$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n = b_n \end{cases}$$

$$Ax = b,$$

$$A = \{a_{ij}\}_{i,j=1}^{n}$$

$$b = (b_1, \ldots, b_n), \qquad x = (x_1, \ldots, x_n)$$

$$C = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} & b_n \\ c_1 & c_2 & \ldots & c_n & \end{pmatrix}.$$

ISBN 978-5-00130-611-5

Ф.Г. АВХАДИЕВ

ОСНОВЫ ЧИСЛЕННЫХ МЕТОДОВ

Ф.Г. АВХАДИЕВ

ОСНОВЫ ЧИСЛЕННЫХ МЕТОДОВ

$$\begin{cases} a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)} \\ a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + \ldots + a_{3n}^{(1)}x_n = b_3^{(1)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n = b_n^{(1)} \end{cases}$$

КАЗАНСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Ф.Г. АВХАДИЕВ

ОСНОВЫ ЧИСЛЕННЫХ МЕТОДОВ

Учебное пособие

КАЗАНЬ
2022

УДК 519.6
ББК 22.19
А22

Печатается по решению Учебно-методической комиссии Института математики и механики им. Н.И. Лобачевского (протокол № 7 от 02.06.2022 г.)

Научный редактор
кандидат физико-математических наук, доцент **А.В. Ожегова**

Рецензенты:
кандидат физико-математических наук, доцент кафедры теории функций и приближений ИМиМ им. Н.И. Лобачевского **Р.Г. Насибуллин;**
кандидат физико-математических наук, доцент кафедры теории функций и приближений КФУ **Ю.Р. Агачев**

Авхадиев Ф.Г.
А22 **Основы численных методов** / Ф.Г. Авхадиев. – Казань: Издательство Казанского университета, 2022. – 444 с.

ISBN 978-5-00130-611-5

Книга представляет собой курс лекций, читаемый автором студентам КФУ. В ней изложены численные методы решения систем алгебраических уравнений, теория интерполяции функций и квадратурных формул, а также алгоритмы приближенного решения дифференциальных, операторных и интегральных уравнений.

Книга предназначена для студентов-бакалавров.

УДК 519.6
ББК 22.19

ISBN 978-5-00130-611-5

Предисловие

Учебное пособие представляет собой обработанный курс лекций, читаемый автором студентам-бакалаврам в Институте математики и механики им. Н.И. Лобачевского Казанского федерального университета.

В книге изложены численные методы решения систем линейных алгебраических уравнений, алгоритмы решения нелинейных уравнений и систем, теория интерполяции функций и квадратурных формул, наилучшие приближения в функциональных пространствах, а также базовые алгоритмы приближенного решения дифференциальных, операторных и интегральных уравнений. Более полно с содержанием книги можно познакомиться по оглавлению.

Некоторые заголовки содержат общепринятые сокращения: СЛАУ — система линейных алгебраических уравнений, ОДУ — обыкновенные дифференциальные уравнения, УЧП — уравнения в частных производных, МКР — метод конечных разностей, НП — наилучшие приближения.

Поскольку лекции предназначены для студентов-математиков, все основные утверждения даны с полными доказательствами. Отметим также, что для понимания содержания книги вполне достаточно знаний, получаемых студентами на первых двух курсах уни-

верситета по линейной алгебре, дифференциальному и интегральному исчислению, начальным фактам теории функциональных пространств.

По большому счету практика вычислений является искусством, творческим процессом, основанном на знании различных алгоритмов. Поэтому при изучении данного курса нужно уделить особое внимание численным алгоритмам решения рассматриваемых задач. Конечно, леммы и теоремы, теоретические оценки погрешности тех или иных алгоритмов играют важную, но вспомогательную роль. Однако знание теоретических основ приобретает первостепенное значение при создании новых численных методов и алгоритмов.

В заключение отметим, что представленные лекции следует рассматривать как вводный курс в обширную область знаний, называемой "Вычислительной математикой". Все мы понимаем, что в настоящее время роль вычислительной математики неуклонно растет в связи с бурным развитием компьютерных технологий.

Выражаю благодарность рецензентам за ряд полезных замечаний, которые были учтены мной при компьютерной верстке книги.

Казанский федеральный университет,
июнь 2022 года,
Ф.Г. Авхадиев

Оглавление

Глава 1

Методы решения СЛАУ

В первой части курса мы будем рассматривать базовые алгоритмы решения трех задач. В основном речь пойдет о решении системы линейных алгебраических уравнений (сокращенно, СЛАУ) вида

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2 \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n = b_n \end{cases}.$$

С использованием матриц и векторов линейную систему можно записать в виде одного уравнения

$$Ax = b,$$

где $A = \{a_{ij}\}_{i,j=1}^{n}$ — квадратная матрица порядка n с вещественными или комплексными коэффициентами, $b = (b_1, \ldots, b_n)$ — заданный вектор, а $x = (x_1, \ldots, x_n)$

— искомый вектор.

Из курса линейной алгебры мы знаем, что если определитель (=детерминант) $\det A$ матрицы

$$A = \{a_{ij}\}_{i,j=1}^{n}$$

отличен от нуля, то решение СЛАУ существует и определяется единственным образом при любом заданном векторе $b = (b_1, \ldots, b_n)$, и это решение можно найти по правилу Крамера.

Применение формул Крамера связано с вычислением определителей, а эта операция является весьма трудоемкой для больших n. А именно, вычислительная сложность, т. е. число арифметических операций, необходимых для вычисления $\det A$, превосходит $n!$, если пользоваться классическим определением детерминанта.

Поэтому для решения СЛАУ разработаны менее трудоемкие, новые методы. Они делятся на два типа: точные и приближенные. "Точным" называют метод, позволяющий найти решение за конечное число шагов (арифметических операций).

Забегая вперед, отметим, что в следующей главе 2 будем рассматривать базовые методы приближенного решения нелинейных алгебраических уравнений и систем нелинейных уравнений. В частности, будем знакомиться с методами нахождения собственных значений

и собственных векторов конечномерных линейных операторов, заданных матрицами.

Для СЛАУ рассмотрим сначала несколько точных методов решения, а именно, правило Крамера, метод Гаусса и его модификации, метод ортогонализации и несколько методов, связанных со специальными разложениями матрицы системы.

1.1 О формулах Крамера

Пусть $A = (a_{ij})$ — квадратная матрица порядка n, $A^{(k)}$ — квадратная матрица, полученная из матрицы A заменой k-того столбца элементов $(a_{1k}, a_{2k}, \ldots, a_{nk})$ на столбец свободных членов $(b_1, b_2, \ldots, b_n)$. Правило Крамера предполагает, что $\det A \neq 0$.

Тогда, как хорошо известно, решение уравнения

$$Ax = b$$

существует, единственно и определяется следующими формулами Габриэля Крамера (1704-1752):

$$x_k = \frac{\det A^{(k)}}{\det A}, \qquad k = 1, 2, \ldots, n.$$

Найдем теперь число арифметических операций N, необходимых для определения решения методом Крамера. Будем учитывать только умножения и деления

(пренебрегаем сложениями и вычитаниями) и пользуемся индуктивным определением детерминантов матриц, равносильным стандартному.

Очевидно, имеем n делений, а количество умножений равно $(n+1)M_n$, где M_n — число умножений при вычислении определителя матрицы порядка n. Таким образом, $N = n + (n+1)M_n$.

Далее применим метод математической индукции. Для матрицы второго порядка определитель равен

$$a_{11}a_{22} - a_{21}a_{12}$$

и содержит $2 = 2!$ умножения. Применяя разложение по элементам третьей строки для определителя матрицы третьего порядка, получаем $M_3 = 2!\,3 = 3!$. Аналогично, если $M_k = k!$ $(k \geq 3)$, то, применяя разложение по элементам последней строки для определителя матрицы порядка $k+1$, немедленно получаем, что $M_{k+1} = (k+1)!$, и, следовательно, $N = n + (n+1)!$ при указанном методе вычисления определителей. Напомним, что $n! \approx \sqrt{2\pi n}\,(n/e)^n$.

На самом деле можно считать, что

$$N = n + (n+1)M_n = O(n^4),$$

так как метод Гаусса, который мы рассмотрим в следующем пункте, позволяет вычислить определитель матрицы порядка n за значительно меньшее число умно-

жений и делений. А именно, $M_n = O(n^3)$ при использовании алгоритма Гаусса.

Замечание Существует иная, более общая, формулировка правила Крамера для совместных систем, не требующая предположения $\det A \neq 0$. А именно, имеет место такое утверждение: *если* $x_1, \ldots, x_n$ *— одно из решений системы, то для любых коэффициентов* $c_1, c_2, \ldots, c_n$ *справедливо равенство*

$$(c_1x_1 + c_2x_2 + \ldots + c_nx_n) \det A = -\det C,$$

где C *— следующая квадратная матрица порядка* $n+1$

$$C = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} & b_n \\ c_1 & c_2 & \ldots & c_n & 0 \end{pmatrix}.$$

1.2 Метод Гаусса

Речь идет об известном с древнейших времен методе последовательного исключения переменных.

В настоящее время этот подход принято называть методом Гаусса в честь Карла Фридрихса Гаусса (1777-1855), систематически использовавшего метод последовательного исключения переменных на практике при

обработке результатов проводимых им геодезических измерений в Нижней Саксонии.

1.2.1 Основные алгоритмы

Рассмотрим систему линейных алгебраических уравнений общего вида

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2 \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n = b_n \end{cases} . \quad (1.1)$$

Применим к этой системе специальные элементарные преобразования следующих трех типов:

(α) умножение или деление элементов какой-либо строки на число, отличное от нуля;

(β) добавление к элементам какой-либо строки чисел, пропорциональных элементам одной из предыдущих строк;

(γ) добавление к элементам какой-либо строки чисел, пропорциональных элементам одной из последующих строк.

Целью этих преобразований является приведение матрицы A к верхнетреугольной матрице, т. е. к матрице, у которой равны нулю все элементы, расположенные ниже главной диагонали.

Основной алгоритм Гаусса, основанный на операциях вида (α) и (β), опишем по шагам.

Шаг 1.1 Предположим, что $a_{11} \neq 0$. Делим на это число коэффициенты первого уравнения, получаем новое первое уравнение вида

$$x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 + \ldots + a_{1n}^{(1)} x_n = b_1^{(1)}.$$

Шаг 1.2 Умножаем новое первое уравнение на число $a_{k1}, \quad (k = 2, \ldots, n)$, и вычитаем из k-го уравнения. Получаем новые уравнения вида

$$\begin{cases} a_{22}^{(1)} x_2 + a_{23}^{(1)} x_3 + \ldots + a_{2n}^{(1)} x_n = b_2^{(1)} \\ a_{32}^{(1)} x_2 + a_{33}^{(1)} x_3 + \ldots + a_{3n}^{(1)} x_n = b_3^{(1)} \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n2}^{(1)} x_2 + a_{n3}^{(1)} x_3 + \ldots + a_{nn}^{(1)} x_n = b_n^{(1)} \end{cases} . \qquad (1.2)$$

Фактически мы имеем новую систему линейных алгебраических уравнений порядка $(n-1)$ с неизвестными $x_2, \ldots, x_n$. С ней поступаем точно так же, как и с исходной системой.

Шаги 2.1 и 2.2 Предполагаем, что $a_{22}^{(1)} \neq 0$, 1-ое уравнение из (1.2) делим на $a_{22}^{(1)}$

$$x_2 + a_{23}^{(2)} x_3 + \ldots + a_{2n}^{(2)} x_n = b_2^{(2)}.$$

Это уравнение умножаем на $a_{k2}^{(1)} \quad (k = 3, \ldots, n)$ и вы-

читаем из k-го уравнения. Уравнения с номерами

$$k = 3, \ldots, n$$

преобразуются к следующему виду

$$\begin{cases} a_{33}^{(2)} x_3 + a_{34}^{(2)} x_4 + \ldots + a_{3n}^{(2)} x_n = b_3^{(2)} \\ a_{43}^{(2)} x_3 + a_{44}^{(2)} x_4 + \ldots + a_{4n}^{(2)} x_n = b_4^{(2)} \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n3}^{(2)} x_3 + a_{n4}^{(2)} x_4 + \ldots + a_{nn}^{(2)} x_n = b_n^{(2)} \end{cases} . \qquad (1.3)$$

Шаги 3.1 и 3.2 аналогичны шагам 2.1 и 2.2. А именно, предполагаем, что $a_{33}^{(2)} \neq 0$, 1-ое уравнение из (1.3) делим на $a_{33}^{(2)}$ и приходим к уравнению вида

$$x_3 + a_{34}^{(3)} x_4 \ldots + a_{3n}^{(3)} x_n = b_3^{(3)},$$

с использованием полученного уравнения исключаем переменную x_3 из всех последующих уравнений. Далее, продолжаем процесс. Понятно, что через $2n - 1$ шаг, в предположении отличности от нуля чисел (ведущих элементов)

$$a_{11},\ a_{22}^{(1)},\ a_{33}^{(2)},\ \ldots,\ a_{nn}^{(n-1)},$$

получаем следующую систему с верхнетреугольной матрицей

$$\begin{cases} x_1 + a_{12}^{(1)} x_2 + \ldots + a_{1n}^{(1)} x_n = b_1^{(1)} \\ x_2 + \ldots + a_{2n}^{(2)} x_n = b_2^{(2)} \\ \ldots\ldots\ldots \\ x_n = b_n^{(n)} \end{cases} . \qquad (1.4)$$

Переход от СЛАУ вида (1.1) к системе вида (1.4) называется **прямым ходом метода Гаусса**.

Обратный ход метода Гаусса заключается в нахождении неизвестных $x_1, \ldots, x_n$ из системы (1.4) в порядке, обратном номеру неизвестной. Согласно (1.4), имеем $x_n = b_n^{(n)}$. Из $(n-1)$-й строки находим x_{n-1}:

$$x_{n-1} = -a_{n-1,n}^{(n-1)} x_n + b_{n-1}^{(n-1)}.$$

Зная x_n, x_{n-1} и используя $(n-2)$-ю строку системы (1.4), определяем x_{n-2}, и т.д. Наконец, находим x_1 по формуле $x_1 = -a_{12}^{(1)} x_2 - a_{13}^{(1)} x_3 \ldots - a_{1n}^{(1)} x_n + b_1^{(1)}$ по известным значениям $x_2, \ldots, x_n$.

Замечание 1. Если $\det A \neq 0$, то за счет перестановки строк в системе (1.1) можно добиться того, что все ведущие элементы в основном алгоритме Гаусса будут отличны от нуля. В этом можно убедиться по индукции.

Действительно, если $\det A \neq 0$, но $a_{11} = 0$, то хотя бы один элемент a_{j1} $(2 \leq j \leq n)$ первого столбца матрицы должен быть отличен от нуля. Мы можем пе-

реставить строки с номерами 1 и j, и проделать шаги 1.1 и 1.2 с новым $a_{11}^{(0)} = a_{1j} \neq 0$. Если $\det A \neq 0$ и $a_{11}^{(0)} a_{22}^{(1)} a_{33}^{(2)} \dots a_{kk}^{(k-1)} \neq 0$, то детерминант системы

$$\begin{cases} a_{k+1\,k+1}^{(k)} x_{k+1} + a_{k+1\,k+2}^{(k)} x_{k+2} + \ldots + a_{k+1\,n}^{(2)} x_n = b_{k+1}^{(k)} \\ a_{k+2\,k+1}^{(k)} x_{k+1} + a_{k+2\,k+2}^{(k)} x_4 + \ldots + a_{k+2\,n}^{(k)} x_n = b_{k+2}^{(k)} \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n\,k+1}^{(k)} x_{k+1} + a_{n\,k+2}^{(k)} x_{k+2} + \ldots + a_{n\,n}^{(k)} x_n = b_n^{(k)} \end{cases}.$$

отличен от нуля. Поэтому среди элементов первого столбца $a_{k+1\,k+1}^{(k)}, a_{k+2\,k+1}^{(k)}, \ldots, a_{n\,k+1}^{(k)}$ имеется хотя бы один элемент $a_{j\,k+1}^{(k)}$, отличный от нуля. Ясно, что перестановка строк с номерами $k+1$ и j позволяет продолжить прямой ход алгоритма Гаусса с ведущим элементом

$$\tilde{a}_{k+1\,k+1}^{(k)} = a_{j\,k+1}^{(k)} \neq 0.$$

Описанный выше основной алгоритм Гаусса связан лишь с операциями типа (α) и (β). Легко видеть, что для перестановки строк, о чем идет речь в замечании, необходимо привлечь и преобразования типа (γ).

Отметим также, что при перестановке двух строк матрицы определитель меняет знак. Этот факт легко доказать по индукции, взяв за базу индукции $n = 2$ и пользуясь разложением по элементам нетронутой перестановкой строки для определителя матрицы порядка $n+1$ при индуктивном переходе.

1.2.2 Число операций. Модификации

Вычислим число арифметических операций N, необходимых для выполнения алгоритма Гаусса. Как и ранее, будем учитывать только операции умножения и деления. При обратном ходе определение x_n не требует затрат, при определении x_{n-1} используется одно умножение, для нахождения x_k требуется $k-1$ умножение, обратный ход заканчивается вычислением x_1 за $n-1$ умножение. В итоге, число умножений для осуществления обратного хода Гаусса равно

$$N_1 = 1 + 2 + \ldots + (n-1) = \frac{(n-1)n}{2} = O(n^2).$$

Рассмотрим прямой ход. В шаге 1.1 имеется n делений на число a_{11}. Шаг 1.2 связан с n умножениями на числа a_{k1} для $k = 2, 3, \ldots, n$, т.е. число умножений в шаге 1.2 равно $n(n-1)$. Итак, шаги 1.1 и 1.2 требуют $n + n(n-1) = n^2$ арифметических операций умножения и деления.

Число умножений и делений для шагов 2.1 и 2.2 вычисляется аналогично и равно $(n-1)^2$, для шагов 3.1, 3.2 — $(n-2)^2$ и т.д.

Очевидно, искомое число операций для прямого хода Гаусса определяется формулой

$$N_2 = n^2 + (n-1)^2 + \ldots + 2^2 + 1^2 = \frac{n(n+1)(2n+1)}{6}.$$

Таким образом, арифметическая сложность алгоритма Гаусса равна

$$N = N_1 + N_2 = \frac{n(n^2 + 3n - 1)}{3} = O(n^3).$$

Метод Гаусса с выбором ведущих элементов В замечании 1 мы уже отметили необходимые изменения основного алгоритма Гаусса в том случае, когда диагональный элемент $a_{kk}^{(k-1)}$ равен нулю. Кроме того, поскольку деление на малое число может привести к большим ошибкам, то неприятной является и ситуация, когда элемент $a_{kk}^{(k-1)}$ отличен от нуля, но является малым числом.

Поэтому рекомендуется следующее усовершенствование (модификация) основного метода Гаусса.

На первом шаге выбирают коэффициент $a_{j_1 1}$, который является максимальным по модулю среди элементов первого столбца и меняют местами первую строку со строкой под номером j_1. Ясно, что в шагах 1.1 и 1.2 коэффициент $a_{j_1 1}$ играет роль a_{11}.

Аналогично поступаем на k-том шаге. В качестве $a_{kk}^{(k-1)}$ берем элемент $a_{j_k k}^{(k-1)}$, максимальный по модулю среди чисел $a_{kk}^{(k-1)}, \ldots, a_{nk}^{(k-1)}$. Меняем местами строки под номерами k и j_k (если $k \neq j_k$) и следуем основному алгоритму Гаусса.

При этом, если детерминант матрицы отличен от

нуля, то, очевидно, все ведущие элементы $\widetilde{a_{kk}^{(k-1)}} = a_{j_k k}^{(k-1)}$ будут отличны от нуля.

Описанный алгоритм называется методом Гаусса с выбором ведущих элементов по столбцам.

Существуют две других разновидности этого алгоритма с выбором ведущего элемента. А именно, в качестве ведущего элемента выбирают коэффициент, максимальный по модулю среди элементов строки, т. е. среди чисел $a_{kk}^{(k-1)}, \ldots, a_{kn}^{(k-1)}$. Практически это связано с соответствующей перестановкой столбцов на k-том шаге основного алгоритма Гаусса. Другая разновидность связана с выбором ведущего элемента, максимального по модулю среди всех элементов матрицы, с которой мы работаем на k-том шаге основного алгоритма Гаусса. Ясно, что метод Гаусса с выбором ведущего элемента по всей матрице связан с возможной перестановкой как строк, так и столбцов.

Метод Гаусса с оптимальным исключением переменных

Этот метод преобразует невырожденную матрицу в единичную. Приведем укрупненные шаги.

Шаг 1. Начало остается таким же, как и раньше. Делим первое уравнение на a_{11} и получаем из 1-го уравнения

$$x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 \ldots + a_{1n}^{(1)} x_n = b_1^{(1)}.$$

Умножаем это уравнение на a_{21} и вычитаем из 2-го уравнения. Затем второе уравнение делим на $a_{22}^{(1)}$. Используя новое 2-е уравнение, из первого исключаем x_2. Первые два уравнения преобразуются к виду

$$\begin{cases} x_1 + a_{13}^{(2)} x_3 + \ldots + a_{1n}^{(2)} x_n = b_1^{(2)} \\ x_2 + a_{23}^{(2)} x_3 + \ldots + a_{2n}^{(2)} x_n = b_2^{(2)} \end{cases} . \qquad (1.5)$$

Шаг 2. Пользуясь (1.5), из k-того уравнения $k \geq 3$ исключаем x_1, x_2, затем, используя преобразованное 3-е уравнение во всех уравнениях (кроме третьего уравнения) исключаем x_3.

Продолжаем процесс. В результате матрица системы преобразуется в единичную матрицу, а система уравнений преобразуется к виду

$$x_1 = b_1^{(2n)}, \quad x_2 = b_2^{(2n)}, \ldots, \quad x_n = b_n^{(2n)}.$$

Можно сочетать этот основной алгоритм оптимального исключения с одним из алгоритмов Гаусса с выбором ведущих элементов. Итог таков: невырожденная квадратная матрица преобразуется в диагональную. Легко увидеть, что число операций увеличивается, но порядок остается тем же. А именно, число умножений и делений для указанных модификаций алгоритма Гаусса равно

$$N = N_1 + N_2 = O(n^3).$$

1.2.3 Вычисление обратных матриц

Пусть $\det A \neq 0$. При применении основного алгоритма Гаусса без перестановок строк будем иметь

$$\det A = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} =$$

$$= a_{11} \begin{vmatrix} 1 & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \dots & \dots & \dots & \dots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ a_{32}^{(1)} & \dots & a_{3n}^{(1)} \\ \dots & \dots & \dots \\ a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{vmatrix}.$$

Продолжая основной алгоритм, получаем формулу

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

т. е. $\det A$ равен произведению ведущих элементов. В общем случае, когда применение метода Гаусса сопровождается перестановками строк, имеем

$$\det A = (-1)^m \, a_{11}^{(0)} \cdot a_{22}^{(1)} \dots a_{nn}^{(n-1)},$$

где m — суммарная перестановка строк.

Рассмотрим вычисление обратной матрицы. Эта задача является более сложной. Пусть $A = (a_{ij})_{i,j=1}^{n}$; предполагаем, что $\det A \neq 0$, так как это условие необходи-

мо и достаточно для существования обратной матрицы A^{-1}. Из курса линейной алгебры известно, что элементы обратной матрицы определяются формулами

$$b_{ij} = \frac{A_{ji}}{\det A},$$

где A_{ij} – алгебраическое дополнение к элементу a_{ij}.

Мы не пользуемся этими формулами. Рассмотрим иной подход к определению A^{-1}. А именно, неизвестную обратную матрицу $X = A^{-1}$ будем искать как решение матричного уравнения

$$AX = E := \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix},$$

где через E обозначена единичная матрица.

Понятно, что нахождение неизвестной матрицы

$$X = \begin{pmatrix} x_1^{(1)} & \dots & x_1^{(k)} & \dots & x_1^{(n)} \\ x_2^{(1)} & \dots & x_2^{(k)} & \dots & x_2^{(n)} \\ \dots & \dots & \dots & \dots & \dots \\ x_n^{(1)} & \dots & x_n^{(k)} & \dots & x_n^{(n)} \end{pmatrix}$$

требует вычисления n^2 чисел $x_j^{(k)}$. Неизвестные числа составляют n столбцов. Для определения k-го столбца неизвестных $x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)}$ имеем следующую систе-

му линейных алгебраических уравнений:

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_n^{(k)} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix},$$

где в правой части единица стоит на k-ой строчке, остальные координаты этого вектора равны нулю.

Следовательно, эту систему можно записать в виде СЛАУ

$$\sum_{j=1}^{n} a_{mj} x_j^{(k)} = \delta_{mk}, \quad m = 1, 2, \dots, n,$$

где δ_{mk} — символ Кронекера.

Полученную систему для определения вектора

$$x^{(k)} = (x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)})$$

решаем методом Гаусса.

Так как число столбцов у матрицы $X = A^{-1}$ равно n, то приходится решать n однотипных систем. Поскольку эти системы уравнений отличаются только правыми частями, то очевидно, что прямой ход метода Гаусса можно проводить одновременно для всех систем. Наиболее простой и эффективный алгоритм возникает при применении метода Гаусса с оптимальным исклю-

чением переменных. На практике новый алгоритм сводится к следующим действиям.

Записываем рядом матрицы A и E, получаем следующую прямоугольную матрицу

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} & 1 & 0 & \dots & 0 & 0 \\ a_{21} & a_{22} & \dots & a_{2n} & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} & 0 & 0 & \dots & 0 & 1 \end{pmatrix},$$

где число строк в два раза меньше, чем число столбцов.

К строкам длины $2n$ применяем преобразования типа (α), (β) и (γ), так же, как и в методе Гаусса с оптимальным исключением переменных. В результате преобразований получаем

$$\begin{pmatrix} 1 & 0 & \dots & 0 & 0 & b_{11} & b_{12} & \dots & b_{1n} \\ 0 & 1 & \dots & 0 & 0 & b_{21} & b_{22} & \dots & b_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 1 & b_{n1} & b_{n2} & \dots & b_{nn} \end{pmatrix}.$$

Сама матрица A преобразовалась в единичную, а на месте единичной матрицы E возникает автоматически некоторая новая матрица, которая и является обратной матрицей.

Этот алгоритм можно обосновать и без привлечения матричного уравнения. Действительно, для заданной матрицы каждое преобразование строк типа (α),

(β) и (γ) равносильно умножению слева на некоторую невырожденную матрицу B_j. Поэтому при применении метода Гаусса с оптимальным исключением переменных мы получаем формулу: $E = BA$, где

$$B = B_1 B_2 \dots B_m,$$

m — число преобразований типа (α), (β) и (γ), которые использовались для преобразования матрицы A в единичную. По определению обратной матрицы из равенства $E = BA$ немедленно получаем, что $B = X = A^{-1}$.

С другой стороны, в приведенном выше алгоритме одновременного преобразования записанных рядом матриц A и E, над строками единичной матрицы проводятся те же преобразования, что и над строками матрицы A. Следовательно, матрица E преобразуется в матрицу $BE = B = X = A^{-1}$.

Таким образом, мы познакомились с двумя доказательствами достоверности указанного алгоритма для вычисления обратной матрицы.

В заключение приведем примеры, показывающие равносильность преобразований строк типа (α), (β) и (γ) умножению слева на некоторую невырожденную матрицу. Для простоты мы выбрали лишь матрицы третьего порядка. Понятно, что эти примеры легко обобщаются на матрицы любого порядка и на действия с любыми строками.

(α) Умножаем вторую строку на число c:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ c\,a_{21} & c\,a_{22} & c\,a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

(β) Умножаем вторую строку на число c и прибавляем к третьей строке:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} + c\,a_{21} & a_{32} + c\,a_{22} & a_{33} + c\,a_{23} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & c & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

(γ) Умножаем третью строку на число c и прибавляем ко второй строке:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} + c\,a_{31} & a_{22} + c\,a_{32} & a_{23} + c\,a_{33} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

1.3 Метод ортогонализации

Метод ортогонализации представляет собой, как и методы Крамера и Гаусса, точный метод, позволяющий найти решение СЛАУ с применением конечного числа арифметических операций.

Рассмотрим СЛАУ вида $Ax = b$, где

$$A = (a_{ij})_{i,j=1}^{n}, \quad x = (x_1, \ldots, x_n), \quad b = (b_1, \ldots, b_n).$$

Предположим, что $\det A \neq 0$. Тогда существует единственное решение $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ этой системы.

Введем новые переменные $y = (x_1, \ldots, x_n, 1)$.

Запишем нашу систему $Ax = b$ в новой форме как однородную систему уравнений для этого $(n+1)$-мерного вектора $y = (x_1, \ldots, x_n, 1)$:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n - b_1 \cdot 1 = 0 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n - b_2 \cdot 1 = 0 \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n - b_n \cdot 1 = 0 \end{cases} . \qquad (1.6)$$

Рассмотрим следующие $(n+1)$-мерные векторы, определенные формулами

$$a^{(i)} = (a_{i1}, a_{i2}, \ldots, a_{in}, -b_i), \quad i = \overline{1, n},$$

и скалярные произведения

$$(a^{(i)}, y) := a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n - b_i \cdot 1.$$

Тогда наша исходная система (1.6) запишется в виде системы уравнений

$$(a^{(i)}, y) = 0, \quad i = \overline{1, n},$$

в терминах скалярного произведения. Таким образом, решение исходной СЛАУ свелось к следующей задаче: найти $(n+1)$-мерный вектор

$$y = (x_1, \ldots, x_n, 1)$$

с последней координатой, равной единице, и ортогональный заданным векторам $a^{(i)}, \quad i = \overline{1, n}$.

Для решения этой новой задачи введем $(n+1)$-мерный вектор

$$a^{(n+1)} = (0, \ldots, 0, 1)$$

с последней координатой, равной единице, и имеющий первые n координат, равные нулю.

Рассмотрим новую систему векторов

$$a^{(1)}, a^{(2)}, \ldots, a^{(n)}, a^{(n+1)}.$$

Эта система является линейно независимой, так как

$\det A \neq 0$ и поэтому

$$\det A_{n+1} := \det \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} & -b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & -b_2 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} & -b_n \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} = 1 \cdot \det A \neq 0.$$

Воспользуемся теперь методом ортогонализации Грама-Шмидта. А именно, построим ортонормированную систему векторов

$$v^{(1)}, v^{(2)}, \dots, v^{(n)}, v^{(n+1)},$$

которая получается из линейно независимой системы $a^{(1)}, a^{(2)}, \dots, a^{(n)}, a^{(n+1)}$ по рекуррентным формулам

$$v^{(1)} = \frac{a^{(1)}}{\|a^{(1)}\|}, \quad v^{(2)} = \frac{a^{(2)} - (a^{(2)}, v^{(1)})\, v^{(1)}}{\|a^{(2)} - (a^{(2)}, v^{(1)})\, v^{(1)}\|}, \quad \dots,$$

$$v^{(k)} = \frac{a^{(k)} - \sum_{j=1}^{k-1} (a^{(k)}, v^{(j)})\, v^{(j)}}{\|a^{(k)} - \sum_{j=1}^{k-1} (a^{(k)}, v^{(j)})\, v^{(j)}\|} \quad (k = 2, 3, \dots, n+1).$$

По построению имеем: вектор

$$v^{(n+1)} = (v_1^{(n+1)}, v_2^{(n+1)}, \dots, v_{n+1}^{(n+1)})$$

ортогонален векторам $v^{(1)}, v^{(2)}, \dots, v^{(n)}$. Следователь-

но, вектор $v^{(n+1)}$ также ортогонален векторам

$$a^{(1)}, a^{(2)}, \ldots, a^{(n)}$$

и, кроме того, $\|v^{(n+1)}\| = 1$.

Утверждение. *$(n+1)$-ая координата вектора*

$$v^{(n+1)} = \left(v_1^{(n+1)}, \ldots, v_{n+1}^{(n+1)}\right)$$

отлична от нуля.

Обоснование: Предположим, что $v_{n+1}^{(n+1)} = 0$. Но тогда скалярное произведение $(v^{(n+1)}, a^{(n+1)})$ равно нулю, так как

$$(v^{(n+1)}, a^{(n+1)}) = v_1^{(n+1)} \cdot 0 + \ldots + v_n^{(n+1)} \cdot 0 + 0 \cdot 1 = 0.$$

Таким образом, вектор $v^{(n+1)}$ ортогонален всем элементам линейно независимой системы $(n+1)$-мерных векторов

$$a^{(1)}, a^{(2)}, \ldots, a^{(n)}, a^{(n+1)}.$$

Следовательно, $v^{(n+1)}$ — нулевой вектор. Это противоречит тому, что по построению $\|v^{(n+1)}\| = 1$.

Теперь легко получить формулы для записи в явном виде искомого решения $y^* = (x_1^*, x_2^*, \ldots, x_n^*, 1)$.

Поскольку $v_{n+1}^{(n+1)} \neq 0$, то определен $(n+1)$-мерный

вектор

$$y^* := \frac{v^{(n+1)}}{v_{n+1}^{(n+1)}} = \left(\frac{v_1^{(n+1)}}{v_{n+1}^{(n+1)}},\ \frac{v_2^{(n+1)}}{v_{n+1}^{(n+1)}},\ \ldots,\ \frac{v_n^{(n+1)}}{v_{n+1}^{(n+1)}},\ 1 \right),$$

ортогональный векторам $a^{(1)}, a^{(2)}, \ldots, a^{(n)}$ и имеющий последнюю координату, равную единице.

Следовательно, решение рассматриваемой системы $Ax = b$ определяется формулами

$$x_1^* = \frac{v_1^{(n+1)}}{v_{n+1}^{(n+1)}},\ \ x_2^* = \frac{v_2^{(n+1)}}{v_{n+1}^{(n+1)}},\ \ \ldots,\ \ x_n^* = \frac{v_n^{(n+1)}}{v_{n+1}^{(n+1)}}.$$

Число арифметических операций для метода ортогонализации сравнимо с числом операций в методе Гаусса, но среди операций имеются непростые вычисления евклидовой нормы векторов $a^{(1)}$ и

$$a^{(k)} - \sum_{j=1}^{k-1} (a^{(k)}, v^{(j)})\, v^{(j)} \quad (k = 2, \ldots, n).$$

1.4 Методы факторизации

Факторизация означает представление в виде произведения. Применительно к функциям или операторам под факторизацией понимают представление в виде суперпозиции.

Пусть A — квадратная матрица порядка n, пред-

ставимая в виде

$$A = BC,$$

где C – верхнетреугольная матрица

$$C = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ 0 & c_{22} & \dots & c_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & c_{nn} \end{pmatrix},$$

и B – нижнетреугольная матрица

$$B = \begin{pmatrix} b_{11} & 0 & \dots & 0 \\ b_{11} & b_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ b_{11} & b_{22} & \dots & b_{nn} \end{pmatrix}.$$

Предположим, что $\det A \neq 0$. Поскольку

$$\det A = \det B \det C,$$

то получаем

$$\det B = \prod_{k=1}^{n} b_{kk} \neq 0, \quad \det C = \prod_{k=1}^{n} c_{kk} \neq 0.$$

Таким образом, обе матрицы B и C являются невырожденными.

Уравнение $Ax = b$ запишется в виде $BCx = b$. Решение этой системы может быть получено последо-

вательным решением двух систем уравнений

$$\begin{cases} By = b \\ Cx = y \end{cases}.$$

Решение каждой из этих систем получается просто в силу того, что их матрицы являются треугольными. Можно записать и явные формулы. Ясно, что решение $By = b$ сводится к последовательному определению неизвестных $y_1, y_2, \ldots, y_n$ по формулам

$$y_1 = \frac{b_1}{b_{11}},\ y_2 = \frac{b_2 - b_{21}y_1}{b_{22}}, \ldots,\ y_n = \frac{b_n - \sum_{j=1}^{n-1} b_{nj}y_j}{b_{nn}}.$$

Зная $y_1, y_2, \ldots, y_n$, последовательно определяем

$$x_n, x_{n-1}, \ldots, x_1$$

по формулам $x_n = y_n/c_{nn}$,

$$x_{n-1} = \frac{y_{n-1} - c_{n-1\,n}x_n}{c_{n-1\,n-1}}, \ldots,\ x_1 = \frac{y_1 - \sum_{j=2}^{n} c_{1j}x_j}{c_{11}}.$$

Нетрудно видеть, что число умножений и делений, необходимых для решения СЛАУ имеет порядок $O(n^2)$.

Рассмотрим базовые методы, основанные на факторизации матриц.

Метод квадратного корня

Требуется решить систему линейных алгебраиче-

ских уравнений

$$Ax = b, \quad a_{ij} \in \mathbb{C}, \quad \det A \neq 0,$$

где $A = A^*$ — самосопряженная матрица, т. е.

$$a_{ij} = \overline{a_{ji}}, \quad \text{в частности,} \quad a_{kk} \in \mathbb{R}.$$

Если элементы матрицы являются вещественными числами, то матрица является самосопряженной тогда и только тогда, когда она совпадает с транспонированной. Иными словами, матрица является симметричной относительно своей главной диагонали.

Самосопряженную матрицу можно представить в виде

$$A = S^* D S, \tag{1.7}$$

где D — диагональная матрица, а S — верхнетреугольная матрица, т. е. имеет вид

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ 0 & s_{22} & \dots & s_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & s_{nn} \end{pmatrix}.$$

Строго говоря, для того чтобы представление (1.7) было возможным, необходимо еще отличие от нуля некоторых коэффициентов, возникающих в ходе преобразований (см. ниже примеры).

Очевидно, решение системы $Ax = b$, т. е. системы

$$S^*DSx = b$$

сводится к последовательному решению двух простых систем

$$\begin{cases} S^*Dy = b \\ Sx = y \end{cases},$$

где

$$S^*D = \begin{pmatrix} s_{11}^* d_{11} & 0 & \dots & 0 \\ s_{21}^* d_{11} & s_{22}^* d_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ s_{n1}^* d_{11} & s_{n2}^* d_{22} & \dots & s_{nn}^* d_{nn} \end{pmatrix}$$

— нижнетреугольная матрица, $s_{kj}^* = \overline{s_{jk}}$.

Понятно, что на практике основная трудность состоит в том, что для заданной матрицы $A = A^*$ нужно найти подходящие матрицы S и D, удовлетворяющие равенству (1.7).

Рассмотрим сначала случай $n = 2$. Дана матрица

$$A = \begin{pmatrix} a_{11} & a_{12} \\ \overline{a_{12}} & a_{22} \end{pmatrix},$$

а матрицы

$$D = \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix},$$

$$S = \begin{pmatrix} s_{11} & s_{12} \\ 0 & s_{22} \end{pmatrix}, \qquad S^* = \begin{pmatrix} \overline{s_{11}} & 0 \\ \overline{s_{12}} & \overline{s_{22}} \end{pmatrix}$$

нужно определить так, чтобы выполнялось равенство $A = S^*DS$. Имеем

$$S^*DS = \begin{pmatrix} d_{11}\overline{s_{11}} & 0 \\ d_{11}\overline{s_{12}} & d_{22}\overline{s_{22}} \end{pmatrix} \begin{pmatrix} s_{11} & s_{12} \\ 0 & s_{22} \end{pmatrix} =$$

$$= \begin{pmatrix} d_{11}s_{11}^2 & d_{11}s_{11}s_{12} \\ d_{11}\overline{s_{12}}s_{11} & d_{11}\overline{s_{12}}s_{12} + d_{22}s_{22}^2 \end{pmatrix}.$$

Для определения неизвестных коэффициентов получаем систему нелинейных уравнений

$$\begin{cases} d_{11}s_{11}^2 = a_{11} \neq 0, \quad a_{11} \in \mathbb{R}, \\ d_{11}s_{11}s_{12} = a_{12}, \\ d_{11}\overline{s_{12}}s_{11} = \overline{a_{12}}, \\ d_{11}\overline{s_{12}}s_{12} + d_{22}s_{22}^2 = a_{22}. \end{cases}$$

Число уравнений меньше, чем число неизвестных. Поэтому, если это система разрешима, то решение не является единственным. Но нам нужно лишь одно из возможных решений, которое можно определить следующим образом.

Из самосопряженности матрицы A следует, что числа a_{11} и a_{22} являются вещественными числами. Дополнительно предположим, что $a_{11} \neq 0$. Тогда можно по-

ложить, что d_{11} равен плюс или минус единице, точнее, полагаем $d_{11} = \operatorname{sign} a_{11}$. Тогда

$$s_{11} = \sqrt{\frac{a_{11}}{d_{11}}}, \qquad s_{12} = \sqrt{\frac{a_{12}}{d_{11}\sqrt{a_{11}/d_{11}}}}.$$

Предположим, далее, что $a_{22} - |s_{12}|^2 d_{11} \neq 0$. Тогда можно взять $d_{22} = \operatorname{sign}(a_{22} - |s_{12}|^2 d_{11})$ и определить

$$s_{22} = \sqrt{\frac{a_{22} - d_{11}\overline{s_{12}}s_{12}}{d_{22}}}.$$

Общий случай, когда $n \geq 3$. Перемножение матриц показывает, что факторизация имеет место тогда и только тогда, когда справедливы следующие равенства

$$a_{ij} = \sum_{k=1}^{i} \overline{s_{ki}} d_{kk} s_{kj}, \qquad i \leq j.$$

Решение этой системы можно определить в явном виде. Элементы d_{ii} будем брать равными 1 или -1.

При $i = j = 1$ уравнение имеет вид $a_{11} = s_{11}^2 d_{11}$, поэтому можно взять

$$d_{11} = \operatorname{sign} a_{11}, \quad s_{11} = \sqrt{\frac{a_{11}}{d_{11}}}.$$

Пусть $i = 1, j \geq 2$, уравнение имеет вид $a_{1j} = s_{11} d_{11} s_{1j}$,

отсюда

$$s_{1j} = \frac{a_{1j}}{s_{11} d_{11}}.$$

Далее рассматриваем случай $i \geq 2$. Непосредственными вычислениями получаем следующие рекуррентные соотношения для последовательного определения остальных элементов матриц S и D:

$$d_{ii} = \operatorname{sign}\left(a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk} \right),$$

$$s_{ii} = \left| a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk} \right|^{1/2},$$

$$s_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} \overline{s}_{ki} s_{kj} d_{kk}}{s_{ii} d_{ii}}, \quad i < j.$$

Для больших n метод квадратного корня требует примерно $n^3/3$ арифметических операций.

Решение системы с ненулевыми главными минорами

Дана квадратная матрица $A = (a_{ij})$, у которой главные миноры отличны от нуля, т. е.

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

Для такой матрицы, как это доказывается в курсе линейной алгебры, справедливо разложение $A = BC$, где

B — нижнетреугольная матрица, C — верхнетреугольная матрица. При определении коэффициентов c_{ij}, b_{ij} имеется произвол. Можно взять $b_{kk} = 1$. Имеются явные формулы для других коэффициентов c_{ij}, b_{ij} (см., например, стр. 26–32 книги [17]).

1.5 Метод прогонки

Для матриц специального вида, часто встречающихся на практике, разработаны упрощенные методы, позволяющие эффективно применять метод Гаусса. Мы проиллюстрируем это на примере алгоритма решения СЛАУ вида

$$Ax = d = (d_1 \dots d_n),$$

когда матрица имеет вид

$$A = \begin{pmatrix} -b_1 & c_1 & 0 & \dots & 0 \\ a_2 & -b_2 & c_2 & \dots & 0 \\ 0 & a_3 & -b_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_n & -b_n \end{pmatrix},$$

т. е. является ленточной 3-х диагональной матрицей.

Излагаемый ниже метод называется методом прогонки и используется при решении ряда задач, в частности, при численном решении краевой задачи для линейного дифференциального уравнения второго поряд-

ка.

Итак, система линейных алгебраических уравнений имеет вид

$$a_i x_{i-1} - b_i x_i + c_i x_{i+1} = d_i \quad (i = 1, 2, \ldots, n),$$

где $a_1 = 0$, $c_n = 0$, т. е. 1-ое и n-ое уравнения имеют только по два слагаемых в левой части уравнений.

Прямой ход метода прогонки связан с определением прогоночных коэффициентов ξ_i, η_i и выводом формул

$$x_{i-1} = \xi_i x_i + \eta_i, \quad i = 2, 3, \ldots,$$

необходимых для реализации последовательного исключения переменных.

Опишем кратко, как возникает прогоночный метод для решения СЛАУ с нашей 3-х диагональной матрицей.

Пусть $b_1 \neq 0$. Тогда из 1-го уравнения

$$-b_1 x_1 + c_1 x_2 = d_1,$$

получаем формулы

$$x_1 = \frac{c_1}{b_1} x_2 - \frac{d_1}{b_1}, \quad \xi_2 = \frac{c_1}{b_1}, \quad \eta_2 = -\frac{d_1}{b_1}.$$

Подставляя выражение для x_1 во второе уравнение, имеем

$$a_2(\xi_2 x_2 + \eta_2) - b_2 x_2 + c_2 x_3 = d_2,$$

отсюда находим x_2 по формуле

$$x_2 = \frac{c_2}{b_2 - a_2\xi_2}x_3 + \frac{-d_2 + a_2\eta_2}{b_2 - a_2\xi_2}.$$

Следовательно, соответствующие прогоночные коэффициенты даны формулами

$$\xi_3 = \frac{c_2}{b_2 - a_2\xi_2}, \qquad \eta_3 = \frac{-d_2 + a_2\eta_2}{b_2 - a_2\xi_2}.$$

Выражение $x_2 = \xi_3 x_3 + \eta_3$ подставляем в 3-е уравнение. Новое 3-е уравнение содержит лишь две неизвестных, x_3 и x_4. Поэтому из нового 3-его уравнения переменная x_3 определяется через x_4 формулой вида $x_3 = \xi_4 x_4 + \eta_4$ и т.д. Закономерность строения прогоночных коэффициентов ясна. Для переменной x_k при $2 \le k \le n-1$ получаем формулу $x_k = \xi_{k+1}x_{k+1} + \eta_{k+1}$ с прогоночными коэффициентами

$$\xi_{k+1} = \frac{c_k}{b_k - a_k\xi_k}, \quad \eta_{k+1} = \frac{-d_k + a_k\eta_k}{b_k - a_k\xi_k}.$$

В частности, из $(n-1)$-го уравнения находим $x_{n-1} = \xi_n x_n + \eta_n$ и подставляем это выражение в последнее уравнение. Новое последнее уравнение будет содержать только одну переменную x_n. Поэтому из него находим $x_n = \eta_{n+1}$, где

$$\eta_{n+1} = \frac{-d_n + a_n\eta_n}{b_n - a_n\xi_n}.$$

Формально мы можем считать, что $\xi_{n+1} = 0$.

