ — приближение по МКЭ к наименьшему собственному

числу λ_1 оператора задачи (4.1), (4.2). Метод Ритца, для которого МКЭ является частным случаем, дает приближенное значение λ_1 с избытком, поэтому $\mu_1^{(n)} \geqslant \lambda_1 h/6$. Далее, $\lambda_1 = \min_{x \in D(A) \setminus \{0\}} (Ax, x)/\|x\|^2$. При этом

$$(Ax, x) = \int_0^1 \{(2+t)[x'(t)]^2 + tx^2(t)\}\, dt,$$

и

$$\lambda_1 \geqslant 2 \min_{x \in D(A) \setminus \{0\}} \frac{\int_0^1 [x'(t)]^2\, dt}{\int_0^1 x^2(t)\, dt}.$$

Выражение справа есть удвоенное значение наименьшего собственного числа оператора $-d^2/dt^2$ при краевых условиях (4.2). Это наименьшее число равно π^2, отсюда $\mu_1^{(n)} \geqslant h\pi^2/6$. Таким образом, введенное выше, в п. 4°, число c' можно принять равным $\pi^2/6$, тогда $c = 1/c' = 6/\pi^2$, и мы приходим к оценке погрешности искажения

$$\hat{\xi}_n \leqslant \frac{48\sqrt{3}}{\pi^2} \varepsilon_1 h^{-5/2} \|f\| \leqslant 8{,}43 \varepsilon_1 h^{-5/2} \|f\|. \tag{4.14}$$

6°. Погрешности алгоритма и округления зависят от того, каким методом решается алгебраическая система МКЭ. Эта система — трехдиагональная; по-видимому, ее целесообразно решать методом прогонки. Можно также преобразовать эту систему по методу Натансона и решать ее итерациями, но число обусловленности матрицы МКЭ довольно велико, и итерации будут сходиться медленно.

§ 5. УРАВНЕНИЕ ФРЕДГОЛЬМА

1°. Рассмотрим уравнение

$$(Ax)(t) := x(t) + \int_0^1 e^{t\tau} x(\tau)\, d\tau = f(t). \tag{5.1}$$

Обозначим через K интегральный оператор Фредгольма

$$(Kx)(t) = \int_0^1 e^{t\tau} x(\tau)\, d\tau,$$

так что $A = I + K$. Оператор K — симметричный в $L_2(0, 1)$.

Докажем, что он в этом же пространстве положителен. Действительно,

$$(Kx, x) = \int_0^1 \int_0^1 e^{t\tau} x(t) x(\tau)\, dt\, d\tau = \sum_{n=0}^{\infty} \frac{1}{n!} \int_0^1 \int_0^1 t^n \tau^n x(t) x(\tau)\, dt\, d\tau =$$

$$= \sum_{n=0}^{\infty} \frac{1}{n!} \left(\int_0^1 t^n x(t)\, dt \right)^2 \geqslant 0;$$

при этом, если $(Kx, x) = 0$, то

$$\forall n \geqslant 0, \quad \int_0^1 t^n x(t)\, dt = 0,$$

и так как система $\{t^n\}$ полна в $L_2(0, 1)$, то $x(t) \equiv 0$.

Как доказано в п. 2° § 3 главы 11, $\|A^{-1}\| \leqslant 1$. Оценим еще $\|A\|$; для этого заметим, что

$$\|K\|^2 \leqslant \int_0^1 \int_0^1 K^2(t, \tau)\, dt\, d\tau = \int_0^1 \int_0^1 e^{2t\tau}\, dt\, d\tau = \int_0^1 \frac{e^{2t} - 1}{2t}\, dt.$$

В силу известного неравенства: $\xi > 0$, $(e^{\xi} - 1)/\xi < e^{\xi}$, находим

$$\|K\|^2 \leqslant \int_0^1 e^{2t}\, dt = (e^2 - 1)/2$$

и, следовательно, $\|A\| \leqslant 1 + \sqrt{(e^2 - 1)/2} < 2{,}8$.

2°. Чтобы решить приближенно уравнение (5.1), воспользуемся методом Бубнова — Галеркина (в данном случае он совпадает с методом Ритца), взяв в качестве координатной систему $\varphi_0(t) = 1$; $n > 0$, $\varphi_n(t) = (1/\sqrt{2}) \cos n\pi t$, ортонормированную и полную в $L_2(0, 1)$. Приближенное решение $x_n(t)$ имеет вид

$$x_n(t) = a_0^{(n)} + \sum_{k=1}^{n} \left(a_k^{(n)} / \sqrt{2} \right) \cos n\pi t. \tag{5.2}$$

Вектор $a^{(n)}$ коэффициентов $a_k^{(n)}$ определяется из алгебраической системы

$$a^{(n)} + M_n a^{(n)} = f^{(n)}, \tag{5.3}$$

в которой M_n — матрица чисел

$$\int_0^1 \int_0^1 e^{t\tau} \varphi_j(t) \varphi_k(\tau)\, dt\, d\tau, \qquad 0 \leqslant j,\ k \leqslant n, \tag{5.4}$$

а $f^{(n)}$ — вектор с составляющими

$$f_k^{(n)} = \int_0^1 f(t) \varphi_k(t)\, dt, \qquad 0 \leqslant k \leqslant n. \tag{5.5}$$

Как показано в п. 2° § 3 главы 11, матрица M_n неотрицательна, и $\|(I_n + M_n)^{-1}\| \leqslant 1$. Здесь I_n — единичная матрица порядка n.

Там же отмечено, что $\|M_n\| \leqslant \|K\|$ и, следовательно, $\|M_n\| \leqslant 1{,}8$, а $\|I_n + M_n\| \leqslant 2{,}8$.