Обратный ход прогонки тривиален: $x_n = \eta_{n+1}$ найден на последнем шаге прямого хода, находим $x_{n-1} = \xi_n \eta_{n+1} + \eta_n$, затем последовательно определяем x_{n-2}, x_{n-3}, $\ldots$, x_1.

Нетрудно подсчитать число операций, точнее, число умножений и делений прямого и обратного хода прогонки. Обратный ход содержит $(n-1)$ умножение. При прямом ходе имеется $(2n-1)$ деление и $2n$ умножений. Таким образом, метод прогонки для трехдиагональной матрицы порядка n требует не более $5n$ умножений и делений.

Может быть так, что $b_k - a_k \xi_k = 0$ для некоторого номера. Тогда приведенный алгоритм не осуществим. Но имеется весьма простое достаточное условие, гарантирующее отличие от нуля знаменателей в формулах для прогоночных коэффициентов.

Определение 1.1 *Говорят, что матрица имеет диагональное преобладание, если для любого номера* $i = 1, 2, \ldots, n$

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

В частности, для трехдиагональной матрицы условие диагонального преобладания сводится к следующим тре-

бованиям

$$|b_1| > |c_1|,\ |b_n| > |a_n|,\ |b_k| > |a_k|+|c_k|,\ (2 \le k \le n-1).$$

Утверждение. *Пусть A — трехдиагональная матрица с диагональным преобладанием. Тогда для всех номеров $|\xi_k| < 1$, следовательно, $b_k - a_k\xi_k \neq 0$ и метод прогонки применим.*

Доказательство. Докажем по индукции, что

$$|\xi_k| < 1.$$

При $k = 2$ имеем

$$\xi_2 = \frac{c_1}{b_1} \Longrightarrow |\xi_2| < 1, \quad \text{так как} \quad |b_1| > |c_1|.$$

Далее, пусть дано, что $|\xi_k| < 1$, тогда оценим $|\xi_{k+1}|$ следующим образом:

$$|\xi_{k+1}| \le \frac{|c_k|}{|b_k| - |a_k| \cdot |\xi_k|} < \frac{|c_k|}{|b_k| - |a_k|} < 1.$$

Последнее строгое неравенство (< 1) следует из того, что $|c_k| < |b_k| - |a_k|$ по определению диагонального преобладания.

Итак, для любого номера $|\xi_k| < 1$. Но тогда

$$|b_k - a_k\xi_k| \ge |b_k| - |a_k| > 0,$$

так как $|b_k| > |c_k| + |a_k| \ge |a_k|$.

Этим и завершается доказательство.

Следствие 1.0.1 *Предположим, что A — трехдиагональная матрица с диагональным преобладанием. Тогда* $\det A \neq 0$.

Это утверждение допускает обобщение. Сформулируем это обобщение в виде задачи на доказательство.

Упражнение. Пусть A — матрица с диагональным преобладанием. Тогда $\det A \neq 0$.

Замечание. Диагональное преобладание является лишь достаточным (но не необходимым) условием для реализации метода прогонки.

1.6 О нормах векторов и матриц

Рассмотрим n-мерное вещественное евклидово пространство $\mathbb{R}^n$ (или $\mathbb{C}^n$), $n \geq 2$. Тогда для любого вектора $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ (или $\mathbb{C}^n$) определена евклидова норма

$$\|x\| = \sqrt{|x_1|^2 + |x_2|^2 + \ldots + |x_n|^2},$$

согласованная со скалярным произведением

$$(x, y) = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$$

векторов $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Напомним, что в $\mathbb{C}^n$ скалярное произведение определе-

но формулой

$$(x, y) = x_1\overline{y_1} + x_2\overline{y_2} + \ldots + x_n\overline{y_n}.$$

При изучения ряда вопросов, в частности, топологических, нет необходимости вводить иные нормы. Тем более, как гласит известная теорема функционального анализа, в конечномерном пространстве все нормы эквивалентны, т. е. для любых двух норм $\|x\|'$, $\|x\|''$ существуют положительные числа c_1, c_2 такие, что выполняются неравенства

$$c_1\|x\|' \le \|x\|'' \le c_2\|x\|'$$

для любого $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ (или $\mathbb{C}^n$).

1.6.1 p-нормы векторов

При изучении ряда прикладных задач, например, при исследовании сходимости итерационных методов решения систем алгебраических уравнений, числовые значения и простота вычисления норм векторов и соответствующих им норм матриц имеют важное значение. В особенности, оказываются полезными следующие p-нормы ($1 \le p \le +\infty$):

$$\|x\|_p := (|x_1|^p + |x_2|^p + \ldots + |x_n|^p)^{1/p}, \quad 1 \le p < +\infty,$$

$$\|x\|_\infty := \max\{|x_1|, |x_2|, \ldots, |x_n|\}, \quad p = +\infty,$$

где $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ (или $\mathbb{C}^n$). Очевидно, если все координаты вектора $x = (x_1, x_2, \ldots, x_n)$ равны нулю или является отличной от нуля только одна из координат, то норма не зависит от p, т. е. $\|x\|_p = \|x\|_q = const$ для любых допустимых p и q.

Теорема 1.1 *Для любого вектора $x \in \mathbb{R}^n$ (или $\mathbb{C}^n$) его p-норма $\|x\|_p$ является невозрастающей функцией от параметра $p \in [1, +\infty]$. В частности, будем иметь неравенства*

$$\|x\|_1 \geq \|x\|_2 \geq \|x\|_\infty.$$

Равенство $\|x\|_p = \|x\|_q$ для любых допустимых различных p и q имеет место тогда и только тогда, когда отличной от нуля является не более, чем одна из координат этого вектора.

Доказательство. Очевидно, что достаточно рассмотреть случай, когда количество ненулевых координат вектора не меньше, чем 2. Кроме того, p-норма не меняется, если мы поменяем номера координат. Поэтому, без ограничения общности предполагаем, что

$$|x_1| = \|x\|_\infty := \max\{|x_1|, |x_2|, \ldots, |x_n|\} > 0,$$

$$x_j \neq 0, \quad 2 \leq j \leq m, \quad 2 \leq m \leq n.$$

Обозначим $\alpha_j = \frac{|x_j|}{|x_1|}$. Простые преобразования дают формулу

$$\|x\|_p = |x_1| \left(1 + \sum_{j=2}^{m} \alpha_j^p\right)^{1/p}, \quad 0 < \alpha_j \leq 1.$$

Пусть

$$y = y(p) := \ln \frac{\|x\|_p}{|x_1|} = \frac{1}{p} \ln \left(1 + \sum_{j=2}^{m} \alpha_j^p\right).$$

Очевидно, нам достаточно убедиться в том, что $y(p)$ — строго убывающая функция при условии, что количество ненулевых координат вектора x не меньше, чем 2. А этот факт проверяется простыми вычислениями. Действительно, имеем

$$y'(p) := -\frac{1}{p^2} \ln \left(1 + \sum_{j=2}^{m} \alpha_j^p\right) + \frac{\sum_{j=2}^{m} \alpha_j^p \ln \alpha_j}{p \left(1 + \sum_{j=2}^{m} \alpha_j^p\right)} < 0$$

с учетом неравенств

$$\ln \alpha_j \leq 0, \quad \ln \left(1 + \sum_{j=2}^{m} \alpha_j^p\right) > 0.$$

Таким образом, мы показали, что норма $\|x\|_p$ является строго убывающей функцией от $p \in [1, +\infty]$, когда количество ненулевых координат вектора не меньше, чем

2.

Этим и завершается доказательство.

1.6.2 Нормы матриц

Пусть A — квадратная матрица порядка n с элементами a_{kj} из $\mathbb{R}$ или $\mathbb{C}$.

Понятно, что такая матрица задает линейный непрерывный оператор $A : \mathbb{R}^n \to \mathbb{R}^n$ или $A : \mathbb{C}^n \to \mathbb{C}^n$, сопоставляющий любому вектору x его образ y, определенный равенством $y = Ax$.

Нормой матрицы A будем называть норму этого линейного оператора A, т. е. норма $\|A\|$ матрицы A определяется равенством

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

В силу конечномерности пространств $\mathbb{R}^n$ и $\mathbb{C}^n$ супремум можно заменить на максимум. Следовательно, число $c \geq 0$ является нормой матрицы A тогда и только тогда, когда выполняются два следующих свойства:

1) $\|Ax\| \leq c\,\|x\|$ для любого вектора x;

2) существует вектор $x_0 \neq \theta$ такой, что

$$\|Ax_0\| = c\,\|x_0\|.$$

В силу линейности оператора при определении нормы и проверке приведенных свойств 1) и 2) можно огра-

ничиться векторами, для которых $\|x\| = 1$.

Ясно также, что норма оператора будет зависеть от того, каким образом задана сама норма векторов. В следующей теореме даны простые формулы для нахождения p-норм матриц

$$\|A\|_p = \sup_{x \neq \theta} \frac{\|Ax\|_p}{\|x\|_p}$$

в трех важных для приложений случаях, когда $p = 1$, $p = 2$ и $p = \infty$.

Через A^* мы будем обозначать матрицу, сопряженную к матрице A. Предполагаем, что в $\mathbb{R}^n$ и $\mathbb{C}^n$ заданы стандартные ортонормированные базисы, тогда

$$(x, Ay) = (A^*x, y) \quad \text{для любых векторов } x \text{ и } y,$$

т. е. A^* определяет сопряженный линейный оператор.

Теорема 1.2 *Для квадратной матрицы A порядка n с элементами a_{kj} из $\mathbb{R}$ или $\mathbb{C}$ имеют место следующие формулы для норм:*

$$1) \qquad \|A\|_1 = \max_{1 \leq j \leq n} \alpha_j, \quad \alpha_j := \sum_{k=1}^{n} |a_{kj}|,$$

т. е. $\|A\|_1$ определяется "максимальным" столбцом;

$$2) \qquad \|A\|_\infty = \max_{1 \leq k \leq n} \beta_k, \quad \beta_k := \sum_{j=1}^{n} |a_{kj}|,$$

т. е. $\|A\|_\infty$ *определяется "максимальной" строкой;*

3) $\|A\|_2 = \max\{\sqrt{\lambda} : \lambda - \textit{собственное значение}\, A^*A\}$.

Доказательство. Пусть $x = (x_1, x_2, \ldots, x_n)$, $y = Ax$, где $y = (y_1, y_2, \ldots, y_n)$. Имеем

$$y_k = \sum_{j=1}^{n} a_{kj}\, x_j.$$

1) Очевидно,

$$\|Ax\|_1 = \|y\|_1 = |y_1| + |y_2| + \ldots + |y_n| = \sum_{k=1}^{n} |y_k| =$$

$$= \sum_{k=1}^{n} \left| \sum_{j=1}^{n} a_{kj}\, x_j \right| \le \sum_{j=1}^{n} |x_j| \sum_{k=1}^{n} |a_{kj}| =$$

$$= \sum_{j=1}^{n} \alpha_j |x_j| \le c \sum_{j=1}^{n} |x_j| = c\|x\|_1,$$

где

$$c = \max_{1 \le j \le n} \alpha_j.$$

С другой стороны, существует номер столбца j_0 такой, что $c = \alpha_{j_0}$. Рассмотрим вектор $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)$, у которой $x_{j_0}^0 = 1$, а все остальные координаты равны нулю, и вектор $Ax^0 = y^0 = (y_1^0, y_2^0, \ldots, y_n^0)$. Но тогда

$\|x^0\|_1 = 1$, $y_k^0 = a_{kj_0}$ и

$$\|y^0\|_1 = \sum_{k=1}^{n} |a_{kj_0}| = c = c\|x^0\|_1.$$

Следовательно,

$$\|A\|_1 = c = \max_{1\le j\le n} \alpha_j,$$

что и требовалось доказать.

2) Величина $\|A\|_\infty$ вычисляется проще. Действительно, имеем

$$\|Ax\|_\infty = \max_{1\le k\le n} |y_k| = \max_{1\le k\le n} \left| \sum_{j=1}^{n} a_{kj}\, x_j \right| \le$$

$$\le \max_{1\le k\le n} \sum_{j=1}^{n} |a_{kj}| \max_{1\le j\le n} |x_j| = \|x\|_\infty \max_{1\le k\le n} \beta_k = c\,\|x\|_\infty,$$

где

$$c = \max_{1\le k\le n} \beta_k.$$

С другой стороны, существует номер строки k_0 такой, что $c = \beta_{k_0}$. Рассмотрим вектор $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)$ такой, что $|x_j^0| = 1$ для любого номера j, причем $a_{k_0 j} x_j^0 = |a_{k_0 j}|$. Последнее условие выбора $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)$ можно выполнить так: полагаем $x_j^0 = 1$, если $a_{k_0 j} = 0$; если же $a_{k_0 j}$ является вещественным или комплексным числом, отличным от нуля, то возьмем $x_j^0 = \overline{a_{k_0 j}}/|a_{k_0 j}|$.

Тогда $\|x^0\|_\infty = 1$, а для $Ax^0 = y^0 = (y_1^0, y_2^0, \ldots, y_n^0)$ получаем

$$\|y^0\|_\infty = \max_{1\le k\le n} |y_k^0| = |y_{k_0}^0| = \beta_{k_0} = c = c\,\|x^0\|_\infty,$$

что и требовалось доказать.

3) Отметим прежде всего, что определение $\|A\|_2$ является корректным, так как собственные числа матрицы A^*A являются неотрицательными числами. Действительно, если λ — собственное число этой матрицы и $x \neq \theta$ — соответствующий ему собственный вектор, то $A^*Ax = \lambda x$ и $(A^*Ax, x) = (Ax, Ax) = \lambda(x, x)$, поэтому $\lambda = (\|Ax\|_2/\|x\|_2)^2 \geq 0$.

Как доказывается в курсе линейной алгебры, для самосопряженной матрицы A^*A существует матрица U порядка n, обладающая свойствами:

1) $A^*A = U^*DU$, где $D = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ — диагональная матрица, λ_j — собственные числа матрицы A^*A,

2) $U^* = U^{-1}$,

3) $\|U^{-1}y\|_2 = \|y\|_2$ для всех векторов y.

Применяя это утверждение, а также замену векторов $Ux = y$ и определение нормы $\|A\|_2$, получаем

$$\|A\|_2 = \max_{\|x\|_2=1} \sqrt{(Ax, Ax)} = \max_{\|x\|_2=1} \sqrt{(A^*Ax, x)} =$$

$$= \max_{\|U^{-1}y\|_2=1} \sqrt{(Dy, y)} = \max_{\|y\|_2=1} \sqrt{(Dy, y)} =$$

$$= \max_{\|y\|_2=1} \sqrt{\sum_{k=1}^{n} \lambda_k |y_k|^2}.$$

Отсюда получаем $\|A\|_2 \leq \sqrt{\lambda_{k_0}}$, где λ_{k_0} — максимальное из чисел λ_k. В достижении равенства легко убедиться, выбрав элемент $y^0 = (y_1^0, y_2^0, \ldots, y_n^0)$ такой, что $y_{k_0}^0 = 1$, а все остальные координаты этого вектора равны нулю.

Теорема доказана.

Отметим, что в кольце матриц порядка n определяют также алгебраические матричные нормы, удовлетворяющие следующим условиям:

1) $\|A\| \geq 0$ и $\|A\| = 0 \iff A = 0$ (т. е. все $a_{kj} = 0$);

2) $\|\lambda A\| = |\lambda| \, \|A\|$ для любого скаляра λ;

3) $\|A + B\| \leq \|A\| + \|B\|$;

4) $\|AB\| \leq \|A\| \, \|B\|$.

Нетрудно проверить, что *определенные выше нормы матрицы как нормы линейного оператора удовлетворяют требованиям 1) – 4).*

Существует ряд других алгебраических норм матрицы, удовлетворяющих этим условиям. Например, тре-

бованиям 1) – 4) удовлетворяет норма Фробениуса

$$\|A\|_F = \sqrt{\sum_{k=1}^{n}\sum_{j=1}^{n}|a_{kj}|^2},$$

которую называют евклидовой нормой матрицы, так как $\|A\|_F$ — евклидова норма вектора размерности n^2.

С применением свойства 4) легко получаем, что при любом определении матричной нормы

$$\|A^{-1}\|\,\|A\| \geq \|E\| \geq 1.$$

Заметим, что при любом определении нормы векторов операторная норма единичной матрицы

$$E = \mathrm{diag}(1, 1, \ldots, 1)$$

равна единице, а норма Фробениуса $\|E\|_F = \sqrt{n}$. Следовательно, норма Фробениуса не является операторной нормой для любого $n \geq 2$.

Мы будем пользоваться только операторными нормами матриц. В дальнейшем нам потребуется также следующее

Определение 1.2 *Пусть $\lambda_k = \lambda_k(B)$ — собственные значения матрицы B. Число*

$$\rho(B) = \max_k |\lambda_k(B)|$$

называется спектральным радиусом матрицы B.

В терминах спектрального радиуса равенство

$$\|A\|_2 = \max\{\sqrt{\lambda} : \lambda - \text{собственное значение } A^*A\}.$$

можно записать так: $\|A\|_2 = \sqrt{\rho(A^*A)}$.

Отметим также, что норма $\|A\|_2$ называется спектральной нормой.

Приведем одно из применений норм матриц, а именно, *итерационное уточнение обратной матрицы, вычисленной приближенно*. Пусть X_0 — решение матричного уравнения $AX = E$. Если $\|E - AX_0\| = 0$, то решение определено точно. Но на правктике обратная матрица определяется приближенно из-за округлений и других погрешностей вычислений, и мы имеем лишь приближенное равенство $X_0 \approx A^{-1}$. Предположим, что $\varepsilon > 0$ — допустимая погрешность и требуется следующая точность вычислений: $\|A^{-1} - X_0\| < \varepsilon$. Понятно, что если $\|A^{-1} - X_0\| > \varepsilon$, то необходимо уточнение приближенной обратной матрицы X_0. Оказывается, что если $\|E - AX_0\| \neq 0$, но $\|E - AX_0\| = q < 1$, то существует простой итерационный метод, приводящий за небольшое число шагов к матрице $X_k \approx A^{-1}$, для которой справедливо неравенство $\|A^{-1} - X_k\| < \varepsilon$. А именно, рассмотрим итерации

$$X_0 \to X_1 = X_0(2E - AX_0),$$

$X_2, X_3, \ldots$, определяемые формулой

$$X_k = X_{k-1}(2E - AX_{k-1}), \qquad k = 1, 2, \ldots \tag{1.8}$$

Утверждение. *Если* $\|E - AX_0\| = q < 1$, *то*

$$\|A^{-1} - X_k\| \leq \|A^{-1}\| \cdot q^{2^k},$$

и, следовательно, последовательность X_k *сходится к* A^{-1}, *т. е.* $\lim_{k\to\infty} \|A^{-1} - X_k\| = 0$.

Доказательство. Подставляя вместо X_k ее выражение из формулы (1.8), получаем

$$E - AX_k = E - A(X_{k-1}(2E - AX_{k-1})) =$$

$$= E - AX_{k-1} - AX_{k-1} + AX_{k-1}AX_{k-1} =$$

$$= E - AX_{k-1} - AX_{k-1}(E - AX_{k-1}) = (E - AX_{k-1})^2,$$

т. е. $E - AX_k = (E - AX_{k-1})^2$. Применяя эту формулу k раз, будем иметь

$$E - AX_k = (E - AX_{k-1})^2 =$$

$$= (E - AX_{k-2})^4 = \ldots = (E - AX_0)^{2^k}.$$

Кроме того, имеем простую формулу

$$A^{-1} - X_k = A^{-1}(E - AX_k),$$

поэтому

$$A^{-1} - X_k = A^{-1}(E - AX_k) = A^{-1}(E - AX_0)^{2^k}.$$

Но тогда $\|A^{-1} - X_k\| \le \|A^{-1}\| \cdot \|E - AX_k\| \le$

$$\le \|A^{-1}\| \cdot \|E - AX_0\|^{2^k} = \|A^{-1}\| \cdot q^{2^k} \to 0$$

при $k \to \infty$.

1.6.3 Число обусловленности матрицы

Пусть A — заданная квадратная матрица порядка n, причем $\det A \neq 0$. Тогда определена величина

$$\nu(A) = \|A^{-1}\| \cdot \|A\|,$$

которая называется **числом обусловленности матрицы** A.

Предположим, что вектор $b \neq \theta$ определяется приближенно как $\tilde{b}$ в результате каких-то измерений или приближенных вычислений. Возникает необходимость сравнения решений

$$x^* = A^{-1}b, \qquad \tilde{x^*} = A^{-1}\tilde{b}$$

двух следующих систем линейных алгебраических уравнений

$$Ax = b, \qquad Ax = \tilde{b}.$$

Обозначим $\delta := b - \tilde{b}$ и $\xi := x^* - \tilde{x^*}$.

Имеем следующие числовые характеристики:

$\|\delta\|$ — абсолютная погрешность правой части;

$\|\xi\|$ — абсолютная погрешность решения;

$\|\delta\|/\|b\|$ — относительная погрешность правой части;

$\|\xi\|/\|x^*\|$ — относительная погрешность решения.

Поделим относительную погрешность решения на относительную погрешность правой части. Понятно, что максимум этого отношения, т. е. величина

$$\mu(A) := \sup_{\delta \neq \theta} \frac{\|\xi\|/\|x^*\|}{\|\delta\|/\|b\|},$$

называемая мерой обусловленности СЛАУ, характеризует устойчивость решения по отношению к изменениям правой части системы уравнений.

Теорема 1.3 *Справедлива следующая оценка*

$$\mu(A) \leq \nu(A) := \|A^{-1}\| \cdot \|A\|.$$

Доказательство. Имеем

$$\mu(A) = \sup_{\delta \neq 0} \frac{\|A^{-1}\delta\|}{\|\delta\|} \cdot \frac{\|Ax^*\|}{\|x^*\|}.$$

Пользуясь соотношениями

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

$$\|Ax^*\| \leq \|A\| \cdot \|x^*\|,$$

немедленно получаем

$$\mu(A) \leq \|A^{-1}\| \cdot \|A\| = \nu(A).$$

Теорема доказана.

Замечание. Существуют термины, пришедшие из практики приближенных вычислений. К этому типу терминов относится и термин **число обусловленности матрицы**. Если число $\nu(A) = \|A^{-1}\| \cdot \|A\|$ намного больше единицы, то говорят, что матрица **плохо обусловлена**. Если $\nu(A)$ является не очень большим числом, то говорят, что матрица **хорошо обусловлена**.

Понятно, что термин "матрица плохо обусловлена" отражает реальные проблемы: если матрица системы $Ax = b$ плохо обусловлена, то погрешности коэффициентов матрицы A и погрешности правых частей b, а также погрешности округления при расчетах могут сильно исказить решение. Напомним, что при любом определении операторной нормы

$$1 = \|E\| = \|A^{-1}A\| \leq \|A^{-1}\| \cdot \|A\|,$$

т. е. число обусловленности матрицы не меньше, чем единица.

1.7 Метод простых итераций

Пусть $A = (a_{ij})_{i,j=1}^{n}$ и $B = (b_{ij})_{i,j=1}^{n}$ — квадратные матрицы порядка n с вещественными или комплексными элементами. Рассмотрим системы линейных алгебраических уравнений вида

$$Ax = b, \tag{1.9}$$

а также вида

$$x = Bx + c, \tag{1.10}$$

где $b = (b_1, \ldots, b_n)$ и $c = (c_1, \ldots, c_n)$ — заданные векторы из $\mathbb{R}^n$ или из $\mathbb{C}^n$. Свести (1.9) к эквивалентной системе вида (1.10) можно множеством разных способов. Опишем простейший прием.

Пусть α — фиксированное число, отличное от нуля, и пусть E — единичная матрица. Тогда СЛАУ вида (1.9) равносильна системе $0 = \alpha(b - Ax)$, следовательно, равносильна СЛАУ

$$x = (E - \alpha A)x + \alpha b.$$

Последняя система имеет вид (1.10) с матрицей

$$B = E - \alpha A$$

и с заданным вектором $c = \alpha b$.

Метод простых итераций применяется для нахож-

дения решения

$$x^* = (x_1^*, x_2^*, \ldots, x_n^*)$$

системы вида (1.10) в предположении $\det(E - B) \neq 0$.

В частности, если спектральный радиус $\rho(B) < 1$, то число 1 не является собственным значением матрицы B. Поэтому определитель $\det(E - B) \neq 0$, следовательно, существует обратная матрица $(E - B)^{-1}$ и решение $x^* = (E - B)^{-1} c$.

Алгоритм метода простых итераций заключается в следующем. Выбираем вектор

$$x^0 = (x_1^0, x_2^0, \ldots, x_n^0) \quad \text{— начальное приближение.}$$

Выбор начального приближения субъективен. Понятно, что в качестве начального (= нулевого) приближения желательно назначить вектор, близкий к решению. Если нет никакой информации о решении, то нулевое приближение берем "с потолка" , т. е. выбираем произвольно. Последующие приближения определяются по правилам:

$$x^1 = Bx^0 + c \quad \text{— первое приближение,}$$

$$x^2 = Bx^1 + c \quad \text{— второе приближение,}$$

$$\ldots\ldots\ldots$$

$$x^k = Bx^{k-1} + c \quad \text{— } k\text{-е приближение.}$$

Здесь $x^k = (x_1^k, x_2^k, \ldots, x_n^k)$, причем k означает номер итерации (это не показатель степени!).

Если последовательность векторов $(x^k)_{k=0}^{\infty}$ сходится, т. е. существует некоторый вектор

$$x^* = (x_1^*, x_2^*, \ldots, x_n^*),$$

такой, что

$$x^* = \lim_{k\to\infty} x^k, \quad \text{т. е. } \lim_{k\to\infty} \|x^* - x^k\| = 0,$$

то $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ — решение системы (1.10). Действительно, с учетом непрерывности операций вида Bx, $\|x\|$, при $k \to \infty$ из соотношения $x^k = Bx^{k-1} + c$ получаем, что

$$x^* = Bx^* + c.$$

Применяя формулу $x^k = Bx^{k-1} + c$ и индукцию, легко получаем следующую формулу для выражения k-того приближения x^k через начальное приближение:

$$x^k = B^k x^0 + (E + B + \ldots + B^{k-1})\, c.$$

Здесь $E = B^0$ — единичная матрица.

При изучении сходимости метода простых итераций оказываются полезными следующие утверждения из теории матриц (см., например, стр. 135–138 книги Д.К. Фаддеева и В.Н. Фаддеевой "Вычислительные методы линейной алгебры").

1) $B^k \to 0$ при $k \to \infty$ тогда и только тогда, когда $\rho(B) < 1$.

2) Для того, чтобы ряд $E + B + B^2 + \ldots$ сходился, необходимо и достаточно, чтобы $B^k \to 0$ при $k \to \infty$.

В этом случае, сумма ряда равна $(E - B)^{-1}$, т. е.

$$(E - B)^{-1} = E + B + B^2 + \ldots = \sum_{k=0}^{\infty} B^k.$$

Справедливо следующее утверждение.

Теорема 1.4 *Пусть* $\det(E - B) \neq 0$. *Метод простых итераций $x^k = Bx^{k-1} + c$ ($k = 1, 2, 3, \ldots$) для СЛАУ вида $x = Bx + c$ сходится при любом выборе нулевого приближения x^0 тогда и только тогда, когда спектральный радиус $\rho(B) < 1$.*

Доказательство. Будем пользоваться указанными выше утверждениями из теории матриц. Пусть спектральный радиус $\rho(B) < 1$. Тогда последовательность

$$x^k = B^k x^0 + (E + B + \ldots + B^{k-1})c$$

сходится как сумма двух сходящихся последовательностей.

Докажем обратное утверждение. Пусть

$$det(E - B) \neq 0$$

и метод простых итераций сходится при любом выборе

нулевого приближения x^0.

Если $c = 0$, то $x^* = 0$. Получаем, что последовательность $x^k = B^k x^0$ сходится к нулевому вектору при любом x^0. Следовательно, $\rho(B) < 1$.

Если $c \neq 0$, то возьмем $x^0 = 0$. Тогда имеем сходящуюся последовательность $x^k = (E + B + \ldots + B^{k-1})\, c$, где $c \neq 0$. Следовательно, $\rho(B) < 1$ в силу указанных выше утверждений 1) и 2) из теории матриц.

Теорема доказана.

Заметим, что для матриц порядка $n \geq 3$ спектральный радиус вычисляется сложно. Поэтому проверка критерия $\rho(B) < 1$ представляет собой непростую задачу. Более простое достаточное условие, обеспечивающее сходимость метода простых итераций, имеет вид $\|B\| < 1$. Он годен при любом определении нормы векторов и соответствующей операторной нормы матриц. Напомним, что операторная норма матрицы B определяется формулой:

$$\|B\| = \sup_{x \neq 0} \frac{\|Bx\|}{\|x\|} = \max_{\|x\|=1} \|Bx\|.$$

Теорема 1.5 *Пусть* $\|B\| < 1$. *Тогда*

1) система $x = Bx + c$ *имеет единственное решение* x^*,

2) метод простых итераций

$$x^k = Bx^{k-1} + c \quad (k = 1, 2, 3, \ldots)$$

сходится при любом выборе нулевого приближения x^0,

3) имеет место оценка

$$\|x^* - x^k\| \le \frac{\|B\|^k}{1 - \|B\|} \cdot \|x^1 - x^0\|, \qquad k \in \mathbb{N}.$$

Доказательство. *Единственность* легко доказывается от противного. Действительно, если существуют по крайней мере два решения x^* и y^*, то

$$\begin{cases} x^* = Bx^* + c \\ y^* = By^* + c \end{cases}$$

отсюда следует равенство $z^* = Bz^*$ для $z^* = x^* - y^*$. Но тогда, пользуясь определением нормы оператора, получаем

$$\|z^*\| = \|Bz^*\| \le \|B\| \cdot \|z^*\|,$$

что влечет равенство $z^* = 0$ с учетом соотношения $\|B\| < 1$.

Докажем теперь *существование* решения.

Пусть $m \ge 1, \quad p \ge 1$ — натуральные числа. Пользуясь правилом

$$x^k = Bx^{k-1} + c, \quad k = 1, 2, \ldots,$$

образования итераций, получаем

$$x^{m+p} - x^m =$$

$$= Bx^{m+p-1} - Bx^{m-1} = \ldots = B^m(x^p - x^0).$$

С другой стороны, элементарные вычисления дают, что

$$x^p - x^0 =$$

$$= x^p - x^{p-1} + x^{p-1} - \ldots + x^1 - x^0 =$$

$$= B^{p-1}(x^1 - x^0) + \ldots + B^0(x^1 - x^0).$$

Следовательно, имеем равенство

$$x^{m+p} - x^m = (B^{m+p-1} + B^{m+p-2} + \ldots + B^m)(x^1 - x^0),$$

откуда следует, что

$$\|x^{m+p} - x^m\| \leq (\|B\|^m + \|B\|^{m+1} + \ldots)\|x^1 - x^0\| =$$

$$= \frac{\|B\|^m}{1 - \|B\|}\|x^1 - x^0\|.$$

Поскольку $\|B\| < 1$ и поэтому $\|B\|^m \to 0$ при $m \to \infty$, то и $\|x^{m+p} - x^m\| \to 0$ при $m \to \infty$, что влечет фундаментальность по Коши последовательности итераций. Поэтому существует предел

$$x^* = \lim_{k\to\infty} x^k, \quad \text{т. е. } \lim_{k\to\infty} \|x^* - x^k\| = 0,$$

где x^* — решение. Далее, имеем неравенство

$$\|x^{k+p} - x^k\| \le \frac{\|B\|^k}{1 - \|B\|}\|x^1 - x^0\|, \qquad k \in \mathbb{N}.$$

Переходя к пределу при $p \to \infty$, получаем отсюда требуемую оценку теоремы 1.5 для $\|x^* - x^k\|$.

Замечание. В качественном плане теорема 1.5 тесно связана с теоремой 1.4, так как спектральный радиус $\rho(B) \le \|B\|$ при любом определении нормы. Действительно, для любого собственного значения λ_i и соответствующего собственного вектора матрицы B мы можем записать равенство $Bx^i = \lambda_i x^i, \quad x^i \ne 0$. Отсюда следует, что $|\lambda_i|\|x^i\| = \|Bx^i\| \le \|B\|\|x^i\|$, поэтому $|\lambda_i| \le \|B\|$ для любого собственного значения, что влечет неравенство $\rho(B) \le \|B\|$ при любом определении нормы матрицы. Следовательно, условие $\|B\| < 1$ влечет неравенство $\rho(B) < 1$.

Случай матрицы с диагональным преобладанием

Применим доказанную выше теорему к специальному случаю системы вида $Ax = b$, когда матрица $A = (a_{ij})$ является матрицей с диагональным преобладанием по строкам. Напомним, что по определению диагонального преобладания имеем неравенства

$$|a_{ii}| > \sum_{j=1, j\ne i}^{n} |a_{ij}|$$

для любого $i = 1, 2, \ldots, n$.

Для применения теоремы нам необходимо преобразовать систему $Ax = b$ к системе вида $x = Bx + c$. С этой целью проделаем следующие элементарные преобразования.

Первое уравнение $a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1$ системы $Ax = b$ поделим на a_{11}, 2-ое уравнение поделим на a_{22}, и так далее.

Для любого $k = 1, 2, \ldots, n$ в k-м уравнении системы в левой части оставим x_k, остальные члены перенесем в правую часть уравнения. Получаем следующую систему:

$$\begin{cases} x_1 = -\frac{a_{12}}{a_{11}}x_2 - \frac{a_{13}}{a_{11}}x_3 - \ldots - \frac{a_{1n}}{a_{11}}x_n + \frac{b_1}{a_{11}} \\ x_2 = -\frac{a_{21}}{a_{22}}x_1 - \frac{a_{23}}{a_{22}}x_3 - \ldots - \frac{a_{2n}}{a_{22}}x_n + \frac{b_2}{a_{22}} \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n = -\frac{a_{n1}}{a_{nn}}x_1 - \frac{a_{n2}}{a_{nn}}x_2 - \ldots - \frac{a_{nn-1}}{a_{nn}}x_{n-1} + \frac{b_n}{a_{nn}} \end{cases}.$$

Таким образом, мы получили эквивалентную $Ax = b$ систему вида: $x = Bx + c$, где

$$c = \left(\frac{b_1}{a_{11}}, \frac{b_2}{a_{22}}, \ldots, \frac{b_n}{a_{nn}}\right),$$

$$B = (b_{ij}), \quad b_{ii} = 0, \quad b_{ij} = -\frac{a_{ij}}{a_{ii}} (i \neq j).$$

Вычислим норму $\|B\|_\infty$ матрицы B. Эта норма меньше единицы в силу условий диагонального преобладания,

так как

$$\|B\|_\infty = \max_i \sum_{j=1}^{n} |b_{ij}| = \max_i \frac{\sum_{j=1, j\neq i}^{n} |a_{ij}|}{|a_{ii}|} < 1.$$

Следовательно, система $x = Bx + c$ имеет единственное решение, которое можно найти методом простой итерации $x^k = Bx^{k-1} + c$, причем итерации сходятся при любом выборе нулевого приближения. В частности, отсюда следует, что система $Ax = b$ имеет единственное решение при любой правой части, так как неравенство $\|B\|_\infty < 1$ не зависит от b. Поэтому как следствие теоремы о сходимости метода простых итераций и наших построений получаем следующее утверждение.

Теорема 1.6 *Пусть A — матрица с диагональным преобладанием по строкам. Тогда* $\det A \neq 0$.

В заключение укажем некоторые другие, употребительные **способы преобразования системы $Ax = b$ к системе вида $x = Bx + c$.**

Способ 1. Возьмем некоторую невырожденную матрицу H. Имеем: $\det H \neq 0$. Тогда система $Ax = b$ эквивалентна системе $0 = H(b - Ax)$, которая в свою очередь эквивалентна системе

$$x = x + H(b - Ax).$$

Таким образом, мы получаем эквивалентную систему

$x = Bx + c$, где

$$c = Hb, \quad B = E - HA.$$

Обычно стремятся подобрать H так, чтобы

$$\|B\| = \|E - HA\| << 1.$$

Если A^{-1} существует, то мы можем взять $H = A^{-1}$, отсюда $B = 0$.

Замечание. Выбор $B = E - \alpha A$ — частный случай этого способа. Известно такое утверждение (докажите!). *Пусть матрица A является эрмитовой и положительно определенной (т. е. $A = A^*$ и справедливо неравенство $(Ax, x) > 0$ для любого $x \neq \theta$). Тогда для всех достаточно малых $\alpha > 0$ матрица $E - \alpha A$ является эрмитовой и спектральный радиус*

$$\rho(E - \alpha A) < 1.$$

Способ 2. Матрицу A представляем в виде

$$A = C + D,$$

причем $\det C \neq 0$. Тогда система $Ax = b$ перепишется в виде $Cx + Dx = b$, что эквивалентно системе

$$x = -C^{-1}Dx + C^{-1}b.$$

Таким образом, получаем равносильную систему линейных алгебраических уравнений вида $x = Bx + c$, где $c = C^{-1}b, \quad B = -C^{-1}D$.

1.8 Методы Зейделя

I вариант метода Зейделя для систем вида

$$x = Bx + c$$

Рассмотрим систему линейных алгебраических уравнений вида

$$x = Bx + c, \quad B = (b_{ij}) \text{ — } n \times n\text{-матрица},$$

где $x = (x_1, \ldots, x_n), \quad c = (c_1, \ldots, c_n)$. В координатной записи система уравнений $x = Bx + c$ имеет вид

$$x_i = \sum_{j=1}^{n} b_{ij}x_j + c_i, \quad i = 1, 2, \ldots, n.$$

Задаем начальное приближение

$$x^0 = (x_1^0, x_2^0, \ldots, x_n^0).$$

В методе простых итераций для любого номера i координаты последующей итерации определялись по фор-

муле

$$x_i^k = \sum_{j=1}^{n} b_{ij} x_j^{k-1} + c_i, \quad k = 1, 2, \ldots.$$

Метод Зейделя представляет собой модификацию метода простых итераций, и приведенная формула сохраняется только для первой координаты. Алгоритм Зейделя таков:

$$x_1^k = \sum_{j=1}^{n} b_{1j} x_j^{k-1} + c_1,$$

но

$$x_2^k = b_{21} x_1^k + \sum_{j=2}^{n} b_{2j} x_j^{k-1} + c_2.$$

Для определения x_3^k используются величины x_1^k, x_2^k, уже известные по двум предыдущим формулам. А именно, полагаем

$$x_3^k = b_{31} x_1^k + b_{32} x_2^k + \sum_{j=3}^{n} b_{3j} x_j^{k-1} + c_3,$$

и далее, для любого $i \geq 2$ алгоритм Зейделя задается формулой

$$x_i^k = \sum_{j=1}^{i-1} b_{ij} x_j^k + \sum_{j=i}^{n} b_{ij} x_j^{k-1} + c_i.$$

Запишем метод Зейделя с использованием матриц.

Полагаем $B = H + F$, где

$$H = \begin{pmatrix} 0 & 0 & \dots & 0 \\ b_{21} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ b_{n1} & b_{n2} & \dots & 0 \end{pmatrix};$$

$$F = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ 0 & b_{22} & \dots & b_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & b_{nn} \end{pmatrix}.$$

Легко проверить, что итерационный метод Зейделя дает следующий алгоритм:

$$x^k = Hx^k + Fx^{k-1} + c.$$

Это эквивалентно алгоритму

$$(E - H)x^k = Fx^{k-1} + c.$$

Поскольку $E - H$ — треугольная матрица и

$$\det(E - H) = 1 \neq 0,$$

то алгоритм Зейделя оказывается эквивалентным алгоритму

$$x^k = (E - H)^{-1}Fx^{k-1} + (E - H)^{-1}c.$$

Таким образом, метод Зейделя эквивалентен методу простых итераций для системы линейных алгебраических уравнений

$$x = \widetilde{B}x + \widetilde{c},$$

где

$$\widetilde{B} = (E - H)^{-1}F, \quad \widetilde{c} = (E - H)^{-1}c.$$

Очевидно, мы можем применить теоремы о сходимости метода простых итераций к вопросу о сходимости метода Зейделя с заменой матрицы B на матрицу $\widetilde{B}$.

Теорема 1.7 *Пусть матрица $B = H + F$, где матрицы H и F определены выше.*

1) Если

$$\|(E - H)^{-1}F\| < 1,$$

то метод Зейделя сходится при любом начальном приближении x^0. Решение уравнения единственное. Для погрешности k-того приближения справедлива оценка

$$\|x^* - x^k\| \le \frac{q^k}{1 - q}\|x^1 - x^0\|,$$

где $q = \|(E - H)^{-1}F\|$.

2) Пусть $det(E - B) \neq 0$. Метод Зейделя (первый вариант) сходится при любом выборе нулевого приближения x^0 тогда и только тогда, когда спектральный радиус

$$\rho((E - H)^{-1}F) < 1.$$

II вариант метода Зейделя для систем вида $Ax = b$

Этот метод является итерационным и применяется к системе $Ax = b$ в предположении, что все диагональные элементы матрицы системы отличны от нуля. Опишем алгоритм. Как обычно, задаем нулевое приближение $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)$. Определяем последовательно координаты вектора $x^1 = (x_1^1, x_2^1, \ldots, x_n^1)$, затем последовательно вычисляем координаты вектора $x^2 = (x_1^2, x_2^2, \ldots, x_n^2)$ и так далее.

Для детального описания метода, а именно, для описания итерационного перехода от x^{k-1} к x^k, запишем систему $Ax = b$ в координатах. Диагональные элементы и те, которые ниже главной диагонали, заменим на x_i^k, а то, что выше диагонали, заменим на x_i^{k-1}. Имеем уравнения

$$\begin{cases} a_{11}x_1^k + a_{12}x_2^{k-1} + \ldots + a_{1n}x_n^{k-1} = b_1 \\ a_{21}x_1^k + a_{22}x_2^k + \ldots + a_{2n}x_n^{k-1} = b_2 \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1^k + a_{n2}x_2^k + \ldots + a_{nn}x_n^k = b_n \end{cases}.$$

Из 1-го уравнения определяем x_1^k, из 2-го уравнения определяем x_2^k. И так далее. Более точно, имеем

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

$$x_2^k = \frac{b_2 - a_{21}x_1^k - a_{23}x_3^{k-1} - \ldots - a_{2n}x_n^{k-1}}{a_{22}}, \ldots,$$

$$x_n^k = \frac{b_n - a_{n1}x_1^k - a_{n2}x_2^k - \ldots - a_{nn-1}x_{n-1}^k}{a_{nn}}.$$

Таким образом, зная вектор x^0, последовательно найдем векторы $x^1 \to x^2 \to \ldots \to x^k \to \ldots$. Очевидно, этот алгоритм в матричной форме можно записать так: пусть $A = B + C$, где

$$B = \begin{pmatrix} a_{11} & 0 & \ldots & 0 \\ a_{21} & a_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix},$$

$$C = \begin{pmatrix} 0 & a_{12} & \ldots & a_{1n} \\ 0 & 0 & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 \end{pmatrix}.$$

Тогда итерационный алгоритм Зейделя для системы $Ax = b$ можно записать следующим образом:

$$Bx^k + Cx^{k-1} = b. \tag{1.11}$$

Поскольку диагональные элементы матрицы A предполагаются отличными от нуля и поэтому

$$\det B = a_{11}a_{22} \cdot \ldots \cdot a_{nn} \neq 0,$$

то существует обратная матрица B^{-1}, и поэтому алгоритм может быть представлен в форме

$$x^k + B^{-1}Cx^{k-1} = B^{-1}b.$$

Ясно, что метод Зейделя (второй вариант) эквивалентен методу простых итераций

$$x^k = \widetilde{B}x^{k-1} + c, \tag{1.12}$$

где

$$\widetilde{B} = -B^{-1}C, \quad c = B^{-1}b.$$

Но тогда теоремы о сходимости метода простых итераций (1.12) позволяют сформулировать теоремы сходимости для изучаемого метода Зейделя.

Теорема 1.8 *1) Пусть* $\det A \neq 0$, $a_{11}a_{22} \cdot \ldots \cdot a_{nn} \neq 0$. *Второй вариант метода Зейделя сходится для любого* x^0 *тогда и только тогда, когда спектральный радиус* $\rho(\widetilde{B}) < 1$, *где* $\widetilde{B} = -B^{-1}C$.

2) Если

$$\|B^{-1}C\| < 1,$$

то метод Зейделя (второй вариант) сходится для любого x^0 *и имеет место оценка*

$$\|x^* - x^k\| \leq \frac{\|\widetilde{B}\|^k}{1 - \|\widetilde{B}\|} \cdot \|x^1 - x^0\|.$$

1.9 Методы градиентного спуска

Нам потребуется простая связь системы линейных алгебраических уравнений с экстремумом квадратичной функции.

Рассмотрим систему $Ax = b$, где $b \in \mathbb{R}^n$, а матрица A порядка n удовлетворяет двум следующим условиям.

Условие 1. Матрица $A = (a_{ij})$ является действительной и симметричной, т. е. $a_{ij} \in \mathbb{R}$ и $a_{ij} = a_{ji}$. Следовательно, $A = A^T = A^*$, т. е. матрица является самосопряженной. В частности, для любых двух векторов $x \in \mathbb{R}^n$ и $y \in \mathbb{R}^n$ имеет место равенство скалярных произведений

$$(Ax, y) = (x, Ay).$$

Условие 2. Матрица $A = (a_{ij})$ является положительно определенной, что означает $(Ax, x) > 0$ для любого $x \neq \theta$. В частности, из этого условия вытекает, что $\det A \neq 0$. Поэтому существует A^{-1}, и наша система имеет единственное решение $x^* = A^{-1}b$.

Рассмотрим квадратичную функцию

$$F(x) = (Ax, x) - 2(b, x) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j - 2 \sum_{i=1}^{n} b_i x_i.$$

Теорема 1.9 *Пусть A — действительная, симметричная, положительно определенная матрица. Тогда справедливы следующие утверждения:*

1) $x^* = A^{-1}b$ доставляет минимум функционалу $F(x)$, т. е.

$$F(x) \geq F(x^*),$$

для любого $x \in \mathbb{R}^n$.

2) Если $\widetilde{x}$ — точка минимума, т. е.

$$F(x) \geq F(\widetilde{x}),$$

для любого $x \in \mathbb{R}^n$, то $\widetilde{x} = x^*$.

Доказательство. Оба утверждения теоремы являются следствиями тождества

$$F(x) - F(x^*) = (A(x - x^*), x - x^*)$$

для квадратичной функции $F(x) = (Ax, x) - 2(b, x)$. Само тождество также легко проверяется: раскрываем скобки в правой и левой частях тождества и убеждаемся в равенстве с использованием соотношений

$$(x, Ay) = (Ay, x), \quad Ax^* = b$$

и коммутативности скалярного произведения в пространстве $\mathbb{R}^n$. Действительно, с одной стороны,

$$F(x) - F(x^*) = (Ax, x) - 2(b, x) - (Ax^*, x^*) + 2(b, x^*) =$$

$$= (Ax, x) - 2(Ax^*, x) + (Ax^*, x^*).$$

С другой стороны, имеем:

$$(A(x-x^*),x-x^*)=(Ax,x)-(Ax^*,x)-(Ax,x^*)+$$

$$+(Ax^*,x^*)=(Ax,x)-2(Ax^*,x)+(Ax^*,x^*).$$

Этим и завершается доказательство теоремы.

На основании этой теоремы поиск решения системы $Ax=b$ сводится к поиску точки минимума функции n переменных, а именно, квадратичной функции, определенной равенством

$$F(x)=F(x_1,\ldots,x_n)=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x_ix_j-2\sum_{i=1}^{n}b_ix_i.$$

Как известно из курса математического анализа, точка минимума этой функции является решением системы уравнений

$$\frac{\partial F(x_1,\ldots,x_n)}{\partial x_i}=0,\quad i=1,2,\ldots,n.$$

Непосредственными вычислениями с учетом равенств $a_{ij}=a_{ji}$, находим

$$\frac{\partial F(x_1,\ldots,x_n)}{\partial x_i}=2\left(\sum_{j=1}^{n}a_{ij}x_j-b_i\right),\quad i=1,2,\ldots,n.$$

Сравнение последних формул наглядно показывает связь между решениями системы $Ax=b$ с действи-

тельной симметричной матрицей и точками экстремума квадратичной функции $F(x) = (Ax, x) - 2(b, x)$.

Метод покоординатного спуска

Пусть $A = (a_{ij})$ — действительная симметричная положительно определенная матрица. Для любого $k = 1, 2, \ldots, n$ имеем $a_{kk} = (Ae_k, e_k) > 0$, где e_k – базисный вектор, k-тая координата которого равна единице, а остальные координаты равны нулю.

Алгоритм определения точки минимума функции $F(x) = (Ax, x) - 2(b, x)$, называемый методом покоординатного спуска, заключается в следующем. Как обычно, берем некоторое начальное приближение

$$x^0 = (x_1^0, \ldots, x_n^0) \in \mathbb{R}^n.$$

Рассмотрим вспомогательную функцию

$$y = F(x_1, x_2^0, \ldots, x_n^0)$$

одной переменной x_1 и находим точку экстремума как корень уравнения

$$\frac{\partial F(x_1, x_2^0, \ldots, x_n^0)}{\partial x_1} = 2\left(a_{11}x_1 + \sum_{j=2}^{n} a_{1j}x_j^0 - b_1\right) = 0.$$

Отсюда находим

$$x_1^1 = \frac{-a_{12}x_2^0 - \ldots - a_{1n}x_n^0 + b_1}{a_{11}}.$$

Для определения x_2^1 рассматриваем функцию

$$y = F(x_1^1, x_2, x_3^0, \ldots, x_n^0),$$

и определяем точку экстремума $x_2 = x_2^1$ как корень уравнения

$$\frac{\partial F(x_1^1, x_2, x_3^0, \ldots, x_n^0)}{\partial x_2} = 0.$$

Имеем

$$x_2^1 = \frac{-a_{21}x_1^1 - a_{23}x_3^0 - \ldots - a_{2n}x_n^0 + b_2}{a_{22}}.$$

Продолжаем процесс. Дальнейшие подробности не приводим, так как ясно, что метод покоординатного спуска в точности совпадает с методом Зейделя (второй вариант) и для него справедлива теорема 1.8 о сходимости.

Метод градиентного спуска

Поясним сначала идею градиентного спуска в общем случае. Пусть $F : \mathbb{R}^n \to \mathbb{R}$ — некоторая непрерывно дифференцируемая функция, имеющая единственную точку минимума $x^* \in \mathbb{R}^n$. Отправляясь от некоторого нулевого приближения $x^0 \in \mathbb{R}^n$ можно организовать "спуск" от точки $x^0 \in \mathbb{R}^n$ к точке минимума $x^* \in \mathbb{R}^n$ с помощью итераций

$$x^k = x^{k-1} - \tau \operatorname{grad} F(x^{k-1}), \quad k = 1, 2, \ldots,$$

где τ — фиксированное положительное число, возможно, достаточно малое. Поскольку градиент функции направлен в сторону возрастания этой функции, то антиградиент $-\mathrm{grad}\, F(x^{k-1})$ задает направление убывания. Интуитивно понятно, что в пределе мы придем к точке минимума. Геометрически понятно также, что если число τ велико, то итерационный процесс будет расходящимся. Возникает важный вопрос о выборе подходящего параметра τ. Рассмотрим простой пример.

Возьмем $n = 2$, $x = (x_1, x_2)$ и $F(x) = x_1^2 + x_2^2$. Тогда $x^* = (0, 0)$, $\mathrm{grad} F(x) = (2x_1, 2x_2) = 2x$ и итерации запишутся так:

$$x^k = x^{k-1} - 2\tau\, x^{k-1} = (1 - 2\tau)\, x^{k-1} = \ldots = (1 - 2\tau)^k\, x^0.$$

Ясно, что для любого $x^0 \neq x^* = (0, 0)$ сходимость имеет место тогда и только тогда, когда $|1-2\tau| < 1$, т. е. когда $0 < \tau < 1$.

Более общий метод градиентного спуска, называемый нестационарным методом градиентного спуска, имеет вид

$$x^k = x^{k-1} - \tau_{k-1}\, \mathrm{grad}\, F(x^{k-1}), \quad k = 1, 2, \ldots,$$

где τ_{k-1} — положительное число, зависящее от k.

Применим описанную идею спуска к поиску решения системы линейных алгебраических уравнений вида $Ax = b$, где $A = (a_{ij})_{i,j=1}^n$ — действительная, сим-

метричная и положительно определенная матрица. Тогда, как мы уже знаем, $x^* = A^{-1}b$ доставляет минимум функции, определенной формулой

$$F(x) = (Ax, x) - 2(b, x) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j - 2 \sum_{i=1}^{n} b_i x_i.$$

Поскольку для нашей функции

$$\operatorname{grad} F(x) = 2[Ax - b],$$

то итерации метода градиентного спуска для нахождения решения системы $Ax = b$ можно определить формулой

$$x^k = x^{k-1} + t\,(b - Ax^{k-1}),$$

где параметр $t = 2\tau > 0$, или более общей формулой

$$x^k = x^{k-1} + t_{k-1}\,(b - Ax^{k-1}),$$

где параметры $t_{k-1} > 0$.

Справедливо следующее утверждение о сходимости стационарного метода градиентного спуска.

Теорема 1.10 *Предположим, что A — действительная, симметричная, положительно определенная матрица, $\rho(A)$ — спектральный радиус этой матрицы.*

Пусть параметр t выбран таким, что

$$t \in \left(0, \frac{2}{\rho(A)}\right).$$

Тогда метод градиентного спуска

$$x^k = x^{k-1} + t\,(b - Ax^{k-1})$$

сходится при любом начальном приближении $x^0 \in \mathbb{R}^n$, причем

$$\|x^* - x^k\|_2 \le q^k\,\|x^* - x^0\|_2,$$

где

$$q = \|E - t\,A\|_2 < 1.$$

Доказательство. Существование и единственность решения $x^* = A^{-1}b$ гарантированы тем, что A — действительная, симметричная, положительно определенная матрица. Рассматриваем соотношения:

$$\begin{cases} x^* = x^* + t\,(b - Ax^*) & \text{— тождество,} \\ x^k = x^{k-1} + t\,(b - Ax^{k-1}) & \text{— заданная итерация.} \end{cases}$$

Вычитаем из первого уравнения второе. Применяя элементарные преобразования и индукцию, получим

$$x^* - x^k = x^* - x^{k-1} + t(-Ax^* + Ax^{k-1}) =$$

$$= (E - t\,A)(x^* - x^{k-1}) =$$

$$= (E - t\,A)(E - t\,A)(x^* - x^{k-2}) = \ldots =$$

$$= (E - tA)^k (x^* - x^0).$$

Обозначим $q = \|E - t\,A\|_2$. Тогда

$$\|x^* - x^k\|_2 \le q^k \, \|x^* - x^0\|_2.$$

Очевидно, если $q = \|E - t\,A\|_2 < 1$, то итерации сходятся. Нам остается оценить q и убедиться, что $q < 1$.

Поскольку матрица $E - t\,A$ является симметричной и действительной, то она будет самосопряженной. Поэтому имеем равенство $q = \|E - t\,A\|_2 = \rho(E - t\,A)$.

Ясно, что собственные значения матрицы $E - t\,A$ имеют вид $1 - t\,\lambda$, где λ — собственное значение самосопряженной, положительно определенной матрицы A. Так как $t \in (0,\, 2/\rho(A))$ и

$$\lambda \in [\lambda_{\min}, \rho(A)] \subset (0, \rho(A)],$$

то получаем, что $1 - t\lambda < 1$ и $1 - t\lambda \ge 1 - t\rho(A) > -1$.

Таким образом, $-1 < 1 - t\lambda < 1$, где λ — любое из собственных значений матрицы A. Следовательно,

$$q = \max_{\lambda} |1 - t\,\lambda| < 1$$

для любого фиксированного числа $t \in (0,\, 2/\rho(A))$.

Теорема доказана.

Рассмотрим теперь один из нестационарных методов градиентного спуска, называемый методом наиско-

рейшего спуска, который связан с параметром

$$t_k = \frac{(r^k, r^k)}{(r^k, Ar^k)}, \tag{1.13}$$

где $r^k = b - Ax^k$. Если на некотором шаге $r^k = 0$, то $0 = b - Ax^k$, следовательно, $x^* = x^k$. Процесс обрывается, так как найдено точное решение. Если же $r^k \neq 0$ для любого k, то итерационный процесс продолжается бесконечно, и его сходимость гарантируется следующей теоремой Канторовича.

Теорема 1.11 *Пусть A — действительная, симметричная, положительно определенная матрица. Тогда метод градиентного спуска с выбором (1.13), т. е. с итерациями*

$$x^k = x^{k-1} + \frac{(r^{k-1}, r^{k-1})}{(r^{k-1}, Ar^{k-1})}(b - Ax^{k-1}),$$

сходится при любом начальном приближении $x^0 \in \mathbb{R}^n$, причем

$$\|x^* - x^k\|_2 \leq \frac{\|b - Ax^0\|_2}{m}\left(\frac{M-m}{M+m}\right)^k,$$

где

$$M = \|A\|_2, \quad m = \frac{1}{\|A^{-1}\|_2}.$$

Схема доказательства такова. Как и в предыдущем случае, существование и единственность решения

$x^* = A^{-1}b$ гарантированы тем, что A — действительная симметричная положительно определенная матрица. Рассматриваем соотношения

$$\begin{cases} x^* = x^* + t_{k-1}(b - Ax^*), \\ x^k = x^{k-1} + t_{k-1}(b - Ax^{k-1}). \end{cases}$$

Вычитаем из первого равенства второе. Применяя элементарные преобразования и индукцию, получаем

$$x^* - x^k = x^* - x^{k-1} + t_{k-1}(-Ax^* + Ax^{k-1}) =$$

$$= (E - t_{k-1}A)(x^* - x^{k-1}) =$$

$$= (E - t_{k-1}A)(E - t_{k-2}A)(x^* - x^{k-2}) = \ldots =$$

$$= \prod_{i=0}^{k-1}(E - t_iA)(x^* - x^0).$$

Кроме того, имеем

$$x^* - x^0 = A^{-1}b - x^0 = A^{-1}(b - Ax^0),$$

отсюда следует, что

$$\|x^* - x^0\|_2 \leq \|A^{-1}\|_2 \cdot \|b - Ax^0\|_2 = \frac{\|b - Ax^0\|_2}{m}.$$

Применяя доводы, аналогичные тем, которые использовались при доказательстве предыдущей теоре-

мы, можно показать, что

$$\left\| \prod_{i=0}^{k-1} (E - t_i A) \right\|_2 \leq \left(\frac{M-m}{M+m} \right)^k \to 0 \quad \text{при} \quad k \to \infty.$$

Этим и завершается доказательство.

Необходимо отметить, что в учебной и научной литературе по численным методам можно найти ряд обобщений изученных нами методов градиентного спуска. Отметим лишь одну плодотворную идею, позволяющего каждому создать и исследовать новый метод градиентного спуска. А именно, возьмем невырожденные квадратные матрицы C_k порядка n и параметры $t_k > 0$. Тогда можно рассмотреть обобщенный метод градиентного спуска, задавая итерации формулой

$$x^k = x^{k-1} + t_{k-1} C_{k-1} (b - Ax^k).$$

Покажите, что специальным выбором t_k и C_k можно получить метод Зейделя как частный случай обобщенного метода градиентного спуска.

В заключение приведем метод итерационного уточнения приближенного решения СЛАУ.

Рассмотрим систему $Ax = b$, где $A = (a_{ij})$ — квадратная матрица порядка n. Пусть $\det A \neq 0$. Теоретически мы можем тогда найти точное решение по формуле $x^* = A^{-1}b$. Часто на практике обратная матрица A^{-1} определяется приближенно. Но тогда $\widetilde{x}^* = \widetilde{A}^{-1}b$ яв-

ляется лишь приближенным решением. Возникает вопрос: как уточнить приближенное решение? Этого можно достичь с помощью итерационного уточнения найденного приближенного решения.

Если $Ax = b$, то $0 = \widetilde{A}^{-1}(b - Ax)$. Следовательно, система $Ax = b$ эквивалентна системе

$$x = x + \widetilde{A}^{-1}(b - Ax),$$

или, что то же самое, системе

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

Обозначим $B = E - \widetilde{A}^{-1}A$, $c = \widetilde{A}^{-1}b$.

Пусть

$$\|B\| = \|E - \widetilde{A}^{-1}A\| = q < 1,$$

тогда метод простых итерации $x^k = Bx^{k-1} + c$ сходится при любом выборе нулевого приближения $x^0 \in R^n$ и справедлива оценка

$$\|x^* - x^k\| \le q^k \|x^* - x^0\| \le \frac{q^k}{1-q} \|x^1 - x^0\|.$$

Следовательно, $\|x^* - x^k\| \to 0$ при $k \to \infty$.

1.10 Задачи и упражнения

1. Пусть A — невырожденная матрица. Покажите, что для любого $\lambda \in \sigma(A)$ справедливы неравенства $1/\|A^{-1}\| \leq |\lambda| \leq \|A\|$.

2. Докажите неравенство $\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty$.

3. Убедитесь в том, что для произвольных матриц A и B спектры матриц AB и BA совпадают.

4. Покажите, что число обусловленности матрицы A не меняется при умножении матрицы A на ненулевое число.

5. Пусть A — симметричная положительно определенная матрица, $A \neq \beta E$ для $\beta \in \mathbb{R}$. Докажите, что число обусловленности

$$\|(A+\alpha E)^{-1}\|_2 \ \|A+\alpha E)\|_2 \quad (\alpha > 0)$$

является монотонно убывающей функцией от α.

Указание. Решения задач 1 – 5 можно найти в [7].

6. Вычислите решение системы

$$10^{-3}x_1 + x_2 = 5, \quad x_1 - x_2 = 6$$

двумя методами: основным методом Гаусса и методом Гаусса с выбором ведущего элемента как максимального по модулю среди элементов столбца. Проведите вычисления с двумя значащими цифрами после запятой и сравните результаты. (Задача взята из книги [12]).

7. Докажите известную формулу: если $n \geq 2$, то

$$\Delta_n = \begin{vmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-1} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_n & x_n^2 & \dots & x_n^{n-1} \end{vmatrix} = \prod_{i>j} (x_i - x_j)$$

для определителя Вандермонда.

Указание. Легко обосновать формулу:

$$\Delta_n(x) = \begin{vmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-1} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n-1} & x_{n-1}^2 & \dots & x_{n-1}^{n-1} \\ 1 & x & x^2 & \dots & x^{n-1} \end{vmatrix} = \Delta_{n-1} \prod_{j=1}^{n-1} (x - x_j).$$

Далее применяем метод математической индукции.

8. Пусть A — квадратная матрица порядка n, норма $\|A\|$ определена как норма линейного оператора. Докажите формулу Бёрлинга

$$\rho(A) = \lim_{k \to \infty} \sqrt[k]{\|A^k\|}.$$

9. Пусть $A = \{a_{kj}\}_{k,j=1}^n$ — квадратная матрица порядка $n \geq 2$. Докажите, что число $n \max_{k,j} |a_{kj}|$ является алгебраической нормой матрицы, но не является операторной нормой.

Глава 2

Нелинейные уравнения и системы

Будем рассматривать уравнение вида: $f(x) = 0$, где $x \in \mathbb{R}$ или $x \in S$, S – некоторый отрезок $[a, b] \subset \mathbb{R}$. Предполагаем, что отображение

$$f : \mathbb{R} \to \mathbb{R}$$

или

$$f : S \to \mathbb{R}$$

является непрерывной функцией. Решение нелинейного уравнения требует предварительного анализа: применяя аналитические и графические методы, нужно исследовать множество корней и определить интервал расположения интересующего нас корня.

Опишем сначала элементарный метод, связанный с последовательным делением пополам интервала расположения корня x^*.

2.1 Метод дихотомии

Основан на теореме Коши о промежуточном значении функции, непрерывной на некотором отрезке. Точнее, нам нужен следующий частный случай теоремы Коши:

если функция f непрерывна на отрезке $[a, b]$ и имеет место неравенство $f(a) \cdot f(b) < 0$, то существует такая точка $c \in (a, b)$, что $f(c) = 0$.

Итак, пусть функция f непрерывна на отрезке $[a, b]$ и справедливо неравенство $f(a) \cdot f(b) < 0$. Корень

$$x^* = c \in (a, b)$$

можно найти с помощью следующего итерационного процесса. Возьмем середину отрезка

$$x_1 = \frac{a + b}{2}.$$

Возможны 3 случая.

Случай 1: $f(x_1) = 0$. Тогда процесс завершен: число $x^* = x_1$ — искомый корень.

Случай 2: $f(x_1) \neq 0$ и $f(a)f(x_1) > 0$. Тогда

$$f(b)f(x_1) < 0,$$

поэтому берем половину исходного отрезка, полагая

$$[a_1, b_1] = [x_1, b].$$

Случай 3: $f(x_1) \neq 0$ и $f(a)f(x_1) < 0$. Тогда берем половину исходного отрезка, полагая $[a_1, b_1] = [a, x_1]$.

На втором шаге возьмем середину отрезка $[a_1, b_1]$:

$$x_2 = \frac{a_1 + b_1}{2}.$$

Снова возможны 3 случая. Имеем: либо $f(x_2) = 0$ (и тогда процесс завершен, так как $x^* = x_2$ — искомый корень), либо существует половина $[a_2, b_2]$ (вида $[x_2, b_1]$ или $[a_1, x_2]$) отрезка $[a_1, b_1]$, обладающая свойством

$$f(a_2)f(b_2) < 0.$$

Далее, продолжаем процесс деления отрезка пополам. На 3-ем шаге рассматриваем середину отрезка $[a_2, b_2]$:

$$x_3 = \frac{a_2 + b_2}{2}.$$

Снова имеем: либо $f(x_3) = 0$ и процесс завершается, либо существует подходящая половина $[a_3, b_3]$ отрезка $[a_2, b_2]$, обладающая свойством $f(a_3)f(b_3) < 0$, и тогда продолжаем процесс деления отрезка пополам. Очевидно, при продолжении процесса деления возможны два исхода: либо на некотором шаге мы найдем точное значение корня $x^* = x_k$, либо существуют бесконечная последовательность точек $x_{k+1} = (a_k + b_k)/2$ и счетная

система отрезков

$$[a_{k+1}, b_{k+1}] \subset [a_k, b_k], \quad (k \in \mathbb{N})$$

со следующим свойством: $f(a_k)f(b_k) < 0$. Ясно, что тогда

$$b_k - a_k = \frac{b-a}{2^k}, \qquad x^* = \lim_{k\to\infty} x_k,$$

и $f(x^*) = 0$ в силу непрерывности функции f.

2.2 Итерационные методы

Уравнение $f(x) = 0$ заменяем на равносильное уравнение вида: $x = \varphi(x)$. Переход от первого уравнения ко второму можно осуществить различными способами. Например, уравнение $f(x) = 0$ равносильно уравнению $x = \varphi(x)$, где $\varphi(x) = x - f(x)$.

2.2.1 Применение простых итераций

Рассмотрим стандартный метод простых итераций для решения нелинейного уравнения вида $x = \varphi(x)$.

А именно, выбираем нулевое приближение $x_0 \in \mathbb{R}$ и рассматриваем итерации, определяемые формулой: $x_k = \varphi(x_{k-1})$, где $k = 1, 2, \ldots$.

Как следствие теоремы Банаха о сжимающих отображениях имеем следующее утверждение.

Теорема 2.1 *Пусть функция* $\varphi : \mathbb{R} \to \mathbb{R}$ *удовлетворяет условию Липшица*

$$|\varphi(x) - \varphi(y)| \leq \alpha|x - y|, \quad \forall x, y \in \mathbb{R},$$

где постоянная $\alpha \in (0, 1)$. *Тогда существует единственный корень* x^* *уравнения* $x = \varphi(x)$, *при любом выборе нулевого приближения* $x_0 \in \mathbb{R}$ *итерационный метод* $x_k = \varphi(x_{k-1})$ *сходится, а именно,*

$$x^* = \lim_{k\to\infty} x_k,$$

причем

$$|x^* - x_k| \leq \frac{\alpha^k}{1 - \alpha}|x_1 - x_0|.$$

При удачном выборе x_0 теорему 2.1 можно распространить на функции, заданные на некотором отрезке.

Теорема 2.2 *Пусть точка* $x_0 \in \mathbb{R}$ *такова, что*

$$|\varphi(x) - \varphi(y)| \leq \alpha|x - y|, \quad \forall x, y \in S = [x_0 - \varepsilon, x_0 + \varepsilon],$$

где $\varepsilon > 0$, $\alpha < 1$. *Пусть число* $m = |\varphi(x_0) - x_0|$ *удовлетворяет условию*

$$\frac{m}{1 - \alpha} \leq \varepsilon.$$

Тогда на S *уравнение* $x = \varphi(x)$ *имеет единственный*

корень x^*, причем

$$x^* = \lim_{k\to\infty} x_k,$$

где

$$x_k = \varphi(x_{k-1}), \quad k = 1, 2, \ldots.$$

Доказательство. Для применения теоремы о сжимающих отображениях нужно показать, что для любого $x \in S$ значение функции также лежит на S, т. е. $\varphi(x) \in S$. Иными словами, мы имеем дело с отображением $\varphi : S \to S$. Этот факт устанавливается просто. Действительно, имеем

$$\varphi(x) - x_0 = \varphi(x) - \varphi(x_0) + \varphi(x_0) - x_0,$$

отсюда следует

$$|\varphi(x) - x_0| \le |\varphi(x) - \varphi(x_0)| + |\varphi(x_0) - x_0| \le$$

$$\le \alpha|x - x_0| + m \le \alpha\varepsilon + (1 - \alpha)\varepsilon = \varepsilon,$$

т. е. $|\varphi(x) - x_0| \le \varepsilon$, что и требовалось доказать.

Замечание 1. В условиях теоремы 2.2

$$|x_1 - x_0| = |\varphi(x_0) - x_0| = m.$$

Поэтому имеет место следующая оценка скорости сходимости:

$$|x^* - x_k| \le \frac{m}{1 - \alpha}\,\alpha^k.$$

2.2.2 Порядок итерационного метода

Предположим, что функция φ дифференцируема достаточное число раз в некоторой окрестности корня x^*.

Определение 2.1 *Пусть $m \in \mathbb{N} \setminus \{1\}$. Число m называется порядком итерационного метода, если в точке $x^* = \varphi(x^*)$ имеют место равенства*

$$\varphi'(x^*) = \varphi''(x^*) = \ldots = \varphi^{(m-1)}(x^*) = 0,$$

но $\varphi^{(m)}(x^*) \neq 0$.

Если порядок итерационного метода $m \geq 2$, то можно получить более точные оценки скорости сходимости итераций. Покажем это.

Запишем формулу Тейлора в окрестности x^*

$$\varphi(x) = \varphi(x^*) + \frac{\varphi'(x^*)}{1!}(x - x^*) + \frac{\varphi'(x^*)}{2!}(x - x^*)^2 + \ldots$$

$$+\frac{\varphi^{(m-1)}(x^*)}{(m-1)!}(x - x^*)^{m-1} + \frac{\varphi^{(m)}(\xi)}{m!}(x - x^*)^m.$$

Положим

$$x = x_{k-1}, \quad \varphi(x_{k-1}) = x_k, \quad \varphi(x^*) = x^*.$$

Так как $\varphi'(x^*) = \varphi''(x^*) = \ldots = \varphi^{(m-1)}(x^*) = 0$, будем иметь:

$$x_k - x^* = \varphi(x_{k-1}) - \varphi(x^*) = \frac{\varphi^{(m)}(\xi)}{m!}(x_{k-1} - x^*)^m.$$

Предположим, что существует такое число $M_m > 0$, что в некоторой окрестности корня справедливо неравенство

$$|\varphi^{(m)}(x)| \le M_m.$$

Тогда получаем оценки

$$|x^* - x_k| \le \frac{M_m}{m!} |x^* - x_{k-1}|^m \le$$

$$\le \left(\frac{M_m}{m!}\right)^{1+m+m^2+\ldots+m^{k-1}} |x^* - x_0|^{m^k} =$$

$$= \left(\frac{M_m}{m!}\right)^{\frac{m^k-1}{m-1}} |x^* - x_0|^{m^k}, \quad m \ge 2.$$

Рассмотрим подробнее важный частный случай, когда порядок итерационного метода $m = 2$. В этом случае имеем

$$|x^* - x_k| \le \left(\frac{M_2}{2}\right)^{2^k-1} |x^* - x_0|^{2^k} = \frac{2}{M_2} q^{2^k},$$

где

$$q = \frac{M_2}{2}|x^* - x_0|.$$

Ясно, что если нулевое приближение x_0 выбрано удачно, а именно, так, чтобы

$$q = \frac{M_2}{2}|x^* - x_0| < 1,$$

то итерационный метод сходится со скоростью

$$|x^* - x_k| \leq \frac{2}{M_2} q^{2^k}.$$

Уместно отметить, что успешное применение формальных методов итераций при решении нелинейных уравнений и систем нелинейных уравнений имеет важный неформальный этап, зависящий от интуиции и опыта вычислителя. А именно, необходим удачный выбор нулевого приближения.

2.3 Метод Ньютона

Рассмотрим уравнение $f(x) = 0$. Метод Ньютона, называемый также методом касательных, является итерационным методом. Алгоритм таков: выбираем нулевое приближение x_0, такое, что $f'(x_0) \neq 0$. Итерации определяются формулой

$$x_k = x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})}, \quad k \in \mathbb{N}.$$

Формально метод Ньютона можно получить следующим образом. Если $f'(x) \neq 0$, то уравнение $f(x) = 0$ равносильно уравнению

$$x = \varphi(x), \quad \text{где} \quad \varphi(x) := x - \frac{f(x)}{f'(x)}.$$

Ясно, что метод простых итераций $x_k = \varphi(x_{k-1})$ с выбранной выше функцией φ порождает метод Ньютона. Порядок итерационного метода Ньютона $m = 2$, так как

$$\varphi'(x^*) = \frac{f(x^*)f''(x^*)}{f'^2(x^*)} = 0.$$

Следовательно, мы можем применить оценку

$$|x^* - x_k| \leq \frac{2}{M_2} q^{2^k},$$

если нулевое приближение выбрано достаточно близким к искомому корню.

Если $f \in C^2(\mathbb{R})$ и имеет место оценка

$$\sup_{x \in \mathbb{R}} |\varphi'(x)| = \sup_{x \in \mathbb{R}} \left| \frac{f(x)f''(x)}{f'^2(x)} \right| = \alpha < 1,$$

то φ является сжимающим отображением, и поэтому справедлива оценка

$$|x^* - x_k| \leq \frac{\alpha^k}{1 - \alpha} |x_1 - x_0|.$$

Метод Ньютона допускает следующую геометрическую интерпретацию. Пусть x_0 — нулевое приближение, и пусть $f'(x_0) \neq 0$. Проведем касательную к графику функции f в точке $(x_0, f(x_0))$. Уравнение касательной имеет вид $y = f'(x_0)(x - x_0) + f(x_0)$. Эта касательная пересекает ось абсцисс в точке $x_1 = x_0 - f(x_0)/f'(x_0)$. Получили первое приближение. Далее, проведем касательную к графику функции f в точке $(x_1, f(x_1))$. Точка пересечения этой касательной с осью абсцисс представляет собой второе приближение

$$x_2 = x_1 - f(x_1)/f'(x_1).$$

Продолжаем процесс.

Употребительной модификацией метода Ньютона является метод хорд. Алгоритм: выбираем две точки a и x_0, удовлетворяющие условию $f(a)f(x_0) < 0$. Итерации строятся по формуле

$$x_k = x_{k-1} - \frac{f(x_{k-1})(x_{k-1} - a))}{f(x_{k-1}) - f(a)}, \quad k \in \mathbb{N}.$$

Более простой модификацией метода Ньютона является следующий алгоритм:

$$x_k = x_{k-1} - \frac{f(x_{k-1})}{f'(x_0)}, \quad k \in \mathbb{N}.$$

2.4 Собственные значения матрицы

Пусть $A = (a_{kj})$ — квадратная матрица порядка $n \geq 2$, элементы которой $a_{kj} \in \mathbb{R}$ или $a_{kj} \in \mathbb{C}$.

Число $\lambda \in \mathbb{C}$ называется собственным значением матрицы A, если существует такой ненулевой вектор x, что $Ax = \lambda x$. Этот ненулевой вектор x называют собственным вектором, соответствующим собственному значению λ.

Таким образом, если $\lambda \in \mathbb{C}$ — собственное значение матрицы A, то однородное уравнение $(A - \lambda E)x = \theta$ имеет ненулевое решение, а это возможно тогда и только тогда, когда $\det(A - \lambda E) = 0$. Следовательно, все собственные значения матрицы A определяются как корни уравнения $\det(A - \lambda E) = 0$. Легко видеть, что

$$P_n(A; \lambda) := \det(A - \lambda E)$$

— алгебраический полином от переменной λ и имеет вид

$$P_n(A; \lambda) = (-1)^n[\lambda^n - p_{n-1}\lambda^{n-1} - \ldots - p_1\lambda - p_0].$$

Полином $P_n(A; \lambda)$ называют характеристическим полиномом матрицы A. Согласно основной теореме алгебры, характеристическое уравнение $P_n(A; \lambda) = 0$ имеет корни $\lambda_k \in \mathbb{C}$ $(k = 1, 2, \ldots, n)$.

Таким образом, мы можем сказать, что спектр мат-

рицы

$$\sigma(A) = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$$

содержит не более, чем n чисел, так как некоторые корни могут оказаться кратными. Напомню, что число

$$\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$$

называется спектральным радиусом и существенно используется при изучении сходимости методов итераций.

В некоторых частных случаях все собственные значения матрицы легко определяются. Приведем два примера.

Пусть D — диагональная матрица порядка n. Тогда характеристический полином имеет вид

$$P_n(D; \lambda) = \prod_{k=1}^{n} (d_{kk} - \lambda).$$

Следовательно, $\lambda_k = d_{kk}$ для любого $k = 1, 2, \ldots, n$.

Пусть A — матрица порядка 2. Тогда

$$P_2(A; \lambda) = (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21}.$$

Поэтому собственные значения легко определяются как корни квадратного уравнения

$$(a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0.$$

В общем случае вычисление собственных значений матрицы представляет собой непростую задачу, для решения которой разработаны специальные методы. Приведем теоремы, лежащие в основе ряда методов вычисления собственных значений матрицы, например, методов Леверье, А.Н. Крылова, А.М. Данилевского и метода вращений.

Теорема 2.3 *Спектры подобных матриц совпадают.*

Доказательство. Пусть A и B — подобные матрицы. Тогда по определению подобия существует такая невырожденная матрица C, что $B = CAC^{-1}$. Пусть λ — одно из собственных значений матрицы B и x — соответствующий собственный вектор. Тогда $y := C^{-1}x \neq \theta$. Имеем соотношения

$$Bx = \lambda x \Leftrightarrow CAC^{-1}x = \lambda x \Leftrightarrow$$

$$\Leftrightarrow AC^{-1}x = \lambda C^{-1}x \Leftrightarrow Ay = \lambda y.$$

Следовательно,

$$\lambda \in \sigma(B) \Leftrightarrow \lambda \in \sigma(A),$$

что и требовалось.

Отметим еще раз, что разработан ряд эффективных методов нахождения всех собственных значений матриц высокого порядка. Многие из них основаны на приведении матрицы преобразованиями подобия к мат-

рице простого вида, для которой собственные значения находятся легко.

Рассмотрим круги, связанные с квадратной матрицей $A = (a_{kj})$ порядка $n \geq 2$, а именно, круги

$$D_k(A) = \left\{ z \in \mathbb{C} : |z - a_{kk}| \leq R_k(A) = \sum_{j=1, j\neq k}^{n} |a_{kj}| \right\},$$

где $k = 1, 2, \ldots, n$. Проблема локализации собственных значений заданной матрицы частично решается следующей теоремой.

Теорема 2.4 *(Первая теорема Гершгорина.) Любое собственное значение λ матрицы $A = (a_{kj})$ лежит в одном из кругов $D_k(A)$.*

Доказательство. Пусть $x = (x_1, x_2, \ldots, x_n) \neq \theta$ — собственный вектор, соответствующий собственному значению λ матрицы A. Пусть $x_k \neq 0$ — максимальная по модулю координата этого вектора. Приравнивая k-тые координаты в векторном равенстве $Ax = \lambda x$, получаем: $\sum_{j=1}^{n} a_{kj} x_j = \lambda x_k$. Отсюда следует, что

$$|\lambda - a_{kk}| = \left| \sum_{j=1, j\neq k}^{n} a_{kj} \frac{x_j}{x_k} \right| \leq \sum_{j=1, j\neq k}^{n} |a_{kj}| = R_k(A).$$

Таким образом, $\lambda \in D_k(A)$, что и требовалось доказать.

В заключение опишем **интерполяционный метод вычисления собственных значений матрицы**.

Пусть $A = (a_{kj})$ — квадратная матрица порядка $n \geq 2$. Собственные значения этой матрицы вычисляются в два этапа.

Этап 1 — нахождение характеристического полинома матрицы. Зададим узлы $x_1, x_2, \ldots, x_n, x_{n+1} \in \mathbb{R}$ и вычислим

$$y_1 = \det(A - x_1 E), \quad y_2 = \det(A - x_2 E), \ldots,$$

$$y_{n+1} = \det(A - x_{n+1} E).$$

Обозначим $f(\lambda) = P_n(A; \lambda) \equiv \det(A - \lambda E)$.

Зная $y_j = f(x_j)$, мы можем построить интерполяционный полином Лагранжа

$$L_{n+1}(f; x) = \sum_{j=1}^{n+1} y_j \frac{\omega_{n+1}(x)}{(x - x_j)\omega'_{n+1}(x_j)},$$

где

$$\omega_{n+1}(x) = \prod_{k=1}^{n+1} (x - x_k).$$

Поскольку степень полинома f меньше числа узлов, то $L_{n+1}(f; x) \equiv f(x)$. Следовательно, характеристический полином степени n определяется явно формулой

$$P_n(A; \lambda) = \sum_{j=1}^{n+1} \frac{y_j \, \omega_{n+1}(\lambda)}{(\lambda - x_j)\omega'_{n+1}(x_j)}.$$

Этап 2. Определяем все собственные значения

$$\lambda_1, \lambda_2, \ldots, \lambda_n,$$

решая уравнение $P_n(A;\lambda) = 0$, т. е. уравнение

$$\sum_{j=1}^{n+1} \frac{y_j\, \omega_{n+1}(\lambda)}{(\lambda - x_j)\omega'_{n+1}(x_j)} = 0.$$

2.5 Метод Ньютона для систем

Рассмотрим систему нелинейных уравнений следующего вида

$$\begin{cases} f_1(x_1, x_2, x_3, \ldots, x_n) = 0 \\ f_2(x_1, x_2, x_3, \ldots, x_n) = 0 \\ \qquad \ldots\ldots\ldots\ldots\ldots \\ f_n(x_1, x_2, x_3, \ldots, x_n) = 0 \end{cases}.$$

Предположим, что функции $f_j : \mathbb{R}^n \to \mathbb{R}$ являются непрерывно дифференцируемыми. Предположим также, что существует решение $x^* = (x_1^*, x_2^*, \ldots, x_n^*) \in \mathbb{R}^n$ этой системы уравнений. Наша цель — построить итерационный метод для нахождения этого решения.

Указанную систему формально можно записать в виде одного уравнения для отображения $F : \mathbb{R}^n \to \mathbb{R}^n$, определяя вектор $F(x)$ равенствами

$$F(x) = (f_1(x), f_2(x), \ldots, f_n(x)) \in \mathbb{R}^n.$$

Очевидно, рассматриваемая нелинейная система уравнений может быть записана как одно уравнение

$$F(x) = \theta,$$

где θ — нулевой вектор.

Рассмотрим матрицу Якоби для $F : \mathbb{R}^n \to \mathbb{R}^n$:

$$F'(x) := \begin{pmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 & \dots & \partial f_1/\partial x_n \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 & \dots & \partial f_2/\partial x_n \\ \dots & \dots & \dots & \dots \\ \partial f_n/\partial x_1 & \partial f_n/\partial x_2 & \dots & \partial f_n/\partial x_n \end{pmatrix}.$$

Предположим, что $\det F'(x) \neq 0$. Тогда существует обратная матрица $[F'(x)]^{-1}$.

Метод Ньютона для решения уравнения $F(x) = \theta$, равносильного системе уравнений, заключается в следующем.

Берем нулевое приближение $x^0 = (x_1^0, x_2^0, \dots, x_n^0)$. Итерации $x^k = (x_1^k, x_2^k, \dots, x_n^k)$ определяются по формуле

$$x^k = x^{k-1} - \left[F'(x^{k-1})\right]^{-1} F(x^{k-1}), \quad k \in \mathbb{N}.$$

Рассмотрим одну теорему о сходимости метода Ньютона для систем уравнений. Пусть a, a_1, a_2 — положительные постоянные.

Обозначим $B_a(x^*) = \{x \in \mathbb{R}^n : \|x - x^*\| \leq a\}$.

Пусть

$$c = a_1 a_2, \quad 0 < b < \min\{a, 1/c\},$$

$$B_b(x^*) = \{x \in \mathbb{R}^n : \|x - x^*\| \le b\}.$$

Теорема 2.5 *Пусть* $x^* = (x_1^*, x_2^*, \ldots, x_n^*) \in \mathbb{R}^n$ *— решение уравнения* $F(x) = \theta$. *Предположим, что для любых точек* $x \in B_a(x^*)$ *и* $y \in B_a(x^*)$ *выполнены условия:*

$$\left\| [F'(x)]^{-1} \right\| \le a_1,$$

$$\|F(x) - F(y) - F'(y)(x - y)\| \le a_2 \|x - y\|^2.$$

Пусть нулевое приближение $x^0 \in B_b(x^*)$. *Тогда*

$$x^k = x^{k-1} - \left[F'(x^{k-1})\right]^{-1} F(x^{k-1}) \in B_b(x^*), \quad \forall k \in \mathbb{N},$$

последовательность итераций сходится, т. е.

$$x^* = \lim_{k \to \infty} x^k.$$

Доказательство. Полагая $x = x^*$, $y = x^k$, можем написать

$$\left\| F(x^*) - F(x^k) - F'(x^k)(x^* - x^k) \right\| \le a_2 \|x^* - x^k\|^2.$$

Так как $F(x^*) = \theta$, то получаем неравенство

$$\left\| F(x^k) + F'(x^k)(x^* - x^k) \right\| \le a_2 \|x^* - x^k\|^2.$$

Далее, пользуясь простым неравенством

$$\left\|[F'(x^k)]^{-1}F(x^k) + (x^* - x^k)\right\| \leq$$

$$\leq \|[F'(x^k)]^{-1}\| \, \|F'(x^k)\{[F'(x^k)]^{-1}F(x^k) + (x^* - x^k)\}\| =$$

$$= \|[F'(x^k)]^{-1}\| \, \|F(x^k) + F'(x^k)(x^* - x^k)\|,$$

с учетом неравенства $\left\|[F'(x)]^{-1}\right\| \leq a_1$, имеем

$$\left\|[F'(x^k)]^{-1}F(x^k) + (x^* - x^k)\right\| \leq$$

$$\leq a_1 a_2 \, \|x^* - x^k\|^2 = c \, \|x^* - x^k\|^2.$$

Поскольку

$$\|x^* - x^{k+1}\| = \|x^* - x^k + \left[F'(x^k)\right]^{-1} \, F(x^k)\|,$$

то по индукции получаем

$$\|x^* - x^{k+1}\| \leq c \, \|x^* - x^k\|^2 \leq \ldots \leq c^{2^k - 1} \|x^* - x^0\|^{2^k}.$$

Индукцией также получаем, что условие $x^0 \in B_b(x^*)$ влечет $x^k \in B_b(x^*)$ для любой итерации, так как

$$\|x^* - x^{k+1}\| \leq c \, \|x^* - x^k\|^2 \leq c \, b^2 < b.$$

Далее, имеем: число $q = c \, b < 1$ в силу выбора b. Следовательно, предыдущая оценка запишется в виде нера-

венства

$$\|x^* - x^{k+1}\| \le \frac{1}{c}\, q^{2^k},$$

что влечет сходимость итераций к точному решению.

Этим и завершается доказательство теоремы.

Нахождение обратных матриц $\left[F'(x^{k-1})\right]^{-1}$ представляет собой трудоемкую задачу. Поэтому рассматривают упрощенную версию метода Ньютона, задавая итерации формулой

$$x^k = x^{k-1} - \left[F'(x^0)\right]^{-1} F(x^{k-1}), \quad \forall k \in \mathbb{N}.$$

Кроме метода Ньютона существуют и **другие итерационные методы**. Опишем один из них.

Рассмотрим систему нелинейных уравнений следующего вида

$$\begin{cases} x_1 = \varphi_1(x_1, x_2, x_3, \ldots, x_n) \\ x_2 = \varphi_2(x_1, x_2, x_3, \ldots, x_n) \\ \qquad \ldots\ldots\ldots\ldots\ldots \\ x_n = \varphi_n(x_1, x_2, x_3, \ldots, x_n) \end{cases}.$$

Предположим, что функции $\varphi_j : \mathbb{R}^n \to \mathbb{R}$ являются непрерывными. Обозначим решение этой системы уравнений как $x^* = (x_1^*, x_2^*, x_3^*, \ldots, x_n^*) \in \mathbb{R}^n$.

Указанную систему формально можно записать в виде одного уравнения для отображения $\Phi : \mathbb{R}^n \to \mathbb{R}^n$,

определяя вектор $\Phi(x)$ равенствами

$$\Phi(x) = (\varphi_1(x), \varphi_2(x), \ldots, \varphi_n(x)) \in \mathbb{R}^n.$$

Тогда наша нелинейная система уравнений может быть записана как одно уравнение

$$x = \Phi(x).$$

Рассмотрим метод прямых итераций. А именно, задаем нулевое приближение $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)$. Итерации $x^k = (x_1^k, x_2^k, \ldots, x_n^k)$ определяются по формуле

$$x^k = \Phi(x^{k-1}), \quad k \in \mathbb{N}.$$

Как следствие теоремы Банаха о сжимающих отображениях получаем следующее утверждение.

Теорема 2.6 *Пусть отображение* $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ *является сжимающим, т. е.*

$$\|\Phi(x) - \Phi(y)\| \leq \alpha \|x - y\|, \quad \forall x, y \in \mathbb{R}^n,$$

где постоянная $\alpha \in (0, 1)$. *Тогда существует единственное решение* x^* *уравнения* $x = \Phi(x)$, *при любом выборе нулевого приближения* $x^0 \in \mathbb{R}^n$ *итерационный метод* $x^k = \Phi(x^{k-1})$ *сходится, а именно,*

$$x^* = \lim_{k \to \infty} x^k,$$

причем

$$\|x^* - x^k\| \le \frac{\alpha^k}{1-\alpha}\|x^1 - x^0\|.$$

Для систем нелинейных уравнений строятся итерационные методы Зейделя по аналогии с методами итераций Зейделя для СЛАУ. Как обычно, задаем нулевое приближение. В обобщении первого варианта метода Зейделя последующие итерации определяются формулами: $k \in \mathbb{N}$ и

$$\begin{cases} x_1^k = \varphi_1(x_1^{k-1}, x_2^{k-1}, x_3^{k-1}, \dots, x_n^{k-1}) \\ x_2^k = \varphi_2(x_1^k, x_2^{k-1}, x_3^{k-1}, \dots, x_n^{k-1}) \\ \qquad \dots\dots\dots\dots\dots\dots \\ x_n^k = \varphi_n(x_1^k, x_2^k, x_3^k, \dots, x_{n-1}^k, x_n^{k-1}) \end{cases}.$$

2.6 Задачи и упражнения

1. Найдите миллионный член последовательности чисел Фибоначчи

$$0,\ 1,\ 1,\ 2,\ 3,\ 5,\ 8,\ 13,\ 21,\ 34,\ 55, \dots .$$

2. Докажите вторую теорему Гершгорина: *если объединение кругов $D_k(A)$ состоит из нескольких связных частей, то каждая связная часть содержит столько собственных значений, сколько кругов ее составляют.*

3. Найдите характеристический полином следующей матрицы Фробениуса

$$F_n = \begin{pmatrix} p_{n-1} & p_{n-2} & \dots & p_2 & p_1 & p_0 \\ 1 & 0 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 & 0 \\ 0 & 0 & \dots & 0 & 1 & 0 \end{pmatrix}.$$

4. Пользуясь методом Ньютона, постройте алгоритм для вычисления числа $\sqrt{7}$, рассматриваемая это число как корень уравнения $x^2 = 7$.