3°. Пусть $f \in C^{(2)}[0, 1]$, тогда и $x \in C^{(2)}[0, 1]$. Разложим $x(t)$ в ряд Фурье по косинусам:

$$x(t) = \sum_{n=0}^{\infty} \alpha_k \cos n\pi t$$

и оценим коэффициенты

$$\alpha_n = (1/\sqrt{2}) \int_0^1 x(t) \cos n\pi t \, dt, \qquad n > 0.$$

Интегрируя дважды по частям, получим

$$\alpha_n = \frac{(-1)^n x'(1) - x'(0)}{n^2\pi^2\sqrt{2}} + \frac{\beta_n}{n^2\pi^2}. \tag{5.6}$$

Здесь β_n есть коэффициент разложения $x''(t)$ в ряд Фурье по косинусам. Далее,

$$|\alpha_n| \leqslant \frac{|x'(0)| + |x'(1)|}{n^2\pi^2\sqrt{2}} + \frac{|\beta_n|}{n^2\pi^2}.$$

Величины $|x'(0)$ и $|x'(1)|$ оцениваются с помощью уравнения (5.1), из которого следует

$$x'(t) = f'(t) - \int_0^1 e^{t\tau}\tau f(\tau)\, d\tau.$$

Отсюда $|x'(0)| \leqslant |f'(0)| + \|f\|/\sqrt{3}$ и $|x'(1)| \leqslant |f'(1)| + (\sqrt{e^2 - 1}/2)\|f\|$; здесь и ниже $\|\cdot\|$ означает норму в $L_2(0, 1)$. Обозначая $c = [|f'(0)| + |f'(1)| + (1/\sqrt{3} + \sqrt{(e^2-1)/2})\|f\|]/\sqrt{2}$, имеем

$$n > 0, \quad |\alpha_n| \leqslant (c + |\beta_n|)/n^2\pi^2. \tag{5.7}$$

Обозначим через $x^{(n)}(t)$ n-ю частную сумму ряда Фурье функции $x_*(t)$ — решения уравнения (5.1).

4°. Оператор A — положительно определенный; введем в рассмотрение его энергетическую норму:

$$|x|^2 = \|x\|^2 + (Kx, x) \leqslant (1 + \|K\|)\|x\|^2 \leqslant 2{,}8\|x\|^2, \tag{5.8}$$

или $|x| \leqslant 1{,}7\,\|x\|$. Так как x_n есть приближенное решение по Ритцу, то $|x_* - x_n| \leqslant |x_* - x^{(n)}| \leqslant 1{,}7\|x_* - x^{(n)}\|$; оценив правую часть этого неравенства, мы найдем погрешность аппроксимации в нашей задаче.

По соотношениям (5.7), (5.8)

$$\|x - x^{(n)}\|^2 = \sum_{k=n+1}^{\infty} \alpha_k^2 \leqslant \sum_{k=n+1}^{\infty} \frac{(c + |\beta_k|)^2}{k^4\pi^4} \leqslant$$

$$\leqslant \frac{2c^2}{\pi^4} \sum_{k=n+1}^{\infty} \frac{1}{n^4} + \frac{2}{\pi^4} \sum_{k=n+1}^{\infty} \frac{\beta_k^2}{k^4}.$$

Последняя сумма просто оценивается по неравенству Бесселя:

$$\sum_{k=n+1}^{\infty} \frac{\beta_k^2}{k^4} \leqslant \frac{1}{(n+1)^4} \sum_{k=0}^{\infty} \|x''\|^2.$$

Далее,

$$x''(t) = f''(t) - \int_0^1 \tau^2 e^{\tau t} x(\tau)\, d\tau;$$

$$\|x''\| \leqslant \|f''\| + \left\| \int_0^1 \tau^2 e^{\tau t} x(\tau)\, d\tau \right\|;$$

$$\left\| \int_0^1 \tau^2 e^{\tau t} x(\tau)\, dt \right\|^2 \leqslant \|x\|^2 \left\{ \int_0^1 \tau^4 e^{2\tau}\, d\tau \right\} = \frac{e^2-3}{4} \|x\|^2 \leqslant \frac{e^2-3}{4} \|f\|^2.$$

Отсюда $\|x''\| \leqslant \|f''\| + (\sqrt{e^2-3}/2)\|f\| \leqslant \|f''\| + 1{,}05\|f\| := 2^{-1}\pi^4 c_1$ и

$$(2/\pi^4) \sum_{k=n+1}^{\infty} \beta_k^2/k^4 \leqslant c_1/(n+1)^4.$$

Чтобы оценить предшествующую сумму, заметим, что

$$k^{-4} < \int_{k-1}^{k} \sigma^{-4}\, d\sigma,$$

отсюда

$$\sum_{k=n+1}^{\infty} n^{-4} < \int_n^{\infty} \sigma^{-4}\, d\sigma = 3^{-1} n^{-3}.$$

Собрав результаты, получим оценку аппроксимации

$$|x - x_n| \leqslant 1{,}7\, \|x - x^{(n)}\| \leqslant 1{,}7 \left[\frac{2c^2}{3\pi^4 n^3} + \frac{c_1}{(n+1)^4} \right]^{1/2}. \qquad (5.9)$$

Отбросив менее существенное второе слагаемое справа, получим несколько более грубую, но более простую оценку

$$|x - x_n| \leqslant 1{,}4c/(\pi^2 n^{3/2}). \qquad (5.10)$$

5°. Перейдем к погрешности искажения. Обозначим через μ_{jk}, $0 \leqslant j,\ k \leqslant n$, элементы матрицы M_n. Если вычисления производятся с повышенной точностью, то $|\delta(\mu_{jk})| \leqslant \varepsilon_1 |\mu_{jk}|$; если Γ_n — погрешность матрицы M_n, то $\Gamma_n \text{ю} \varepsilon_1 M_n$ и $\|\Gamma_n\| \leqslant \varepsilon_1 \|\widehat{M}_n\| \leqslant \varepsilon_1 \sqrt{n}\, \|M_n\| \leqslant \varepsilon_1 \sqrt{n}\, \sqrt{(e^2-1)/2} \leqslant 1{,}8\varepsilon_1 \sqrt{n}$; точно так же $|\delta(f_k^{(n)})| \leqslant \varepsilon_1 |f_k^{(n)}|$ и $\|\delta^{(n)}\| \leqslant \varepsilon_1 \|f^{(n)}\|$, где $\delta^{(n)} = \delta(f^{(n)})$.