5. Найдите решения системы уравнений

$$\begin{cases} x_1^2 + 4x_2^2 = 1 \\ x_1^4 + x_2^4 = 0,5 \end{cases}$$

с пятью верными знаками. (Задача взята из книги [12]).

6. Рассмотрите пример верхнетреугольной матрицы порядка n, все собственные значения которой являются простыми (т. е. спектр состоит из n различных чисел). Найдите все собственные векторы такой матрицы.

Глава 3

Интерполяция функций

Математические модели многих проблем естествознания используют функции и операции, содержащие предельный переход. При расчетах мы можем использовать лишь конечное число значений функции, поэтому нужно построить приближенные дискретные аналоги используемых операций.

Прошедший проверку временем и ставший стандартным способ перехода к дискретным аналогам основных операций анализа состоит в следующем. Функцию приближают либо полиномами, либо тригонометрическими суммами, либо сплайнами, используя при этом лишь конечное число значений функции. И основные операции проводят над этими приближениями.

Для заданной непрерывной функции можно определить полином, значения которого совпадают со значениями выбранной функции в нескольких точках. Удовлетворяющий такому условию полином наименьшей степени называется интерполяционным. Замена функ-

ции ее интерполяционным полиномом используется при приближенных вычислениях интегралов.

Наиболее употребительной является интерполяционная формула, открытая Лагранжем в 1795 году, хотя сама интерполяция использовалась задолго до него. По-видимому, описание первой интерполяционной формулы принадлежит Ньютону (приведено в его труде "Метод разностей", опубликованном в 1736 году).

Более общие интерполяционные формулы были найдены в 19 веке Коши, Эрмитом и другими математиками. Наиболее трудные вопросы по оценкам погрешности при полиномиальной интерполяции были решены лишь в 20 веке А. Лебегом, С.Н. Бернштейном, Д. Джексоном, С.Б. Стечкиным и рядом других математиков. При этом существенно использовались фундаментальные результаты Вейерштрасса и П.Л. Чебышева.

Отметим также, что интерполяция представляет собой лишь один из разделов обширной теории приближения функций, развитие которой продолжается и в настоящее время.

3.1 Полиномы Лагранжа

Пусть на отрезке $[a, b]$ заданы точки $x_1, x_2, \ldots, x_n$. Предполагаем, что $x_k \neq x_j$ при $k \neq j$. Для непрерывной функции f будем рассматривать следующую задачу.

Задача. Найти алгебраический полином $L_n(f; x)$

наименьшей степени и такой, что

$$L_n(f; x_j) = f(x_j), \quad j = 1, 2, \ldots, n.$$

Функцию $L_n(f; x)$ называют интерполяционным полиномом Лагранжа, а точки x_j $(j = 1, \ldots, n)$ — узлами интерполяционного полинома Лагранжа или узлами интерполяции.

Рассмотрим пример. Пусть функция $g : [a, b] \to \mathbb{R}$ обладает свойством

$$g(x_n) = 1, \quad g(x_j) = 0 \quad (j = 1, \ldots, n-1).$$

Понятно, что полином $L_n(g; x)$ должен иметь степень $\geq n-1$, так как он должен делиться без остатка на полином $p_n(x) := \prod_{j=1}^{n-1}(x - x_j)$. Поэтому полагаем $L_n(g; x) = c_n\, p_n(x)$, где константа $c_n = 1/p_n(x_n)$.

Теорема 3.1 *Для любой функции $f : [a, b] \to \mathbb{R}$ и узлов $x_1, x_2, \ldots, x_n$ интерполяционный полином Лагранжа $L_n(f; x)$ степени не выше $n-1$ существует и определяется единственным образом.*

Доказательство. Если $n = 1$, то $L_1(f; x) \equiv f(x_1)$.

Пусть $n \geq 2$. Надеясь на то, что существует полином Лагранжа степени $\leq n-1$, искомый полином можем записать в виде

$$L_n(f; x) = \sum_{k=1}^{n} a_k x^{k-1} = a_1 + a_2 x + \ldots a_n x^{n-1}.$$

Коэффициенты этого полинома должны определяться из условий:

$$L_n(f; x_j) = f(x_j) \qquad (j = 1, 2, \ldots, n) \Longleftrightarrow$$

$$\Longleftrightarrow \begin{cases} a_1 + a_2 x_1 + \ldots + a_n x_1^{n-1} = f(x_1) \\ a_1 + a_2 x_2 + \ldots + a_n x_2^{n-1} = f(x_2) \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots \\ a_1 + a_2 x_n + \ldots + a_n x_n^{n-1} = f(x_n) \end{cases} .$$

Для нахождения неизвестных $a_1, a_2, \ldots a_n$ получаем систему уравнений, определитель которой

$$\Delta_n = \begin{vmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \ldots & x_2^{n-1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 & \ldots & x_n^{n-1} \end{vmatrix} = \prod_{i>j} (x_i - x_j)$$

является определителем Вандермонда и отличен от нуля. Следовательно, система имеет единственное решение, которое можно определить по правилу Крамера

$$a_k = \frac{\Delta_{n,k}}{\Delta_n},$$

где $\Delta_{n,k}$ — определитель матрицы, полученной из матрицы Вандермонда заменой k-го столбца на столбец свободных членов $f(x_1), f(x_2), \ldots, f(x_n)$. Поэтому ин-

терполяционный полином Лагранжа запишется в виде:

$$L_n(f;x) = \sum_{k=1}^{n} \frac{\Delta_{n,k}}{\Delta_n} \cdot x^{k-1}.$$

По построению $L_n(f;x)$ — полином степени $\leq n-1$ и $f(x) \approx L_n(f;x)$ вблизи узлов для $f \in C[a,b]$.

Приведем второе доказательство единственности, показывающее, в частности, что $L_n(f;x) \equiv f(x)$ для любого полинома f степени не выше $n-1$.

Предположим, что для $f \in C[a,b]$ имеется еще один интерполяционный полином $Q(x)$ степени $\leq n-1$:

$$Q(x) = \sum_{k=1}^{n} b_k x^{k-1}, \quad Q(x_j) = f(x_j), \quad j = 1, 2, \ldots, n.$$

Рассмотрим разность

$$p(x) = L_n(f;x) - Q(x) = \sum_{k=1}^{n} (a_k - b_k) x^{k-1}.$$

Очевидно, $p(x)$ — полином степени $\leq n-1$. Для любого $j = 1, \ldots, n$ имеем:

$$p(x_j) = L_n(f;x_j) - Q(x_j) = f(x_j) - f(x_j) = 0.$$

Таким образом, получаем, что полином $p(x)$ степени не выше $n-1$ имеет n различных корней $x_1, x_2, \ldots, x_n$.

Согласно основной теореме алгебры корней должно

быть не больше $n-1$ за исключением случая, когда $p(x) \equiv 0$. Поэтому имеем

$$p(x) \equiv 0 \Rightarrow L_n(f;x) \equiv Q(x).$$

Полученное противоречие и доказывает единственность. В частности, справедливо

Следствие 3.1.1 *Если $Q(x)$ — алгебраический полином степени $\leq n-1$, то*

$$L_n(Q;x) \equiv Q(x).$$

Представление Лагранжа для интерполяционного полинома

Приведем представление для полинома Лагранжа $L_n(f;x)$ в виде явной формулы, включающей узлы интерполяции $x_1, x_2, \ldots, x_n$ и значения функции f в этих точках. Одновременно мы получим второе доказательство существования интерполяционного полинома. Потребуются следующие полиномы степени $n-1$: если $n=1$, то $l_1(x) \equiv 1$, а при $n \geq 2$ и $k = 1, \ldots, n$

$$l_k(x) = \prod_{j=1, j\neq k}^{n} \frac{x - x_j}{x_k - x_j} =$$

$$= \frac{(x-x_1)\ldots(x-x_{k-1})(x-x_{k+1})\ldots(x-x_n)}{(x_k-x_1)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_n)},$$

которые называются **фундаментальными полино-**

мами Лагранжа. В узлах интерполяции получаем

$$l_k(x_j) = \delta_{kj} = \begin{cases} 1, & \text{если } k = j; \\ 0, & \text{если } k \neq j. \end{cases}$$

Рассмотрим полином

$$Q(x) = \sum_{k=1}^{n} f(x_k) l_k(x).$$

Имеем: степень $Q \leq n-1$, кроме того,

$$Q(x_j) = \sum_{k=1}^{n} f(x_k) l_k(x_j) = \sum_{k=1}^{n} f(x_k)\delta_{kj} = f(x_j)$$

для любого $j = 1, \ldots, n$.

В силу единственности интерполяционного полинома получаем $Q(x) \equiv L_n(f; x)$, следовательно,

$$L_n(f; x) = \sum_{k=1}^{n} f(x_k) l_k(x).$$

Эта формула и есть основное представление интерполяционного полинома Лагранжа.

Часто удобнее пользоваться другой записью основного представления. Рассмотрим произведение

$$\omega_n(x) = (x - x_1)(x - x_2)\ldots(x - x_n) = \prod_{j=1}^{n}(x - x_j).$$

Легко видеть, что

$$l_k(x) = \frac{A_{nk}(x)}{B_{nk}},$$

где

$$A_{nk}(x) = \frac{\omega_n(x)}{x - x_k}, \qquad B_{nk} = \prod_{j=1, j\neq k}^{n} (x_k - x_j) = \omega_n'(x_k),$$

так как

$$\omega_n'(x) = (x-x_2)\dots(x-x_n)+(x-x_1)(x-x_3)\dots(x-x_n)+$$

$$\dots + (x - x_1)(x - x_2)\dots(x - x_{n-1}).$$

Следовательно, получаем видоизмененное, равносильное основному, 3-е представление

$$L_n(f; x) = \sum_{k=1}^{n} f(x_k) \frac{\omega_n(x)}{(x - x_k)\omega_n'(x_k)}.$$

Таким образом, справедливо утверждение.

Теорема 3.2 *Для любой функции $f \in C[a, b]$ и заданных узлов $x_1, x_2, \dots, x_n$ справедливо следующее представление Лагранжа*

$$L_n(f; x) = \sum_{k=1}^{n} f(x_k) l_k(x) = \sum_{k=1}^{n} f(x_k) \frac{\omega_n(x)}{(x - x_k)\omega_n'(x_k)}.$$

Заметим, что при доказательстве этой теоремы и при доказательстве теорем существования и единственности интерполяционного полинома непрерывность функции f никак не используется. Однако непрерывность или гладкость функции f необходимы, как только мы начинаем оценивать погрешность интерполяции.

3.2 О погрешности интерполяции

Будем рассматривать снова узлы $x_1, x_2, \ldots, x_n \in [a, b]$.

Нас будет интересовать остаточный член интерполяции

$$r_n(x) = f(x) - L_n(f; x),$$

называемый также погрешностью интерполяции.

Теорема 3.3 *Пусть $f \in C^{(n-1)}[a, b]$ и во всех точках интервала (a, b) существует производная $f^{(n)}(x)$. Тогда для любого $x \in [a, b]$ существует точка $\xi \in (a, b)$ такая, что*

$$r_n(x) = \frac{f^{(n)}(\xi)}{n!}\,\omega_n(x),$$

где $\omega_n(x) = \prod_{k=1}^{n}(x - x_k)$.

Доказательство. Ясно, что $x = x_j$ — тривиальный случай. Так как в этом случае $r_n(x_j) = 0 = \omega_n(x_j)$, т. е. доказываемое равенство выполняется автоматически.

Фиксируем $x \neq x_j,\ j = 1, \ldots, n, \quad x \in [a, b]$, и рассмотрим вспомогательную функцию

$$\varphi(t) = f(t) - L_n(f; t) - C\omega_n(t) \quad a \leq t \leq b.$$

Постоянную C выбираем из условия $\varphi(x) = 0$, пользуясь тем, что $\omega_n(x) \neq 0$, т. е. полагаем

$$C = \frac{f(x) - L_n(f; x)}{\omega_n(x)} = \frac{r_n(x)}{\omega_n(x)}.$$

Заметим теперь, что уравнение $\varphi(t) = 0$ имеет на отрезке $[a, b]$ не менее $(n + 1)$ корня, так как

$$\left\{ \begin{array}{l} f(x_j) - L_n(f; x_j) - C\omega_n(x_j) = 0,\ j = 1, 2, \ldots, n \\ \varphi(x) = 0 \end{array} \right\}.$$

По теореме Ролля между двумя нулями φ имеется корень уравнения $\varphi'(t) = 0$, следовательно, $\varphi'(t) = 0$ имеет не менее n корней. Если $n > 1$, продолжим этот процесс. Получаем: $\varphi''(t) = 0$ имеет $\geq (n - 1)$ корень. Если $(n-1) > 1$, то продолжаем процесс. По индукции находим, что уравнение $\varphi^{(n)}(t) = 0$ имеет хотя бы один корень $\xi \in (a, b)$. Но тогда

$$\varphi^{(n)}(\xi) = f^{(n)}(\xi) - Cn! = 0,$$

так как $L_n^{(n)}(f;x) \equiv 0$ и $\omega_n^{(n)}(x) \equiv n!$. Поэтому

$$\frac{r_n(x)}{\omega_n(x)} = C = \frac{f^{(n)}(\xi)}{n!},$$

что и требовалось показать.

Следствие 3.3.1 *Если* $|f^{(n)}(x)| \le M_n = \text{const}$ *для всех* $x \in [a,b]$, *то*

$$|r_n(x)| = |f(x) - L_n(f;x)| \le \frac{M_n}{n!}(b-a)^n$$

для любого $x \in [a,b]$.

Доказательство. Действительно, имеем

$$|r_n(x)| \le \frac{M_n}{n!}\,|\omega_n(x)|,$$

кроме того,

$$|\omega_n(x)| = \left|\prod_{k=1}^{n}(x - x_k)\right| \le (b-a)^n.$$

Следствие 3.3.2 *Пусть функция* f *имеет производные любого порядка. Обозначим*

$$\max_{x\in[a,b]} |f^{(n)}(x)| = M_n < \infty.$$

Если

$$\frac{\sqrt[n]{M_n}}{n} \to 0 \quad \textit{при} \quad n \to \infty,$$

то $r_n(x) \to 0$ *равномерно на* $[a,b]$ *при* $n \to \infty$.

Доказательство. Методом математической индукции с использованием определения числа e легко получаем неравенство

$$n! > \left(\frac{n}{e}\right)^n.$$

Поэтому для любого $x \in [a,b]$

$$|r_n(x)| \le \frac{M_n}{n!}(b-a)^n \le \left(\frac{\sqrt[n]{M_n}}{n} e(b-a)\right)^n \to 0$$

при $n \to \infty$. Здесь мы учли, что из условия

$$\varepsilon_n = \frac{\sqrt[n]{M_n}}{n} \to 0$$

следует сходимость $\varepsilon_n\, e\,(b-a) \to 0$ при $n \to \infty$, а значит и $\{\varepsilon_n\, e\,(b-a)\}^n \to 0$. Таким образом,

$$\max_{x\in[a,b]} |r_n(x)| := \|r_n\|_{C[a,b]} \le \{\varepsilon_n\, e\,(b-a)\}^n \to 0$$

при $n \to \infty$.

Пример. Пусть $f_0(x) = e^{-x}, \quad x \in [0,1]$. Рассмотрим вопрос о числе узлов n, гарантирующих следующее неравенство для погрешности: $|r_n(x)| < \varepsilon = 0,01$ для всех $x \in [0,1]$.

Простые выкладки

$$M_n = \max_{x\in[0,1]} \left| \frac{d^n(e^{-x})}{dx^n} \right| = \max_{x\in[0,1]} e^{-x} = 1$$

и применение предыдущей теоремы

$$|r_n(x)| \leq \frac{M_n}{n!}(1-0)^n = \frac{1}{n!}$$

показывают, что неравенство

$$|r_n(x)| < 0,01$$

будет выполняться наверняка при $n \geq 5$.

3.3 Полиномы Чебышева

Рассмотрим функции, определяемые формулами:

$$T_0(t) = 1,\ T_1(t) = t,\ T_n(t) = \cos(n \arccos t),\ n \geq 2.$$

Как показывают результаты П.Л. Чебышева, эти функции оказываются полиномами, наименее отклоняющимися от нуля. Они называются полиномами Чебышева первого рода, и для них справедлива следующая рекуррентная формула:

$$T_{n+1}(t) = 2tT_n(t) - T_{n-1}(t).$$

Лемма 3.1 *Пусть* $n \in \mathbb{N}$. *Функция* $T_n : [-1,1] \to \mathbb{R}$

является полиномом степени n со старшим коэффициентом 2^{n-1} и с нулями

$$t_k^0 = \cos\left(\frac{2k-1}{2n}\pi\right), \quad k = 1, 2, \ldots, n,$$

т. е.

$$T_n(t) = 2^{n-1}\prod_{k=1}^{n}\left(t - \cos\frac{2k-1}{2n}\pi\right),$$

причем максимум и минимум $T_n(t)$ достигаются в точках $t_k^* = \cos\frac{k\pi}{n}$, $T_n(t_k^*) = (-1)^k$ $(k = 0, \ldots, n)$.

Доказательство. Обозначим $\arccos t = \alpha$. Имеем

$$T_1(t) = \cos(\arccos t) = t,$$

и

$$T_2(t) = \cos 2\alpha = 2\cos^2\alpha - 1 = 2t^2 - 1.$$

Получим теперь рекуррентную формулу.

Пусть $T_1, T_2, \ldots, T_n$ известны, найдем

$$T_{n+1}(t) = \cos[(n+1)\alpha] = \cos n\alpha\cos\alpha - \sin n\alpha\sin\alpha =$$

$$= T_n(t)\cdot t - \frac{\cos(n-1)\alpha - \cos(n+1)\alpha}{2} =$$

$$= tT_n(t) - \frac{1}{2}T_{n-1}(t) + \frac{1}{2}T_{n+1}(t).$$

Таким образом,

$$T_{n+1}(t) = 2tT_n(t) - T_{n-1}(t).$$

Зная $T_1 = t$, $T_2 = 2t^2 - 1$, мы можем найти T_3, затем T_4, T_5 и т.д. Рекуррентная формула показывает, что $T_n(t)$ — полином степени n со старшим членом $2^{n-1}t^n$.

Найдем корни уравнения $T_n(t) = 0$, т. е. уравнения $\cos(n \arccos t) = 0$. Имеем

$$n \arccos t = \frac{2k-1}{2}\pi,\ t_k^0 = \cos\left(\frac{2k-1}{2n}\pi\right),$$

где $k = 1, \ldots, n$. Зная старший член и все корни полинома, получаем, что

$$T_n(t) = 2^{n-1}\prod_{k=1}^{n}\left(t - \cos\frac{2k-1}{2n}\pi\right).$$

Очевидно, максимальное и минимальное значения

$$T_n(t) = \cos(n \arccos t)$$

равны ± 1. Точки экстремума легко определяются из соотношений $n \arccos t_k^* = \pi k$, $T_n(t_k^*) = \cos(\pi k) = (-1)^k$, где $k = 0, 1, \ldots, n$, т. е. экстремумы достигаются в точках t_k^* $(k = 0, \ldots, n)$. Лемма доказана полностью.

Теорема 3.4 *(Теорема Чебышева) Для любого натурального числа n имеет место формула*

$$\inf_{t_1,t_2,\ldots,t_n\in[-1,1]}\left\|\prod_{k=1}^{n}(t-t_k)\right\|_{C[-1,1]} = \frac{1}{2^{n-1}},$$

причем инфимум достигается на узлах Чебышева

$$t_k^0 = \cos\left(\frac{2k-1}{2n}\pi\right) \qquad k = 1, \ldots, n.$$

Доказательство. Полином

$$\frac{T_n(t)}{2^{n-1}} = \prod_{k=1}^{n}\left(t - \cos\frac{2k-1}{2n}\pi\right)$$

удовлетворяет условиям теоремы и норма

$$\frac{\|T_n(t)\|_{C[-1,1]}}{2^{n-1}} = \frac{1}{2^{n-1}}.$$

Нужно доказать, что это искомый инфимум.

Предположим обратное: существует полином

$$Q_n(t) = \prod_{k=1}^{n}(t - t_k) = t^n + b_{n-1}t^{n-1} + \ldots + b_0$$

такой, что

$$\|Q_n(t)\|_{C[-1,1]} < \frac{1}{2^{n-1}}.$$

Рассмотрим разность

$$q(t) = \frac{T_n(t)}{2^{n-1}} - Q_n(t) = \prod_{k=1}^{n}(t - t_k^0) - \prod_{k=1}^{n}(t - t_k) =$$

$$= a_{n-1}t^{n-1} + \ldots + a_0 - (b_{n-1}t^{n-1} + \ldots + b_0).$$

Видно, что $q(t)$ — полином степени $\leq (n-1)$. С дру-

гой стороны, в точках экстремума полинома Чебышева получаем

$$q(t_k^*) = \frac{(-1)^k}{2^{n-1}} - Q(t_k^*), \quad |Q_n(t_k^*)| < \frac{1}{2^{n-1}}.$$

Следовательно,

$$q(t_0^*) = \frac{1}{2^{n-1}} - Q(t_0^*) > 0,$$

$$q(t_1^*) = \frac{-1}{2^{n-1}} - Q(t_1^*) < 0,$$

$$q(t_2^*) > 0, \ldots.$$

Продолжая процесс, получаем, что полином $q(t)$ меняет знак не менее, чем n раз.

Отсюда следует, что $q(t)$ имеет не менее n корней, и эти корни $\tau_1, \tau_2, \ldots, \tau_n$ лежат между точками t_k^* из интервала $(-1, 1)$. Поскольку степень $q(t)$ не выше, чем $(n-1)$, то $q(t) \equiv 0$. Пришли к противоречию. Этим и завершается доказательство.

Общая задача оптимального выбора узлов Дано некоторое семейство

$$F \subset C[a, b].$$

Нужно найти величину

$$V_n(F) = \inf_{x_1, x_2, \ldots, x_n \in [a,b]} \sup_{f \in F} \max_{a \le x \le b} |r_n(x)|.$$

Иными словами, необходимо подобрать узлы $x_1, \ldots, x_n$ на отрезке $[a, b]$ так, чтобы полученная сетка узлов была бы оптимальной для выбранного семейства F.

Рассмотрим эту задачу для следующего семейства функций

$$W^n M = \{f \in C[a,b] \,:\, \exists f^{(m)}(x)\,(x \in [a,b],$$

$$m = 1, \ldots, n), |f^{(n)}(x)| \le M\},$$

где M — некоторая положительная постоянная.

Оказывается, что можно найти $V_n(W^n M)$ с применением теоремы Чебышева.

Теорема 3.5 *Имеет место формула*

$$V_n(W^n M) = \frac{M}{n!}\frac{(b-a)^n}{2^{2n-1}},$$

причем оптимальными являются узлы Чебышева

$$x_k = \frac{a+b}{2} + \frac{b-a}{2}\cos\left(\frac{2k-1}{2n}\pi\right), \quad k = 1, 2, \ldots, n.$$

Доказательство. Напомним: условие $|f^{(n)}(x)| \le M$ влечет оценку

$$|r_n(x)| \le \frac{M}{n!}|\omega_n(x)|$$

для любого $x \in [a, b]$, где $\omega_n(x) = (x - x_1)\ldots(x - x_n)$.

Рассмотрим сначала специальный частный случай

$$f_0(x) := \frac{M}{n!}\omega_n(x).$$

Поскольку

$$f_0^{(n)}(x) \equiv \frac{M}{n!}n! = M,$$

то получаем, что $f_0 \in W^n M$. Очевидно, интерполяционный полином Лагранжа по узлам $x_1, x_2, \ldots, x_n$ для функции $f_0(x)$ тождественно равен нулю. Поэтому

$$|r_{0n}(x)| := |f_0(x) - L_n(f_0; x)| \equiv |f_0(x)| = \frac{M}{n!} \cdot |\omega_n(x)|$$

для любого $x \in [a, b]$.

Таким образом,

$$|r_n(x)| \le \frac{M}{n!}|\omega_n(x)| = |r_{0n}(x)|.$$

Отсюда следует

$$\sup_{f \in W^n M} \max_{x \in [a,b]} |r_n(x)| = \frac{M}{n!} \max_{x \in [a,b]} |\omega_n(x)|,$$

и нам необходимо минимизировать эту величину за счет выбора узлов $x_1, x_2, \ldots, x_n \in [a, b]$.

Сделаем замену переменной

$$x = \frac{a+b}{2} + \frac{b-a}{2}\,t, \quad t \in [-1, 1], \quad x \in [a, b].$$

Тогда

$$x - x_k = \frac{b-a}{2}(t - t_k),$$

где

$$x_k = \frac{a+b}{2} + \frac{b-a}{2}t_k,$$

$$\omega_n(x) = \frac{(b-a)^n}{2^n}\prod_{k=1}^{n}(t - t_k).$$

Следовательно, искомая величина определяется формулой

$$V_n(W^n M) = \frac{M}{n!}\frac{(b-a)^n}{2^n}\inf_{t_1,t_2,\ldots,t_n\in[-1,1]}\prod_{k=1}^{n}|t - t_k|.$$

По теореме Чебышева для любого n искомый инфимум равен $\frac{1}{2^{n-1}}$ и достигается для узлов

$$t_k = \cos\frac{2k-1}{2n}\pi.$$

Поэтому

$$V_n(W_n M) = \frac{M}{n!}\frac{(b-a)^n}{2^{2n-1}}.$$

Обратная замена переменных $t_k \to x_k$ дает

$$x_k = \frac{a+b}{2} + \frac{b-a}{2}\cos\frac{2k-1}{2n}\pi, \quad k = 1, \ldots, n.$$

Этим и завершается доказательство теоремы.

3.4 Лебеговы оценки погрешности

Оценки Лебега для остаточного члена зависят от двух констант: от наилучшего равномерного приближения $E_n(f)$ и константы Лебега Λ_n.

Величина $E_n(f)$, называемая наилучшим равномерным приближением $f \in C[a,b]$ алгебраическими полиномами степени $\leq n-1$, определяется следующим образом

$$E_n(f) = \inf_{a_0,\ldots,a_{n-1}\in\mathbb{R}} \left\| f(x) - \sum_{k=0}^{n-1} a_k x^k \right\|_{C[a,b]} .$$

Для любой функции $f \in C[a,b]$ введенная величина $E_n(f) \to 0$ при $n \to \infty$. Этот факт является простым следствием одной теоремы Вейерштрасса, которую мы рассмотрим в следующем пункте.

3.4.1 Теорема Вейерштрасса

Теорема 3.6 *Всякая непрерывная функция на конечном отрезке допускает равномерную аппроксимацию с любой наперед заданной точностью алгебраическими полиномами, т.е. для любой функции $f \in C[a,b]$ и для любого $\varepsilon > 0$ существует такой алгебраический полином $p_n(x)$, что для всех $x \in [a,b]$*

$$|f(x) - p_n(x)| < \varepsilon.$$

Доказательство. Кроме доказательства Вейерштрасса, имеется несколько новых доказательств этой теоремы. Приведем доказательство Лебега, рассуждения которого легко запоминаются.

Лебег выводит утверждение теоремы Вейерштрасса из трех простых фактов.

Шаг 1. Согласно теореме Кантора, непрерывная на отрезке функция является равномерно непрерывной, поэтому она может быть равномерно аппроксимирована ломаными, т. е. непрерывными кусочно-линейными функциями.

Шаг 2. Всякая ломаная из m звеньев представима в виде

$$y = a_0 + \sum_{j=1}^{m} a_j |x - x_{j-1}|,$$

где $x_0 = a < x_1 < \ldots < x_{m-1} < x_m = b$ — абсциссы вершин ломаной. Это утверждение устанавливается элементарными рассуждениями, так как указанное представление задает непрерывную кусочно-линейную функцию при любом выборе $a_0, a_1, \ldots, a_m$, а эти коэффициенты для заданной ломаной однозначно определяются.

Действительно, если $y = k_j x + b_j$ — уравнение ломаной на j-том отрезке $[x_{j-1}, x_j]$, то коэффициенты

$$a_1, a_2, \ldots, a_m$$

явно определяются из системы линейных уравнений

$$a_1 - \sum_{j=2}^{m} a_j = k_1,$$

$$\sum_{j=1}^{s} a_j - \sum_{j=s+1}^{m} a_j = k_s, \quad s = 2, \ldots, m-1,$$

$$\sum_{j=1}^{m} a_j = k_m.$$

Затем можно определить коэффициент a_0 равенством

$$a_0 = y(a) - \sum_{j=1}^{m} a_j |a - x_{j-1}|.$$

В силу первых двух шагов достаточно показать, что функция $|x - x_j|$ равномерно аппроксимируется алгебраическими полиномами на отрезке

$$[x_j - (b-a), x_j + (b-a)].$$

Заменой переменных $x - x_j = (b-a)t$ вопрос сводится к следующему шагу.

Шаг 3. Функция $|t|$ равномерно аппроксимируется алгебраическими полиномами на отрезке $[-1, 1]$. Действительно, имеем

$$|t| = \sqrt{1 - (1 - t^2)} = (1 - \alpha)^{1/2}, \quad \alpha = 1 - t^2 \in [0, 1].$$

Ряд Тейлора

$$(1-\alpha)^{1/2} = 1 - \frac{1}{2}\alpha - \sum_{j=2}^{\infty} \frac{(2j-3)!!}{(2j)!!}\alpha^j$$

сходится равномерно на $[-1,1]$ по признаку Вейерштрасса, так как для всех $\alpha \in [-1,1]$

$$\frac{(2j-3)!!}{(2j)!!}|\alpha|^j \le \frac{(2j-3)!!}{(2j)!!} \le \frac{1}{j\sqrt{j}}.$$

Последнее неравенство легко доказывается методом математической индукции, а ряд

$$\sum_{j=1}^{\infty} \frac{1}{j\sqrt{j}},$$

как известно, является сходящимся. Из равномерной сходимости ряда Тейлора для функции $(1-\alpha)^{1/2}$ следует, что разность

$$|t| - \left(1 - \frac{1}{2}(1-t^2) - \sum_{j=2}^{N} \frac{(2j-3)!!}{(2j)!!}(1-t^2)^j\right)$$

равномерно стремится к 0 при $N \to \infty$. Таким образом, функция $|t|$ равномерно аппроксимируется на отрезке $[-1,1]$ алгебраическими полиномами четной степени.

Этим и завершается доказательство.

3.4.2 Оценки погрешности, поведение остаточного члена при $n \to \infty$

Получим сначала формулу для погрешности r_n без предположения дифференцируемости функции f.

Теорема 3.7 *Для любой $f \in C[a,b]$ и n узлов интерполяции $x_1, \ldots, x_n$ ($n \geq 2$)*

$$r_n(x) = f(x) - L_n(f;x) = \sum_{k=1}^{n} [f(x) - f(x_k)]\, l_k(x),$$

где $L_n(f;x)$ — интерполяционный полином Лагранжа, а

$$l_k(x) = \frac{\omega_n(x)}{(x - x_k)\omega_n'(x_k)}$$

— фундаментальный полином Лагранжа, $k = 1, \ldots, n$.

Доказательство. Рассмотрим некоторый полином

$$Q(x) = \sum_{k=1}^{n} b_k x^{k-1}$$

степени $\leq n-1$. Поскольку он совпадает со своим интерполяционным полиномом Лагранжа $L_n(Q;x)$, полученным интерполяцией по n точкам, будем иметь

$$Q(x) \equiv L_n(Q;x) = \sum_{k=1}^{n} Q(x_k) l_k(x).$$

Применяя эту формулу к частному случаю $Q(x) \equiv 1$, получаем следующее тождество для фундаментальных полиномов Лагранжа:

$$1 = \sum_{k=1}^{n} l_k(x).$$

Умножаем обе части тождества на $f(x)$ и заносим этот множитель под знак суммы. Будем иметь

$$f(x) = \sum_{k=1}^{n} f(x) l_k(x).$$

С другой стороны,

$$L_n(f; x) = \sum_{k=1}^{n} f(x_k) l_k(x).$$

Вычитая второе равенство из первого, получаем искомую формулу. Таким образом, теорема доказана.

Определение 3.1 *Функция*

$$\Lambda_n(x) = \sum_{k=1}^{n} |l_k(x)|$$

называется функцией Лебега для узлов

$$x_1, x_2, \ldots, x_n \in [a, b],$$

а число $\Lambda_n = \max_{x \in [a,b]} \Lambda_n(x)$ *— константой Лебега.*

Имеем простые неравенства

$$1 \le \Lambda_n(x) \le \Lambda_n, \qquad \forall x \in [a,b].$$

Легко видеть, что первое неравенство является простым следствием тождества $\sum l_k(x) = 1$, а второе неравенство — следствие определения Λ_n.

Теорема 3.8 *Пусть $f \in C[a,b]$. Тогда справедливы следующие оценки Лебега:*

$$|r_n(x)| \le E_n(f)[1 + \Lambda_n(x)] \le 2\Lambda_n(x)E_n(f)$$

и

$$\|r_n(x)\|_{C[a,b]} \le 2\Lambda_n \cdot E_n(f).$$

Следовательно,

а) если $\overline{x} \in [a,b]$, $\Lambda_n(\overline{x})E_n(f) \to 0$ при $n \to \infty$, то

$$r_n(\overline{x}) = f(\overline{x}) - L_n(f;\overline{x}) \to 0 \quad \textit{при} \quad n \to \infty.$$

б) если $\Lambda_n E_n(f) \to 0$ при $n \to \infty$, то равномерно на отрезке $[a,b]$

$$r_n(x) \to 0 \quad \textit{при} \quad n \to \infty.$$

Доказательство. Запишем равенство

$$r_n(x) = f(x) - L_n(f;x) = f(x) - Q(x) + Q(x) - L_n(f;x),$$

где $Q(x) = \sum_{k=1}^{n} a_k x^{k-1}$ — произвольный полином степени $\leq n-1$. Следовательно,

$$L_n(Q; x) \equiv Q(x).$$

Поэтому

$$|r_n(x)| \leq |f(x) - Q(x)| + \left| \sum_{k=1}^{n} [f(x_k) - Q(x_k)] l_k(x) \right| \leq$$

$$\leq \|f(x) - Q(x)\|_{C[a,b]} + \|f(x) - Q(x)\|_{C[a,b]} \sum_{k=1}^{n} |l_k(x)| =$$

$$= \|f(x) - Q(x)\|_{C[a,b]} (1 + \Lambda_n(x)).$$

В силу произвольности $Q(x)$ отсюда следует

а)

$$|r_n(\overline{x})| \leq E_n(f)[1 + \Lambda_n(\overline{x})] \leq$$

$$\leq 2\Lambda_n(\overline{x}) E_n(f) \to 0 \qquad \text{при} \qquad n \to \infty,$$

и, аналогично,

б)

$$\|r_n(x)\|_{C[a,b]} \leq 2\Lambda_n \cdot E_n(f) \to 0 \qquad \text{при} \qquad n \to \infty.$$

Таким образом, теорема Лебега доказана.

Замечания. Понятно, что сходимость или расходимость интерполяционного процесса зависит как от выбора последовательности точек интерполяции, т. е.

последовательности сеток

$$\Omega_n = \{x_{n1}, x_{n2}, \ldots, x_{nn}\},$$

так и от гладкости интерполируемой функции.

Существуют примеры очень простых по виду функций, для которых интерполяционный процесс по сеткам с равноотстоящими узлами расходится.

Пример 1 (пример С.Н. Бернштейна): последовательность интерполяционных полиномов $L_n(f;x)$, построенных по равноотстоящим узлам $x_{nk} = -1 + 2k/n$, $k = 0, 1, \ldots, n$, для функции $f(x) = |x|$ на отрезке $[-1, 1]$, не сходится к функции $|x|$ при $n \to \infty$ ни в одной точке этого отрезка, кроме трех точек -1, 0, 1.

Пример 2 (пример Рунге (Runge C.)): последовательность интерполяционных полиномов $L_n(f;x)$, построенных по равноотстоящим узлам $x_{nk} = -1 + 2k/n$, $k = 0, 1, \ldots, n$, для гладкой функции

$$f(x) = \frac{1}{25x^2 + 1}, \quad -1 \le x \le 1,$$

не сходится равномерно к $f(x)$ на отрезке $[-1, 1]$.

Как показывают эти примеры, наиболее простые и естественные сетки с равноотстоящими узлами оказываются неэффективными, т. е. приводят к расходящемуся интерполяционному процессу даже для некоторых гладких функций. Но не следует думать, что сетки

с равноотстоящими узлами занимают исключительное положение при приближении непрерывных функций. Пример Бернштейна является лишь усилением частного случая следующей теоремы существования.

Теорема Фабера: *для любой последовательности сеток*

$$\Omega_n = \{x_{n1}, x_{n2}, \ldots, x_{nn}\} \subset [a, b]$$

существует непрерывная на этом отрезке функция f такая, что последовательность интерполяционных полиномов Лагранжа $L_n(f; x)$ не сходится равномерно к этой функции на отрезке $[a, b]$ при $n \to \infty$.

Известно также, что для каждой непрерывной функции существует своя оптимальная последовательность сеток. А именно, имеет место следующий положительный результат.

Теорема Марцинкевича: *для любой функции f, непрерывной на отрезке $[a, b]$, существует такая последовательность сеток $\Omega_n = \Omega_n(f) \subset [a, b]$, для которой соответствующий интерполяционный процесс сходится равномерно на отрезке $[a, b]$ при $n \to \infty$.*

Для гладких функций аналог теоремы Фабера неверен, и нет необходимости пользоваться теоремой Марцинкевича, так как существуют универсальные для всего класса гладких функций оптимальные последовательности сеток. К оптимальным относится, например, последовательность сеток с узлами Чебышева. Имеет

место следующая **теорема В.И. Крылова**, доказанная им в 1956 г. (см. [10]).

Теорема 3.9 *Для любой абсолютно непрерывной функции* $f : [a, b] \to \mathbb{R}$ *её полиномы Лагранжа* $L_n(f; x)$, *построенные на последовательности сеток с узлами Чебышева*

$$x_{nk} = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\frac{2k-1}{2n}\pi\right), \quad k = 1, 2, \ldots, n,$$

равномерно сходятся к $f(x)$ *на* $[a, b]$ *при* $n \to \infty$.

3.4.3 Константы Лебега

Для фиксированной сетки

$$\Omega_n = \{x_{n1}, x_{n2}, \ldots, x_{nn}\} \subset [a, b]$$

процесс интерполирования можно рассматривать как применение линейного оператора P_n, действующего из банахова пространства $C[a, b]$ в себя и определенного равенством $(P_n f)(x) = L_n(f; x)$. Очевидно, P_n является линейным оператором, так как

$$L_n(f+g; x) = L_n(f; x) + L_n(g; x),\ L_n(cf; x) = cL_n(f; x)\,,$$

$(c = const)$, и, кроме того, P_n является оператором проектирования, т. е.

$$P_n^2 f := P_n(P_n f) = P_n f.$$

Теорема 3.10 *Норма линейного оператора P_n, определяемого равенством*

$$(P_n f)(x) = L_n(f; x), \quad x \in [a, b],$$

равна константе Лебега, т. е.

$$||P_n|| = \Lambda_n := \max_{x\in[a,b]} \Lambda_n(x) := \max_{x\in[a,b]} \sum_{k=1}^{n} |l_k(x)|.$$

Доказательство. Из представления Лагранжа

$$(P_n f)(x) = L_n(f; x) = \sum_{k=1}^{n} f(x_k) l_k(x)$$

вытекает, что

$$|(P_n f)(x)| \le \max_{x_{nk}\in[a;b]} |f(x_k)| \sum_{k=1}^{n} |l_k(x)| \le$$

$$\le \Lambda_n(x) ||f||_{C[a,b]}.$$

Следовательно,

$$||P_n|| \le \max_{x\in[a,b]} \Lambda_n(x) = \Lambda_n.$$

С другой стороны, возьмем одну из точек x_0, где достигается максимум функции Лебега $\Lambda_n(x)$. Очевидно, существует непрерывная на отрезке $[a, b]$ функция f_0 такая, что

$$f_0(x_k) = sign\, l_k(x_0)$$

для всех $k = 1, 2, \ldots, n$ и $||f_0||_{C[a,b]} = 1$. Тогда

$$(P_n f_0)(x_0) = L_n(f_0; x_0) = \sum_{k=1}^{n} |l_k(x_0)| =$$

$$= \Lambda_n(x_0) = \Lambda_n ||f_0||_{C[a,b]},$$

что влечет неравенство

$$||P_n|| \geq \Lambda_n,$$

завершающее доказательство теоремы.

Заменой переменных $t = cx + d$ $(c \neq 0)$ легко убедиться в том, что константа Лебега не зависит от длины отрезка интерполирования, а зависит только от относительного расположения узлов. А именно, константа Λ_n, вычисленная для отрезка $[a, b]$ и сетки узлов

$$\Omega_n = \{x_{n1}, x_{n2}, \ldots, x_{nn}\} \subset [a, b],$$

совпадает с константой Λ_n, определенной для отрезка $[ca + d, cb + d]$ и сетки

$$\Omega'_n = \{cx_{n1} + d, cx_{n2} + d, \ldots, cx_{nn} + d\} \subset [ca + d, cb + d].$$

Понятно, что зависимость константы Лебега от числа и взаимного расположения узлов имеет большое значение, так как через эту константу оценивается погрешность интерполяции.

Для больших значений числа узлов n определение точных значений константы Λ_n представляет собой трудную задачу. Поэтому в следующей теореме мы ограничимся двусторонними оценками Λ_n.

Теорема 3.11 *Для равноотстоящих узлов при интерполяции алгебраическими полиномами константа Лебега удовлетворяет неравенствам*

$$\frac{2^{n-3}}{n^2} < \Lambda_n < 2^{n-1}.$$

Доказательство. Без ограничения общности рассмотрим отрезок $[a,b] = [1,n]$, т. е. $a = 1, b = n$, с узлами $x_1 = 1, x_2 = 2, \ldots, x_n = n$. Тогда

$$\Lambda_n = \max_{1\le x\le n} \sum_{k=1}^{n} \prod_{j\neq k} \left|\frac{x-j}{k-j}\right| =$$

$$= \max_{1\le x\le n} \sum_{k=1}^{n} \frac{1}{(n-k)!(k-1)!} \prod_{j\neq k} |x-j|.$$

Для любого $x \in [1,n]$ имеем оценку

$$\prod_{j\neq k} |x-j| < (n-1)!,$$

поэтому верхняя оценка легко следует из тождества для биномиальных коэффициентов:

$$\Lambda_n < \sum_{k=1}^{n} \frac{(n-1)!}{(n-k)!(k-1)!} = 2^{n-1}.$$

Нижняя оценка для константы Лебега получается следующим образом. Имеем простые неравенства

$$\Lambda_n \geq \Lambda_n(3/2) = \sum_{k=1}^{n} \frac{1}{(n-k)!(k-1)!} \prod_{j \neq k} |3/2 - j|$$

и

$$\prod_{j \neq k} |3/2 - j| = \frac{\prod_{j=1}^{n} |3/2 - j|}{|k - 3/2|} \geq \frac{(n-2)!}{4n} > \frac{(n-1)!}{4n^2}.$$

Применение тождества для биномиальных коэффициентов завершает доказательство.

Для узлов Чебышева

$$t_{nk} = \cos \frac{\pi(2k+1)}{2n}, \quad k = 0, 1, \ldots, n-1,$$

можем записать

$$\Lambda_n = \max_{-1 \leq t \leq 1} \sum_{k=0}^{n-1} \frac{|T_n(t)|}{|t - t_{nk}||T_n'(t_{nk})|} =$$

$$= \max_{-1 \leq t \leq 1} \sum_{k=0}^{n-1} \frac{|\cos(n \arccos t)| \sqrt{1 - t_{nk}^2}}{n\,|t - t_{nk}|}.$$

Ясно, что в этом случае получение оценок точных значений константы Λ_n представляет собой более трудную задачу по сравнению со случаем равноотстоящих узлов. Тем не менее, эти трудности были преодолены, и имеет место следующая теорема С.Н. Бернштейна.

Теорема 3.12 *Для узлов Чебышева константа Лебега имеет логарифмический рост, в частности, можно записать*

$$\Lambda_n = O(\ln n), \quad n \to \infty.$$

В середине 20-го столетия усилиями ряда математиков было доказано, что логарифмический рост для константы Лебега является минимальным из всех возможных: *доказано существование постоянной* $c > 0$ *такой, что* $\Lambda_n \geq c \ln n$ *для любой сетки из* n *узлов.*

Таким образом, узлы Чебышева относятся к семейству оптимальных узлов, для которых константа Лебега имеет логарифмический рост. Этот факт имеет не только теоретическое, но и практическое значение, так как узлы Чебышева для отрезка $[a, b]$ определены явно простыми формулами:

$$x_{nk} = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\frac{2k-1}{2n}\pi\right), \quad k = 1, 2, \ldots, n.$$

3.5 Формулы Ньютона для интерполяционного полинома

Для $f \in C[a,b]$ и точек $x_1, x_2, \ldots, x_n \in [a,b]$ интерполяционный полином $L_n(f;x)$ по этим n узлам записывается по формуле

$$L_n(f;x) = \sum_{k=1}^{n} f(x_k) l_k(x),$$

где

$$l_k(x) = \frac{\omega_n(x)}{(x-x_k)\omega_n'(x_k)},$$

$\omega_n(x) = (x-x_1)(x-x_2)\ldots(x-x_n)$. Если добавить новый узел x_{n+1} и строить интерполяционный полином по узлам $x_1, x_1, \ldots, x_n, x_{n+1} \in [a,b]$, то получаем следующее представление Лагранжа

$$L_{n+1}(f;x) = \sum_{k=1}^{n+1} f(x_k) \frac{\omega_{n+1}(x)}{(x-x_k)\omega_{n+1}'(x_k)},$$

где

$$\omega_{n+1}(x) = (x-x_1)(x-x_2)\ldots(x-x_{n+1}).$$

Ясно, что при добавлении нового узла приходится пересчитывать все слагаемые в представлении Лагранжа.

3.5.1 Полиномы Ньютона

Формула для интерполяционного полинома, которая не требует пересчета всех слагаемых при добавлении нового узла, была известна еще Ньютону. Такая формула называется формулой Ньютона для интерполяционного полинома или интерполяционным полиномом Ньютона. Она получается следующим образом.

Для $f \in C[a,b]$ и узлов $x_1, x_2, \ldots, x_n \in [a,b]$ интерполяционный полином Ньютона имеет вид

$$L_n(f;x) = A_0 + A_1(x-x_1) + A_2(x-x_1)(x-x_2) +$$

$$+ \ldots + A_{n-1}(x-x_1)\ldots(x-x_{n-1}),$$

т. е.

$$L_n(f;x) = \sum_{j=1}^{n} A_{j-1}\omega_{j-1}(x),$$

где $\omega_0(x) = 1$, $\omega_k(x) = (x-x_1)\ldots(x-x_k)$ при $k \geq 1$.