Воспользуемся результатами § 2 главы 3; в формулах этого параграфа надо в данном случае заменить A_n на $I_n + M_n$. Условие $\|A_n^{-1}\Gamma_n\| \leqslant \beta < 1$ можно заменить чуть более жестким условием $\|A_n^{-1}\| \cdot \|\Gamma_n\| \leqslant \beta < 1$ и принять $\beta = 1{,}8\sqrt{n}\,\varepsilon_1$; это допустимо, потому что в обычной вычислительной практике

$\sqrt{n} \ll \varepsilon_1^{-1}$. Теперь формула (3.1.7) дает оценку погрешности искажения

$$\hat{\xi}_n \leqslant (1-\beta)^{-1}\left[\|(I_n + M_n)^{-1}\| \cdot \|\Gamma_n\| \cdot \|f^{(n)}\| + \|(I_n + M_n)^{-1}\|\varepsilon_1\|f^{(n)}\|\right.$$

Если заменить $\|f^{(n)}\|$ большей величиной $\|f\|$ и отбросить в знаменателе малое число β, то получится искомая оценка:

$$\hat{\xi}_n \leqslant \varepsilon_1 (1{,}8\sqrt{n} + 1)\|f\|. \tag{5.11}$$

6°. Матрица $I_n + M_n$ — положительно определенная (ее нижняя грань больше единицы), поэтому систему (5.3) можно решать по методу Холецкого, общая оценка погрешности которого дается формулой (6.2.15); в этой формуле следует заменить A на $I_n + M_n$. О погрешности этого метода сказано в главе 6, § 2.

УКАЗАТЕЛЬ ЛИТЕРАТУРЫ

1. Агмон С. Дуглис А., Ниренберг Л. Оценки решений эллиптических уравнений вблизи границы/Пер. с англ. Л. В. Волевича. М., 1962.
2. Алексиева И. В. О приближенном решении трехмерных задач теории пластичности//Вестн. Ленингр. ун-та, 1978. № 1.
3. Атакишиева Р. Х. Устойчивость метода Бубнова — Галеркина для уравнения с вполне непрерывным оператором в банаховом пространстве//Докл. АН АзССР. 1966. Т. 22, № 1.
4. Бабушка И., Витасек Э., Прагер М. Численные методы решения дифференциальных уравнений/Пер. с англ. В. Л. Каткова. М., 1969.
5. Балакришнан А. В. Прикладной функциональный анализ/Пер. с англ. В. И. Благодатских. М., 1980.
6. Бахвалов Н. С. Численные методы. М., 1973.
7. Березин И. С., Жидков Н. П. Методы вычислений. Т. 1. М., 1966.
8. Бирман М. Ш. О методе Фридрихса расширения положительно определенного оператора до самосопряженного//Зап. Ленингр. горн. ин-та. 1956. Т. 33, № 3.
9. Вазов В., Форсайт Дж. Разностные методы решения дифференциальных уравнений в частных производных/Пер. с англ. М., 1963.
10. Вайникко Г. М. Необходимое и достаточное условие устойчивости метода Галеркина — Петрова//Уч. зап. Тартуск. ун-та. 1956. Вып. 177.
11. Вайникко Г. М. О сходимости и устойчивости метода коллокации//Дифф. ур-ния. 1965. Т. I, № 2.
12. Вайникко Г. М. О сходных операторах//Докл. АН СССР. 1968. Т. 179, № 5.
13. Вайникко Г. М. Анализ дискретизационных методов. Тарту, 1976.
14. Вайникко Г. М., Михлин С. Г. Резольвента Фредгольма и обращение матрицы, линейно зависящей от параметра//Уч. зап. Тартуск. ун-та. 1978. Вып. 448.
15. Варга Р. С. Функциональный анализ и теория аппроксимации в численном анализе/Пер. с англ. Ю. А Кузнецова и А. М. Мацокина. М., 1974.
16. Велиев М. А. Исследование устойчивости метода Бубнова — Галер-

кина для нестационарных задач//Докл. АН СССР. 1964. Т. 157, № 1.
17 Велиев М. А. Об устойчивости метода Бубнова — Галеркина для нестационарных задач//Вопросы вычисл. мат. и вычисл. техн. Баку, 1965.
18. Велиев М. А. Об устойчивости метода Бубнова — Галеркина для уравнения первого порядка в гильбертовом пространстве//Сиб. мат. журн. 1968. Т. 9, № 4.
19 Велиев М. А. Об устойчивости метода Бубнова — Галеркина для уравнений первого порядка с переменными коэффициентами в гильбертовом пространстве//Дифф. ур-ния. 1969. Т. 5, № 3.
20. Велиев М. А. Некоторые достаточные условия устойчивости метода Бубнова — Галеркина для уравнений второго порядка в гильбертовом пространстве//Уч. зап. Азерб. ун-та. Серия физ.-мат. наук. 1972. № 3.
21. Велиев М. А. К устойчивости метода Бубнова — Галеркина для одного класса вырождающихся параболических уравнений//Научн. труды Азерб. ун-та. 1979, № 1.
22. Велиев М. А. Об устойчивости метода конечных элементов для параболических уравнений//Докл. АН АзССР. 1981. Т. 37, № 4.
23 Велиев М. А. К устойчивости метода Бубнова — Галеркина для линейных уравнений первого порядка в гильбертовом пространстве//Докл. АН АзССР. 1981. Т. 37, № 5.
24. Велиев М. А. Устойчивость метода Бубнова — Галеркина для линейных уравнений с переменными коэффициентами//Докл. АН АзССР. 1981. Т. 37, № 7.
25. Воеводин В. В. Ошибки округления и устойчивость в прямых методах линейной алгебры. М., 1969.
26. Воеводин В. В., Фаддеева В. Н., Колотилина Л. Ю. Материалы по математическому обеспечению ЭВМ. Вычислительные методы линейной алгебры. Набор матриц для тестирования. В 3 частях. Л., 1982.
27. Вулих Б. З. Краткий курс теории функций вещественной переменной. М., 1973.
28. Габдулхаев Б. Г. Об одном прямом методе решения интегральных уравнений//Изв. вузов. Математика. 1965. № 3.
29. Габдулхаев Б. Г. Приближенное решение сингулярных интегральных уравнений методом механических квадратур//Докл. АН СССР. 1968. Т. 179, № 2.
30. Габдулхаев Б Г. Оптимальные аппроксимации решений линейных задач. Казань, 1980.
31. Габдулхаев Б. Г., Душков П. Н. Метод механических квадратур для сингулярных интегральных уравнений//Изв. вузов. Математика. 1974. № 12(151).
32. Гавурин М. К. Лекции по методам вычислений. М., 1971.
33. Галин Л. А. Упругопластическое кручение стержней полигонального сечения//Прикл. мат. и мех. 1944. Т. 8, № 4
34. Галин Л. А. Упругопластическое кручение призматических стержней// Прикл. мат. и мех 1949. Т. 13, № 3.
35. Гантмахер Ф. Р Теория матриц. 3-е изд. М., 1967.
36. Гельман И. В. К задаче о минимуме нелинейного функционала//Уч. зап. Ленингр. пед. ин-та им. А. И. Герцена. 1958. Т. 166.
37. Гельфанд И. М, Локуциевский О. В. Метод прогонки для решения разностных уравнений/В кн.: Годунов С. К, Рябенький В. С. Введение в теорию разностных схем. М, 1962. Добавление II.
38. Гловинский Р., Лионс Ж.-Л., Тремольер Р. Численное исследование вариационных неравенств/Пер. с фр. А. С. Кравчука. М., 1979.
39 Годунов С. К. Решение систем линейных уравнений. Новосибирск, 1980.
40. Годунов С. К., Рябенький В. С. Введение в теорию разностных схем. М., 1962.
41. Головкин К. К. О приближении функций в произвольных нормах//

Труды Мат. ин-та им. В. А. Стеклова АН СССР. 1964. Т. 70.
42. Гохберг И. Ц., Фельдман И. А. Уравнения в свертках и проекционные методы их решения. М., 1971.
43 Гурса Э. Курс математического анализа. В 3 т.//Пер. с фр. М. Г. Шестопал. М., 1934. Т. 3, ч. 2.
44 Гюнтер Н. М. Теория потенциала и ее приложения к основным задачам математической физики. М., 1953.
45 Демьянович Ю. К. Устойчивость метода сеток для эллиптических задач//Докл. АН СССР. 1965. Т. 164, № 1.
46 Демьянович Ю. К. Об оценке скорости сходимости проекционных методов решения некоторых вариационных задач. 1980. — Деп. в ВИНИТИ, № 1247.
47 Демьянович Ю. К., Михлин С. Г. О сеточной аппроксимации функций соболевских классов//Зап. науч. семин. Ленингр. отд-ния Мат. ин-та АН СССР им В. А. Стеклова. 1973. Т. 35.
48. Джишкариани А. В. О быстроте сходимости метода Бубнова — Галеркина//Журн. вычисл. мат. и мат. физ. 1964. Т. 4, № 2.
49 Джишкариани А. В. О методе Ритца для одного нелинейного уравнения. Сообщ. АН ГССР. 1968. Т. 51, № 1.
50. Джишкариани А. В. О приближенном методе Ритца для одного нелинейного уравнения//Труды Тбилисск. мат. ин-та. 1969. Т. 36.
51. Джишкариани А. В. Устойчивость метода Ритца для одного нелинейного уравнения//Журн. вычисл. мат. и мат. физ. 1970. Т. 10, № 4.
52. Джишкариани А. В. К решению сингулярных интегральных уравнений приближенными проекционными методами//Журн. вычисл. мат. и мат. физ. 1979. Т. 19, № 5.
53. Джишкариани А. В. К решению сингулярных интегральных уравнений проекционными методами//Журн. вычисл. мат. и мат. физ. 1981. Т. 21, № 2.
54. Довбыш Л. Н. Устойчивость метода Ритца для задач спектральной теории операторов//Труды Мат. ин-та им. В. А. Стеклова АН СССР. 1965. Т. 84.
55. Дюво Г., Лионс Ж.-Л. Неравенства в механике и физике/Пер. с фр. С. Ю. Прищепионка и Т. Н. Рожковской. М., 1980.
56. Запрудский Я. М. Стійкість методу Гальоркіна — Петрова у банаховому просторі//Докл. АН УССР. Сер. А. 1972. № 4.
57 Ильин В. П. О сходимости вариационных процессов//Докл. АН СССР. 1951. Т. 81, № 2.
58. Ильин В. П. Некоторые неравенства в функциональных пространствах и их применение к исследованию сходимости вариационных процессов// Труды Мат. ин-та им. В. А. Стеклова АН СССР. 1959. Т. 53.
59. Ильин В. П. Свойства некоторых классов дифференцируемых функций многих переменных, заданных в n-мерной области//Труды Мат. ин-та им. В. А. Стеклова АН СССР. 1962. Т. 66.
60. Ильин В. П. Некоторые замечания о сходимости последовательностей функций полиномиального типа в пространствах $W_p^{(l)}(G)$//Зап. науч. семин Ленингр. отд-ния Мат. ин-та им. В. А. Стеклова АН СССР. 1968. Т. 11.
61. Канторович Л. В. Об одном методе приближенного решения дифференциальных уравнений в частных производных//Докл АН СССР. 1934 Т. 2.
62. Канторович Л. В. Об одном эффективном методе решения экстремальных задач для квадратичных функционалов//Докл. АН СССР. 1945. Т. 48, № 7.
63 Канторович Л. В. Функциональный анализ и прикладная математика//Успехи мат. наук. 1948. Т. 3, № 6(28).
64 Канторович Л. В., Акилов Г. П. Функциональный анализ. 2-е изд М., 1977.
65 Канторович Л. В., Крылов В. И. Приближенные методы высшего анализа. 3-е изд. Л.; М., 1949.