Для определения неизвестных коэффициентов A_0, $A_1, A_2, \ldots, A_{n-1}$ имеем n уравнений

$$L_n(f;x_1) = f(x_1),\ \ldots,\ L_n(f;x_n) = f(x_n).$$

Легко показать, что A_k $(k = 0, 1, \ldots, n-1)$ однозначно определяется этими уравнениями, зависит лишь от значений функции в точках $x_1, x_2, \ldots, x_k$, следовательно, не меняется при добавлении нового узла x_{n+1}.

Для первых двух коэффициентов вычисления весь-

ма просты: из первых двух уравнений имеем

$$f(x_1) = A_0, \quad f(x_2) = A_0 + A_1(x_2 - x_1),$$

отсюда

$$f(x_2) - f(x_1) = A_1(x_2 - x_1) \quad \Rightarrow \quad A_1 = \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

Из третьего уравнения

$$f(x_3) = A_0 + A_1(x_3 - x_1) + A_2(x_3 - x_1)(x_3 - x_2)$$

простыми выкладками определяется A_2:

$$f(x_3) - f(x_1) - \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x_3 - x_1) =$$

$$= A_2(x_3 - x_1)(x_3 - x_2),$$

$$f(x_3) - f(x_1) - \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x_3 - x_2) - f(x_2) + f(x_1) =$$

$$= A_2(x_3 - x_1)(x_3 - x_2),$$

$$A_2(x_3 - x_1) = \frac{f(x_3) - f(x_2)}{x_3 - x_2} - \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

По индукции легко получаем, что A_k однозначно определяется и зависит лишь от значений функции в точках $x_1, x_2, \ldots, x_k$.

3.5.2 Применение разделенных разностей

Выражения

$$\frac{f(x_2)-f(x_1)}{x_2-x_1},\quad \frac{f(x_3)-f(x_2)}{x_3-x_2}$$

называются разделенными разностями 1-го порядка и обозначаются через $f(x_1;x_2)$ и $f(x_2;x_3)$, соответственно. Разделенные разности высоких порядков определяются индуктивно. А именно, разделенная разность 2-го порядка $f(x_1;x_2;x_3)$ задается формулой

$$f(x_1;x_2;x_3)=\frac{f(x_2;x_3)-f(x_1;x_2)}{x_3-x_1},$$

а разделенная разность $f(x_1;x_2;\ldots;x_k)$ порядка $k-1$ определяется так:

$$f(x_1;x_2;\ldots;x_k)=\frac{f(x_2;x_3;\ldots;x_k)-f(x_1;x_2;\ldots;x_{k-1})}{x_k-x_1}.$$

Для полноты картины значения f в узлах, т. е. числа $f(x_1),f(x_2),\ldots,f(x_n)$ называют разделенными разностями порядка 0.

Теорема 3.13 *Справедлива следующая формула*

$$f(x_1;x_2;\ldots;x_k)=\sum_{j=1}^{k}\frac{f(x_j)}{\omega_k'(x_j)}= \tag{3.1}$$

$$= \sum_{j=1}^{k} f(x_j) \prod_{i=1, i\neq j}^{k} \frac{1}{x_j - x_i},$$

где

$$\omega_k(x) = \prod_{j=1}^{k} (x - x_j).$$

Доказательство. Утверждение тривиально при $k = 1$. Для случая $k = 2$

$$f(x_1; x_2) = \frac{f(x_2) - f(x_1)}{x_2 - x_1} =$$

$$= \frac{f(x_1)}{x_1 - x_2} + \frac{f(x_2)}{x_2 - x_1} = \sum_{j=1}^{2} \frac{f(x_j)}{\omega_2'(x_j)}.$$

Применим метод математической индукции. Предположим, что формула верна до порядка $k - 1$ и выведем ее для разделенных разностей порядка k. Можем записать: $f(x_1; x_2; \ldots; x_{k+1}) =$

$$= \frac{f(x_2; x_3; \ldots; x_{k+1}) - f(x_1; x_2; \ldots; x_k)}{x_{k+1} - x_1},$$

тогда по предположению индукции $f(x_1; x_2; \ldots; x_{k+1}) =$

$$= \frac{1}{x_{k+1} - x_1} \sum_{j=2}^{k+1} f(x_j) \prod_{i=2, i\neq j}^{k+1} \frac{1}{x_j - x_i} -$$

$$-\frac{1}{x_{k+1}-x_1}\sum_{j=1}^{k} f(x_j)\prod_{i=1,i\neq j}^{k}\frac{1}{x_j-x_i}.$$

Значения $f(x_1)$ и $f(x_{k+1})$ входят лишь в одну из сумм и коэффициенты при них вычисляются просто. Коэффициент при $f(x_1)$ равен

$$-\frac{1}{x_{k+1}-x_1}\prod_{i=2}^{k}\frac{1}{x_1-x_i}=\prod_{i=2}^{k+1}\frac{1}{x_1-x_i}=\frac{1}{\omega'_{k+1}(x_1)},$$

и коэффициент при $f(x_{k+1})$ дается формулой

$$\frac{1}{x_{k+1}-x_1}\prod_{i=2}^{k}\frac{1}{x_{k+1}-x_i}=\prod_{i=1}^{k}\frac{1}{x_{k+1}-x_i}=\frac{1}{\omega'_{k+1}(x_{k+1})}.$$

Коэффициент при $f(x_m)$ для случая $2\leq m\leq k$ также нетрудно вычисляется и равен

$$\frac{1}{x_{k+1}-x_1}\left(\prod_{i=2,i\neq m}^{k+1}\frac{1}{x_m-x_i}-\prod_{i=1,i\neq m}^{k}\frac{1}{x_m-x_i}\right)=$$

$$=\frac{1}{x_{k+1}-x_1}\left(\frac{1}{x_m-x_{k+1}}-\frac{1}{x_m-x_1}\right)\prod_{i=2,i\neq m}^{k}\frac{1}{x_m-x_i}=$$

$$=\frac{1}{(x_m-x_{k+1})(x_m-x_1)}\prod_{i=2,i\neq m}^{k}\frac{1}{x_m-x_i}=$$

$$= \prod_{i=1, i\neq m}^{k+1} \frac{1}{x_m - x_i} = \frac{1}{\omega'_{k+1}(x_m)}.$$

Таким образом, коэффициенты при $f(x_m)$ имеют требуемую форму для всех допустимых значений m, этим и завершается доказательство теоремы.

В качестве следствий теоремы получаем следующие **свойства разделенных разностей**.

Свойство 1. Разделенная разность является линейным функционалом от f, т. е. для любых постоянных C_1 и C_2

$$(C_1\, f + C_2\, g)(x_1, x_2, \ldots, x_n) =$$

$$= C_1\, f(x_1; x_2; \ldots; x_n) + C_2\, g(x_1; x_2; \ldots; x_n).$$

Свойство 2. Разделенная разность является симметричной функцией своих аргументов, т. е. инвариантна относительно перестановки аргументов, например, $f(x_1; x_2) = f(x_2; x_1)$.

Как уже отмечалось выше, в формуле

$$L_n(f; x) = A_0 + A_1(x - x_1) + A_2(x - x_1)(x - x_2) + \ldots$$

$$+A_{n-1}(x - x_1)(x - x_2)\ldots(x - x_{n-1}) = \sum_{k=1}^{n} A_{k-1}\, \omega_{k-1}(x)$$

для функции $f \in C[a,b]$ и узлов $x_1, \ldots, x_n$ первые три коэффициента имеют вид

$$A_0 = f(x_1),\ A_1 = f(x_1; x_2),\ A_2 = f(x_1; x_2; x_3).$$

Покажем, что для любого k, $1 \le k \le n$,

$$A_{k-1} = f(x_1; x_2; \ldots; x_k).$$

Теорема 3.14 *Интерполяционный полином для функции $f \in C[a,b]$ по узлам $x_1, x_2, \ldots, x_n$ можно представить формулой Ньютона*

$$L_n(f; x) = \sum_{k=1}^{n} f(x_1; x_2; \ldots; x_k)\ \omega_{k-1}(x).$$

Доказательство. Через $L_m(f; x)$ обозначим интерполяционный полином Лагранжа, построенный по узлам $x_1, x_2, \ldots, x_m$, $1 \le m \le n$. Согласно представлению Лагранжа, имеем: $L_1(f; x) = f(x_1)$,

$$L_{m-1}(f; x) = \sum_{j=1}^{m-1} f(x_j) \prod_{i=1, i\neq j}^{m-1} \frac{x - x_i}{x_j - x_i} \quad (m \ge 2).$$

В силу простого равенства

$$L_n(f; x) = f(x_1) + \sum_{m=2}^{n} [L_m(f; x) - L_{m-1}(f; x)],$$

достаточно показать, что разность

$$p(x) = L_m(f; x) - L_{m-1}(f; x)$$

равна $f(x_1; x_2; \ldots; x_m)\ \omega_{m-1}(x)$. С одной стороны, эта разность является полиномом степени не выше $m-1$ и обращается в нуль в точках $x_1, x_2, \ldots, x_{m-1}$. Поэтому $p(x) = A_{m-1}\omega_{m-1}(x)$, где A_{m-1} — некоторая постоянная.

С другой стороны, $p(x_m) =$

$$= L_m(f; x_m) - L_{m-1}(f; x_m) = f(x_m) - L_{m-1}(f; x_m)] =$$

$$= f(x_m) - \sum_{j=1}^{m-1} f(x_j) \prod_{i=1, i\neq j}^{m-1} \frac{x_m - x_i}{x_j - x_i} =$$

$$= f(x_m) + \sum_{j=1}^{m-1} f(x_j) \frac{x_m - x_j}{x_j - x_m} \prod_{i=1, i\neq j}^{m-1} \frac{x_m - x_i}{x_j - x_i} =$$

$$= A \left[f(x_m) \prod_{i=1}^{m-1} \frac{1}{x_m - x_i} + \sum_{j=1}^{m-1} f(x_j) \prod_{i=1, i\neq j}^{m} \frac{1}{x_j - x_i} \right],$$

где $A = \prod_{i=1}^{m-1}(x_m - x_i)$. Согласно предыдущей теореме выражение в квадратных скобках равно разделенной разности $f(x_1; x_2; \ldots; x_m)$. Таким образом,

$$A_{m-1}\omega_{m-1}(x_m) = p(x_m) = \omega_{m-1}(x_m) f(x_1; x_2; \ldots; x_m).$$

Следовательно, $A_{m-1} = f(x_1; x_2; \ldots; x_m)$, что и требовалось доказать.

Из доказанной теоремы непосредственно следует, что при добавлении к узлам $x_1, x_2, \ldots, x_n$ нового узла x_{n+1} будем иметь

$$L_{n+1}(f; x) = L_n(f; x) + f(x_1; x_2; \ldots; x_n; x_{n+1})\, \omega_n(x),$$

т. е. приходится вычислять только одно дополнительное слагаемое.

Из последней формулы можно получить полезное тождество. Учитывая равенство $L_{n+1}(f; x_{n+1}) = f(x_{n+1})$ и пользуясь формальной заменой $x_{n+1} = x$, будем иметь

$$f(x) \equiv L_n(f; x) + f(x_1; x_2; \ldots; x_n; x)\, \omega_n(x).$$

Приведем еще одно следствие. Речь идет о свойствах разделенных разностей высоких порядков для полиномов.

Свойство 3. Если Q — полином степени n, то разделенные разности порядка $(n + 1)$ и выше для этого полинома Q равны 0.

Действительно, пусть $m \geq n + 1$, тогда имеем: $Q(x) \equiv L_m(Q; x)$ и

$$\sum_{k=1}^{m} f(x_1; x_2; \ldots; x_k)\ \omega_{k-1}(x) = Q(x).$$

Из условия совпадения степеней полиномов в этом равенстве получаем, что

$$f(x_1; x_2; \ldots; x_m) = 0$$

при $m \geq n + 2$.

3.5.3 Применение конечных разностей

В этом пункте мы запишем формулу Ньютона для интерполяционного полинома с заменой разделенных разностей на конечные разности.

Рассмотрим узлы $x_1, \ldots, x_n \in [a, b]$. Для функции $f \in C[a, b]$, обозначим

$$y_k = f(x_k),\ k = 1, 2, \ldots, n.$$

По определению, конечная разность 1-го порядка равна

$$\Delta^1 y_k = y_{k+1} - y_k = \Delta y_k$$

(как и при определении дифференциалов функций принято отождествлять Δ^1 и Δ).

Конечная разность 2-го порядка $\Delta^2 y_k = \Delta^1(\Delta^1 y_k) = = \Delta(y_{k+1} - y_k) = y_{k+2} - y_{k+1} - (y_{k+1} - y_k)$ выражается формулой

$$\Delta^2 y_k = y_{k+2} - 2y_{k+1} + y_k,$$

и конечная разность 3-го порядка − формулой

$$\Delta^3 y_k = \Delta(\Delta^2 y_k) = y_{k+3} - 2y_{k+2} + y_{k+1} - y_{k+2} + 2y_{k+1} - y_k =$$

$$= y_{k+3} - 3y_{k+2} + 3y_{k+1} - y_k.$$

Индуктивно определяем конечную разность порядка m. Получаем

$$\Delta^m y_k = \Delta(\Delta^{m-1} y_k) = \sum_{j=0}^{m} (-1)^j C_m^j y_{k+m-j},$$

где C_m^j — биномиальные коэффициенты.

На отрезке $[a, b]$ возьмем равноотстоящие узлы

$$a \le x_1, x_2 = x_1 + h, \ \ldots, \ x_n = x_1 + (n-1)h \le b,$$

с шагом $h > 0$ и поменяем разделенные разности на конечные разности в формуле

$$L_n(f; x) = \sum_{k=1}^{n} f(x_1; x_2; \ldots; x_k)\, \omega_{k-1}(x).$$

Имеем: $f(x_1) = y_1$,

$$f(x_1, x_2) = \frac{f(x_2) - f(x_1)}{x_2 - x_1} = \frac{y_2 - y_1}{h} = \frac{\Delta^1 y_1}{h},$$

$$f(x_1; x_2; x_3) = \frac{f(x_2, x_3) - f(x_1, x_2)}{x_3 - x_1} = \frac{\frac{\Delta y_2}{h} - \frac{\Delta y_1}{h}}{2h} = \frac{\Delta^2 y_1}{2h^2}$$

и по индукции

$$f(x_1; x_2; \ldots; x_k) = \frac{\Delta^{k-1} y_1}{(k-1)! h^{k-1}}.$$

С учетом естественного соглашения $\Delta^0 y_1 = y_1$, получаем формулу

$$L_n(f; x) = \sum_{k=1}^{n} \frac{\Delta^{k-1} y_1 \ \omega_{k-1}(x)}{(k-1)! h^{k-1}} = \sum_{k=0}^{n-1} \frac{\Delta^k y_1}{k! h^k} \ \omega_k(x).$$

Эта формула приобретает универсальный вид

$$L_n(f; x) = y_1 + \sum_{k=1}^{n-1} \frac{\Delta^k y_1}{k!} t(t-1) \ldots (t-k+1)$$

при следующей замене переменных

$$x = x_1 + ht, \ 0 \le t \le n-1.$$

Выведенная формула называется формулой Ньютона для интерполирования вперед. Это название имеет естественное объяснение.

Напомним прежде всего, что при выводе основного представления Лагранжа (или Ньютона) для интерполяционного полинома не было требований на взаимное расположение узлов, кроме условия: $x_k \neq x_j$ при $k \neq j$. Далее, если интерполяционные полиномы используются для приближенного определения значений функции, заданной таблично, то наибольший вклад в

значение $L_n(f;x)$ в фиксированной точке x вносят узлы, ближайшие к точке x. Поэтому полученная выше формула с узлами

$$x_k = x_1 + kh, \ (h > 0, \ \ k = 0, 1, \ldots, n-1)$$

считается полезной для интерполирования в начале таблицы.

Если интерполируется значение функции в конце таблицы, то применяют другие формулы. Для шага $h > 0$ берутся узлы

$$x_1, x_1 - h, x_1 - 2h, x_1 - 3h, \ldots$$

и снова можно пользоваться формулой Ньютона через разделенные разности.

При интерполировании в середине таблицы в качестве первых узлов выгодно брать узлы, ближайшие к точке x и удовлетворяющие, например, неравенствам $x < x_{2k}$, $x > x_{2k-1}$.

Подобные идеи являются классическими и плодотворно реализованы рядом математиков. Читатель найдет замечательные формулы Гаусса, Бесселя, Стирлинга и других классиков для интерполяционного полинома в ряде книг, например, в учебнике И.С. Березина и Н.П. Жидкова (см. [4], том 1, стр.125-142).

3.6 Кратная интерполяция

При построении интерполяционного полинома Лагранжа мы требовали совпадения значений функции со значениями полинома в узлах. Понятно, что если дополнительно требовать совпадения значений и некоторых производных в узлах, то интерполяционный полином будет лучше приближать функцию.

Наиболее простым является следующий частный случай. Рассмотрим узлы $x_1, x_2, \ldots, x_n \in [a, b]$ и непрерывно дифференцируемую функцию f на этом отрезке.

Интерполяционный полином $H_n(f; x)$ ищется как полином наименьшей степени, удовлетворяющий следующим условиям

$$H_n(f; x_1) = f(x_1), \ \ldots, H_n(f; x_n) = f(x_n);$$

$$H'_n(f; x_1) = f'(x_1), \ldots, H'_n(f; x_n) = f'(x_n).$$

Для определения $H_n(f; x)$ получаем $2n$ уравнений. Естественно искать его как полином степени $2n - 1$:

$$H_n(f; x) = a_0 + a_1 x + \ldots + a_{2n-1} x^{2n-1}.$$

Оказывается, что такой полином, называемый интерполяционным полиномом Эрмита-Фейера, существует и находится единственным образом. Мы получим этот факт из более общего утверждения.

3.6.1 Полиномы Эрмита

Пусть f — непрерывная, достаточное число раз дифференцируемая функция на отрезке $[a, b]$. Заданы узлы интерполяции

$$x_1, x_2, \ldots, x_n \in [a, b]$$

и их кратности (натуральные числа)

$$a_1, a_2, \ldots, a_n.$$

Требуется найти полином наименьшей степени $H(x)$, называемый интерполяционным полиномом Эрмита, по следующим условиям:

в каждой узловой точке x_j $(j = 1, 2, \ldots, n)$ должны выполняться равенства

$$H^{(k)}(x_j) = f^{(k)}(x_j) \tag{3.2}$$

для всех

$$k = 0, 1, \ldots, a_j - 1.$$

Очевидно, для записи системы уравнений (3.2) достаточно, чтобы функция f была бы непрерывно дифференцируемой $(a_j - 1)$-раз в некоторой окрестности точки x_j, где $j = 1, 2, \ldots, n$.

Число уравнений для определения $H(x)$ равно

$$m = a_1 + a_2 + \ldots + a_n,$$

поэтому естественно искать полином $H(x)$ как полином степени $\leq m-1$.

Теорема 3.15 *Интерполяционный полином Эрмита степени $\leq m-1$ существует и определяется единственным образом, причем его можно представить в следующей форме*

$$H(x) = P_1(x) + (x-x_1)^{a_1}P_2(x) + \\ +(x-x_1)^{a_1}(x-x_2)^{a_2}P_3(x) + \ldots \\ +(x-x_1)^{a_1}(x-x_2)^{a_2}\ldots(x-x_{n-1})^{a_{n-1}}P_n(x), \tag{3.3}$$

где $P_j(x)$ — полином степени $\leq a_j - 1$.

Доказательство. Покажем сначала, что для каждого полинома $Q(x)$ степени не выше $m-1$ справедливо представление формулой (3.3) с указанными оценками на степени полиномов P_k. Действительно, имеем

$$Q(x) = (x-x_1)^{a_1}(x-x_2)^{a_2}\ldots(x-x_{n-1})^{a_{n-1}}P_n(x) + q(x),$$

степень $q(x) \leq a_1+a_2+\ldots+a_{n-1}-1$, а степень $P_n(x) \leq m-1-(a_1+a_2+\ldots+a_{n-1}) = a_n - 1$. Далее, можем записать

$$q(x) = (x-x_1)^{a_1}\ldots(x-x_{n-2})^{a_{n-2}}P_{n-1}(x) + q_1(x),$$

где степень P_{n-1} не превосходит $a_{n-1}-1$. Продолжаем процесс и в итоге получаем представление (3.3).

Поэтому полином Эрмита степени $\leq m-1$ можно искать в виде (3.3). Остается доказать, что полиномы P_1, P_2, $\ldots$, P_n в этом представлении определяются последовательно из условий интерполирования.

Полином P_1 однозначно определяется из условий интерполирования в точке x_1. Действительно, так как

$$H(x) - P_1(x) = (x - x_1)^{a_1} Q_1(x),$$

где $Q_1(x)$ — некоторый полином, то

$$[(x - x_1)^{a_1} Q_1(x)]^{(k)} \mid_{x=x_1} = 0 \quad \text{для} \quad k = 0, 1, \ldots, a_1 - 1.$$

Поэтому

$$[H(x) - P_1(x)]^{(k)} \mid_{x=x_1} = 0 \quad \text{для} \quad k = 0, 1, 2, \ldots, a_1 - 1,$$

а значит

$$P_1^{(k)}(x_1) = H^{(k)}(x_1) = f^{(k)}(x_1)$$

для $k = 0, 1, \ldots, a_1 - 1$. Степень полинома $P_1(x)$ не превосходит $a_1 - 1$, поэтому $P_1^{(k)}(x) = 0$ для $k \geq a_1$. Этими условиями P_1 определяется в полной мере. Например, можно воспользоваться формулой Тейлора

$$P_1(x) = b_0 + \frac{b_1}{1!}(x - x_1) + \ldots + \frac{b_{a_1-1}}{(a_1 - 1)!}(x - x_1)^{a_1 - 1}.$$

Зная P_1 и условия интерполяции в точке x_2, определяем

$P_2(x)$. Из (3.3) следует

$$\frac{H(x)-P_1(x)}{(x-x_1)^{a_1}}-P_2(x)=(x-x_2)^{a_2}Q_2(x),$$

где $Q_2(x)$ — некоторый полином, поэтому производная функции

$$(x-x_2)^{a_2}Q(x)$$

до порядка (a_2-1) в точке x_2 обращается в нуль, т. е.

$$\left[\frac{H(x)-P_1(x)}{(x-x_1)^{a_1}}-P_2(x)\right]^{(k)}\Bigg|_{x=x_2}=0$$

для $k=0,1,\ldots,a_2-1$. Отсюда следует

$$P_2^{(k)}(x_2)=\left[\frac{H(x)-P_1(x)}{(x-x_1)^{a_1}}\right]^{(k)}\Bigg|_{x=x_2}$$

для $k=0,1,\ldots,a_2-1$. Значения функции P_1 и ее производных в точке x_2 известны, а числа

$$H(x_2),H'(x_2),\ldots,H^{(a_2-1)}(x_2)$$

заданы условиями интерполяции

$$H(x_2)=f(x_2),\quad \ldots,H^{(a_2-1)}(x_2)=f^{(a_2-1)}(x_2).$$

Кроме того, $P_2(x)$ — полином степени $\le a_2-1$, поэтому $P_2^{(k)}(x)\equiv 0$ для $k\ge a_2$. По формуле Тейлора можем найти P_2.

Продолжая процесс, по индукции находим все P_k $(k = 1, 2, \ldots, n)$ по той же схеме, причем P_k определяется единственным образом условиями интерполяции в точке x_k. Этим и завершается доказательство теоремы.

Приведем прямое доказательство единственности $H(x)$ в форме (3.3). Предположим, что существует $\widetilde{H}(x)$ — полином степени $\leq m-1$, удовлетворяющий всем условиям интерполирования по Эрмиту. Рассмотрим разность

$$q(x) = H(x) - \widetilde{H}(x).$$

Степень $q(x)$ не превосходит $m-1$, но уравнение

$$q(x) = 0$$

имеет n корней суммарной кратности m, т. е. $q(x)$ можно представить в виде

$$q(x) = (x - x_1)^{a_1} \ldots (x - x_n)^{a_n} q_1(x),$$

где $q_1(x)$ — некоторый полином. Если q_1 не обращается тождественно в нуль, то степень полинома $q(x)$ не ниже $a_1 + a_2 + \ldots + a_n = m$, что невозможно. Следовательно, $q_1(x) \equiv q(x) \equiv 0$. Этим и завершается доказательство единственности.

Получим теперь формулу для остаточного члена при интерполяции с кратными узлами для функции $f \in C^m[a, b]$.

Теорема 3.16 *Пусть* $m = a_1 + a_2 + \ldots + a_n$. *Если* $f \in C^m[a,b]$, *то существует точка* $\xi \in (a,b)$ *такая, что*

$$r(x) = f(x) - H(x) = \frac{f^{(m)}(\xi)}{m!}\ \Omega(x),$$

где $\Omega(x) = (x - x_1)^{a_1}(x - x_2)^{a_2} \ldots (x - x_n)^{a_n}$.

Доказательство аналогично доказательству формулы для остаточного члена интерполяционного полинома Лагранжа.

Достаточно рассмотреть случай, когда $x \neq x_j$. Пусть

$$\varphi(t) = f(t) - H(t) - C\ \Omega(t), \quad a \leq t \leq b.$$

Для фиксированной точки x из $[a,b]$, $x \neq x_j$, постоянная C определяется из условия $\varphi(x) = 0$, т.е.

$$C = \frac{r(x)}{\Omega(x)}.$$

В точках $x, x_1, x_2, \ldots, x_n$ функция $\varphi(t)$ обращается в нуль кратности a_1, a_2, $\ldots$, a_n, соответственно. По теореме Ролля $\varphi'(t) = 0$ в некоторых промежуточных точках ξ_1, ξ_2, $\ldots$, $\xi_n \in (a,b)$. Кроме того, если $a_j \geq 2$, то $\varphi'(x_j) = 0$, причем x_j будет для производной нулем порядка $a_j - 1$. Таким образом, функция $\varphi'(t)$ имеет нули суммарной кратности m. Аналогично получаем, что суммарная кратность нулей второй производной функции $\varphi(t)$ равна $m - 1$. Продолжаем процесс. В итоге

получаем, что $\varphi^{(m)}(\xi) = 0$ по крайней мере для одной точки $\xi \in (a, b)$. Тогда

$$0 = \varphi^{(m)}(\xi) = f^{(m)}(\xi) - H^{(m)}(\xi) - C\Omega^{(m)}(\xi),$$

следовательно, $Cm! = f^{(m)}(\xi)$. Поэтому

$$\frac{f^{(m)}(\xi)}{m!} = C = \frac{r(x)}{\Omega(x)},$$

этим и завершается доказательство.

Выделим два частных случая полиномов Эрмита.

1) Пусть кратности всех узлов равны единице. Тогда мы должны получить, что $H(x) = L_n(f; x)$, и в этом легко убедиться. Действительно, в силу равенств $a_1 = a_2 = \ldots = a_n = 1$ все полиномы P_k в представлении (3.3) имеют нулевую степень, т. е. являются константами. Поэтому формула (3.3) сводится к формуле Ньютона для интерполяционного полинома Лагранжа с коэффициентами $P_k = A_{k-1} = f(x_1; x_2; \ldots; x_k)$.

2) Пусть $n = 1$, $m = a_1 \geq 2$. Тогда в представлении (3.3) необходимо положить $P_k(x) \equiv 0$ при $k \geq 2$. Из доказательства теоремы следует, что

$$P_1(x) = f(x_1) + \frac{f'(x_1)}{1!}(x - x_1) + \frac{f''(x_1)}{2!}(x - x_1)^2 + \ldots$$

$$+\frac{f^{(m-1)}(x_1)}{m!}(x - x_1)^{m-1}.$$

Очевидно, эта формула в сочетании с полученной выше формулой для остаточного члена при интерполяции с кратными узлами равносильна формуле Тейлора для функции $f \in C^m[a,b]$ с остаточным членом в форме Лагранжа.

3.6.2 Полиномы Эрмита-Фейера

Пусть $n \geq 2$ и все узлы имеют одинаковую кратность, равную двум, т. е. $a_1 = a_2 = \ldots = a_n = 2$. Тогда $m = 2n$. В этом случае мы получаем интерполяционный полином $H(x) = H_n(f;x)$ Эрмита-Фейера, для которого можно получить другое явное представление типа формулы Лагранжа для $L_n(f;x)$.

Мы будем пользоваться стандартными обозначениями $l_k(x)$ для фундаментальных полиномов Лагранжа. Напомним, что

$$l_k(x) = \frac{\omega_n(x)}{(x-x_k)\omega'(x_k)},$$

где

$$\omega_n(x) = (x-x_1)(x-x_2)\ldots(x-x_n).$$

Теорема 3.17 *Справедлива следующая формула для полинома Эрмита-Фейера:*

$$H_n(f;x) = \sum_{k=1}^{n} y_k\, l_k^2(x)\,[1 - c_k(x-x_k)] +$$

$$+\sum_{k=1}^{n} y'_k\, l_k^2(x)\,(x-x_k), \tag{3.4}$$

где

$$y_k = f(x_k),\ y'_k = f'(x_k),\ c_k = \frac{\omega''_n(x_k)}{\omega'_n(x_k)}.$$

Доказательство. Легко проверить, что степень полинома, представленного формулой (3.4), не превосходит $2n-1$, так как степени квадратов фундаментальных полиномов Лагранжа равны $2n-2$. С учетом единственности полинома кратной интерполяции нам достаточно проверить выполнение условий

$$f(x_1) = H_n(f; x_1),\ \ldots, f(x_n) = H_n(f; x_n);$$

$$f'(x_1) = H'_n(f; x_1), \ldots, f'(x_n) = H'_n(f; x_n).$$

Поскольку

$$l_k(x_j) = \delta_{kj} = \begin{cases} 1, \text{ если } k = j, \\ 0, \text{ если } k \neq j, \end{cases}$$

для каждого $j = 1, 2, \ldots, n$, будем иметь

$$H_n(f; x_j) = y_j\, 1^2[1 - c_j \cdot 0] + y'_j \cdot 1^2 \cdot 0 = y_j,$$

так как в суммах остаются лишь слагаемые с индексами $k = j$.

Теперь проверим равенства для производных, т. е.

$H_n'(x_j) = y_j'$. Имеем: $H_n'(f;x) = \sum_{k=1}^{n} l_k^2(x)(y_k' - y_k c_k) +$

$$+ \sum_{k=1}^{n} 2 l_k(x) l_k'(x) \{ y_k + (y_k' - y_k c_k)\,(x - x_k) \},$$

отсюда следует

$$H_n'(f;x_j) = y_j' - y_j\, c_j + 2 l_j'(x_j)\, y_j.$$

Пользуясь определением производной и правилом Лопиталя, найдем величины $2l_j'(x_j)$:

$$2l_j'(x_j) = 2 \lim_{x \to x_j} l_j'(x) = 2 \lim_{x \to x_j} \frac{\omega_n'(x)(x - x_j) - \omega_n(x)}{(x - x_j)^2 \omega_n'(x_j)} =$$

$$= 2 \lim_{x \to x_j} \frac{\omega_n''(x)(x - x_j) + \omega_n'(x) - \omega_n'(x)}{2(x - x_j)\omega_n'(x_j)} =$$

$$= \frac{\omega_n''(x_j)}{\omega_n'(x_j)} = c_j.$$

Следовательно,

$$H_n'(f;x_j) = y_j' + [2l_j'(x_j) - c_j]\, y_j = y_j',$$

что и требовалось доказать.

3.7 Тригонометрическая интерполяция

Рассмотрим 2π-периодическую функцию $f \in C(\mathbb{R})$ с вещественными значениями, сетку с $2n+1$ узлами

$$x_0, x_1, \ldots, x_{2n} \in [0, 2\pi],$$

удовлетворяющими условиям

$$0 < |x_i - x_j| < 2\pi, \ i \neq j.$$

Выражение

$$\frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos kx + b_k \sin kx$$

будем называть тригонометрическим полиномом степени n, если $a_n^2 + b_n^2 \neq 0$.

Естественной является следующая задача: построить тригонометрический полином $T_n(f; x)$ степени не выше n, удовлетворяющий условиям

$$T_n(f; x_0) = f(x_0), \ \ldots, \ T_n(f; x_{2n}) = f(x_{2n}).$$

Таким образом, для определения неизвестных коэффициентов $a_0, a_1, \ b_1, \ldots, a_n, b_n$ имеем систему линейных алгебраических уравнений

$$T_n(f; x_j) = f(x_j) \quad (j = 0, \ldots, 2n)$$

порядка $2n+1$. Можно показать, что определитель матрицы

$$\begin{pmatrix} 1/2 & \cos x_0 & \sin x_0 & \dots & \cos nx_0 & \sin nx_0 \\ 1/2 & \cos x_1 & \sin x_1 & \dots & \cos nx_1 & \sin nx_1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1/2 & \cos x_{2n} & \sin x_{2n} & \dots & \cos nx_{2n} & \sin nx_{2n} \end{pmatrix} \tag{3.5}$$

отличен от нуля, поэтому интерполяционный тригонометрический полином определится единственным образом.

Но мы выбираем другой, более плодотворный путь доказательства. А именно, предъявим аналог формулы Лагранжа для $T_n(f;x)$ и проверим лишь выполнение условий интерполирования и единственности.

3.7.1 Аналог формулы Лагранжа

Нам потребуются фундаментальные тригонометрические полиномы Лагранжа $t_k(x)$, заданные формулами

$$t_k(x) = \frac{\prod_{j=0, j\neq k}^{2n} \sin\frac{x-x_j}{2}}{\prod_{j=0, j\neq k}^{2n} \sin\frac{x_k-x_j}{2}},\ k = 0, 1, \dots, 2n.$$

Понятно, что обозначение $t_k(x)$ взято для простоты вместо требуемого громоздкого обозначения

$$t_k(x, x_0, x_1, \dots, x_{2n}).$$

Теорема 3.18 *Для каждой непрерывной 2π-периодической функции f ее тригонометрический интерполяционный полином степени не выше n существует и определяется единственным образом, причем его можно представить в форме Лагранжа*

$$T_n(f;x) = \sum_{k=0}^{2n} f(x_k) t_k(x).$$

Доказательство. Шаг 1. Так как

$$t_k(x_j) = \delta_{kj} = \begin{cases} 0, & j \neq k \\ 1, & j = k \end{cases},$$

равенства значений функции и $T_n(f;x)$ в узлах получаются легко:

$$T_n(f;x_j) = \sum_{k=0}^{2n} f(x_k)\delta_{kj} = f(x_j)\delta_{jj} = f(x_j).$$

Шаг 2. Нам нужно убедиться в том, что функция

$$\sum_{k=0}^{2n} f(x_k) t_k(x)$$

— тригонометрический полином степени не выше n. Очевидно, достаточно показать, что $t_k(x)$ являются тригонометрическими полиномами степени не выше n. Этот факт доказывается методом математической индукции

с применением формул тригонометрии. Функцию $t_k(x)$ можно представить в следующем виде

$$t_k(x) = C \prod_{j=0, j\neq k}^{2n} \sin \frac{x - x_j}{2} \ \ (k = 0, 1, \ldots, 2n),$$

где C — величина, не зависящая от x. Так как произведение содержит $2n$ сомножителей при любом k, то наша задача сводится к следующей: требуется доказать, что функция вида

$$g_n(x) = \prod_{j=1}^{2n} \sin \frac{x - t_j}{2}$$

является тригонометрическими полиномами степени не выше n. Имеем: $g_1(x) =$

$$= \sin \frac{x - t_1}{2} \sin \frac{x - t_2}{2} =$$

$$= \frac{1}{2} \left[\cos \frac{t_2 - t_1}{2} - \cos \frac{2x - t_1 - t_2}{2} \right] =$$

$$= \frac{1}{2} \left[\cos \frac{t_2 - t_1}{2} - \cos x \cos \frac{t_1 + t_2}{2} - \sin x \sin \frac{t_1 + t_2}{2} \right].$$

Таким образом, $g_1(x)$ можно представить в виде

$$g_1(x) = a_0 + a_1 \cos x + b_1 \sin x,$$

где постоянные a_0, a_1, b_1 явно выражаются через t_1, t_2.

Пусть утверждение верно для $n = m$. Тогда

$$g_{m+1}(x) = g_n(x) \cdot \sin\frac{x - t_{2m+1}}{2} \sin\frac{x - t_{2m+2}}{2}.$$

По аналогии с g_1 произведение двух последних множителей приводится к виду $c_0 + c_1 \cos x + d_1 \sin x$. Поэтому можем записать: $g_{m+1}(x) =$

$$= (c_0 + c_1 \cos x + d_1 \sin x) \left(\frac{a_0}{2} + \sum_{j=1}^{m} a_j \cos jx + b_j \sin jx \right).$$

Перемножая и преобразуя произведения синусов и косинусов в суммы, легко убеждаемся в том, что $g_{m+1}(x)$ — тригонометрический полином степени не выше $m+1$.

Шаг 3. Докажем единственность $T_n(f; x)$. Предположим, что существует другой интерполяционный тригонометрический полином $T_n(x)$ степени не выше n. Рассмотрим разность $q(x) = T_n(f; x) - T_n(x)$, которая также является тригонометрическим полиномом степени $\leq n$ и обращается в нуль в узлах сетки:

$$q(x_j) = T_n(f; x_j) - T_n(x_j) = f(x_j) - f(x_j) = 0.$$

Отсюда будет следовать $q(x) \equiv 0$. Действительно, в формуле

$$q(x) = \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos kx + b_k \sin kx$$

можно заменить независимую переменную $x \in [0, 2\pi]$ на комплексную переменную $z = e^{ix}$, где i — мнимая единица. С учетом равенства $z = 1/\overline{z}$ и формул Эйлера получаем

$$\cos kx = \frac{z^k + \overline{z}^k}{2} = \frac{z^k + 1/z^k}{2} = \frac{z^{n+k} + z^{n-k}}{2z^n},$$

$$\sin kx = \frac{z^k - \overline{z}^k}{2i} = \frac{z^k - 1/z^k}{2i} = -i\frac{z^{n+k} - z^{n-k}}{2z^n},$$

поэтому

$$q(x) = \frac{a_0 z^n + \sum_{k=1}^{n}[a_k(z^{n+k} + z^{n-k}) - b_k i(z^{n+k} - z^{n-k})]}{2z^n}.$$

Числитель последней дроби равен нулю тождественно, так как он является алгебраическим полиномом степени $\leq 2n$ относительно переменной $z = e^{ix}$ и обращается в нуль в $2n + 1$ точке $z = e^{ix_j}(j = 0, \ldots, 2n)$. Следовательно, $q(x) \equiv 0$.

Замечание. Нетрудно видеть, что *справедлив следующий аналог теоремы Лебега для оценки погрешности в точке* $x \in [0, 2\pi]$*:*

$$|r_n(x)| := |f(x) - T_n(f; x)| \leq 2\, E_n^{(T)}(f)\ \Lambda_n^{(T)}(x),$$

где величина $E_n^{(T)}(f)$ *— наилучшее равномерное приближение функции* f *тригонометрическими полиномами степени* $\leq n$ *и* $\Lambda_n^{(T)}(x)$ *— тригонометрическая*

функция Лебега, определенная равенством

$$\Lambda_n^{(T)}(x) = \sum_{k=0}^{2n} |t_k(x)|,\ x \in [0, 2\pi].$$

3.7.2 Случай равноотстоящих узлов

Рассмотрим равноотстоящие узлы

$$x_0 = 0, \quad x_1 = h,\ x_2 = 2h,\ \ldots,\ x_{2n} = 2nh = \frac{4n\pi}{2n+1}$$

с шагом $h = \frac{2\pi}{2n+1}$.

В этом случае формулы для фундаментальных тригонометрических полиномов значительно упрощаются. Более того, можно найти явные формулы для коэффициентов a_k и b_k для тригонометрического интерполяционного полинома

$$T_n(f; x) = \frac{a_0}{2} + \sum_{m=1}^{n} a_m \cos mx + b_m \sin mx.$$

Нам потребуется известная функция из теории тригонометрических рядов Фурье, а именно, ядро Дирихле

$$D_n(t) = \frac{1}{2} + \cos t + \ldots + \cos nt \equiv \frac{\sin(n + \frac{1}{2})t}{2 \sin \frac{t}{2}}.$$

Теорема 3.19 *Для каждой непрерывной 2π-периоди-*

ческой функции f и равноотстоящих узлов

$$x_k = \frac{2\pi}{2n+1}k,\ \ k = 0, 1, \ldots, 2n,$$

справедлива формула

$$T_n(f; x) = \frac{2}{2n+1}\sum_{k=0}^{2n} f(x_k)D_n(x - x_k),$$

$$m.\ e.\quad t_k(x) = \frac{2}{2n+1}D_n(x - x_k),$$

а коэффициенты Фурье для тригонометрического полинома $T_n(f; x)$ определяются формулами

$$a_m = \frac{2}{2n+1}\sum_{k=0}^{2n} f(x_k)\cos mx_k, \quad m = 0, 1, \ldots, n,$$

$$b_m = \frac{2}{2n+1}\sum_{k=0}^{2n} f(x_k)\sin mx_k, \quad m = 1, \ldots, n.$$

Доказательство. Пусть

$$\tau_k(x) = \frac{2}{2n+1}D_n(x - x_k).$$

Представление для $T_n(f; x)$ посредством ядра Дирихле немедленно следует из предыдущей теоремы, если мы покажем, что $\tau_k(x)$ равен фундаментальному тригонометрическому полиному Лагранжа $t_k(x)$. В силу теоремы единственности тригонометрического интерполя-

ционного полинома равенство $\tau_k(x) = t_k(x)$ будет верно для любого $k = 0, 1, 2, \ldots, 2n$, если $\tau_k(x)$ являются тригонометрическими полиномами степени не выше n и, кроме того, имеют место равенства

$$\tau_k(x_j) = \delta_{kj} = \begin{cases} 1, & k = j, \\ 0, & k \neq j. \end{cases}$$

Пользуясь первой формулой для ядра Дирихле

$$D_n(x - x_k) = \frac{1}{2} + \cos(x - x_k) + \ldots + \cos n(x - x_k)$$

и формулами элементарной математики

$$\cos m(x - x_k) = \cos mx_k \cdot \cos mx + \sin mx_k \cdot \sin mx,$$

мы легко убеждаемся, что $\tau_k(x)$ — тригонометрический полином степени $\leq n$, так как D_n содержит слагаемые $\cos m(x - x_k)$ с $m \leq n$.

Для вычисления $\tau_k(x_j)$ удобнее пользоваться второй формулой ядра Дирихле, в силу которой

$$\tau_k(x) = \frac{1}{2n+1} \cdot \frac{\sin(n + \frac{1}{2})(x - x_k)}{\sin \frac{x - x_k}{2}}.$$

Для $j \neq k$ непосредственно получаем

$$\tau_k(x_j) = \frac{1}{2n+1} \cdot \frac{\sin\left[\frac{2n+1}{2} \cdot \frac{2\pi}{2n+1}(j - k)\right]}{\sin \frac{2\pi}{2n+1}(j - k)} = 0,$$

а $\tau_k(x_k)$ определяется как предел $\tau_k(x)$ при $x \to x_k$. Привлекая первый замечательный предел, легко получаем:

$$\tau_k(x_k) = \lim_{x \to x_k} \frac{1}{2n+1} \cdot \frac{\frac{2n+1}{2}(x - x_k)}{\frac{x - x_k}{2}} = 1.$$

Нам остается получить формулы для коэффициенты a_m, b_m. С этой целью запишем полученное представление для $T_n(f; x)$ посредством ядра Дирихле с заменой этого ядра соответствующей суммой косинусов. Имеем

$$T_n(f; x) = \frac{2}{2n+1} \sum_{k=0}^{2n} f(x_k) \left(\frac{1}{2} + \sum_{m=1}^{n} \cos m(x - x_k) \right) =$$

$$= \frac{1}{2n+1} \sum_{k=0}^{2n} f(x_k) +$$

$$+ \frac{2}{2n+1} \sum_{m=1}^{n} \sum_{k=0}^{2n} f(x_k)[\cos mx_k \cos mx + \sin mx_k \sin mx].$$

Не зависящее от переменной x слагаемое в этой сумме равно

$$\frac{1}{2n+1} \sum_{k=0}^{2n} f(x_k),$$

а коэффициенты при $\cos mx$ и $\sin mx$ равны, соответ-

ственно, выражениям

$$\frac{2}{2n+1}\sum_{k=0}^{2n} f(x_k)\cos mx_k, \qquad \frac{2}{2n+1}\sum_{k=0}^{2n} f(x_k)\sin mx_k,$$

что и требовалось доказать.

Замечание. Так как

$$h = \Delta x_k = x_k - x_{k-1} = \frac{2\pi}{2n+1},$$

то можем записать коэффициенты a_m и b_m в виде следующих сумм

$$a_m = \frac{1}{\pi}\sum_{k=0}^{2n} f(x_k)\cos mx_k \cdot \Delta x_k,$$

$$b_m = \frac{1}{\pi}\sum_{k=0}^{2n} f(x_k)\sin mx_k \cdot \Delta x_k.$$

Тогда нетрудно заметить, что коэффициенты Фурье тригонометрического интерполяционного полинома для равноотстоящих узлов являются интегральными суммами для коэффициентов Фурье самой функции $f(x)$. Следовательно, для $n \to \infty$

$$a_m \to \frac{1}{\pi}\int_0^{2\pi} f(x)\cos mx\,dx,$$

$$b_m \to \frac{1}{\pi}\int_0^{2\pi} f(x)\sin mx\,dx.$$

3.8 Сплайн-интерполяция

Как мы видели выше, вопрос о равномерной сходимости интерполяционных полиномов к интерполируемой функции при неограниченном росте числа точек интерполяции является сложным. Во-первых, наилучшее равномерное приближение $E_n(f)$ должно иметь "хорошие" свойства. Во-вторых, в общем случае успеха можно добиться лишь специальным подбором узлов.

Было обнаружено, что вопросы сходимости значительно упрощаются, если в качестве приближающих функций используются кусочно-полиномиальные функции. Такие функции, удовлетворяющие некоторым дополнительным требованиям, называются сплайнами.

Теория сплайн-интерполяции бурно развивается с сороковых годов 20-го столетия. Мы рассмотрим лишь основы этой богатой теории.