66 Карпиловская Э. Б. О сходимости интерполяционного метода для обыкновенных дифференциальных уравнений//Успехи мат. наук. 1953. Т. 8, № 3(55).
67. Карпиловская Э. Б. О сходимости метода коллокации//Докл. АН СССР. 1963. Т. 151, № 3.
68. Карпиловская Э. Б. О сходимости метода коллокации для некоторых граничных задач математической физики//Сиб. мат. журн. 1963. Т. 4, № 3.
69. Качанов Л. М. Основы теории пластичности. М., 1969.
70. Качуровский Р. И. Нелинейные монотонные операторы в банаховых пространствах//Усп. мат. наук. 1968. Т. 33, № 2(140).
71. Колмогоров А. Н., Фомин С. В. Элементы теории функций и функционального анализа. 5-е изд. М., 1981.
72. Красносельский М. А. Сходимость метода Галеркина для нелинейных уравнений//Докл. АН СССР. 1950. Т. 73, № 6.
73. Красносельский М. А., Вайникко Г. М, Забрейко П. П., Рутицкий Я. Б., Стеценко В. Я. Приближенное решение операторных уравнений. М., 1969.
74. Купрадзе В. Д., Гегелиа Т. Г., Башелейшвили М. О., Бурчуладзе Т. В. Трехмерные задачи математической теории упругости и термоупругости. М., 1976.
75. Ладыженская О. А., Уральцева Н. Н. Линейные и квазилинейные уравнения эллиптического типа 2-е изд. М., 1973.
76. Лионс Ж.-Л. Оптимальное управление системами, описываемыми уравнениями с частными производными/Пер. с фр. Н. Х. Розова. М., 1972.
77. Лионс Ж.-Л. Некоторые методы решения нелинейных краевых задач/Пер. с фр. Л. Р. Волевича. М., 1972.
78. Маматов А. З. Применение метода Галеркина к некоторому квазилинейному уравнению параболического типа//Вестн. Ленингр. ун-та. 1981. № 13.
79. Маматов А. З. О погрешности и устойчивости метода Бубнова — Галеркина для квазилинейных параболических задач с производной по времени в граничном условии. 1981. 35 с. — Деп. в ВИНИТИ 20.08.1981, № 146.
80. Марчук Г. И. Методы вычислительной математики. Новосибирск, 1973.
81. Михлин С. Г. Интегральные уравнения и их приложения. М.; Л., 1949.
82. Михлин С. Г. Проблемам минимума квадратичного функционала. М., 1952.
83. Михлин С. Г. По поводу метода Ритца//Докл. АН СССР. 1956. Т. 106, № 3.
84. Михлин С. Г. Вариационные методы в математической физике. 2-е изд. М., 1970.
85. Михлин С. Г. Об устойчивости метода Ритца//Докл. АН СССР. 1960. Т. 135, № 1.
86. Михлин С. Г. Некоторые условия устойчивости метода Ритца//Вестн. Ленингр. ун-та. 1961. № 13.
87. Михлин С. Г. Многомерные сингулярные интегралы и интегральные уравнения. М., 1962.
88. Михлин С. Г. О методе Ритца в нейлинейных задачах//Докл. АН СССР. 1962. Т. 142, № 4.
89. Михлин С. Г. Об устойчивости некоторых вычислительных процессов//Докл. АН СССР. 1964. Т. 157, № 2.
90. Михлин С. Г. Численная реализация вариационных методов. М, 1966.
91. Михлин С. Г. О функциях Коссера//Проблемы мат. анализа. Л., 1966.
92. Михлин С. Г. Курс математической физики. М., 1968.