Можно отметить, что сплайны возникли уже на заре математического анализа в работах Лейбница и особенно в трудах Эйлера при разработке прямых методов вариационного исчисления.

Английское слово "сплайн" переводится как рейка. Оно стало математическим термином по праву: инженеры и чертежники издавна использовали гибкие рейки для ручной интерполяции функций, заданных значениями на конечном числе точек.

3.8.1 Определение сплайнов степени m

Перейдем к точным определениям. Непрерывная функция

$$g \,:\, [a,b] \to \mathbb{R}$$

называется полиномиальным сплайном, если существует разбиение

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b$$

такое, что на каждом частичном отрезке $[x_{k-1}, x_k]$ функция

$$g\,|_{[x_{k-1},x_k]} := g_k \,:\, [x_{k-1}, x_k] \to \mathbb{R}.$$

является некоторым полиномом определенной степени.

Определение 3.2 *Пусть $f \in C[a,b]$, и пусть заданы узлы*

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b, \qquad n \in \mathbb{N}.$$

Говорят, что функция $g(x) = S_n^m(f;x)$ является для f интерполяционным сплайном степени $m \geq 1$, если выполняются условия:

1) g непрерывна на $[a,b]$, а на каждом частичном отрезке $[x_{k-1}, x_k]$

$$g(x) = g_k(x),$$

где $g_k(x)$ — некоторый полином степени $\leq m$, т. е.

имеет вид

$$g_k(x) = \sum_{j=0}^{m} a_{kj} x^j;$$

2) для каждого узла x_j $(j = 0, \ldots, n)$

$$g(x_j) = f(x_j);$$

3) если $m \geq 2$, *то* $g \in C^{(m-1)}[a, b]$.

Отметим, что в специальной литературе, где рассматриваются более общие вопросы, определенные нами сплайны называются сплайнами порядка m и дефекта 1.

Сплайны предоставляют удобный аппарат приближения функций конечной гладкости. Мы рассмотрим подробнее лишь наиболее употребительные на практике сплайны первой степени ($m = 1$) и кубические сплайны ($m = 3$).

При исследовании порядка приближения нам потребуется понятие модуля непрерывности для функции $f \in C[a, b]$. Напомним определение и некоторые свойства. *Модуль непрерывности* $\omega(f, \delta)$ определяется следующим образом: для фиксированного положительного числа $\delta \in (0, b - a]$

$$\omega(f, \delta) := \sup_{x', x'' \in [a,b], |x' - x''| \leq \delta} |f(x') - f(x'')|.$$

Из определения непосредственно следует, что модуль непрерывности является монотонно неубывающей функцией переменной δ, $\delta \in (0, b-a]$. Кроме того, условие $f \in C[a,b]$ равносильно равенству

$$\lim_{\delta \to 0} \omega(f;\delta) = 0$$

в силу теоремы Кантора о равномерной непрерывности функции, непрерывной на отрезке.

Принято выделять подпространства непрерывных функций посредством фиксации свойства модуля непрерывности. Одним из наиболее употребительных подпространств является класс $\operatorname{Lip}\alpha$ (Липшиц-альфа), где $\alpha \in (0,1]$ — фиксированное число.

По определению, $f \in \operatorname{Lip}\alpha$ означает существование некоторой постоянной $M > 0$ такой, что для всех $x', x'' \in [a,b]$ имеет место неравенство

$$|f(x') - f(x'')| \le M|x' - x''|^{\alpha}.$$

Очевидно, условие $f \in \operatorname{Lip}\alpha$ равносильно неравенству

$$\omega(f;\delta) \le M\delta^{\alpha}$$

с некоторой постоянной $M > 0$. Отметим также, что если $f \in C^1[a,b]$, то $f \in \operatorname{Lip} 1$, но обратное утверждение, вообще говоря, неверно.

Действительно, для любого отрезка $[x', x''] \subset [a, b]$ по формуле Лагранжа о конечных приращениях можно записать: $\exists \xi \in (x', x'')$ такое, что

$$f(x'') - f(x') = f'(\xi)(x'' - x'),$$

поэтому

$$|f(x'') - f(x')| \leq M|x'' - x'|$$

с постоянной

$$M = \max_{x \in [a,b]} |f'(x)| < \infty.$$

С другой стороны, функция $f(x) = |x|$, $x \in [-1, 1]$, не имеет производной в точке нуль, т. е. не является непрерывно дифференцируемой, но она удовлетворяет условию Липшица с постоянной $M = 1$, так как

$$|f(x'') - f(x')| = ||x''| - |x'|| \leq |x'' - x'|.$$

3.8.2 Сплайны первой степени

Рассмотрим сплайн первой степени $g(x) = S_n^1(f; x)$ для функции

$$f \in C[a, b], \quad a = x_0 < \ldots < x_n = b.$$

По определению интерполяционного сплайна $g \in C[a, b]$, $g(x_k) = f(x_k)$, $k = 0, 1, \ldots, n$, кроме того, на любом

частичном отрезке $[x_{k-1}, x_k]$

$$g(x) = g_k(x) = a_k x + b_k.$$

Таким образом, речь идет об аппроксимации $f \in C[a, b]$ непрерывными, кусочно-линейными функциями.

Существование и единственность интерполяционного сплайна 1-ой степени получаются тривиально. Действительно, нахождение $g_k(x) = a_k x + b_k$ геометрически сводится к построению отрезка прямой, проходящей через 2 точки с координатами $(x_{k-1}, f(x_{k-1}))$, $(x_k, f(x_k))$.

Кроме того, мы можем интерпретировать

$$g_k(x) = a_k x + b_k$$

как интерполяционный полином Лагранжа, построенный по двум узлам x_{k-1}, x_k. По доказанному ранее, такой полином существует, определяется единственным образом и может быть представлен по формуле Лагранжа на отрезке $[x_{k-1}, x_k]$ в явном виде как

$$g(x) = g_k(x) = f(x_{k-1})\frac{x - x_k}{x_{k-1} - x_k} + f(x_k)\frac{x - x_{k-1}}{x_k - x_{k-1}}.$$

Равенства $g(x_k) = f(x_k)$ и $g(x_{k-1}) = f(x_{k-1})$ очевидны. Рассмотрим аппроксимационные свойства сплайнов первой степени. Отметим прежде всего *представ-*

ление типа Лагранжа

$$S_n^1(f;x) = \sum_{j=0}^{n} f(x_j)\ s_j(x),$$

где $s_j(x)$ — *фундаментальные сплайны первой степени* со стандартным свойством $s_j(x_k) = \delta_{kj}$. Мы можем написать их в явном виде. Для крайних узлов

$$s_0(x) = \begin{cases} \frac{x_1 - x}{x_1 - a} & \text{при } \ a \le x \le x_1, \\ 0 & \text{при } \ x_1 \le x \le b; \end{cases}$$

$$s_n(x) = \begin{cases} 0 & \text{при } \ a \le x \le x_{n-1}, \\ \frac{x - x_{n-1}}{b - x_{n-1}} & \text{при } \ x_{n-1} \le x \le b; \end{cases}$$

и при любом $1 \le j \le n-1$, т. е. для внутренних узлов,

$$s_j(x) = \begin{cases} 0 & \text{при } \ a \le x \le x_{j-1}, \\ \frac{x - x_{j-1}}{x_j - x_{j-1}} & \text{при } \ x_{j-1} \le x \le x_j, \\ \frac{x_{j+1} - x}{x_{j+1} - x_j} & \text{при } \ x_j \le x \le x_{j+1}, \\ 0 & \text{при } \ x_{j+1} \le x \le b. \end{cases}$$

Норма оператора $S_n^1 : C[a,b] \to C[a,b]$ легко вычисляется и равна 1 при любом n, так как

$$\sum_{j=0}^{n} |s_j(x)| \equiv \sum_{j=0}^{n} s_j(x) \equiv 1.$$

В силу ограниченности нормы оператор S_n^1 должен обладать хорошими аппроксимационными свойствами.

Мы получим оценки погрешности интерполяции с использованием модуля непрерывности интерполируемой функции или ее производной, а также диаметра разбиения $x_0 = a < x_1 < x_2 < \ldots < x_n = b$, определяемого стандартно как

$$\delta_n = \max_{k=1,\ldots,n} |x_k - x_{k-1}|.$$

Теорема 3.20 *Для каждой функции $f \in C[a,b]$ ее интерполяционный сплайн $S_n^1(f;x)$, построенный по сетке $x_0 = a < x_1 < x_2 < \ldots < x_n = b$ с диаметром разбиения δ_n, обладает следующими свойствами:*

$$1) \ \left\| f(x) - S_n^1(f;x) \right\|_{C[a,b]} \le \omega(f, \delta_n);$$

$$2) \ S_n^1(f;x) \rightrightarrows f(x) \quad \textit{при} \quad \delta_n \to 0.$$

Доказательство. Утверждение 2) следует из 1) в силу свойств модуля непрерывности. Поэтому достаточно доказать 1).

Пусть $x \in [a,b]$, тогда x попадает в один из частичных отрезков, т. е. $x \in [x_{k-1}, x_k]$ для некоторого k. Тогда

$$f(x) - S_n^1(f;x) = f(x) - g_k(x) = f(x)\frac{x_k - x + x - x_{k-1}}{x_k - x_{k-1}} -$$

$$-\frac{f(x_{k-1})(x_k - x)}{x_k - x_{k-1}} - \frac{f(x_k)(x - x_{k-1})}{x_k - x_{k-1}} =$$

$$= [f(x) - f(x_{k-1})]\frac{x_k - x}{x_k - x_{k-1}} + [f(x) - f(x_k)]\frac{x - x_{k-1}}{x_k - x_{k-1}}.$$

Из соотношений

$$0 \le x - x_{k-1} \le x_k - x_{k-1} \le \delta_n,\ 0 \le x_k - x \le x_k - x_{k-1} \le \delta_n$$

следует

$$|f(x) - f(x_{k-1})| \le \omega(f, \delta_n),$$

$$|f(x) - f(x_k)| \le \omega(f, \delta_n).$$

Таким образом, приходим к соотношениям

$$|f(x) - S_n(f; x)| \le \omega(f; \delta_n)\frac{x - x_{k-1}}{x_k - x_{k-1}} +$$

$$+\omega(f; \delta_n)\frac{x_k - x}{x_k - x_{k-1}} = \omega(f; \delta_n).$$

Теорема доказана.

Отметим простое следствие теоремы.

Если $\alpha \in (0, 1]$ *и* $f \in Lip\,\alpha$, *то существует постоянная* M *такая, что* $\omega(f, \delta_n) \le M\delta_n^\alpha$. *Поэтому*

$$\|f(x) - S_n(f; x)\|_{C[a,b]} = O(\delta_n^\alpha).$$

Для непрерывно дифференцируемых функций погрешность интерполяции допускает более сильную оценку.

Теорема 3.21 *Пусть* $f \in C^1[a,b]$, $S_n^1(f;x)$ — *ее интерполяционный сплайн 1-ой степени, построенный по узлам* $x_0 = a < x_1 < x_2 < \ldots < x_n = b$ *с диаметром* δ_n. *Тогда*

$$\left\| f(x) - S_n^1(f;x) \right\|_{C[a,b]} \le \frac{\delta_n}{4} \omega(f', \delta_n).$$

Доказательство. Как и в теореме 3.20 получаем формулы

$$f(x) - S_n^1(f;x) = f(x) - g_k(x) =$$

$$= [f(x) - f(x_{k-1})] \frac{x_k - x}{x_k - x_{k-1}} + [f(x) - f(x_k)] \frac{x - x_{k-1}}{x_k - x_{k-1}}$$

для $x \in [x_{k-1}, x_k]$. По формуле Лагранжа о конечных приращениях существуют $\xi \in (x_{k-1}, x)$ и $\eta \in (x, x_k)$ такие, что

$$f(x) - f(x_{k-1}) = f'(\xi)(x - x_{k-1}),$$

$$f(x) - f(x_k) = -f'(\eta)(x_k - x).$$

Следовательно,

$$f(x) - S_n^1(f;x) = [f'(\xi) - f'(\eta)] \frac{(x_k - x)(x - x_{k-1})}{x_k - x_{k-1}}.$$

Оценим сверху модуль правой части. Из соотношений

$$|\xi - \eta| \le x_k - x_{k-1} \le \delta_n$$

и определения модуля непрерывности следует неравенство

$$|f'(\xi) - f'(\eta)| \le \omega(f', \delta_n),$$

которое вместе с элементарным неравенством

$$\frac{(x_k - x)(x - x_{k-1})}{x_k - x_{k-1}} \le \frac{(x_k - x_{k-1})}{4} \le \frac{\delta_n}{4}$$

влечет искомый факт:

$$\|f(x) - S_n^1(f; x)\|_{C[a;b]} \le \omega(f', \delta_n)\frac{\delta_n}{4}.$$

Можно выделить 2 важных следствия доказанной теоремы.

Следствие 3.21.1 *Если* $\alpha \in (0, 1]$ *и* $f' \in Lip\,\alpha$, *то*

$$\|f(x) - S_n^1(f; x)\|_{C[a,b]} = O(\delta_n^{1+\alpha}).$$

Следствие 3.21.2 *Для любой функции* $f \in C^2[a, b]$

$$\|f(x) - S_n^1(f; x)\|_{C[a,b]} = O(\delta_n^2).$$

В частности, если интерполяционный полином построен по равноотстоящим узлам с шагом $h = \delta_n = \frac{b-a}{n}$, *то*

$$\|f(x) - S_n^1(f; x)\|_{C[a,b]} = O\left(\frac{1}{n^2}\right).$$

Отметим так называемое "свойство насыщаемости" сплайна первой степени, которое заключается в следу-

ющем: дальнейшее увеличение порядка гладкости интерполируемой функции, например, требование

$$f \in C^r[a,b], \quad r \geq 3,$$

не приводит к лучшим оценкам погрешности аппроксимации, чем оценка $O(\delta_n^2)$ для дважды непрерывно дифференцируемых функций.

Невозможность дальнейшего повышения порядка малости погрешности за счет порядка гладкости интерполируемой функции можно демонстрировать на простом примере.

Пример. Рассмотрим сколь угодно гладкую функцию $f_0(x) = x^2$ на отрезке $[-1,1]$ и сетку с равноотстоящими узлами

$$x_k = -1 + kh, h = 2/n, \;\; k = 0,1,\ldots,n.$$

Пусть n — нечетное число. Тогда один из частичных отрезков имеет вид $[-h/2, h/2]$, и на этом отрезке, очевидно, $S_n^1(f_0,x) \equiv h^2/4$. Поэтому

$$\|f_0(x) - S_n^1(f;x)\|_{C[a;b]} \geq |f_0(0) - S_n^1(f;0)| = h^2/4.$$

Если n — четное число, то полученная оценка снизу для погрешности интерполяции также верна (покажите!).

Замечание. Обратите внимание, что в предыдущих рассуждениях речь идет об оценках погрешности,

гарантированных для всех функций из заданных классов функций. Понятно, что для конкретной функции аппроксимация может быть намного лучше. Например, если взять непрерывную, кусочно-линейную функцию, то погрешность тождественно равна нулю при подходящем выборе сетки.

Рассмотрим теперь вариационное свойство сплайнов первой степени. Нам потребуется *пространство Соболева* $W_2^1[a,b]$*, определяемое как пространство абсолютно непрерывных функций* $F : [a,b] \to \mathbb{R}$*, для которых существует интеграл* $\int_a^b F'^2(x)dx$ *и норма определена равенством*

$$\|F\|_{W_2^1} = \|F\|_{C[a,b]} + \|F'\|_{L_2[a,b]}.$$

Известно, что $W_2^1[a,b]$ — полное линейное нормированное (т. е. банахово) пространство. Производная функции F понимается как обобщенная производная в смысле Соболева, т. е. существует некоторая интегрируемая в смысле Лебега функция F', удовлетворяющая равенству

$$\int_a^b F(x)\varphi'(x)\,dx = -\int_a^b F'(x)\varphi(x)\,dx$$

для любой пробной функции $\varphi \in C^1[a,b]$ такой, что $\varphi(a) = \varphi(b) = 0$. Пространство $W_2^1[a,b]$ содержит в себе все непрерывные, кусочно-гладкие функции, определенные на отрезке $[a,b]$, в частности, сплайны $S_n^1(F;x)$.

Пусть $f : [a,b] \to \mathbb{R}$ — заданная непрерывная функция, и

$$a = x_0 < x_1 < \ldots < x_n = b$$

— некоторая фиксированная сетка.

Рассмотрим задачу минимизации функционала

$$\Phi(F) = \int_a^b F'^2(x)\,dx$$

при следующих условиях:

1) $F \in W_2^1[a,b]$,

2) имеют место равенства $F(x_k) = f(x_k)$ для всех $k = 0, \ldots, n$.

Очевидно, сплайн $g(x) = S_n^1(f;x)$ является одной из функций, удовлетворяющей обоим условиям.

Теорема 3.22 *Для любой функции F, удовлетворяющей условиям 1) и 2), имеет место неравенство*

$$\int_a^b F'^2(x)\,dx \geq \int_a^b \left(\frac{dS_n^1(f;x)}{dx}\right)^2 dx,$$

где равенство достигается тогда и только тогда, когда $F(x) \equiv S_n^1(f;x)$.

Доказательство. Пусть $g(x) = S_n^1(f;x)$. Имеем

$$\Phi(F-g) = \int_a^b (F'-g')^2 dx = \int_a^b (F'^2 - 2F'g' + g'^2)dx =$$

$$= \int_a^b F'^2 dx - \int_a^b g'^2 dx - 2\int_a^b (F'g' - g'^2)dx.$$

Вычисления показывают, что третий интеграл равен нулю. Действительно, пользуясь аддитивностью интеграла и формулой интегрирования по частям, получаем

$$A = \int_a^b g'(x)[F'(x) - g'(x)]dx =$$

$$= \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} g'(x)d[F(x) - g(x)].$$

Так как

$$g'(x) = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} = C_k$$

не зависит от x, то

$$A = \sum_{k=1}^{n} C_k \int_{x_{k-1}}^{x_k} d[F(x) - g(x)] = 0$$

в силу того, что

$$\int_{x_{k-1}}^{x_k} d[F(x) - g(x)] =$$

$$= [F(x_k) - g(x_k)] - [F(x_{k-1}) - g(x_{k-1})] = 0.$$

Таким образом, мы доказали, что

$$\int_a^b F'^2(x)dx = \int_a^b g'^2(x)dx + \Phi(F - g) \geq \int_a^b g'^2(x)dx.$$

Следовательно, с учетом обозначения $g(x) = S_n^1(f;x)$

$$\min \Phi(F) = \Phi(S_n^1(f;x)).$$

Докажем теперь единственность экстремальной функции. Предположим, что существует еще одна экстремальная функция F_1. Но тогда

$$\int_a^b g'^2(x)dx = \int_a^b F_1'^2 dx = \int_a^b g'^2(x)dx + \Phi(F_1 - g),$$

отсюда следует

$$\Phi(F_1 - g) = \int_a^b (F_1' - g')^2 dx = 0,$$

значит $F_1'(x) = g'(x)$ почти всюду на $[a,b]$, отсюда

$$F_1(x) = g(x) + \text{Const} \equiv S_n^1(f;x) + \text{Const}.$$

Константа равна нулю в силу равенств $F_1(x_k) = g(x_k)$, поэтому $F_1(x) \equiv S_n^1(f;x)$, что и требовалось доказать.

3.8.3 Кубические сплайны

Для заданной функции $f \in C[a,b]$ и узлов $a = x_0 < x_1 < x_2 < \ldots < x_n = b$ сплайн третьей степени, т. е. кубический сплайн

$$g(x) = S_n^3(f;x)$$

определяется тремя условиями:

I) на каждом отрезке $[x_{k-1}, x_k]$ $(k = 1, 2, \ldots, n)$

$$g(x) = g_k(x) = a_{k0} + a_{k1}x + a_{k2}x^2 + a_{k3}x^3$$

— полином степени ≤ 3;

II) для каждого $k = 0, 1, \ldots, n$

$$g(x_k) = f(x_k);$$

III) $g \in C^2[a, b]$, т. е. g, g', g'' непрерывны на $[a, b]$. Это условие фактически сводится к дважды гладкой склейке на внутренних узловых точках полиномов g_k из соседних частичных отрезков: для каждого

$$k = 1, 2, \ldots, n - 1$$

должны выполняться равенства

$$g_k(x_k) = g_{k+1}(x_k),\ g'_k(x_k) = g'_{k+1}(x_k),\ g''_k(x_k) = g''_{k+1}(x_k).$$

Условиями I—III кубический сплайн определяется не единственным образом, поскольку число неизвестных коэффициентов a_{kj} равно $4n$, а число уравнений для их определения равно $4n - 2$. А именно, $n + 1$ уравнение дано условиями интерполирования и $3(n-1)$ уравнений предоставлены условиями дважды гладкой склейки на внутренних узловых точках.

Таким образом, нужны еще 2 условия. Дополни-

тельные условия вида $g'(a) = g'(b)$, $g''(a) = g''(b)$ обычно применяются для периодических функций с периодом $T = b - a$.

Для непериодических функций наиболее употребительными являются так называемые естественные кубические сплайны, они определяются присоединением следующих дополнительных условий:

$$g''(a) = g''(b) = 0.$$

Теорема 3.23 *Для каждой функции $f \in C[a,b]$ ее естественный кубический сплайн $g(x) = S_n^3(f;x)$, построенный по сетке $x_0 = a < x_1 < x_2 < \ldots < x_n = b$, существует и определяется единственным образом.*

Доказательство. Матрица системы из $4n$ линейных алгебраических уравнений для прямого определения неизвестных коэффициентов a_{kj} оказывается громоздкой. Поэтому используется такой "трюк". В дополнение к числам $y_0 = g''(x_0) = 0$, $y_n = g''(x_n) = 0$ вводятся неизвестные заранее параметры (моменты):

$$y_1 = g''(x_1),\ y_2 = g''(x_2), \ldots, y_{n-1} = g''(x_{n-1}).$$

Покажем, что по этим параметрам однозначно определяются $g_k(x)$, а сами числа y_k $(k = 1, 2, \ldots, n-1)$ находятся как решение несложной системы линейных алгебраических уравнений порядка $n - 1$.

На каждом частичном отрезке $[x_{k-1}, x_k]$ функция $g''(x) \equiv g_k''(x)$ является линейной, поэтому

$$g''(x) = (1-t)y_{k-1} + ty_k, \;\; t = \frac{x - x_{k-1}}{\Delta x_k}, \;\; \Delta x_k = x_k - x_{k-1}.$$

Интегрированием по переменной t с учетом равенства $dx = \Delta x_k dt$ получаем

$$g'(x) = g'(x_{k-1}) + \frac{\Delta x_k}{2}(1 - (1-t)^2)y_{k-1} + \frac{\Delta x_k}{2}t^2 y_k,$$

$$g(x) = g(x_{k-1}) + \Delta x_k t g'(x_{k-1}) +$$
$$+ \frac{(\Delta x_k)^2}{6}(3t + (1-t)^3 - 1)y_{k-1} + \frac{(\Delta x_k)^2}{6}t^3 y_k.$$

Полагая $t = 1$ в выражении для $g(x)$ и учитывая равенства $g(x_{k-1}) = f(x_{k-1}), \;\; g(x_k) = f(x_k)$, находим

$$g'(x_{k-1}) = \frac{f(x_k) - f(x_{k-1})}{\Delta x_k} - \frac{\Delta x_k}{3}y_{k-1} - \frac{\Delta x_k}{6}y_k.$$

Подставляя это значение $g'(x_{k-1})$ в выражение для $g'(x)$ и полагая $t = 1$, получаем

$$g'(x_k) = \frac{f(x_k) - f(x_{k-1})}{\Delta x_k} + \frac{\Delta x_k}{6}y_{k-1} + \frac{\Delta x_k}{3}y_k.$$

Равенства $\quad g_k'(x_k) = g_{k+1}'(x_k) \qquad (k = 1, 2, \ldots, n-1),$ т. е. условия непрерывной склейки первых производных, приводят к линейной системе для моментов y_k

$(k = 1, 2, \ldots, n-1)$:

$$\frac{f(x_k) - f(x_{k-1})}{\Delta x_k} + \frac{\Delta x_k}{6} y_{k-1} + \frac{\Delta x_k}{3} y_k =$$

$$= \frac{f(x_{k+1}) - f(x_k)}{\Delta x_{k+1}} - \frac{\Delta x_{k+1}}{3} y_k - \frac{\Delta x_{k+1}}{6} y_{k+1}$$

или, что то же самое, к системе

$$\Delta x_k \; y_{k-1} + 2\,(\Delta x_k + \Delta x_{k+1}) \; y_k + \Delta x_{k+1} \; y_{k+1} = b_k,$$

где $k = 1, 2, \ldots, n-1$, $y_0 = y_n = 0$, а свободные члены даны равенствами

$$b_k = 6\,\frac{f(x_{k+1}) - f(x_k)}{\Delta x_{k+1}} - 6\;\frac{f(x_k) - f(x_{k-1})}{\Delta x_k}.$$

Нетрудно показать, что полученная система однозначно разрешима: матрица системы относится к типу "трехдиагональной с диагональным преобладанием".

Отметим также, что кубический сплайн можно построить иным выбором вспомогательных параметров, а именно, исходя из величин $z_k = g'(x_k)$ $(k = 0, 1, \ldots, n)$. При таком подходе получается формула (докажите!)

$$g_k(x) = (1-t)^2(1+2t) f(x_{k-1}) + t^2(3-2t) f(x_k) +$$

$$+ t(1-t)\Delta x_k \,[(1-t)\, z_{k-1} - t\, z_k],$$

и система для определения параметров z_k также оказывается трехдиагональной.

Опишем теперь кратко вариационное свойство естественных сплайнов. Пусть $f : [a,b] \to \mathrm{R}$ — заданная непрерывная функция, и $a = x_0 < x_1 < \ldots < x_n = b$ — некоторая фиксированная сетка. Рассмотрим задачу минимизации функционала энергии

$$E(F) = \int_a^b F''^2(x)\,dx$$

при следующих условиях:

1) $F \in W_2^2[a,b] = \{F \in C[a,b]$: существует обобщенная производная F'' и $F'' \in L_2[a,b]\}$;

2) $F(x_k) = f(x_k)$, где $k = 0, \ldots, n$.

Очевидно, кубический сплайн $g(x) = S_n^3(f;x)$ является одной из функций, удовлетворяющей обоим условиям.

Теорема 3.24 *Для любой функции F, удовлетворяющей условиям 1) и 2),*

$$\int_a^b F''^2(x)\,dx \geq \int_a^b \left(\frac{d^2 S_n^3(f;x)}{dx^2}\right)^2 dx,$$

где равенство достигается тогда и только тогда, когда $F(x) \equiv S_n^3(f;x)$ — естественный кубический сплайн.

Доказательство аналогично доказательству теоремы 3.22. В силу равенства

$$F''^2 - g''^2 = (F'' - g'')^2 + 2g''(F'' - g''),$$

можем написать

$$E(F) - E(g) = E(F - g) + 2\int_a^b g''(F'' - g'')\,dx.$$

Интеграл от функции $2g''(F'' - g'')$ для $g(x) = S_n^3(f; x)$ равен нулю, в чем легко убедиться интегрированием по частям.

3.9 Задачи и упражнения

1. Пользуясь точными значениями $\sin 0$, $\sin \frac{\pi}{6}$, $\sin \frac{\pi}{2}$ и интерполяционным полиномом Лагранжа, найдите приближенное значение $\sin \frac{\pi}{7}$ и дайте оценку погрешности.

2. Пользуясь полиномом Лагранжа, найдите приближенное значение $\log 70$ и дайте оценку погрешности в двух случаях: заданы а) $\log 1$, $\log 10$; b) $\log 1$, $\log 10$, $\log 100$.

3. Найдите приближенное значение $\operatorname{arctg} \frac{1}{2}$ и дайте оценку погрешности.

4. Для полиномов Чебышева первого рода докажите тождество:

$$\frac{1 - xt}{1 - 2xt + x^2} = \sum_{n=0}^{\infty} x^n T_n(t), \qquad |x| < 1, \quad |t| \le 1.$$

5. Покажите, что для любого $n \ge 1$ полином Че-

бышева $T_n(t)$ удовлетворяет следующему дифференциальному уравнению

$$(1-t^2)T_n''(t) - tT_n'(t) + n^2T_n(t) = 0.$$

6. Для функции $f(x) = \sin \pi x$ и узлов

$$\{0, 1/4, 1/3, 1/2\}$$

запишите интерполяционный полином Ньютона.

7. Пусть $f(x) = 3x^3 + 2x^2 + x + 1$, и заданы узлы $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$. Найдите разделенную разность $f(x_1, x_2, x_3, x_4)$ и конечную разность $\Delta^3 f_1$.

8. Найдите интерполяционный полином в форме Ньютона для функции $f(x) = x^4$ и узлов $\{0, 1, 2, 3\}$.

9. Для функции $f(x) = x^4$ и двух узлов $\{0, 1\}$ запишите интерполяционный полином Эрмита-Фейера.

10. Аппроксимируйте полином Чебышева $T_3(x)$ на отрезке $[-1, 1]$ интерполяционным полиномом Эрмита с одним узлом $x_0 = 0$ кратности 3. Дайте оценку погрешности приближения.

11. Рассмотрите полином Лагранжа для равноотстоящих узлов $x_0, x_1, \ldots, x_n$, где $a = x_0, b = x_n$ и

$$x_1 - x_0 = x_2 - x_1 = \ldots = x_n - x_{n-1} = h = \frac{b-a}{n}.$$

Преобразуйте интерполяционный полином

$$L_{n+1}(f;x)=\sum_{k=0}^{n}f(x_k)l_k(x)$$

степени $\leq n$ с помощью замены переменной

$$t=\frac{x-a}{h}\quad(x=a+ht).$$

Покажите, что для выбранной сетки из равноотстоящих узлов $x_0,x_1,\ldots,x_n$ имеет место формула

$$L_{n+1}(f;x)=\frac{(-1)^n t(t-1)\ldots(t-n)}{n!}\sum_{k=0}^{n}f(x_k)\frac{(-1)^k C_n^k}{(t-k)},$$

где

$$C_n^k=\frac{n!}{k!(n-k)!}$$

— биномиальные коэффициенты.

12. Найдите разность между интерполяционным полиномом Лагранжа по узлам $x_0=a$, $x_1=c=(a+b)/2$, $x_2=b$ и интерполяционным полиномом Эрмита по тем же узлам, но разной кратности: x_0,x_2 — простые узлы, а x_1 — узел кратности 2.

13. Функцию $f(x)=e^{\sin x}$ аппроксимируйте тригонометрическим интерполяционным полиномом по узлам $x_k=\frac{2\pi k}{3}$, $k=0,1,2$.

14. Функцию $f(x)=x^2$ аппроксимируйте на отрезке $[0,1]$ сплайном первой степени с выбором узлов:

$x_k = kh$, $h = 1/n$, $k = 0, 1, \ldots, n$. Дайте оценку погрешности приближения.

15. При доказательстве теоремы Вейерштрасса по методу Лебега нам встретилась система линейных алгебраических уравнений

$$a_1 - \sum_{j=2}^{m} a_j = k_1,$$

$$\sum_{j=1}^{s} a_j - \sum_{j=s+1}^{m} a_j = k_s, \quad s = 2, \ldots, m-1,$$

$$\sum_{j=1}^{m} a_j = k_m$$

относительно неизвестных a_j. Покажите, что решение этой системы можно записать в явном виде.

16. Вы знакомы с доказательством существования и единственности естественного кубического сплайна $g(x) = S_n^3(x; f)$, $a \le x \le b$, определяемого дополнительными условиями: $g''(a) = g''(b) = 0$. Докажите два аналога этой теоремы, когда дополнительные условия имеют вид

$$g''(a) = \alpha, \ g''(b) = \beta$$

или

$$g'(a) = \alpha, \ g'(b) = \beta,$$

где α и β — заданные числа.

Глава 4

Наилучшие приближения

Пусть F — линейное нормированное пространство над полем вещественных чисел. Рассмотрим некоторую систему $\{l_1, l_2, \ldots, l_n\}$ линейно-независимых элементов из F . Их линейные комбинации, т. е. элементы вида

$$f_n = \sum_{k=1}^{n} \alpha_k l_k \quad (\alpha_k \in \mathbb{R})$$

образуют замкнутое подпространство $F_n = \{f_n\}$. Для любого $f \in F$ ставится задача минимизации функционала $\Phi : \mathbb{R}^n \to \mathbb{R}$, определенного равенством

$$\Phi(\alpha_1, \ldots, \alpha_n) = \left\| f - \sum_{k=1}^{n} \alpha_k l_k \right\|_F .$$

Инфимум этой нормы, т. е. неотрицательная величина

$$E_n(f) = \inf_{\alpha_1,\ldots,\alpha_n} \Phi(\alpha_1, \ldots, \alpha_n)$$

называется наилучшим приближением $f \in F$ (элементами $f_n \in F_n \subset F$). Существование и единственность наилучшего приближения легко следуют из определения и классических теорем анализа. Остается открытым лишь вопрос о нахождении этой величины.

Далее, если существует элемент

$$f_n^0 = \sum_{k=1}^{n} \alpha_k^0 l_k \in F_n,$$

на котором достигается этот инфимум, то его называют элементом наилучшего приближения. Возникают естественные **вопросы:**

1) существует ли элемент наилучшего приближения f_n^0;

2) определяется ли единственным образом;

3) каков алгоритм практического построения f_n^0.

Забегая вперед, укажем, что существование элемента наилучшего приближения имеет место при самых общих предположениях.

Для единственности и алгоритма построения элемента наилучшего приближения необходимы дополнительные предположения о структуре пространства F.

Вопрос 3) мы рассмотрим в двух случаях, когда пространство F является гильбертовым или F — банахово пространство $C[a, b]$.

4.1 НП в пространствах с нормой

Докажем сначала теорему существования.

Теорема 4.1 *Пусть F — линейное нормированное пространство над полем вещественных чисел. Тогда для любого $f \in F$ существует элемент наилучшего приближения $f_n^0 \in F_n$.*

Доказательство. Если $f \in F_n$, то ясно, что $f_n^0 = f$ и $E_n(f) = 0$. Таким образом, этот случай является простым.

Рассмотрим нетривиальный случай, когда

$$f \notin F_n, \qquad E_n(f) = \inf\{\|f - f_n\|, \quad f_n \in F_n\} > 0.$$

По определению инфимума существует последовательность $u_m \in F_n$ ($m \in \mathbb{N}$) такая, что

$$\|f - u_m\| \le E_n(f) + \frac{1}{m}.$$

Применяя неравенство треугольника, получаем

$$\|u_m\| \le \|u_m - f\| + \|f\| \le E_n(f) + 1 + \|f\|,$$

т. е. последовательность u_m ограничена. Поскольку в конечномерном пространстве F_n из любой ограниченной последовательности можно выделить сходящуюся подпоследовательность, то существует, в частности, под-

последовательность u_{m_k} такая, что

$$\lim_{k\to\infty} u_{m_k} = u_0 \in F_n.$$

Переходя к пределу при $k \to \infty$ в неравенстве

$$E_n(f) \le \|f - u_{m_k}\| \le E_n(f) + \frac{1}{m_k}$$

будем иметь

$$E_n(f) = \|f - u_0\|.$$

Так как $u_0 \in F_n$, элемент $f_n^0 = u_0$ является элементом наилучшего приближения по определению.

Для формулировки теоремы единственности нам потребуется следующее важное определение.

Определение 4.1 *Норма пространства F называется строго выпуклой, если для каждой пары линейно-независимых элементов $f, g \in F$ выполнено строгое неравенство треугольника:*

$$\|f + g\| < \|f\| + \|g\|.$$

Ясно, что строгую выпуклость нормы по-иному можно охарактеризовать следующим свойством (равносильным приведенному определению):

если $\|f + g\| = \|f\| + \|g\|$, то существует число $\lambda \ge 0$ такое, что либо $f = \lambda g$, либо $g = \lambda f$.

Строгая выпуклость нормы оказывается достаточным (хотя и не необходимым) условием единственности элемента наилучшего приближения.

Теорема 4.2 *Пусть F — линейное нормированное пространство со строго выпуклой нормой. Тогда для каждого $f \in F$ элемент наилучшего приближения определяется единственным образом.*

Доказательство. Если $f \in F_n$, то $E_n(f) = 0$ и, очевидно, элемент наилучшего приближения совпадает с f, т. е. определяется единственным образом. Для нетривиального случая докажем единственность от противного. А именно, предположим обратное:

существуют $f \in F \setminus F_n$, $f_n^0 \in F_n$ и $f_n^1 \in F_n$ такие, что $f_n^0 \neq f_n^1$ и

$$E_n(f) = \|f - f_n^0\| = \|f - f_n^1\|.$$

Для среднего арифметического

$$g = \frac{f_n^0 + f_n^1}{2}$$

элементов f_n^0 и f_n^1 имеем: $g \in F_n$ и

$$E_n(f) \leq \|f - g\| = \left\| f - \frac{f_n^0 + f_n^1}{2} \right\| =$$

$$= \frac{\|f - f_n^0 + f - f_n^1\|}{2} \leq$$

$$\leq \frac{\|f - f_n^0\| + \|f - f_n^1\|}{2} = E_n(f).$$

Отсюда следует, что $\|f - g\| = E_n(f)$ и $\|f - f_n^0 + f - f_n^1\| = \|f - f_n^0\| + \|f - f_n^1\|$.

Первое из этих равенств означает, что среднее арифметическое элементов f_n^0 и f_n^1 также является элементом наилучшего приближения. А из второго равенства в силу строгой выпуклости нормы следует, что элементы $f - f_n^0$ и $f - f_n^1$ являются линейно-зависимыми. Следовательно, существует число λ такое, что либо $f - f_n^0 = \lambda(f - f_n^1)$, либо $f - f_n^1 = \lambda(f - f_n^0)$.

Рассмотрим два случая: $\lambda = 1$ и $\lambda \neq 1$. Если $\lambda = 1$, то $f - f_n^0 = f - f_n^1$, т.е. $f_n^0 = f_n^1$, Получили противоречие.

Пусть теперь $\lambda \neq 1$. Тогда $f(1 - \lambda) = f_n^0 - \lambda f_n^1$ или $f(1 - \lambda) = f_n^1 - \lambda f_n^0$. Поделив на $1 - \lambda$, получаем, что $f \in F_n$ как линейная комбинация элементов f_n^0 и f_n^1, что противоречит выбору f.

Этим и завершается доказательство.

Примеры пространств со строго выпуклыми нормами

1) *Норма в любом гильбертовом пространстве является строго выпуклой.*

Доказательство. Если элементы $f, g \in F$ гильбертова пространства F являются линейно-независимыми, то для их скалярного произведения имеет место строгое неравенство Коши $|(f, g)| < \|f\| \cdot \|g\|$. С учетом

этого получаем

$$\|f+g\|^2 = (f+g, f+g) = \|f\|^2 + \|g\|^2 + (f,g) + (g,f) <$$

$$< \|f\|^2 + \|g\|^2 + 2\|f\| \cdot \|g\| = (\|f\| + \|g\|)^2.$$

2) *Для любого* $p \in (1, \infty)$ *строго выпуклую норму имеет пространство Лебега* $L^p_\rho(a,b)$ ($\rho(x) > 0$ п. в. на $[a,b]$) с нормой

$$\|f\| = \left(\int_a^b \rho(x) |f(x)|^p dx \right)^{1/p}.$$

Для случая $p = 2$ это пространство является гильбертовым. Для остальных значений параметра утверждение следует из того, что для линейно-независимых функций известные интегральные неравенства Гельдера и Минковского являются строгими.

Популярные банаховы пространства, нормы в которых не являются строго выпуклыми

1) *Норма пространства* $C[a,b]$ *не является строго выпуклой.*

Достаточно рассмотреть случай, когда $[a,b] = [0,1]$. Возьмем два линейно-независимые элементы этого пространства: $f(x) = 1$ и $g(x) = x$. Имеем

$$\|f+g\|_{C[0,1]} = \max_{x \in [0,1]} (1+x) = 2,$$

$$\|f\|_{C[0,1]} = 1, \quad \|g\| = \max_{x\in[0,1]} x = 1,$$

следовательно,

$$\|f+g\|_{C[0,1]} = \|f\|_{C[0,1]} + \|g\|_{C[0,1]}.$$

2) *Норма пространства Лебега L^1 также не является строго выпуклой.*

Действительно, для любой пары функции $f(x) \geq 0$ и $g(x) \geq 0$ из этого пространства

$$\|f+g\|_{L^1} = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx = \|f\|_{L^1} + \|g\|_{L^1}$$

в силу линейности интеграла. Легко выбрать f и g линейно-независимыми. Можно, например, взять

$$f(x) = 1, \;\; g(x) = x^2.$$

Утверждение распространяется и на случай весовых пространств $L^1_\rho[a,b]$ с нормой

$$\|f\| = \int_a^b \rho(x)|f(x)|dx$$

и с весом $\rho(x) > 0$ почти всюду на $[a,b]$.

4.2 Случай гильбертова пространства

Пусть F — гильбертово пространство, $l_1, l_2, \ldots, l_n$ — система линейно-независимых элементов из F.

Ясно, что для любого $f \in F$ элемент наилучшего приближения f_n^0 существует и определяется единственным образом, так как норма гильбертова пространства является строго выпуклой.

Оказывается, что в случае гильбертова пространства легко вычислить наилучшее приближение $E_n(f)$ и найти явно f_n^0.

Поскольку любая система линейно-независимых элементов $l_1, l_2, \ldots, l_n$ может быть преобразована в ортонормированную применением процесса ортогонализации Грама-Шмидта и этот процесс описывается явными формулами, то нам необходимо в первую очередь рассматривать наилучшие приближения элементами ортонормированной системы.

Теорема 4.3 *Пусть F — гильбертово пространство, система $\{l_1, l_2, \ldots, l_n\} \subset F$ является ортонормированной. Тогда для любого $f \in F$ наилучшее приближение по этой системе определяется формулой*

$$E_n(f) = \sqrt{\|f\|^2 - \sum_{k=1}^{n} |c_k^0|^2},$$

а элемент наилучшего приближения f_n^0 — формулой

$$f_n^0 = \sum_{k=1}^{n} c_k^0 l_k,$$

где числа c_k^0 определяются равенствами $c_k^0 = (f, l_k)$ и называются коэффициентами Фурье.

Доказательство. Пусть F_n — подпространство, натянутое на систему $\{l_1, l_2, \ldots, l_n\} \subset F$. Рассмотрим произвольный элемент

$$f_n = \sum_{k=1}^{n} \alpha_k l_k$$

этого подпространства. Пользуясь определением нормы в гильбертовом пространстве, можем записать

$$\|f - f_n\|^2 = (f - f_n, f - f_n) = (f - \sum_{k=1}^{n} \alpha_k l_k, f - \sum_{k=1}^{n} \alpha_k l_k) =$$

$$= (f, f) - \sum_{k=1}^{n} (f, \alpha_k l_k) - \sum_{k=1}^{n} (\alpha_k l_k, f) + \sum_{k=1}^{n} \sum_{j=1}^{n} (\alpha_k l_k, \alpha_j l_j).$$

Простыми выкладками, с учетом обозначения $(f, l_k) = c_k^0$, получаем

$$\|f - f_n\|^2 = \|f\|^2 - \sum_{k=1}^{n} (\overline{\alpha_k} c_k^0 + \alpha_k \overline{c_k^0}) +$$

$$+\sum_{k=1}^{n}(|\alpha_k|^2+|c_k^0|^2)-\sum_{k=1}^{n}c_k^0\overline{c_k^0}=$$

$$=\|f\|^2-\sum_{k=1}^{n}|c_k^0|^2+\sum_{k=1}^{n}|\alpha_k-c_k^0|^2.$$

Отсюда следует, что

$$\|f-f_n\|^2\geq\|f\|^2-\sum_{k=1}^{n}|c_k^0|^2,$$

причем это неравенство превращается в равенство тогда и только тогда, когда

$$\sum_{k=1}^{n}|\alpha_k-c_k^0|^2=0,$$

т. е. когда $\alpha_k=c_k^0$ для всех $k=1,2,\ldots,n$. В силу произвольности $f_n\in F_n$ немедленно получаем

$$(E_n(f))^2=\|f\|^2-\sum_{k=1}^{n}|c_k^0|^2=\|f-\sum_{k=1}^{n}c_k^0l_k\|^2.$$

Эти равенства показывают, в частности, что элемент

$$f_n^0=\sum_{k=1}^{n}c_k^0l_k$$

является элементом наилучшего приближения.

Теорема доказана.

Теорема 4.4 *Пусть F–гильбертово пространство. Если $l_1, l_2, \ldots, l_n$ линейно-независимы, то элемент наилучшего приближения f_n^0 для любого $f \in F$ определяется по формуле*

$$f_n^0 = \sum_{k=1}^{n} \alpha_k^0 l_k,$$

где α_k^0 $(k = 1, 2, \ldots, n)$ — решение системы уравнений

$$\sum_{k=1}^{n} \alpha_k (l_k, l_j) = (f, l_j), \qquad j = 1, 2, \ldots, n.$$

Доказательство. Применяя процесс ортогонализации Грама-Шмидта, получаем ортонормированную систему $g_1, g_2, \ldots, g_n$.

Ясно, что элементы наилучшего приближения по исходной системе $l_1, l_2, \ldots, l_n$ и по ортонормированной системе $g_1, g_2, \ldots, g_n$ совпадают. Поэтому элемент наилучшего приближения для $f \in F$ по системе $l_1, l_2, \ldots, l_n$ имеет вид

$$f_n^0 = \sum_{k=1}^{n} c_k^0 g_k,$$

где $c_k^0 = (f, g_k)$ — коэффициенты Фурье. Поскольку

$$g_j = \sum_{k=1}^{n} \alpha_{kj} l_k$$

с некоторыми коэффициентами α_{kj}, то элемент наилучшего приближения может быть представлен в виде

$$f_n^0 = \sum_{k=1}^{n} \alpha_k l_k.$$

Равенства $c_k^0 = (f, g_k) = (f_n^0, g_k)$ означают, что элемент $f - f_n^0$ ортогонален всем g_k, а значит и всем l_k. Поэтому $(f - f_n^0, l_k) = 0$ или, что то же самое, $(f, l_k) = (f_n^0, l_k)$ для всех $k = 1, 2, \ldots, n$. Умножая скалярно обе части выражения для f_n^0 на l_j, с учетом равенства $(f, l_j) = (f_n^0, l_j)$ получаем систему линейных алгебраических уравнений

$$\sum_{k=1}^{n} \alpha_k (l_k, l_j) = (f, l_j) \qquad (j = 1, 2, \ldots, n)$$

для определения неизвестных коэффициентов α_k.