93. Михлин С. Г. Неравенства типа А. А. Маркова и полиномиальные приближения по Ритцу//Дифференциальные уравнения с частными производными. М., 1970.
94. Михлин С. Г. О вариационно-разностном методе для многомерных краевых задач//Зап. науч. семин. Ленингр. отд-ния Мат. ин-та им. В. А. Стеклова АН СССР. 1971. Т. 23.
95. Михлин С. Г. Спектр пучка операторов теории упругости//Успехи мат. наук. 1973. Т. 28, № 3(171).
96. Михлин С. Г. Вариационно-сеточная аппроксимация//Зап. науч. семин. Ленингр. отд-ния Мат. ин-та им. В. А. Стеклова АН СССР. 1974. Т. 48.
97. Михлин С. Г. О постоянных множителях в оценках погрешности вариационно-сеточной аппроксимации//Зап. науч. семин. Ленингр. отд-ния Мат. ин-та им. В. А. Стеклова АН СССР. 1974. Т. 50.
98. Михлин С. Г. О методе наименьших квадратов для многомерных сингулярных интегральных уравнений//Комплексный анализ и его приложения (к 70-летию акад. И. Н. Векуа). М.; Л., 1978.
99. Михлин С. Г. О приближенном решении односторонних вариационных задач//Успехи мат. наук. 1979. Т. 34, № 4(208).
100. Михлин С. Г. О приближенном решении односторонних вариационных задач//Изв. вузов. Математика. 1980. № 2(213).
101. Михлин С. Г. О погрешности приближенного решения односторонних вариационных задач//Вестн. Ленингр. ун-та. 1980, № 13.
102. Михлин С. Г. Об устойчивости решений односторонних вариационных задач: приложения к теории пластичности//Зап. науч. семин. Ленингр. отд-ния Мат. ин-та им. В. А. Стеклова АН СССР. 1980. Т. 102.
103. Михлин С. Г. Погрешность искажения в простейшей односторонней вариационной задаче//Вестн. Ленингр. ун-та. 1981, № 13.
104. Михлин С. Г. О погрешности метода Бубнова — Галеркина для эллиптических краевых задач//Зап. науч. семин. Ленингр. отд-ния Мат. ин-та им. В. А. Стеклова АН СССР. 1981. Т. 111.
105. Михлин С. Г. О погрешностях вычислительных процессов. I//Изв. вузов. Математика 1981. № 7(230).
106. Михлин С. Г. О погрешностях вычислительных процессов. II//Изв. вузов. Математика. 1981, № 8(231).
107. Михлин С. Г. Некоторые новые теоремы об устойчивости вычислительных процессов//Вестн. Ленингр. ун-та. 1982. № 19.
108. Михлин С. Г. Погрешности вычислительных процессов. Тбилиси, 1983.
109. Михлин С. Г. Об одном методе приближенного решения интегральных уравнений//Вестн. Ленингр. ун-та. 1974. № 13.
110 Михлин С. Г. О новом методе приближенного решения интегральных уравнений//Вестн. Ленингр. ун-та. 1976. № 1.
111 Михлин С. Г. Замечания о резольвентном методе//Зап. науч. семин. Ленингр. отд-ния Мат. ин-та им. В. А Стеклова АН СССР. 1976. Т. 58.
112. Михлин С. Г. О приближенном решении интегральных уравнений со слабой особенностью//Вестн. Ленингр. ун-та. 1976. № 13.
113. Михлин С. Г. Об одном варианте резольвентного метода//Вестн. Ленингр. ун-та. 1978. № 1.
114. Михлин С. Г. О погрешностях метода Гаусса решения линейных алгебраических систем//Изв. вузов. Математика. 1984. № 2(261).
115. Михлин С. Г. О погрешностях вычислительных процессов//Вариационно-разностные методы. Ч. I. 1984.
116. Михлин С. Г. О погрешностях метода сопряженных направлений// Вестн. Ленингр. ун-та. 1984. № 19.
117. Михлин С. Г. О погрешностях решений линейных алгебраических систем//Зап. науч. семин. Ленингр. отд-ния Мат. ин-та им. В. А. Стеклова АН СССР. 1984. № 139.
118 Михлин С. Г., Радева Р. К. О приближенном решении сингулярных интегральных уравнений//Изв. вузов. Математика. 1974. № 5(144).
119. Михлин С. Г., Смолицкий Х. Л. Приближенные методы решения дифференциальных и интегральных уравнений. М., 1965.

120. Мусхелишвили Н. И. Некоторые задачи теории упругости. 3-е изд. М., 1940.
121. Надаи А. Пластичность/Пер. с англ. под ред. Л. С. Лейбензона. М.; Л., 1936.
122. Наймарк М. А. Линейные дифференциальные операторы. М., 1954.
123. Натансон И. П. Конструктивная теория функций. М.; Л., 1949.
124. Натансон И. П. К теории приближенного решения уравнений//Уч. зап. Ленингр. пед. ин-та им. А. И. Герцена. 1948. Т. 64.
125. Обэн Ж.-П. Приближенное решение эллиптических краевых задач/Пер. с англ. Ю. А. Кузнецова и А. М. Мацокина. М., 1977.
126. Оганесян Л. А., Руховец Л. А. Вариационно-разностные методы решения эллиптических уравнений. Ереван, 1979.
127. Перлин П. И. Упругопластическое кручение стержней овального поперечного сечения//Инж. сб. 1961. Т. 31.
128. Работнов Ю. Н. (ред.). Теория пластичности. Сб. статей/Пер. с фр., нем. и англ. Л. А. Телешовой и Ю. А. Цвибак. М., 1948.
129. Самарский А. А., Николаев Е. С. Методы решения сеточных уравнений. М., 1978.
130. Смирнов В. И. Курс высшей математики. В 5 т. М.; Л., 1949. Т. 3, ч. 1.
131. Смирнов В. И. Курс высшей математики. В 5 т. М.; Л., 1950, Т. 3, ч. 2.
132. Соболев С. Л. Введение в теорию кубатурных формул. М., 1974.
133. Суворов С. Г. К вопросу об устойчивости вычислительного процесса//Вестн. Ленингр. ун-та. 1982. № 7.
134. Сьярле Ф. Метод конечных элементов для эллиптических задач/Пер. с англ. В. И. Квасова. М., 1980.
135. Тополянский Д. Б., Запрудский Я. М. Исследование устойчивости метода Бубнова — Галеркина для некоторых операторных уравнений с переменными коэффициентами//Укр. мат. журн. 1974. Т. 26, № 5.
136 Тюхтин В. Б. Об оценке погрешности приближенных решений односторонних вариационных задач//Вестн. Ленингр. ун-та. 1981. № 13.
137. Тюхтин В. Б. О скорости сходимости приближенных методов решения односторонних вариационных задач//Вестн. Ленингр. ун-та. 1982. № 13.
138. Уилкинсон Дж. Х. Алгебраическая проблема собственных значений/Пер. с англ. В. В. Воеводина и В. Н. Фаддеевой. М., 1970.
139. Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры. 2-е изд. М.; Л., 1963.
140. Фаддеева В. Н. Метод прямых в применении к некоторым краевым задачам//Труды Мат. ин-та им. В. А. Стеклова АН СССР. 1949. Т. 28.
141. Фаддеева В. Н. Вычислительные методы линейной алгебры. М.: Л., 1950.
142. Форсайт Дж., Молер К. Численное решение систем линейных алгебраических уравнений/Пер. с англ. В. П. Ильина и Ю. И. Кузнецова. М., 1969.
143. Харрик И. Ю. О приближении функций, обращающихся на границе области в нуль, функциями особого вида//Мат. сб. 1955. Т. 37.
144. Харрик И. Ю. О приближении функций, обращающихся вместе с градиентом в нуль на границе области, функциями особого вида//Мат. сб. 1959. Т 47.
145. Харрик И. Ю. О приближении функций, обращающихся в нуль на границе области вместе с частными производными, функциями особого вида//Сиб. мат. журн. 1963. Т. 4, № 2.
146. Черноусько Ф. Л., Баничук Н. В. Вариационные задачи механики и управления. М., 1973.
147. Чистов Е. К. Вариационно-сеточная аппроксимация на регулярной треугольной сетке//Изв. вузов. Математика. 1981. № 11.
148. Чистов Е. К. Устойчивость вариационно-сеточного метода на симплектических сетках//Изв. вузов. Математика. 1982. № 4.

149. Шапошникова Т. О. Некоторые оценки сходимости вариационных методов//Вестн. Ленингр. ун-та. 1971. № 19.
150. Шапошникова Т. О. Априорные оценки погрешности некоторых вариационных методов: Канд. дис. Л., 1973.
151. Шапошникова Т. О. Априорные оценки погрешности некоторых вариационных методов в соболевских пространствах//Вариационно-разностные методы в математической физике. Новосибирск, 1976.
152. Шапошникова Т. О. Априорные оценки погрешности вариационных методов в банаховых пространствах//Журн. вычисл. мат. и мат. физ. 1977. Т. 17, № 5.
153. Яскова Г. Н., Яковлев М. Н. Некоторые условия устойчивости метода Петрова — Галеркина//Труды Мат. ин-та им. В. А. Стеклова АН СССР. 1962. Т. 66.
154. Agmon S., Douglis A., Nierenberg L. Estimates near the boundary for soutions of elliptic partial differential equations satisfying general boundary conditions. II. Communications on pure and applied mathematics. 1964. Vol. 17.
155. Anderssen R. S., de Hoog F. R., Lucas M. A. (ed.). The application and numerical solution of integral equations. Sijthoff & Noordhoff. Alphen aan den Rijn, the Netherland, 1980.
156. Anselone Ph. M. Collectively compact operator approximation theory and application to integral equations. New York, 1971.
157. Brezis H. Problèmes unilatéraux//Journ. de Math. pure et appliquée. 1972. Vol. 51.
158 Brezis H., Stampacchia G. Sur la régularité de la solution d'unéquation elliptique//Bull. Soc Math. de France. 1968. Vol. 96.
159. Brezzi F., Hager W. W., Raviart P. A. Error estimates for finite element solution of variational inequalities. Pt. 1. Primal theory//Numerische Math. 1977. Vol. 28.
160. Calderon A. P. Lebesgue spaces of differentiable functions and distributions//Amer. Math. Soc. Proc. Symp. Pure Math. 1961. Vol. 4.
161. Courant R. Variational methods for the solution of equilibrium and vibrations//Bull. Amer. Math. Soc. 1943. Vol. 49, N 1.
162. Delves L. M., Walsh J. (ed.). Numerical solution of integral equations. Oxford, 1974.
163. El Kolli A. Régularité de la solution et majoration de l'erreur d'approximation pour un problème de Dirichle non linéaire//Comptes Rendus de l'Acad. d. Sci. Française, Ser. A. 1973. Vol. 277.
164. Elliott C. M. On the finite element appoximation of an elliptic variational inequality arising from an implicit time discretisation of the Stefan Problem//IMA J. Numer. Analysis. 1981. Vol. 1, N 1.
165. Falk R. S. Error estimates for the approximation of a class of variational inequalities//Math. Comput. 1974. Vol. 28.
166. Falk R. S. Approximation of an elliptic boundary value problem with unilateral constraints//RAIRO. Anal. Numer. 1975. Vol. 9.
167. Gajewski H. Zur Lösung einer Klasse konvexer Minimalprobleme der Plastizitätstheorie//Z. f. angew. Math. und Mech. 1969. H. 1/2.
168. Gajewski H. Stabilitätsaussagen für einige Aufgaben mit freier Grenze//Z. f. angew. Math. und Mech. 1977. H. 57.
169. Gajewski H. Ein Verfahren zur Ermittlung elastisch-plastische Spannungsfelder//Beiträge zur Spannung- und Dehnungsanalyse/Hrsg. v. K. Schröder. Berlin, 1968.
170. Glowinski R., Lanchon H. Torsion élasto-plastique d'une barre cylindrique de section multi-connexe//J. Mech. 1973. Vol. 12.
171. Glowinski R., Marocco A. Sur l'approximation par éléments finis d'ordre un. et al rélation par pénalisation-duatilté d'une classe de problèmes de problèmes de Dirichle non-linéaires//RAIRO, Anal. Numér. N 2. 1975