В силу существования и единственности элемента наилучшего приближения полученная система должна быть однозначно разрешимой. Итак, определитель этой системы, называемый определителем Грама, отличен от нуля, т. е.

$$\Delta_n = \det((l_k, l_j)) \neq 0.$$

И решение системы имеет вид

$$\alpha_k^0 = \frac{\Delta_n^{(k)}}{\Delta_n},$$

следовательно, мы можем найти элемент наилучшего приближения по формуле

$$f_n^0 = \sum_{k=1}^{n} \frac{\Delta_n^{(k)}}{\Delta_n} l_k.$$

Этим и завершается доказательство теоремы.

4.3 Примеры

Приведем несколько примеров применения доказанных теорем о наилучших применениях.

Пример 1. Наилучшее приближение тригонометрическими полиномами можно построить следующим образом.

В гильбертовом пространстве $F = L^2(0, 2\pi)$ со скалярным произведением

$$(f, g) = \frac{1}{2\pi} \int_0^{2\pi} f(x)\overline{g(x)}dx$$

рассмотрим ортогональную систему

$$\{e^{-irx}, \ldots, e^{-ix}, 1, e^{ix}, \ldots, e^{irx}\}.$$

Элемент наилучшего приближения для любого

$$f \in L^2(0, 2\pi)$$

по указанной системе определяется формулой

$$f_n^0(x) = \sum_{k=-r}^{r} \alpha_k^0 e^{ikx},$$

где

$$\alpha_k^0 = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx} dx.$$

Пример 2. Наилучшее приближение алгебраическими полиномами степени $\leq n$ в пространстве L_ρ^2 с весом ρ ($\rho(x) > 0$ почти всюду на $[a, b]$).

В этом случае естественно рассмотреть систему

$$1, x, x^2, \ldots, x^n.$$

Соответствующая ортонормированная система является системой ортогональных (с весом $\rho(x)$) полиномов

$$P_0(x), P_1(x), \ldots, P_n(x).$$

Элемент наилучшего приближения для любой функции $f \in L_\rho^2(a, b)$ представим в виде

$$f_n^0 = \sum_{k=0}^{n} c_k^0 P_k(x),$$

где

$$c_k^0 = (f, P_k) = \int_a^b \rho(x) f(x) P_k(x) dx.$$

Если система ортогональных полиномом $P_k(x)$ неиз-

вестна, то полином наилучшего приближения ищется в виде

$$f_n^0 = \sum_{k=0}^{n} \alpha_k^0 x^k,$$

неизвестные коэффициенты определяются решением системы линейных алгебраических уравнений

$$\sum_{k=0}^{n} a_{kj}\alpha_k^0 = b_j, \qquad j = 1, 2, \ldots, n,$$

где

$$a_{kj} = \int_a^b \rho(x) x^{k+j} dx, \quad b_j = \int_a^b \rho(x) f(x) x^j dx.$$

Примеры 3.1 и 3.2 (Случай среднеквадратичных приближений на дискретном множестве точек).

На $[a, b]$ возьмем точки $x_1, x_2, \ldots, x_n$ ($x_j \neq x_k$ при $j \neq k$). Рассмотрим определенные на этих узлах функции $f : \{x_1, \ldots, x_n\} \to \mathbb{R}$. Множество всех таких функций образуют конечномерное пространство $F = \{f\}$ со скалярным произведением

$$(f, g) = \sum_{l=1}^{n} f(x_l) g(x_l)$$

и нормой

$$\|f\| = \sqrt{\sum_{l=1}^{n} |f(x_l)|^2}.$$

Далее, в F рассмотрим систему линейно-независимых функций

$$l_1(x), l_2(x), \ldots, l_m(x).$$

Понятно, что должно выполняться неравенство

$$n \geq m.$$

Для любой функции $f \in F$ рассмотрим задачу минимизации квадратичного функционала

$$\Phi(\alpha_1, \alpha_2, \ldots, \alpha_m) = \sum_{l=1}^{n} |f(x_l) - f_m(x_l)|^2$$

на функциях вида

$$f_m(x) = \sum_{k=1}^{m} \alpha_k l_k(x).$$

Такую задачу можно попытаться исследовать методами классического дифференциального исчисления, взяв за отправную точку систему необходимых условий экстремума:

$$\frac{\partial \Phi}{\partial \alpha_j} = 0, \qquad j = 1, \ldots, m.$$

Но нам проще интерпретировать эту задачу как частный случай задачи о наилучшем приближении в

гильбертовом пространстве.

Пример 3.1. Алгебраические полиномы наилучшего среднеквадратичного приближения на дискретном множестве точек получаются так. Для узлов

$$x_1, x_2, \ldots, x_n \in [a, b]$$

и линейно-независимой системы

$$1, x, x^2, \ldots, x^{m-1} \quad (\text{т. е.} \quad l_k(x) = x^{k-1})$$

элемент наилучшего приближения можно представить в виде

$$f_m^0 = \sum_{k=1}^{m} \alpha_k^0 x^{k-1}.$$

Согласно общей теории, неизвестные коэффициенты определяются из системы линейных алгебраических уравнений

$$\sum_{k=1}^{m} \alpha_k (l_k, l_j) = (f, l_j), \qquad j = 1, \ldots, m,$$

где

$$(l_k, l_j) = \sum_{l=1}^{n} x_l^{k+j-2}, \quad (f, l_j) = \sum_{k=1}^{n} f(x_l) x_l^{j-1}.$$

Пример 3.2. Рассмотрим теперь среднеквадратичное приближение тригонометрическими полиномами на

дискретном множестве точек.

Для n узлов

$$x_k = \frac{2k\pi}{n}, \qquad k = 0, \ldots, n-1,$$

рассмотрим пространство функций

$$f \,:\, \{x_l\}_{l=0}^{n-1} \to \mathbb{C}$$

со скалярным произведением

$$(f, g) = \frac{1}{n} \sum_{l=0}^{n-1} f(x_l)\overline{g(x_l)}.$$

Система функций e^{ijx}, $j = 0, 1, \ldots, m-1$ ($n \geq m$) является ортонормированной в этом пространстве. Действительно, имеем

$$(l_k, l_j) = \frac{1}{n} \sum_{l=0}^{n-1} e^{ikx_l} e^{-ijx_l} = \frac{1}{n} \sum_{l=0}^{n-1} e^{i(k-j)\frac{2\pi}{n}l}.$$

Поэтому, если $k = j$, то

$$(l_k, l_k) = \frac{1}{n} \sum_{l=0}^{n-1} \cdot 1 = 1;$$

если же $k \neq j$, то с учетом формул

$$u = e^{i(k-j)\frac{2\pi}{n}} \neq 1, \quad u^n = 1,$$

получаем

$$(l_k, l_j)n = \sum_{l=0}^{n-1} u^l = \frac{u^n - 1}{u - 1} = \frac{e^{2\pi i(k-j)} - 1}{u - 1} = 0.$$

Согласно общей теории элемент наилучшего приближения является отрезком ряда Фурье для заданного элемента f, т. е.

$$f_m^0 = \sum_{k=0}^{m-1} \alpha_k^0 e^{ikx},$$

где

$$\alpha_k^0 = (f, e^{ikx}) = \frac{1}{n} \sum_{l=0}^{n-1} f(x_l) \cdot e^{-ikx_l}.$$

4.4 О чебышевском альтернансе

Рассмотрим подробнее задачу о наилучших приближениях алгебраическими полиномами в банаховом пространстве $C[a, b]$ над полем вещественных чисел. Более точно, для любой функции $f \in C[a, b]$ рассматривается величина — наилучшее приближение f в метрике $C[a, b]$ алгебраическими полиномами степени $\leq n$:

$$E_n(f) = \inf_{P_n} \|f - P_n\|_{C[a,b]},$$

где

$$P_n(x) = a_0 + a_1 x + \ldots + a_n x^n$$

— полиномы степени $\leq n$ с вещественными коэффициентами.

Поскольку $C[a,b]$ — линейное нормированное пространство, то согласно общей теории существует хотя бы один полином наилучшего равномерного приближения, т. е. существует

$$P_n^0(x) = a_0^0 + a_1^0 x + \ldots + a_n^0 x^n$$

такой, что

$$E_n(f) = \|f - P_n^0\|_{C[a,b]}.$$

Норма пространства $C[a,b]$ не является строго выпуклой, поэтому необходим иной подход для доказательства единственности полинома наилучшего равномерного приближения $P_n^0(x)$.

Наилучшие равномерные приближения непрерывных функций алгебраическими полиномами описываются теоремами П.Л. Чебышева. Но прежде всего мы напомним классическую теорему Вейерштрасса.

Теорема 4.5 *Для любой функции $f \in C[a,b]$ и любого $\varepsilon > 0$ существует алгебраический полином $P(x)$ такой, что*

$$\|f - P\|_{C[a,b]} < \varepsilon.$$

Из определения наилучшего приближения непосред-

ственно следует, что $E_n(f) \geq 0$ для любого n и

$$E_0(f) \geq E_1(f) \geq \ldots \geq E_n(f) \geq \ldots \quad (n \geq 1).$$

Легко доказывается и следующее утверждение.

Теорема 4.6 *Для любой функции* $f \in C[a,b]$

$$\lim_{n\to\infty} E_n(f) = 0.$$

Доказательство. Пусть $f \in C[a,b]$, зададимся произвольным $\varepsilon > 0$. По теореме Вейерштрасса существует полином P степени n_0 такой, что $\|f - P\|_{C[a,b]} < \varepsilon$. Следовательно, для всех номеров $n \geq n_0$ с учетом определения наилучшего приближения как инфимума будем иметь

$$E_n(f) \leq E_{n_0}(f) \leq \|f - P\|_{C[a,b]} < \varepsilon.$$

Теорема доказана.

С целью подготовки к пониманию основной теоремы этого параграфа — теоремы о чебышевском альтернансе — рассмотрим задачу нахождения наилучшего приближения в простейших случаях, когда n равно нулю или единице.

Пусть $n = 0$, для непостоянной функции $f \in C[a,b]$ необходимо найти постоянную a_0^0, реализующую следу-

ющий минимум

$$\min_{a_0} \|f - a_0\|_{C[a,b]} = E_0(f).$$

Геометрически очевидно

$$P_0^0(x) = a_0^0 = \frac{M+m}{2}, \qquad E_0(f) = \frac{M-m}{2},$$

где

$$M = \max_{a \le x \le b} f(x) = f(x_1), \quad m = \min_{a \le x \le b} f(x) = f(x_2).$$

Ясно, что существуют по крайней мере 2 различных точки $x_1, x_2 \in [a,b]$ такие, что для остаточного члена $r_0(x) = P_0^0(x) - f(x)$ справедливы равенства

$$r_0(x_1) = -E_0(f), \qquad r_0(x_2) = +E_0(f).$$

Если $n = 1$, то наилучшее приближение

$$E_1(f) = \min_{a_0, a_1} \|f - (a_0 + a_1 x)\|_{C[a,b]}$$

легко определяется геометрически для случая, когда f — выпуклая функция. Имеем

$$P_1^0(x) = a_0^0 + a_1^0 x, \qquad a_1^0 = \frac{f(b) - f(a)}{b - a},$$

а постоянная a_0^0 такова, что для $r_0(x) = P_0^0(x) - f(x)$

справедливы равенства

$$r_0(x_j) = \alpha(-1)^j E_1 f, \qquad \alpha = \pm 1,\ j = 1, 2, 3,$$

где $x_1 = a$, $x_2 \in (a, b)$, $x_3 = b$.

Оказывается верным естественное обобщение этих примеров для любых $n \in \mathbb{N}$: если P_n^0 — полином наилучшего равномерного приближения для $f \in C[a, b]$, то существует не менее $n + 2$ точек

$$x_1 < x_2 < x_3 < \ldots < x_{n+2}, \quad x_k \in [a, b],$$

таких, что

$$P_n^0(x_j) - f(x_j) = \alpha(-1)^j \cdot E_n(f), \qquad j = 1, 2, \ldots, n + 2,$$

где $\alpha = \text{const}$, причем либо $\alpha = 1$, либо $\alpha = -1$.

Теорема 4.7 *(о чебышевском альтернансе) Для любой функции $f \in C[a, b]$ полином $P_n(x)$ степени $\leq n$ является полиномом наилучшего равномерного приближения f тогда и только тогда, когда на $[a, b]$ существует не менее $n + 2$ точек*

$$x_1 < x_2 < x_3 < \ldots < x_{n+2}$$

таких, что для любого $j = 1, 2, \ldots, n + 2$

$$P_n(x_j) - f(x_j) = \alpha(-1)^j \|P_n - f\|_{C[a,b]}, \tag{4.1}$$

где $\alpha = const$, *причем либо* $\alpha = 1$, *либо* $\alpha = -1$.

Доказательство. Необходимость. Пусть $P_n^0(x)$ — полином наилучшего равномерного приближения.

Легко видеть, что для функции

$$r_n(x) = P_n^0(x) - f(x)$$

должны существовать по крайней мере 2 точки x_1 и x_2 такие, что $r_n(x_j) = \alpha(-1)^j \cdot E_n(f)$.

Предположим, что условие альтернанса Чебышева выполняется самое большее на m точках, причем $m \leq n+1$, т. е. на $[a,b]$ существует лишь $m \leq n+1$ точек

$$x_1 < x_2 < x_3 < \ldots < x_m$$

таких, что

$$r_n(x_j) = \alpha(-1)^j E_n(f), \quad j = 1, 2, \ldots, m.$$

Подчеркнем, что число m выбрано максимальным из всех возможных.

Замкнутое множество

$$S = \{x \in [a,b] : \, |r_n(x)| = E_n(f)\}$$

представим в виде

$$S = \bigcup_{j=1}^{m} S_j,$$

где замкнутые множества S_j определены следующим образом:

$$S_1 = \{x \in [a, x_2) : r_n(x) = r_n(x_1)\},$$

$$S_j = \{x \in (x_{j-1}, x_{j+1}) : r_n(x) = r_n(x_j)\},\ 2 \le j \le m-1,$$

$$S_m = \{x \in (x_{m-1}, b] : r_n(x) = r_n(x_m)\}.$$

Легко проверить (с учетом максимальности m), что определения множеств S_j корректны и эти множества не пусты, так как $x_j \in S_j$ и, кроме того,

$$a_{k+1} := \min\{x : x \in S_{k+1}\} > \max\{x : x \in S_k\} =: b_k$$

для всех $k = 1, 2, \ldots, m-1$. Следовательно, существуют точки

$$\xi_1 < \xi_2 < \ldots < \xi_{m-1},$$

удовлетворяющие условиям

$$b_k < \xi_k < a_{k+1}\ (k = 1, 2, \ldots, m-1).$$

Рассмотрим полином

$$s_{m-1}(x) = \lambda(x - \xi_1)(x - \xi_2)\ldots(x - \xi_{m-1}),$$

выбрав знак постоянной λ из условия совпадения знаков $r_n(x_1)$ и $s_{m-1}(x_1)$. Тогда $r_n(x)s_{m-1}(x) > 0$ для лю-

бого $x \in S$, и для достаточно малого $|\lambda| > 0$

$$\|r_n - s_{m-1}\|_{C[a,b]} = \|P_n^0 - s_{m-1} - f\|_{C[a,b]} < E_n(f),$$

а это противоречит тому, что P_n^0 — полином наилучшего равномерного приближения.

Докажем теперь от противного достаточность условия (4.1). Предположим, что P_n удовлетворяет (4.1), но не является полиномом наилучшего равномерного приближения. Возьмем полином наилучшего равномерного приближения P_n^0 и рассмотрим разность

$$q_n(x) = P_n(x) - P_n^0(x).$$

По определению наилучшего приближения

$$\|P_n - f\|_{C[a,b]} > \|P_n^0 - f\|_{C[a,b]} = E_n(f),$$

в частности, во всех узловых точках

$$|P_n(x_j) - f(x_j)| > E_n(f) \geq |P_n^0(x_j) - f(x_j)|.$$

Поэтому значение разности $P_n(x) - P_n^0(x)$, т. е.

$$q_n(x) = [P_n(x) - f(x)] + [f(x) - P_n^0(x)],$$

в любой узловой точке x_j не равно нулю и имеет тот же знак, что и

$$A(x_j) = P_n(x_j) - f(x_j) = \alpha(-1)^j \|P_n - f\|_{C[a,b]}.$$

Таким образом, знаки $q_n(x_j)$ чередуются, следовательно, полином $q_n(x)$ обращается в нуль в некоторых точках $y_1, \ldots, y_{n+1}$ таких, что

$$x_1 < y_1 < x_2 < y_2 < \ldots < y_{n+1} < x_{n+2}.$$

Поскольку $q_n(x)$ является полиномом степени не выше n и обращается в нуль в $n+1$ точке, то $q_n(x) \equiv 0$, т. е. $P_n(x) \equiv P_n^0(x)$. Пришли к противоречию.

Этим и завершается доказательство.

Теорема об альтернансе позволяет установить единственность полинома наилучшего равномерного приближения.

Теорема 4.8 *Для любой функции $f \in C[a,b]$ и любого n полином наилучшего равномерного приближения P_n^0 определяется единственным образом.*

Доказательство. Предположим обратное: пусть имеются два различных полинома наилучшего равномерного приближения $P_n^1(x)$ и $P_n^0(x)$. Тогда для любого $x \in [a,b]$ можем написать неравенства:

$$-E_n(f) \le f(x) - P_n^0(x) \le E_n(f),$$

$$-E_n(f) \le f(x) - P_n^1(x) \le E_n(f).$$

Сложим эти неравенства и поделим на 2. В результате

получим

$$-E_n(f) \le f(x) - \frac{P_n^0(x) + P_n^1(x)}{2} \le E_n(f),$$

следовательно, функция

$$Q(x) = \frac{P_n^0(x) + P_n^1(x)}{2}$$

также является полиномом наилучшего равномерного приближения. По теореме 4.7 о чебышевском альтернансе, примененной к этой функции, на отрезке $[a, b]$ существуют точки

$$x_1 < x_2 < x_3 < \ldots < x_{n+2}$$

такие, что

$$Q(x_j) - f(x_j) = \alpha(-1)^j \|Q - f\| = \alpha(-1)^j E_n(f),$$

где $j = 1, 2, \ldots, n+2$, ($\alpha = 1$, либо $\alpha = -1$). Записав эти равенства в узловых точках в виде

$$2[Q(x_j) - f(x_j)] = P_n^0(x_j) - f(x_j) + P_n^1(x_j) - f(x_j) =$$

$$= 2\alpha(-1)^j \, E_n(f),$$

мы обнаруживаем, что они возможны лишь в том случае, когда

$$P_n^0(x_j) - f(x_j) = P_n^1(x_j) - f(x_j) = \alpha(-1)^j E_n(f).$$

Как следствие получаем, что

$$P_n^0(x_j) = P_n^1(x_j) \quad \text{для} \quad j = 1, 2, \ldots, n+2.$$

Отсюда немедленно следует

$$P_n^0(x) \equiv P_n^1(x),$$

так как степени этих полиномов не превосходят n. Получили противоречие, завершающее доказательство.

Следствие 4.8.1 *Пусть $f \in C[-a, a]$, $a > 0$.*

1) Если f — четная функция, то ее полином наилучшего равномерного приближения P_n^0 также является четным.

2) Если f нечетна, то P_n^0 также нечетный.

Доказательство. Пусть $P_n^0(x)$ — полином наилучшего равномерного приближения $f \in C[-a, a]$.

1) Пусть f — четная функция, т. е. $f(x) = f(-x)$ для любого $x \in [-a; a]$. Тогда для всех $t = -x \in [-a, a]$

$$|P_n^0(-x) - f(x)| =$$

$$= |P_n^0(-x) - f(-x)| = |P_n^0(t) - f(t)| \leq E_n(f).$$

Следовательно, $P_n^0(-x)$ также является полиномом наилучшего равномерного приближения. В силу теоремы единственности

$$P_n^0(-x) = P_n^0(x), \quad \text{для любого} \quad x \in [-a, a].$$

2) Для нечетной функции f имеем

$$| - P_n^0(-x) - f(x)| = | - P_n^0(-x) + f(-x)| =$$

$$= |f(t) - P_n^0(t)| \le E_n(f) \quad \forall x = -t \in [-a, a].$$

Следовательно, $-P_n^0(-x)$ — полином наилучшего равномерного приближения. В силу теоремы единственности получаем

$$-P_n^0(-x) = P_n^0(x).$$

Опишем теперь задачу, показывающую связь полиномов Чебышева первого рода с теоремой о чебышевском альтернансе.

Задача Чебышева. Найти $P_{n-1}^0(x)$ — полином наилучшего равномерного приближения степени $\le n-1$ для функции $f(x) = x^n$, $x \in [-1, 1]$.

Введем в рассмотрение функцию

$$\widetilde{P_n}(x) = \frac{T_n(x)}{2^{n-1}},$$

где $T_n(x) = \cos(n \arccos x)$ — полином Чебышева первого рода. Покажем, что искомый полином определяется по формуле: $P_{n-1}^0(x) = x^n - \widetilde{P_n}(x)$.

Для этого достаточно проверить условие альтернанса Чебышева. Поскольку рассматривается задача для полиномов степени $\le n-1$, это условие должно

выполняться в $n+1$ точке. Пусть

$$x_k = \cos\frac{k\pi}{n}, \qquad k = 0, 1, \ldots, n.$$

Имеем: $x_k^n - P_{n-1}^0(x_k) =$

$$= \frac{T_n(x_k)}{2^{n-1}} = \frac{\cos k\pi}{2^{n-1}} = \frac{(-1)^k}{2^{n-1}} \| \cos(n \arccos x) \|_{C[-1,1]}.$$

Тогда по теореме Чебышева об альтернансе искомый полином наилучшего равномерного приближения дается формулой

$$P_{n-1}^0(x) = x^n - \frac{T_n(x)}{2^{n-1}}.$$

Следствие 4.8.2 *Для любого полинома $P_{n-1}(x)$ степени не выше, чем $n-1$*

$$\|x^n + P_{n-1}(x)\|_{C[-1;1]} \geq \frac{1}{2^{n-1}}.$$

В заключение отметим, что заменой переменной $x = \cos\theta, \quad 0 \leq \theta \leq \pi$, система полиномов Чебышева первого рода

$$\{T_n(x)\}_{n=0}^{\infty}$$

преобразуется в тригонометрическую систему косинусов

$$\{1, \cos\theta, \cos 2\theta, \ldots\}, \quad 0 \leq \theta \leq \pi.$$

С учетом этого легко показать, что $\{T_n(x)\}_{n=0}^{\infty}$ — полная ортогональная система в L_ρ^2 с весовой функцией

$$\rho(x) = \frac{1}{\sqrt{1-x^2}}.$$

Доказательство. Замена переменной $x = \cos\theta$ в интеграле показывает, что ортогональность полиномов Чебышева первого рода

$$\int_{-1}^{1} \frac{T_k(x)T_j(x)}{\sqrt{1-x^2}}dx = 0, \qquad k \neq j,$$

равносильна хорошо известным равенствам

$$\int_0^{\pi} \cos k\theta \cos j\theta \, d\theta = \frac{1}{2}\int_{-\pi}^{\pi} \cos k\theta \cos j\theta \, d\theta = 0, \quad k \neq j.$$

А полнота $\{T_n(x)\}_{n=0}^{\infty}$ вытекает из полноты тригонометрической системы косинусов в пространстве $L^2[0;\pi]$.

Задачи и упражнения

1. Для функции $f(x) = x^3$ постройте полином наилучшего равномерного приближения степени n на отрезке $[0,1]$ для $n = 0, 1, 2, 3$.

2. Для функции $f(x) = x^3$ постройте полином наилучшего приближения первой и второй степени в пространстве $L_2[0,1]$.

3. Пусть функция $f(x) = x^3$ задана в точках $1, 2, 3$. Найдите полином наилучшего среднеквадратичного приближения.

Глава 5

Квадратурные формулы

Для каждой непрерывной функции $f : [a, b] \to \mathbb{R}$ интеграл Римана

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

сколь угодно точно аппроксимируется интегральными суммами вида

$$\sum_{k=1}^{n} f(x_k)\Delta x_k.$$

Но интегральные суммы могут сходиться к значению интеграла очень медленно. Поэтому разработаны оригинальные, эффективные методы численного интегрирования. Важное место среди них занимают классические квадратурные формулы.

Как это принято в теории меры Жордана символом $< a, b >$ мы будем обозначать промежуток от a до b, чтобы охватить одним символом 4 возможных варианта: $[a, b]$, $(a, b]$, $[a, b)$, (a, b).

Пусть $f \in C < a, b >$, заданы узлы

$$x_1, \ldots, x_n \in < a, b > .$$

Будем рассматривать задачу приближенного вычисления интеграла

$$\int_a^b \rho(x) f(x)\, dx,$$

где $\rho = \rho(x)$ — фиксированная весовая функция. Предполагаем, что

$$\rho(x) \in L_1[a, b],\ \rho(x) \geq 0,\ \int_a^b \rho(x)\, dx > 0.$$

Квадратурной принято называть формулу вида

$$\int_a^b \rho(x) f(x)\, dx \approx \sum_{k=1}^{n} A_k f(x_k), \tag{5.1}$$

где A_k — некоторые вещественные числа. Предполагается, что коэффициенты A_k не зависят от f. Точки x_k в формуле (5.1) принято называть узлами.

Определение 5.1 *Пусть M — некоторое семейство функций, непрерывных на промежутке $< a, b >$. Говорят, что квадратурная формула (5.1) точна на множестве M, если для каждой функции $F \in M$*

$$\int_a^b \rho(x) F(x)\, dx = \sum_{k=1}^{n} A_k F(x_k),$$

т. е. приближенное равенство превращается в обычное. В частности, говорят, что квадратурная формула (5.1) точна на множестве алгебраических полиномов степени $\leq m$, если имеют место равенства

$$\int_a^b \rho(x)x^j\,dx = \sum_{k=1}^{n} A_k x_k^j$$

для любого $j = 0, 1, \ldots, m$.

Сам термин "квадратура" восходит к древнегреческой цивилизации. А именно, античными математиками был поставлен вопрос о квадратуре круга (т. е. вопрос о возможности построения с помощью линейки и циркуля квадрата, равновеликого кругу по площади). А вычисление площадей, как вы хорошо знаете, равносильно интегрированию подходящих функций.

Простейшие квадратурные формулы для вычисления интегралов создавались и использовались уже во времена Ньютона и Лейбница.

Прием, лежащий в основе всех классических квадратурных формул, состоит в замене подынтегральной функции некоторым ее приближением (например, интерполяционным полиномом или сплайном).

5.1 Применение полиномов Лагранжа

Пусть $f \in C < a, b >$, рассмотрим интерполяционный полином Лагранжа $L_n(f; x)$, построенный по сетке узлов $\{x_1, x_2, \ldots, x_n\} \subset < a, b >$. Заменяя подынтегральную функцию ее интерполяционным полиномом в форме Лагранжа, получаем приближенную формулу

$$\int_a^b \rho(x) f(x)\, dx \approx \int_a^b \rho(x) L_n(f; x)\, dx =$$

$$= \int_a^b \rho(x) \sum_{k=1}^{n} f(x_k) l_k(x)\, dx = \sum_{k=1}^{n} p_k f(x_k),$$

где

$$p_k = \int_a^b \rho(x) l_k(x)\, dx,$$

или, что то же самое,

$$p_k = \int_a^b \rho(x) \frac{\omega_n(x)}{(x - x_k)\omega_n'(x_k)}\, dx,$$

где

$$\omega_n(x) = (x - x_1)(x - x_2) \ldots (x - x_n).$$

Полученная таким образом квадратурная формула

$$\int_a^b \rho(x) f(x)\, dx \approx \sum_{k=1}^{n} p_k f(x_k)$$

называется интерполяционной квадратурной формулой.

Теорема 5.1 *Квадратурная формула (5.1) с коэффициентами A_k является точной для любого алгебраического полинома степени $\leq n-1$ тогда и только тогда, когда она совпадает с интерполяционной квадратурной формулой, т. е. когда имеет место равенство $A_k = p_k$ для всех $k = 1, 2, \ldots, n$.*

Доказательство. Предположим, что (5.1) точна для каждого полинома степени $\leq n-1$. Тогда эта формула должна быть точной для всех фундаментальных полиномов Лагранжа $l_j(x)$, поскольку они являются полиномами степени $n-1$. Таким образом, для всех $j = 1, \ldots, n$, должны выполняться равенства

$$\int_a^b \rho(x) l_j(x)\, dx = \sum_{k=1}^{n} A_k l_j(x_k) = \sum_{k=1}^{n} A_k \delta_{kj} = A_j.$$

С другой стороны,

$$p_j = \int_a^b \rho(x) l_j(x)\, dx$$

по определению интерполяционной квадратурной формулы. Следовательно, $A_j = p_j$ для всех $j = 1, \ldots, n$.

Обратное утверждение о том, что интерполяционная квадратурная формула является точной для каждого полинома степени $\leq n-1$, является тривиальным. Действительно, если F — полином степени $\leq n-1$, то

$L_n(F;x) \equiv F(x)$, поэтому

$$\int_a^b \rho(x)\,F(x)\,dx = \int_a^b \rho(x)\,L_n(F;x)\,dx =$$

$$= \int_a^b \rho(x) \sum_{k=1}^{n} F(x_k)\,l_k(x)\,dx =$$

$$= \sum_{k=1}^{n} F(x_k) \int_a^b \rho(x)\,l_k(x)\,dx = \sum_{k=1}^{n} p_k\,F(x_k).$$

Теорема доказана.

Погрешность интерполяционной квадратурной формулы

$$R_n(f) = \int_a^b \rho(x) f(x)\,dx - \sum_{k=1}^{n} p_k f(x_k)$$

может быть эффективно оценена для $f \in C^n[a,b]$, где n — число узлов сетки.

Теорема 5.2 *Пусть* $\omega_n(x) = (x-x_1)\dots(x-x_n)$, *и пусть* $n \geq 1$. *Если* $f \in C^n[a,b]$, *то существует точка* $\eta \in [a,b]$ *такая, что для погрешности интерполяционной квадратурной формулы справедлива оценка*

$$|R_n(f)| \leq \frac{|f^{(n)}(\eta)|}{n!} \int_a^b \rho(x)|\omega_n(x)|\,dx.$$

А в частном случае, когда $n = 2$, $x_1 = a$, $x_2 = b$,

имеет место равенство

$$R_2(f) = \frac{f''(\eta)}{2} \int_a^b \rho(x)\omega_2(x)\,dx.$$

Доказательство. Имеем

$$R_n(f) = \int_a^b \rho(x)[f(x) - L_n(f;x)]dx = \int_a^b \rho(x) r_n(x)\,dx.$$

Как было установлено для остаточного члена интерполяции, существует точка $\xi = \xi(x) \in (a,b)$ такая, что

$$r_n(x) = \frac{f^{(n)}(\xi(x))}{n!}\omega_n(x).$$

Следовательно,

$$R_n(f) = \frac{1}{n!} \int_a^b \rho(x)\ f^{(n)}(\xi(x))\ \omega_n(x)\,dx.$$

Отсюда получаем

$$|R_n(f)| \le \frac{1}{n!} \int_a^b \rho(x)\ |f^{(n)}(\xi(x))|\ |\omega_n(x)|\,dx.$$

Утверждение теоремы получается теперь по теореме о среднем для интегралов с учетом непрерывности $f^{(n)}$. В частном случае мы пользуемся знакопостоянством $\omega_2(x) = (x-a)(x-b)$ и применяем теорему о среднем до перехода к абсолютным величинам. Этим и завершается доказательство теоремы.

Для равномерной сетки шага $h = (b-a)/n$ с узлами

$$a = x_0,\ x_1 = a + h,\ \ldots,\ x_n = a + nh = b$$

интерполяционную квадратурную формулу принято называть **формулой Ньютона-Котеса**.

Поскольку число узлов равно $(n+1)$, то в этом случае интерполяционная квадратурная формула имеет вид

$$\int_a^b \rho(x)\, f(x)\, dx \approx \sum_{k=0}^{n} c_k\ f(a + kh),$$

где

$$c_k = \int_a^b \rho(x) \frac{\omega_{n+1}(x)}{(x - x_k)\omega'_{n+1}(x_k)}\, dx,$$
$$\omega_{n+1}(x) = (x - x_0)(x - x_1)\ldots(x - x_n).$$

В этих формулах сделаем замену $x = a + ht$, $0 \le t \le n$. Простые вычисления приводят к следующей универсальной формуле

$$c_k = \frac{(-1)^{n-k} h}{k!(n-k)!} \int_0^n \rho(a + ht) \frac{t(t-1)\ldots(t-n)}{t-k}\, dt.$$

Отметим важное свойство коэффициентов c_k в формуле Ньютона-Котеса. Поскольку квадратурная формула Ньютона-Котеса точна для функции $f(x) \equiv 1$, то

имеем

$$\sum_{k=0}^{n} c_k = \int_a^b \rho(x)\,dx.$$

Отсюда следует, что если все коэффициенты $c_k \geq 0$, то все они ограничены числом, не зависящим от n, поэтому погрешность квадратурной формулы не превосходит по порядку погрешности при вычислении функции. Такая устойчивость в вычислениях может быть нарушена, если коэффициенты c_k имеют разные знаки, так как оценка погрешности зависит от суммы

$$\sum_{k=0}^{n} |c_k|,$$

а эта сумма может неограниченно возрастать с ростом числа n.

5.2 Базовые квадратурные формулы

На практике поступают следующим образом: разбивают промежуток интегрирования на несколько частичных промежутков и на каждом из них применяют интерполяционную квадратурную формулу с небольшим числом узлов.

Получаемые на этом пути формулы называются **составными квадратурными формулами**.

Отметим, что популярные приближенные *форму-*

лы прямоугольников и трапеций, а также формула Симпсона являются составными квадратурными формулами.

5.2.1 Формула трапеций

Рассмотрим сначала малую формулу трапеций для приближенного вычисления интеграла $\int_a^b f(x)\,dx$. Применим формулу Ньютона-Котеса на отрезке $[a,b]$ для сетки с узлами $x_0 = a,\ x_1 = b$. Имеем

$$\int_a^b f(x)\,dx \approx c_0 f(a) + c_1 f(b),$$

где

$$c_0 = \int_a^b \frac{x-b}{a-b}\,dx = \frac{b-a}{2},\ c_1 = \int_a^b \frac{a-x}{b-a}\,dx = \frac{b-a}{2}.$$

Получаем приближенную формулу

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2}\ [f(a)+f(b)],$$

которую принято называть **малой формулой трапеций**, так как правая часть представляет собой площадь трапеции с вершинами в точках $(a,0)$, $(b,0)$, $(a,f(a))$, $(b,f(b))$. **Большая формула трапеций** строится так: сегмент $[a,b]$ делим на $n \geq 2$ равных частей точками

$$a = x_0,\ x_1 = a+h,\ \ldots,\quad x_k = a+kh,\ \ldots,\ x_n = b,$$

и представляем искомый интеграл в виде суммы:

$$\int_a^b f(x)\,dx = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} f(x)\,dx.$$

Применяя на каждом частичном отрезке малую формулу трапеций, находим

$$\int_a^b f(x)\,dx \approx \sum_{k=1}^{n} \frac{x_k - x_{k-1}}{2}[f(x_{k-1} + f(x_k)] =$$

$$= \frac{b-a}{n}\left[\frac{f(a)+f(b)}{2} + f(x_1) + f(x_2) + \ldots + f(x_{n-1})\right].$$

Обозначив

$$f_k = f(x_k), \quad h = (b-a)/n,$$

мы можем записать **большую формулу трапеций** в традиционной форме

$$\int_a^b f(x)\,dx \approx \frac{b-a}{n}\left[\frac{f_0+f_n}{2} + f_1 + f_2 + \ldots + f_{n-1}\right].$$

Из общей теоремы об оценке погрешности интерполяционных квадратурных формул вытекает следующее утверждение: если $f \in C^2[a,b]$, то для малой формулы трапеций существует точка $\eta \in [a,b]$ такая, что погреш-

ность

$$R_2(f) = \int_a^b f(x)\,dx - \frac{b-a}{2}[f(a)+f(b)]$$

определяется формулой

$$R_2(f) = \frac{f''(\eta)}{2}\int_a^b (x-a)(x-b)dx = -\frac{f''(\eta)}{12}(b-a)^3.$$

Оценка погрешности $R_n(f) =$

$$= \int_a^b f(x)\,dx - \frac{b-a}{n}\left[\frac{f_0+f_n}{2} + f_1 + f_2 + \ldots + f_{n-1}\right]$$

большой формулы трапеций дается в следующей теореме.

Теорема 5.3 *Если $f \in C^2[a,b]$, то существует точка $\eta \in [a,b]$ такая, что погрешность большой формулы трапеций равна*

$$R_n(f) = -\frac{(b-a)^3}{12n^2}f''(\eta) = O\left(\frac{1}{n^2}\right).$$

Доказательство. Для произвольного частичного отрезка погрешность малой формулы трапеций определяется формулой

$$-\frac{f''(\eta_k)}{12}\,(x_k - x_{k-1})^3, \;\; \eta_k \in [x_{k-1}, x_k].$$

Поэтому

$$R_n(f) = \sum_{k=1}^{n} \left[-\frac{f''(\eta_k)}{12} \, (x_k - x_{k-1})^3 \right] =$$

$$= -\frac{(b-a)^3}{12n^2} \cdot \left(\frac{1}{n} \sum_{k=1}^{n} f''(\eta_k) \right).$$

Среднее арифметическое чисел $f''(\eta_k)$ лежит между минимальным и максимальным значениями второй производной. Отсюда следует, что

$$\frac{1}{n} \sum f''(\eta_k) = f''(\eta)$$

для некоторой точки $\eta \in [a, b]$. Этим и завершается доказательство.

Следующий простой пример явно показывает невозможность дальнейшего повышения порядка погрешности $O(1/n^2)$ для формулы трапеций за счет повышения порядка гладкости интегрируемой функции.

Пример. Рассмотрим на отрезке $[0, 1]$ сколь угодно гладкую функцию $f(x) = x^2$. Возьмем сетку узлов $x_k = kh, \ k = 0, 1, \ldots, n$, с шагом $h = 1/n$. Пользуясь известной формулой

$$1^2 + 2^2 + 3^2 + \ldots + (n-1)^2 = \frac{(n-1)n(2n-1)}{6},$$

легко вычисляем погрешность формулы трапеций для интеграла $\int_0^1 x^2\,dx$:

$$\int_0^1 x^2\,dx - \frac{1}{n}\left[\frac{1}{2} + \frac{1}{n^2} + \frac{2^2}{n^2} + \ldots + \frac{(n-1)^2}{n^2}\right] = -\frac{1}{6n^2}.$$

Можно получить оценки погрешности формулы трапеций и в случае, когда на функцию накладываются менее жесткие ограничения, чем $f \in C^2[a,b]$. Для этого удобнее пользоваться иной трактовкой большой формулы трапеций, а именно, геометрически очевидной формулой

$$\int_a^b f(x)\,dx \approx \int_a^b S_n^1(f;x)\,dx,$$

где $S_n^1(f;x)$ — сплайн 1-ой степени. Тогда погрешность формулы трапеций определяется равенством

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

где $r_n(x) = f(x) - S_n^1(f;x)$. Понятно, что оценки погрешности $R_n(f)$ без труда следуют из известных неравенств для $r_n(x)$.

Поскольку в формуле трапеций используется равномерная сетка, то диаметр разбиения равен шагу сетки, т. е. $\delta_n = h = (b-a)/n$. Опишем кратко несколько новых *оценок погрешности $R_n(f)$ для формулы трапеций.*

1) Пусть $f \in C[a,b]$, тогда

$$|r_n(x)| \le \omega\left(f, \frac{b-a}{n}\right),$$

поэтому погрешность формулы трапеций можно оценить следующим образом

$$|R_n(f)| \le \omega\left(f, \frac{b-a}{n}\right) \cdot \int_a^b dx \le (b-a)\ \omega\left(f, \frac{b-a}{n}\right).$$

В частности, если $\alpha \in (0,1]$, функция $f \in \operatorname{Lip}\alpha$ с постоянной M, то погрешность формулы трапеций можно оценить так:

$$|R_n(f)| \le \frac{M(b-a)^{1+\alpha}}{n^\alpha} = O\left(\frac{1}{n^\alpha}\right).$$

2) Пусть $\alpha \in (0,1]$, функция $f' \in \operatorname{Lip}\alpha$ с постоянной M_1. Тогда с учетом неравенства

$$|r_n(x)| \le \frac{b-a}{4n}\omega\left(f'; \frac{b-a}{n}\right),$$

получаем следующую оценку для погрешности формулы трапеций:

$$|R_n(f)| \le \frac{b-a}{4n} \cdot \frac{M_1(b-a)^\alpha}{n^\alpha} \cdot (b-a) =$$

$$= \frac{(b-a)^{2+\alpha} M_1}{4n^{1+\alpha}} = O\left(\frac{1}{n^{1+\alpha}}\right).$$

5.2.2 Формулы прямоугольников

Малая формула прямоугольников для $f \in C[a,b]$ имеет вид: $\int_a^b f(x)\,dx \approx (b-a)f(\xi), \quad \xi \in [a,b]$. Она получается из интерполяционной квадратурной формулы в случае, когда полином Лагранжа строится для одной узловой точки $x_1 = \xi$. Наиболее употребительными являются три случая, когда $\xi = a$ или $\xi = b$, т. е. берутся левый или правый концы промежутка интегрирования, или же

$$\xi = c = (a+b)/2,$$

т. е. выбирается средняя точка.

Таким образом, принято различать три различных малых формул прямоугольников. А именно, рассматривают малые формулы левых прямоугольников

$$\int_a^b f(x)\,dx \approx (b-a)f(a)$$

или правых прямоугольников

$$\int_a^b f(x)\,dx \approx (b-a)f(b),$$

а также малую формулу средних прямоугольников

$$\int_a^b f(x)\,dx \approx (b-a)f(c).$$

Пусть

$$h = \frac{b-a}{n}, \quad n \geq 2, \quad x_k = a + kh, \quad k = 0, \ldots, n,$$

и обозначим

$$f(x_k) = f_k, \quad f\left(\frac{x_k + x_{k-1}}{2}\right) = f_{k-\frac{1}{2}}.$$

Большие формулы прямоугольников получаем как составные

$$\int_a^b f(x)\,dx = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} f(x)\,dx,$$

т. е. получаем суммированием малых формул для частичных отрезков. Таким образом возникают большая формула левых прямоугольников

$$\int_a^b f(x)\,dx \approx \frac{b-a}{n}[f_0 + f_1 + \ldots + f_{n-1}],$$

большая формула правых прямоугольников

$$\int_a^b f(x)\,dx \approx \frac{b-a}{n}[f_1 + f_2 + \ldots + f_n],$$

и наконец, большая формула средних прямоугольников

$$\int_a^b f(x)\,dx \approx \frac{b-a}{n}[f_{1/2} + f_{3/2} + \ldots + f_{n-1/2}].$$

Правые части во всех трех формулах прямоуголь-

ников представляют собой интегральную сумму, поэтому мы можем утверждать следующее: *если f интегрируема в смысле Римана на отрезке $[a, b]$, то погрешность приближения для всех трех формул прямоугольников стремится к нулю при $n \to \infty$.*

Зная модуль непрерывности подынтегральной функции, мы можем получить порядковые оценки погрешности $R_n(f)$ для формул прямоугольников.

Теорема 5.4 *Если $f \in C^1[a, b]$ или даже $f \in Lip\,1$, то*

$$R_n(f) = O\left(\frac{1}{n}\right)$$

для всех трех формул прямоугольников.

Доказательство. Имеем

$$R_n(f) = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} [f(x) - f(\xi_k)]dx,$$

где

$$\xi_k = \begin{cases} x_{k-1} & \text{для случая левых прямоугольников,} \\ x_k & \text{для случая правых прямоугольников,} \\ x_{k-1/2} & \text{для случая средних прямоугольников.} \end{cases}$$

Имеет место неравенство

$$|f(x) - f(\xi_k)| \le \omega\left(f; \frac{b-a}{n}\right)$$

для каждого $x \in [x_{k-1}, x_k]$, поэтому

$$|R_n(f)| \le \sum_{k=1}^{n} \omega\left(f; \frac{b-a}{n}\right) \int_{x_{k-1}}^{x_k} dx =$$

$$= \omega\left(f; \frac{b-a}{n}\right)(b-a).$$

Отсюда легко следует утверждение теоремы.

Как показывает следующий пример, для формул левых или правых прямоугольников усилить эту теорему невозможно.

Пример. Рассмотрим функцию $f(x) = x$ на отрезке $[0, 1]$. Точное значение интеграла $\int_0^1 x\,dx$ равно $1/2$, приближенное значение по формуле левых прямоугольников

$$\frac{1}{n}\sum_{k=1}^{n} f(x_{k-1}) = \frac{1}{n}\left(\frac{1}{n} + \frac{2}{n} + \ldots + \frac{n-1}{n}\right) = \frac{1}{2} - \frac{1}{2n},$$

и по формуле правых прямоугольников

$$\frac{1}{n}\sum_{k=1}^{n} f(x_k) = \frac{1}{n}\left(\frac{1}{n} + \frac{2}{n} + \ldots + \frac{n}{n}\right) = \frac{1}{2} + \frac{1}{2n}.$$

Если $f \in C^2[a, b]$, то погрешность для формулы средних прямоугольников допускает улучшенную оценку. В случае $f \in C^2[a, b]$ для формулы средних прямоугольников справедлив удивительный факт: оценка погрешности по порядку оказывается такой же, какой она яв-

ляется для формулы трапеций.

Теорема 5.5 *Если $f \in C^2[a,b]$, то погрешность для формулы средних прямоугольников можно оценить следующим образом: существует точка η такая, что*

$$R_n(f) = \frac{f''(\eta)}{24n^2}(b-a)^3 = O\left(\frac{1}{n^2}\right).$$

Доказательство. Рассмотрим сначала случай малой формулы средних прямоугольников. Имеем

$$R_1(f) = \int_a^b f(x)\,dx - f(c)(b-a) = \int_a^b [f(x) - f(c)]dx.$$

Поскольку $f \in C^2[a,b]$, то существует $\xi = \xi(x) \in (a,b)$ такая, что

$$f(x) = f(c) + \frac{f'(c)}{1!}(x-c) + \frac{f''(\xi)}{2!}(x-c)^2.$$

Интегрируя и применяя теорему о среднем, получаем

$$R_1(f) = f'(c)\int_a^b (x-c)dx + \frac{f''(\eta)}{2!}\int_a^b (x-c)^2 dx =$$

$$= \frac{f''(\eta)}{24}(b-a)^3.$$

Эффект средней точки проявился на этом этапе тем, что $\int_a^b (x-c)\,dx = 0$. Общий случай получается суммированием и применением стандартных рассуждений об

арифметических средних по цепочке равенств:

$$R_n(f) = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} [f(x) - f(x_{k-1/2})]dx =$$

$$= \frac{(b-a)^3}{24n^2} \left[\frac{1}{n} \sum_{k=1}^{n} f''(\eta_k) \right] = \frac{(b-a)^3}{24n^2} f''(\eta_{cp}) = O\left(\frac{1}{n^2}\right).$$

Таким образом, теорема доказана.