172. Golub G. M. Bounds for the round-off errors in the Richardson second order method//Nordisk Tidscriftt for Informationsbehandlung (BIT) 1962 Bd 2, H. 4.
173. Haar A.. Kármán T. von Zur Theorie der Spannungzustände in plastischen und sandartigen Zuständen//Nachr. d. Königlichen Gesellschaft d. Wissenschaft. zu Göttingen. 1909. H. 2.
174. Hille E., Szegö G., Tamarkin J. D. On some generalization of a theorem of A. Markoff//Duke Math. J. 1937. Vol. 3.
175. Kozjakin V. S., Krasnosel'skii M. A. Some remarks on the method of minimal residues//Numer. Funct. Analysis and Optimization. 1981—82. Vol. 4, N 3.
176. Langenbach A. Die Regularisierung nichtlinearer Gleichungen//Math. Nachr. 1962. Bd 24, H. 1.
177. Langenbach A. Monotone Potentialoperatoren. Berlin. 1976.
178. Lanchon H. Solution du problème de torsion élasto-plastique d'une barre cylindrique de section quelconque//Comptes Rendus de l'Acad. d. Sci. Française, 1969. T. 269.
179. Lévy M. Mémoire sur les équation générales des mouvements intérieurs des corps solides ductiles au delà des limites ou l'élasticité pourrait les ramener à leur premier état//Comptes Rendus dc l'Acad. d. Sci. Française. 1870. T. 70.
180. Michlin S. G. Konstanten in einige Ungleichungen der Analysis. Leipzig, 1981.
181. Michlin S. G. Some theorems on the stability of numerical processes// Atti d. Accad. d. Lincei. Classe fis., mat. e nat. 1982. 32, fasc. 2.
182. Michlin S. G. Approximation auf dem Kubischen Gitter. Berlin, 1976.
183. Michlin S. G., Prößdorf S., Singuläre Integraloperatoren. Berlin, 1980.
184. Mises R. von. Mechanik der festen Körper in plastisch-deformabilen Zustand//Nachr. v. d. Königlichen Gesellschaft d. Wissenschaften zu Göttingen. 1913. H. 4.
185. Mittelman H. D. On the approximate solution of nonlinear variatonal inequalities//Numerische Math. 1978. Vol. 29.
186. Morry Ch. B., jr. Multiple integrals in the Calculus of variations. New York, 1966.
187. Mosko U. Implicit variational problems and quasi variational inequalities//Lecture Notes in Math. 1976. Vol. 543.
188. Nitsche J. L-convergence of finite element approximation//Lecture Notes in Math. 1977. Vol. 606.
189. Omodei B. J. Stability of the Rayleigh — Ritz — Galerkin procedure for elliptic boundary value problems. A thesis submitted to the Australien National University. 1975.
190. Pires E. B., Oden J. T. Error estimates for the approximation of a class of variational inequalities arising in unilateral problems with friction//Numer. Funct. Analysis and Optimization. 1981—1982. Vol. 4, N 4.
191. Prößdorf S.. Ratsfeld A. A spline collocation method for singular integral equations with peacewise continuous coefficients//Akad. d. Wissenschaften d DDR. Inst. f. Math. Preprint p-math. 01/84. 1984.
192. Prößdorf O., Silbermann B. Projektionsverfahren und die Näherungsweise Lösung singulärer Gleichungen. Leipzig, 1977.
193. Prößdorf S., Schmidt G A finite element collocation method for singular integral equations//Math. Nachrichten. 1981. Bd. 100.
194. Ritz W Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik//J reine und angewandte Math., 1908. Bd. 135, H. 1.
195. B. Saint-Venant de Sur l'établissement des equiations des mouvements intérieurs opérés dans les corps ductiles au delà des limites où l'élastisité pourrait les ramener à leur premier état//Comptes Rendus d. l'Acad. d. Sci. Francaise. 1870. Vol. 70.

196. Scarpini F. Some nonlinear complementary systems algorithms and applications to unilaternal boundary value problems//Ann. Math Pure Appl. 1980. Vol. 123.
197. Scarpini F., Vivaldi M. A. Error estimates for the approximation of some unilateral problems//RAIRO. Anal. Numer. 1977. Vol. 11, N 2.
198. Sobolev S. L. The problem of propagation of plastical state//Труды Сейсмологич. ин-та АН СССР. 1935. № 49.
199. Strang G., Fix G. A Fourier analysis of the finite element method. Препринт; год издания не указан.
200 Temple G. The general theory of relaxation methods applied to linear systems//Proc. of the Royal Soc. Ser. A. 1939. Vol. 169.
201. Tucker T. S. Stability of nonlinear computing schemes//SIAM. J. Numer. Analysis. 1969. Vol. 6.
202 Wilkinson J. H. Rounding errors in algebraic processes. National Physical Laboratory. Notes on applied sciences. New York, 1963.

Учебное издание

Соломон Григорьевич Михлин

Некоторые вопросы теории погрешностей

Редактор *З. И. Царькова*
Обложка художника *Ю. Н. Васильева*
Художественный редактор *С. В. Алексеев*
Технический редактор *А. В. Борщева*
Корректоры *Е. К. Терентьева, Т. Г. Павлова*

ИБ № 2812

Сдано в набор 14.09.87. Подписано в печать 19.04.88. М-33729. Формат 60×90$^1/_{16}$. Бумага тип. № 2. Гарнитура литературная. Печать высокая. Усл. печ. л. 21. Усл. кр.-отт. 21,19. Уч.-изд. л. 19,75 Тираж 2704 экз. Заказ 789. Цена 4 руб.

Издательство ЛГУ им. А. А. Жданова. 199034, Ленинград, Университетская, наб., 7/9.

Ленинградская типография № 2 головное предприятие ордена Трудового Красного Знамени Ленинградского объединения «Техническая книга» им Евгении Соколовой Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли 198052, г. Ленинград, Л-52, Измайловский проспект, 29.