5.2.3 Формула Симпсона

Стандартный путь построения формулы Симпсона состоит в замене подынтегральных функций параболическими сплайнами, т. е. сплайнами второй степени.

Малая квадратурная формула Симпсона для функции $f \in C[a,b]$ строится по трем узлам:

$$x_1 = a, \quad x_2 = c := \frac{a+b}{2}, \quad x_3 = \ b.$$

Пусть $L_3(f;x)$ — интерполяционный полином Лагранжа. Будем искать его в форме Ньютона

$$L_3(f;x) = A + B(x-a) + C(x-a)(x-b).$$

Приближенное равенство

$$\int_a^b f(x)\,dx \approx \int_a^b L_3(f;x)\,dx$$

будем называть малой формулой Симпсона.

Имеем: $A = f(a)$ в силу условия $L_3(f;a) = f(a)$. Далее, равенство $L_3(f;b) = f(b)$ приводит к уравнению $f(b) = f(a) + B(b-a)$ для определения B, а затем из условия $L_3(f;c) = f(c)$ можно найти постоянную C. Непосредственные вычисления (мы их пропускаем) коэффициентов A, B, C и суммы интегралов

$$A\int_a^b dx + B\int_a^b (x-a)\,dx + C\int_a^b (x-a)(x-b)\,dx$$

приводят к малой формуле Симпсона в привычной форме:

$$\int_a^b f(x)\,dx \approx \frac{b-a}{6}\left[f(a) + 4f(c) + f(b)\right].$$

Оценим теперь погрешность малой формулы Симпсона при условии $f \in C^3[a,b]$. Пользуясь оценкой погрешности интерполяционных квадратурных формул, получаем

$$|R_3(x)| \le \frac{|f'''(\eta)|}{3!}\int_a^b |(x-a)(x-b)(x-c)|dx =$$

$$= \frac{|f'''(\eta)|}{192}(b-a)^4,$$

где $\eta \in [a,b]$.

Большая формула Симпсона составляется из

малых. Полагаем

$$h = \frac{b-a}{n}, \quad n \geq 2, \quad x_k = a + kh, \quad k = 0, 1, \ldots, n,$$

записываем равенство

$$\int_a^b f(x)\,dx = \sum_{k=1}^{n} \int_{k-1}^{x_k} f(x)\,dx$$

и применяем малую формулу Симпсона на каждом из частичных отрезков. Так как $x_k - x_{k-1} = (b-a)/n$, то итоговая формула $\int_a^b f(x)\,dx \approx$

$$\approx \sum_{k=1}^{n} \frac{x_k - x_{k-1}}{6} \left[f(x_{k-1}) + 4f\left(\frac{x_{k-1} - x_k}{2}\right) + f(x_k) \right]$$

с учетом обозначений

$$f_k = f(x_k), \quad f_{k-\frac{1}{2}} = f\left(\frac{x_{k-1} + x_k}{2}\right)$$

имеет вид

$$\frac{3n}{b-a} \int_a^b f(x)\,dx \approx$$

$$\approx \frac{f_0 + f_n}{2} + 2(f_{1/2} + \ldots + f_{n-1/2}) + f_1 + \ldots + f_{n-1}.$$

Это и есть классическая формула Симпсона.

Суммируя погрешности малых формул для частичных отрезков, получаем погрешность большой форму-

лы Симпсона: существуют точки

$$\eta_k \in [x_{k-1}, x_k], \quad \eta \in [a, b]$$

такие, что

$$|R_n(f)| \le \frac{(b-a)^4}{192n^3} \cdot \left(\frac{1}{n} \sum_{k=1}^{n} |f'''(\eta_k)| \right) =$$

$$= \frac{(b-a)^4}{192n^3} |f'''(\eta)| = O\left(\frac{1}{n^3} \right).$$

Формуле Симпсона можно дать другую интерпретацию, позволяющую получить наилучшие оценки погрешности для этой квадратурной формулы при условии $f \in C^4[a, b]$. Снова рассмотрим узлы

$$x_1 = a, \quad x_2 = c := \frac{a+b}{2}, \quad x_3 = \ b.$$

Пусть $H_4(f; x)$ — интерполяционный полином Эрмита, построенный по условиям:

$$H_4(f; a) = f(a),\ H_4(f; b) = f(b),$$

$$H_4(f; c) = f(c),\ H_4'(f; c) = f'(c).$$

Существует единственный интерполяционный полином $H_4(f; x)$ степени ≤ 3 вида: $H_4(f; x) =$

$$= A + B(x-a) + C(x-a)(x-b) + D(x-a)(x-b)(x-c) =$$

$$= L_3(f;x) + D(x-a)(x-b)(x-c).$$

Несмотря на то, что $H_4(f;x) \not\equiv L_3(f;x)$ при $D \neq 0$, приближенное равенство

$$\int_a^b f(x)\,dx \approx \int_a^b H_3(f;x)\,dx$$

совпадает с малой формулой Симпсона при любом D, поскольку $c = (a+b)/2$ и как следствие

$$\int_a^b (x-a)(x-b)\left(x - \frac{a+b}{2}\right)dx = 0.$$

Оценим теперь погрешность малой формулы Симпсона при условии $f \in C^4[a,b]$. Как было показано при рассмотрении интерполяционных полиномов Эрмита, существует $\xi \in (a,b)$ такая, что

$$r(x) = f(x) - H_4(f;x) = \frac{f^{(4)}(\xi)}{4!}(x-a)(x-b)(x-c)^2.$$

Поэтому погрешность малой формулы Симпсона представима в виде

$$R_3(x) = \int_a^b r(x)\,dx =$$

$$= \frac{f^{(4)}(\eta)}{4!}\int_a^b (x-a)(x-b)(x-c)^2 dx = -\frac{f^{(4)}(\eta)}{2880}(b-a)^5,$$

где $\eta \in [a,b]$, а постоянная 2880 — результат вычисле-

ния произведения $6! \cdot 4$. Суммируя погрешности малых формул для частичных отрезков, получаем погрешность большой формулы Симпсона: если $f \in C^4[a,b]$, то существуют точки $\eta_k \in [x_{k-1}, x_k]$ и $\eta \in [a,b]$ такие, что

$$R_n(f) = -\frac{(b-a)^5}{2880n^4} \cdot \left(\frac{1}{n} \sum_{k=1}^{n} f^{(4)}(\eta_k) \right) =$$

$$= -\frac{(b-a)^5}{2880n^4} f^{(4)}(\eta) = O\left(\frac{1}{n^4}\right).$$

Обратим внимание на одно из важных следствий этой оценки погрешности: *формулы Симпсона точны для любого полинома степени, меньшей или равной трем.* Действительно, четвертая производная полинома степени ≤ 3 тождественно равна нулю, а значит, равна нулю и погрешность формулы Симпсона для него.

В заключение отметим, что имеет место утверждение. *Пусть* $r = 1, 2, 3$ *или* 4 *и* $f \in C^r[a,b]$. *Тогда для погрешности формулы Симпсона справедлива формула*

$$R_n(f) = O\left(\frac{1}{n^r}\right).$$

При $r = 3$ или 4 этот факт доказан выше. При $r = 1$ или 2 такое утверждение было доказано выше для формулы трапеций и для формулы средних прямоугольников, что влечет доказываемое утверждение и для формулы Симпсона в силу следующего наблюдения.

Напомним сначала формулы трапеций и средних прямоугольников: $\int_a^b f(x)\,dx \approx$

$$\approx \Phi_t(f;n) := \frac{b-a}{n}\left[\frac{f_0+f_n}{2} + f_1 + f_2 + \ldots + f_{n-1}\right],$$

$$\int_a^b f(x)\,dx \approx \Phi_{mr}(f;n) := \frac{b-a}{n}[f_{1/2} + f_{3/2} \ldots + f_{n-1/2}].$$

Очевидно, формула Симпсона имеет вид

$$\int_a^b f(x)\,dx \approx \frac{\Phi_t(f;n) + 2\Phi_{mr}(f;n)}{3}.$$

Это дает нам возможность легко запомнить формулу Симпсона: *если взять треть формулы трапеций, две третьих формулы средних прямоугольников и просуммировать, то получим формулу Симпсона.* Аналогичное утверждение верно и для погрешности формулы Симпсона. В частности, *для* $f \in C^2[a,b]$ *погрешность формулы Симпсона можно оценить следующим образом: существуют точки* $\eta_1, \eta_2 \in [a,b]$ *такие, что*

$$R_n(f) = \frac{-f''(\eta_1) + f''(\eta_2)}{36n^2}(b-a)^3 = O\left(\frac{1}{n^2}\right).$$

5.3 Квадратурные формулы Гаусса

До сих пор мы рассматривали квадратурные формулы с произвольными узлами. При любом выборе узлов

интерполяционная квадратурная формула

$$\int_a^b \rho(x) f(x)\, dx \approx \sum_{k=1}^{n} A_k\, f(x_k) \tag{5.2}$$

с коэффициентами

$$A_k = p_k := \int_a^b \rho(x) \frac{\omega_n(x)\, dx}{(x - x_k)\omega_n'(x_k)} \tag{5.3}$$

является точной для полиномов степени не выше $n-1$. Гаусс предложил выбирать узлы x_k специальным образом, чтобы эта формула оказалась точной на полиномах наибольшей степени. Он доказал, что интерполяционная квадратурная формула

$$\int_{-1}^{1} f(x)\, dx \approx \sum_{k=1}^{n} p_k\, f(x_k)$$

будет точной для любого полинома степени не выше $2n-1$, если узлы $x_k \in [-1, 1]$ являются нулями полинома Лежандра степени n. Оказалось, что идея Гаусса легко распространяется и на общий случай, т. е. узлы можно выбрать таким образом, что

$$\int_a^b \rho(x) x^m dx = \sum_{k=1}^{n} p_k\ x_k^m \tag{5.4}$$

для любого $m = 0, 1, \ldots, 2n-1$.

Интерполяционные квадратурные формулы, точ-

ные на полиномах степени не выше $2n-1$, называются квадратурными формулами Гаусса или **квадратурными формулами наивысшего алгебраического порядка точности.** Слово "наивысшего" здесь не является случайным, так как

ни при каком выборе узлов $x_1, \ldots, x_n$ и коэффициентов A_k квадратурная формула вида (5.2) не может быть точной для всех полиномов степени $2n$.

Доказательство. Предположим противное. Если существует квадратурная формула вида (5.2), точная на полиномах степени $2n$, то для функции

$$f(x) = \omega_n^2(x), \qquad \omega_n(x) = \prod_{j=1}^{n}(x - x_j),$$

являющейся полиномом степени $2n$, мы получаем противоречивое соотношение

$$0 < \int_a^b \rho(x)\omega_n^2(x)\,dx = \sum_{k=1}^{n} A_k \omega_n^2(x_k) = 0.$$

5.3.1 Структура формул Гаусса

Полиномы P и Q будем называть ортогональными с весом $\rho(x)$, если

$$\int_a^b \rho(x)P(x)Q(x)dx = 0.$$

Напомним: всюду в дальнейшем предполагаем, что весовая функция является интегрируемой и удовлетворяет условиям

$$\rho(x) \geq 0, \quad \int_a^b \rho(x)\,dx > 0.$$

Теорема 5.6 *Квадратурная формула (5.2) точна для любого полинома степени* $\leq 2n-1$ *тогда и только тогда, когда выполняются следующие два условия:*

1) $\omega_n(x) = \prod_{k=1}^n (x - x_k)$ *ортогонален с весом* $\rho(x)$ *любому полиному* $q(x)$ *степени* $\leq n-1$*, т. е.*

$$\int_a^b \rho(x)\omega_n(x)q(x)dx = 0;$$

2) квадратурная формула является интерполяционной, т. е. ее коэффициенты A_k *выражаются формулой (5.3).*

Доказательство. Необходимость. Пусть квадратурная формула является точной для любого полинома степени $\leq 2n-1$. Поскольку $2n-1 \geq n-1$, то формула должна быть интерполяционной, следовательно, условие 2) выполнено.

Проверим условие 1). Возьмем полином $q(x)$ степени $\leq n-1$. Тогда полином $Q(x) = q(x)\omega_n(x)$ имеет степень $\leq 2n-1$, поэтому условие точности дает ра-

венство

$$\int_a^b \rho(x)q(x)\omega_n(x)dx = \sum_{k=1}^{n} A_k\, q(x_k)\ \omega_n(x_k) = 0.$$

Значит, $\omega_n(x)$ удовлетворяет условию 1).

Достаточность. Пусть условия 1) и 2) выполнены. Рассмотрим произвольный полином $Q(x)$ степени $\leq 2n-1$. Его можно представить в виде

$$Q(x) = q(x)\omega_n(x) + r(x),$$

где q и r — полиномы степени $\leq n-1$. Но тогда

$$\int_a^b \rho(x)Q(x)dx = \int_a^b \rho(x)q(x)\omega_n(x)dx + \int_a^b \rho(x)r(x)dx,$$

причем первое слагаемое в правой части этого равенства равно нулю в силу условия 1). Поэтому с учетом условия 2) и равенств $Q(x_k) = r(x_k)$ получаем

$$\int_a^b \rho(x)Q(x)dx = \int_a^b \rho(x)r(x)dx =$$

$$= \sum_{k=1}^{n} A_k\, r(x_k) = \sum_{k=1}^{n} A_k Q(x_k),$$

что и требовалось доказать.

Далее мы покажем, что существует единственная сетка узлов $x_1, x_2, \ldots, x_n$, для которой $\omega_n(x)$ удовлетворяет условию 1) этой теоремы. Окончательное утвер-

ждение вытекает из двух последующих теорем.

Теорема 5.7 *Для любого натурального числа n существует полином $P_n(x)$ степени n, ортогональный с весом $\rho(x)$ любому полиному степени $\leq n-1$.*

Первое доказательство. Для искомого полинома

$$P_n(x) = b_0 + b_1 x + \ldots + b_{n-1}x^{n-1} + x^n$$

требуемое условие ортогональности можно записать в виде равенств: $-\int_a^b \rho(x)\, x^{n+j}\, dx =$

$$= \int_a^b \rho(x)\, (b_0 + b_1 x + \ldots + b_{n-1}x^{n-1})\, x^j\, dx$$

для всех $j = 0, 1, \ldots, n-1$. Очевидно, эти интегральные равенства представляют собой систему линейных алгебраических уравнений относительно неизвестных коэффициентов b_0, b_1, $\ldots$, b_{n-1}. Достаточно показать, что соответствующая однородная система уравнений

$$\int_a^b \rho(x)\, (a_0 + \ldots + a_{n-1}x^{n-1})\, x^j\, dx = 0, \qquad (5.5)$$

где $j = 0, 1, \ldots, n-1$, имеет единственное решение

$$a_0 = a_1 = \ldots = a_{n-1} = 0.$$

С этой целью умножим j-тое уравнение (5.5) на a_j и просуммируем по $j = 0, 1, \ldots, n-1$. Будем иметь ра-

венства

$$\sum_{j=0}^{n-1} a_j \int_a^b x^j \rho(x) \sum_{k=0}^{n-1} a_k x^k \, dx =$$

$$= \int_a^b \rho(x) \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} a_k a_j x^k x^j \, dx =$$

$$= \int_a^b \rho(x) \left| \sum_{k=0}^{n-1} a_k x^k \right|^2 dx = 0.$$

Отсюда с учетом неотрицательности подынтегральной функции следует, что для почти всех $x \in [a, b]$

$$\rho(x) \left| \sum_{k=0}^{n-1} a_k x^k \right|^2 = 0.$$

Если хотя бы один из коэффициентов a_k отличен от нуля, то полином $a_0 + a_1 x + \ldots + a_{n-1} x^{n-1}$ может быть равным нулю лишь в конечном числе точек. Но тогда получили бы $\rho(x) = 0$ почти всюду на промежутке интегрирования, а значит

$$\int_a^b \rho(x) \, dx = 0,$$

что противоречит требованиям на весовую функцию.

Второе доказательство. Над полем вещественных чисел рассмотрим линейное пространство

$$H_n((a, b), \rho)$$

алгебраических полиномов степени $\leq n$ с вещественными коэффициентами, со скалярным произведением

$$(F,G)=\int_a^b \rho(x)F(x)G(x)dx \quad (F,G\in H_n((a,b),\rho))$$

и соответствующей нормой

$$\|F\|=\sqrt{\int_a^b \rho(x)|F(x)|^2dx}.$$

В этом пространстве система элементов

$$\{1,x,x^2,\ldots,x^n\}$$

является линейно-независимой. Действительно, если эта система была бы линейно-зависимой, то найдутся вещественные числа $a_0,a_1,\ldots,a_n$ такие, что хотя бы один из коэффициентов a_k отличен от нуля, но полином

$$a_0+a_1x+\ldots+a_{n-1}x^{n-1}$$

равен нулю как элемент $L_2((a,b),\rho)$, т. е.

$$\int_a^b \rho(x)\left|\sum_{k=0}^{n} a_k\,x^k\right|^2 dx=0,$$

что невозможно.

Применяя процесс ортогонализации Грама-Шмидта к линейно-независимой системе

$$\{1, x, x^2, \ldots, x^n\},$$

получаем ортонормированную систему

$$\{P_0(x), P_1(x), \ldots, P_n(x)\}.$$

По построению $P_n(x)$ является линейной комбинацией элементов $\{1, x, x^2, \ldots, x^n\}$, в которую входит с ненулевым коэффициентом элемент x^n, и ортогонален элементам $\{1, x, x^2, \ldots, x^{n-1}\}$. Таким образом, $P_n(x)$ — полином степени n с вещественными коэффициентами, ортогональный с весом $\rho(x)$ всем полиномам степени $\leq n-1$.

Этим и завершается доказательство.

Процесс ортогонализации Грама-Шмидта приводит к полиному $P_n(x)$ со старшим членом вида $c\,x^n,\ c \neq 0$. Поэтому в дальнейшем полагаем

$$P_n(x) = c\omega_n(x) = c(x - x_1)\ldots(x - x_n).$$

Но для того, чтобы иметь возможность использовать нули ортогонального полинома $P_n(x)$ в качестве узлов квадратурной формулы, нам нужно доказать следующее утверждение.

Теорема 5.8 *Все нули ортогонального полинома P_n ве-*

щественны, просты (т. е. нет кратных корней) и лежат в интервале (a, b).

Доказательство. Пусть ξ — вещественный нуль полинома $P_n(x)$. Тогда функция

$$q(x) = \frac{P_n(x)}{x - \xi}$$

является отличным от тождественного нуля полиномом степени $n - 1$ с вещественными коэффициентами, поэтому

$$0 = \int_a^b \rho(x) q(x) P_n(x) dx = \int_a^b \rho(x) |q(x)|^2 (x - \xi) dx,$$

отсюда

$$\xi = \frac{\int_a^b x\rho(x)|q(x)|^2 dx}{\int_a^b \rho(x)|q(x)|^2 dx} \in (a, b).$$

Если $n = 1$, то $P_1(x) = c(x - \xi)$, где c, ξ — вещественные числа, $c \neq 0$, и доказательство завершено. В общем случае, когда $n \geq 2$, остается показать, что уравнение $P_n(x) = 0$ не имеет ни комплексных, ни кратных корней.

Предположим, что $P_n(\xi) = 0$, где $\xi = \xi_1 + i\xi_2$ — комплексное число (т. е. $\xi_2 \neq 0$). Поскольку $P_n(x)$ — полином с вещественными коэффициентами, то комплексно сопряженное число $\overline{\xi} = \xi_1 - i\xi_2$ также является корнем

уравнения $P_n(x) = 0$ и

$$(x - \xi)q(x) = P_n(x) = \overline{P_n(x)} = (x - \overline{\xi})\overline{q(x)}.$$

Поэтому из условия ортогональности $P_n(x)$ степенным функциям x^j $(j = 0, 1, \ldots, n-1)$ получаем

$$0 = (q, P_n) = \int_a^b \rho(x)q(x)(x - \overline{\xi})\overline{q(x)}\, dx =$$

$$= \int_a^b \rho(x)|q(x)|^2(x - \overline{\xi})\, dx,$$

и, как следствие, равенство

$$\overline{\xi} = \frac{\int_a^b x\rho(x)|q(x)|^2 dx}{\int_a^b \rho(x)|q(x)|^2 dx}.$$

Правая часть этого равенства является вещественным числом, таким образом, пришли к противоречию. Остается доказать отсутствие вещественных кратных корней. Предположим, что ξ — вещественный кратный корень, тогда функция

$$q(x) = \frac{P_n(x)}{(x - \xi)^2}$$

является полиномом с вещественными коэффициентами степени $n - 2$. Снова условие ортогональности при-

водит к противоречию:

$$0 = (q, P_n) = \left(\frac{P_n}{(x-\xi)^2}, P_n\right) =$$

$$= \left(\frac{P_n}{x-\xi}, \frac{P_n}{x-\xi}\right) = \left\|\frac{P_n}{x-\xi}\right\|^2 > 0.$$

Доказательство завершено.

5.3.2 Две оценки погрешности

Приведем две различных оценки погрешности квадратурной формулы

$$\int_a^b \rho(x) f(x)\, dx \approx \sum_{k=1}^{n} p_k\, f(x_k)$$

в предположении, что эта формула точна на полиномах степени $\leq 2n-1$, т. е. является квадратурной формулой Гаусса. Как мы уже знаем, это предположение равносильно следующим условиям:

полином $\omega_n(x) = (x-x_1)\ldots(x-x_n)$ ортогонален с весом $\rho(x)$ любому полиному степени $\leq n-1$, а коэффициенты p_k вычисляются по формулам

$$p_k = \int_a^b \rho(x) \frac{\omega_n(x)\, dx}{(x-x_k)\omega_n'(x_k)}.$$

Напомним, что при любой сетке узлов для любой

интерполяционной квадратурной формулы

$$\sum_{k=0}^{n} p_k = \int_a^b \rho(x)\,dx.$$

Дополнительным **свойством формулы Гаусса является положительность всех коэффициентов** p_k. Убедиться в этом можно так: для любого индекса k функция

$$f_k(x) = \left(\frac{\omega_n(x)}{x - x_k}\right)^2 \quad (f_k(x_k) := \omega_n'^2(x_k) > 0)$$

является полиномом степени $2n-2$, поэтому формула Гаусса для нее точна:

$$\int_a^b \rho(x) f_k(x) dx = \sum_{j=1}^{n} p_j\, f_k(x_j) = p_k\, f_k(x_k).$$

Отсюда следует

$$p_k = \frac{\int_a^b \rho(x) f_k(x) dx}{f_k(x_k)} > 0.$$

Таким образом, при любом числе узлов сетки

$$0 < p_k \le \int_a^b \rho(x)\,dx,$$

т. е. коэффициенты ограничены числом, не зависящим от n, и вычисления по квадратурной формуле наивыс-

шего алгебраического порядка точности оказываются устойчивыми при повышении числа узлов. Эксперты по вычислениям отмечают, что на практике квадратурные формулы Гаусса применяются с числом узлов до 100.

В двух следующих теоремах через

$$\psi_n(f) = \int_a^b \rho(x) f(x) dx - \sum_{k=1}^{n} p_k \, f(x_k)$$

мы будем обозначать погрешность квадратурной формулы Гаусса.

Теорема 5.9 *Для любой функции* $f \in C[a, b]$

$$|\psi_n(f)| \leq 2E_{2n-1}(f) \int_a^b \rho(x) dx,$$

где $E_{2n-1}(f)$ *— наилучшее равномерное приближение* f *полиномами степени* $\leq 2n - 1$.

Доказательство. Пусть $Q(x)$ — произвольный полином степени $\leq 2n - 1$. Имеем

$$\int_a^b \rho(x) Q(x) dx = \sum_{k=1}^{n} p_k \, Q(x_k).$$

Поэтому погрешность квадратурной формулы Гаусса может быть оценена следующим образом: $|\psi_n(f)| =$

$$= \left| \int_a^b \rho(x)[f(x) - Q(x)]dx - \sum_{k=1}^{n} p_k \, [f(x_k) - Q(x_k)] \right| \leq$$

$$\leq \|f(x) - Q(x)\|_{C[a,b]} \left\{ \int_a^b \rho(x)dx + \sum_{k=1}^{n} p_k \right\} =$$

$$= 2\|f(x) - Q(x)\|_{C[a,b]} \int_a^b \rho(x)dx.$$

Отсюда в силу произвольности полинома $Q(x)$ степени $\leq 2n-1$ вытекает утверждение теоремы.

Теорема 5.10 *Для любой функции $f \in C^{2n}[a,b]$ справедливо представление*

$$\psi_n(f) = \frac{f^{(2n)}(\eta)}{(2n)!} \int_a^b \rho(x)\, \omega_n^2(x)dx,$$

где $\eta \in [a,b]$.

Доказательство. Рассмотрим интерполяционный полином Эрмита-Фейера $H_n(f;x)$, построенный по условиям

$$H_n(f;x_k) = f(x_k), H_n'(f;x_k) = f'(x_k)\,(k = 1,2,\ldots,n).$$

Так как $H_n(f;x)$ — полином степени $\leq 2n-1$, то для него формула Гаусса точна, и поэтому

$$\int_a^b \rho(x) f(x)\, dx \approx \sum_{k=1}^{n} p_k\, f(x_k) =$$

$$= \sum_{k=1}^{n} p_k\, H_n(f;x_k) = \int_a^b \rho(x)\, H_n(f;x)\, dx.$$

Отсюда следует

$$\psi_n(f) = \int_a^b \rho(x)\,[f(x) - H_n(f;x)]\,dx.$$

Пользуясь доказанным ранее представлением

$$f(x) - H_n(f;x) = \frac{f^{(2n)}(\xi(x))}{(2n)!}\,\omega_n^2(x) \quad (\xi(x) \in (a,b))$$

для остаточного члена при кратной интерполяции и теоремой о среднем для интегралов, легко получаем требуемую формулу для $\psi_n(f)$.

5.3.3 Важные частные случаи

Лишь при малых значениях числа узлов n мы можем построить явно ортогональные полиномы $P_n(x)$ для произвольного промежутка и допустимого веса, пользуясь, например, процессом ортогонализации Грама-Шмидта. Явные выражения для $P_n(x)$ при любом числе узлов получены лишь в специальных случаях.

Приведем наиболее употребительные ортогональные полиномы и соответствующие им квадратурные формулы Гаусса.

1) Полиномы Лежандра

$$L_n(x) = \frac{d^n(1-x^2)^n}{dx^n}$$

ортогональны с весом $\rho(x) \equiv 1$ на отрезке $[-1, 1]$. Нули L_n еще "вручную" были табулированы до $n = 512$. Соответствующая квадратурная формула

$$\int_{-1}^{1} f(x)dx \approx \sum_{k=1}^{n} p_k\, f(x_k),\; p_k = \int_{-1}^{1} \frac{L_n(x)dx}{(x - x_k)L_n'(x_k)}\, dx,$$

первая среди квадратурных формул наивысшего алгебраического порядка точности, была получена Гауссом.

2) Ортогональными полиномами на отрезке $[-1, 1]$ с весом

$$\rho(x) = \frac{1}{\sqrt{1 - x^2}}$$

оказываются уже знакомые нам полиномы Чебышева первого рода:

$$T_n(x) = \cos(n \arccos x)$$

с нулями

$$x_k = \cos\left(\frac{(2k-1)\pi}{2n}\right) \quad (k = 1, \ldots, n).$$

Легко вычисляются коэффициенты: $p_k = \pi/n$ для любого k. Соответствующая квадратурная формула наивысшего алгебраического порядка точности — формула

Эрмита — имеет вид

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}} dx \approx \frac{\pi}{n} \sum_{k=1}^{n} f\left(\cos \frac{2k-1}{2n}\pi\right).$$

3) Для случая $\rho(x) = \sqrt{1-x^2}$ на отрезке $[-1, 1]$ ортогональные полиномы — полиномы Чебышева второго рода — определены формулами

$$U_n(x) = \frac{\sin(n+1)\theta}{\sin\theta}, \quad \theta = \arccos x.$$

Полином $U_n(x)$ обращается в нуль в точках $x_k = \cos\frac{k\pi}{n+1}$ $(k = 1, \ldots, n)$, а квадратурная формула также имеет явный вид:

$$\int_{-1}^{1} \sqrt{1-x^2} f(x) dx \approx$$

$$\approx \frac{\pi}{n+1} \sum_{k=1}^{n} \sin^2 \frac{k\pi}{n+1} f\left(\cos \frac{k\pi}{n+1}\right).$$

4) Пусть $\rho(x) = (1-x)^\alpha (1+x)^\beta$ на отрезке $[-1, 1]$. Параметры удовлетворяют неравенствам

$$\alpha > -1, \quad \beta > -1,$$

вытекающим из условия интегрируемости весовой функции. Соответствующие ортогональные полиномы

$$P_n^{(\alpha,\beta)}(x) = \frac{1}{(1-x)^\alpha (1+x)^\beta} \cdot \frac{d^n[(1-x)^{n+\alpha}(1+x)^{n+\beta}]}{dx^n}$$

называются полиномами Якоби. Можно показать, что коэффициенты p_k выражаются в явном виде в терминах Γ — функции Эйлера.

5) Для построения квадратурных формул можно также использовать полиномы Лагерра

$$P_n(x) = \frac{1}{x^\alpha e^{-x}} \frac{d^n[x^{n+\alpha} e^{-x}]}{dx^n}.$$

Система полиномов Лагерра является ортогональной с весом $\rho(x) = x^\alpha e^{-x}$ на полуоси $(0, +\infty)$. Имеется естественное условие для параметра: $\alpha > -1$.

6) На числовой прямой $(-\infty, +\infty)$ положительная функция $\rho(x) = e^{-x^2}$ является весовой, поскольку

$$\int_{-\infty}^{\infty} e^{-x^2} dx < \infty.$$

Ортогональные полиномы, соответствующие этому случаю, называются полиномами Эрмита и выражаются формулой $H_n(x) = e^{x^2} \left(e^{-x^2}\right)^{(n)}$.

5.4 Дополнительные вопросы

5.4.1 Интегралы с весом от периодических функций

Рассмотрим 2π-периодическую, непрерывную функцию $f(x)$. Понятно, что в этом случае для вычисления ин-

теграла

$$\int_0^{2\pi} \rho(x)\, f(x)\, dx$$

можно использовать приведенные ранее квадратурные формулы. Для периодических функций естественной является также приближенная формула, получаемая заменой функции ее тригонометрическим интерполяционным полиномом. А именно, полагаем

$$\int_0^{2\pi} \rho(x)\, f(x)\, dx \approx \int_0^{2\pi} \rho(x)\, T_n(f;x)\, dx,$$

где $T_n(f;x)$ — тригонометрический полином степени не выше n, удовлетворяющий условиям

$$T_n(f;x_0) = f(x_0),\ \ldots,\ T_n(f;x_{2n}) = f(x_{2n})$$

на сетке с $2n+1$ узлами

$$x_0, x_1, \ldots, x_{2n} \in [0, 2\pi], \quad 0 < |x_i - x_j| < 2\pi,\ i \neq j.$$

Для получения квадратурной формулы необходимо использовать представление в форме Лагранжа

$$T_n(f;x) = \sum_{k=0}^{2n} f(x_k) t_k(x),$$

где

$$t_k(x) = \frac{\prod_{j=0, j\neq k}^{2n} \sin\frac{x-x_j}{2}}{\prod_{j=0, j\neq k}^{2n} \sin\frac{x_k-x_j}{2}},\ k = 0, 1, \ldots, 2n.$$

Будем иметь

$$\int_0^{2\pi} \rho(x)\, f(x)\, dx \approx \sum_{k=0}^{2n} q_k\, f(x_k), \tag{5.6}$$

где

$$q_k = \int_0^{2\pi} \rho(x)\, t_k(x)\, dx\ \ (k = 0, 1, \ldots, 2n).$$

Поскольку $T_n(F; x) \equiv F(x)$ для любой функции вида

$$F(x) = \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos kx + b_k \sin kx,$$

то построенная квадратурная формула (5.6) будет точна для любого тригонометрического полинома F степени $\leq n$.

5.4.2 О формулах Филона

Пусть $f \in C[a, b]$ и $\omega >> b - a$, т. е. число ω намного больше длины отрезка $[a, b]$. Тогда функции $\cos\omega x$ и $\sin\omega x$, $x \in [a, b]$, многократно меняют знак. Такие функции называют быстро осциллирующими.

Рассмотрим интегралы

$$\int_a^b f(x)\cos\omega x\,dx, \quad \int_a^b f(x)\sin\omega x\,dx.$$

Интегралы такого типа часто встречаются в прикладных задачах, для решения которых используются преобразования Фурье или ряды Фурье. Например, для разложения заданной функции в ряд Фурье необходимо для любого числа $k \in \mathbb{N}$ вычислять интегралы

$$a_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin kx\,dx, \quad b_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos kx\,dx.$$

Очевидно, достаточно рассмотреть интегралы с косинусами, так как интегралы с синусами сводятся к ним заменой переменных.

Применение обычных квадратурных формул может привести к ошибочным результатам. Например, пусть узлы $x_1, x_2, \ldots, x_n$ выбраны так, что они совпадают с корнями уравнения $\cos kx = 0$, т. е. $\cos kx_j = 0$. Применяя к функции $g(x) = f(x)\cos kx$ квадратурную формулу вида

$$\int_0^{2\pi} g(x)\,dx \approx \sum_{j=1}^{n} A_j g(x_j),$$

при любом выборе параметров A_j приходим к неудо-

влетворительному результату: коэффициенты Фурье

$$b_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos kx\,dx \approx \frac{1}{\pi}\sum_{j=1}^{n} A_j f(x_j)\cos kx_j = 0$$

для любой функции f.

На практике для вычисления интегралов от быстро осциллирующих функций рекомендуют формулы Филона (Луи Наполеон Жорж Филон — английский математик французского происхождения, специалист по прикладной математике).

Формулы Филона для приближенного вычисления интегралов $\int_a^b f(x)\cos\omega x\,dx$ можно найти в любом справочнике. Объясним здесь лишь исходную идею Филона.

Пусть f — непрерывная, плавно меняющаяся функция, а функция $\varphi(x)$ является быстро осциллирующей на отрезке $[a,b]$. Построим интерполяционный полином Лагранжа $L_n(f;x)$ по узлам $x_1, x_2, \ldots, x_n \in [a,b]$. Полагаем

$$\int_a^b f(x)\varphi(x)\,dx \approx \int_a^b L_n(f;x)\varphi(x)\,dx =$$

$$= \sum_{j=1}^{n} f(x_j)\int_a^b l_j(x)\varphi(x)\,dx,$$

где $l_j(x)$ — фундаментальные полиномы Лагранжа.

В том случае, когда $\varphi(x) = \cos\omega x$, интегралы вида $\int_a^b x^m\varphi(x)\,dx$ можно вычислить точно интегрированием

по частям. Следовательно, в этом случае явно определяются и интегралы вида $\int_a^b l_j(x)\varphi(x)\,dx$.

5.4.3 О несобственных интегралах

Пусть функция $f : (a, b] \to \mathbb{R}$ непрерывна, но

$$\limsup_{x \to a^+} |f(x)| = \infty,$$

т. е. точка a является особой. Предположим, что сходится несобственный интеграл

$$J(f) = \int_a^b f(x)\,dx := \lim_{\delta \to 0^+} \int_{a+\delta}^b f(x)\,dx.$$

Непосредственное применение квадратурных формул для вычисления такого интеграла может привести к сколь угодно большим ошибкам.

Для корректного вычисления $J(f)$ можно поступить следующим образом: с применением квадратурных формул вычисляем интеграл

$$\int_{a+\delta}^b f(x)\,dx.$$

с достаточно малым $\delta > 0$, сопровождая вычисления с оценкой выбрасываемого интеграла по промежутку $[a, a + \delta]$. Наряду с этим стандартным подходом существуют и искусственные приемы для приближенного вычисления несобственных интегралов. Укажем два

распространенных приема: 1) сведение несобственного интеграла к собственному путем замены переменной или интегрированием по частям с последующим применением одной из квадратурных формул; 2) аддитивное или мультипликативное выделение особенности с последующим комбинированием аналитических и численных методов. Проиллюстрируем эти рекомендации на примере интеграла

$$J = \int_0^1 (1 + x^2)^{-1} \ln x \, dx.$$

Заменой переменных $x = t^k$ с постоянной $k > 1$ получаем $J = k^2 \int_0^1 (1 + t^{2k})^{-1} t^{k-1} \ln t \, dt$. Новая подынтегральная функция $g(t) = (1 + t^{2k})^{-1} t^{k-1} \ln t$ непрерывна на $[-1, 1]$, поэтому интеграл $J = \int_0^1 g(t)dt$ можно вычислять приближенно по известным квадратурным формулам. **При интегрировании по частям** с функциями $u = \ln x$ и

$$v = \int_0^x \frac{dt}{1 + t^2},$$

мы также получаем интеграл от непрерывной функции. **Аддитивное выделение особенностей**: простые преобразования

$$\frac{(\ln x)(1 + x^2 - x^2)}{1 + x^2}$$

позволяют представить наш интеграл в виде суммы

$$J = \int_0^1 \ln x \, dx - \int_0^1 \frac{x^2 \ln x}{1+x^2} \, dx,$$

где первый интеграл легко вычисляется аналитически и равен единице, а ко второму интегралу можно применить одну из квадратурных формул. **Мультипликативное выделение особенностей**: запишем подынтегральную функцию в виде произведения

$$f(x) = \frac{\ln x}{1+x^2} = \rho(x) g(x),$$

где

$$g(x) = -\frac{1}{1+x^2}, \qquad \rho(x) = -\ln x, \qquad g \in C[0,1].$$

К полученному интегралу можно применить квадратурную формулу вида

$$\int_0^1 \rho(x) g(x) dx \approx \sum_{k=1}^{n} A_k \, g(x_k).$$

Для вычисления несобственных интегралов $\int_0^\infty f(x)\,dx$ можно рекомендовать один из следующих путей: либо заменой переменных получить несобственные интегралы по конечному промежутку, например, по формуле $\int_0^\infty f(x)\,dx = \int_0^1 f(x)\,dx - \int_0^1 \frac{f(1/t)}{t^2}\,dt$, далее пользоваться рекомендованными выше приемами; либо вычислять

интеграл по отрезку $[0, A]$ с достаточно большим A, сопровождая вычисления с оценкой выбрасываемого интеграла по лучу $[A, +\infty)$.

5.5 Задачи и упражнения

1. Для интеграла

$$\int_0^1 xf(x)\,dx$$

постройте квадратурную формулу с двумя узлами, точную для всех полиномов: a) первой степени, b) второй степени.

2. Найдите алгебраический порядок точности квадратурной формулы

$$\int_0^1 f(x)\,dx \approx \frac{f(0) + 4f(1/2) + f(1)}{6}.$$

3. Для интеграла

$$\int_0^1 xf(x)\,dx$$

постройте квадратурную формулу Гаусса с двумя узлами.

4. Вычислите с точностью $\varepsilon = 0,1$ интегралы

$$\int_0^1 \frac{dx}{x+2}, \quad \int_0^1 \frac{dx}{x^4+2}$$

с помощью: a) квадратурной формулы прямоугольников, b) квадратурной формулы трапеций.

5. Вычислите интеграл

$$\int_{-1}^1 \frac{dx}{\sqrt{1-x^4}}$$

с помощью формулы Гаусса с двумя узлами.

6. С точностью $\varepsilon = 0,01$ вычислите интеграл

$$\int_0^1 e^x \sin 100x \, dx.$$

7. С точностью $\varepsilon = 0,01$ вычислите несобственный интеграл

$$\int_0^1 \frac{dx}{\sqrt{x(1-x)(x+1)}}.$$

8. Покажите, что следующая квадратурная формула прямоугольников

$$\int_0^{2\pi} f(x)dx \approx \frac{2\pi}{n} \sum_{k=1}^{n} f\left(\frac{2\pi k}{n}\right)$$

является формулой наивысшего тригонометрического порядка точности.

Глава 6

Задачи для ОДУ и УЧП

Из общего курса дифференциальных уравнений известно, что существует большое число обыкновенных дифференциальных уравнений (ОДУ), решения которых выражаются через известные функции, и эти решения можно найти в явном виде. Существуют хорошие справочники, содержащие те ОДУ, для которых известны формулы их общих решений.

Нередко дифференциальные уравнения приходится решать приближенно, т. е. искать то или иное приближение точного решения. Для решения различных задач, связанных с дифференциальными уравнениями, разработан ряд эффективных приближенных методов. Их можно условно разделить на аналитические и численные методы. Часто эти подходы переплетаются, и возникают численно-аналитические методы.

Основным методом численного решения дифференциальных уравнений является метод конечных разностей (МКР). При изучении метода конечных разностей

нужно обратить внимание на 3 составные части метода:

построение разностной схемы и алгоритма для вычисления сеточной функции,

теоретическое обоснование сходимости к точному решению и

оценки погрешности в зависимости от шага сетки и условий гладкости рассматриваемых функций.

Большая часть данного раздела посвящена различным методам приближенного решения одной и той же задачи, а именно, задачи Коши для ОДУ первого порядка, разрешенного относительно производной. Будут рассмотрены также задача Коши для системы дифференциальных уравнений первого порядка, а также задача Коши для дифференциальных уравнений высших порядков, разрешенных относительно старшей производной.

Мы опишем также подробно применение метода конечных разностей к краевой задаче для линейного дифференциального уравнения второго порядка и обоснование сходимости метода.

По сравнению с ОДУ для дифференциальных уравнений в частных производных (УЧП) усложняются как сами постановки задач, так и методы их решений.

Из курса уравнений в частных производных (УЧП) известно, что для каждого типа уравнений существует хорошо разработанная теория о свойствах решений, о

корректных постановках краевых и начально краевых задач и методах их решения. Эти теории существенно зависят от типа уравнений и отличаются друг от друга.

Для УЧП, которые имеют применения на практике, хорошо разработаны методы аналитического представления решений и разнообразные численные методы. Мы рассмотрим типичные задачи для нескольких уравнений математической физики, когда численное решение может быть найдено методом конечных разностей (МКР). Познакомимся также с несколькими приближенными методами решения операторных и интегральных уравнений.

6.1 Задача Коши для ОДУ

Рассмотрим следующую задачу Коши для обыкновенного дифференциального уравнения первого порядка: нужно найти решение $y \in C^1[x_0, b]$ уравнения

$$y' = f(x, y) \tag{6.1}$$

на отрезке $[x_0, b]$, удовлетворяющее условию

$$y|_{x=x_0} = y_0, \tag{6.2}$$

где y_0 — заданное число. Функция f также считается заданной на множестве

$$\Omega = \{(x, y) \in \mathbb{R}^2 : x_0 \le x \le b,\ |y| < \infty\},$$

причем $f \in C(\Omega)$.

При поиске приближенного решения будем предполагать существование и единственность точного решения $y \in C^1[x_0, b]$. Как известно из курса дифференциальных уравнений, для этого правая часть должна обладать некоторыми свойствами. Например, задача (6.1) — (6.2) гарантированно имеет единственное решение на достаточно малом отрезке $[x_0, b]$, если функция f удовлетворяет условию Липшица по второму аргументу: существует число $K > 0$ такое, что

$$|f(x, y) - f(x, z)| \le K|y - z|$$

для любых $y, z \in \mathbb{R}$.

Для приближенного решения задачи (6.1) — (6.2) можно использовать следующие три классических метода:

- метод последовательных приближений Пикара;

- метод Коши (метод степенных рядов с неопределенными коэффициентами);

- метод Эйлера (=метод ломаных).

Рассмотрим их подробнее.

6.1.1 Метод Пикара

Задача (6.1) – (6.2) эквивалентна интегральному уравнению

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))dt. \tag{6.3}$$

Эквивалентность легко показывается: из (6.1) – (6.2) интегрированием получаем (6.3), и обратно, если интегральное уравнение (6.3) имеет непрерывное решение, то оно окажется непрерывно дифференцируемым в силу свойств правой части (6.3).

Начальное условие $y(x_0) = y_0$ для решения интегрального уравнения (6.3) выполнено автоматически, и, дифференцируя (6.3), получаем (6.1).

Метод Пикара — итерационный метод решения интегрального уравнения (6.3). В простейшем варианте в качестве нулевого приближения берется функция, тождественно равная постоянной:

$$y_0(x) \equiv y_0.$$

Последовательно определяются первое приближение

$$y_1(x) = y_0 + \int_{x_0}^{x} f(t, y_0)dt,$$

затем второе приближение

$$y_2(x) = y_0 + \int_{x_0}^{x} f(t, y_1(t))dt,$$

и наконец, общее правило для определения n-ого приближения задается равенством

$$y_n(x) = y_0 + \int_{x_0}^{x} f(t, y_{n-1}(t))dt.$$

Естественно ожидать, что для достаточно большого $n \in \mathbb{N}$

$$y_n(x) \approx y(x),$$

т. е. n-е приближение близко к точному решению задачи. Этот факт легко доказать, если правая часть уравнения (6.1) удовлетворяет некоторым условиям. Приведем точные утверждения.

Сначала установим простой факт: *если итерационный процесс приводит к равномерно сходящейся последовательности, то предельная функция является искомым решением.*

Теорема 6.1 *Пусть $f \in C(\Omega)$. Если последовательность Пикара $y_n(x)$ равномерно сходится к некоторой функции $y(x)$ на $[x_0, b]$, то $y(x)$ — решение задачи (6.1) — (6.2).*

Доказательство. По построению $y_0 = \text{const}$, y_1 – непрерывная функция как интеграл с переменным верхним пределом от непрерывной функции. Последовательно

получаем, что $y_{n-1} \in C[x_0, b]$. И поэтому

$$y_n(x) = y_0 + \int_{x_0}^{x} f(t, y_{n-1}(t))dt \in C[x_0, b].$$

По условию теоремы $y_n(x) \rightrightarrows y(x)$, и, как мы установили, функция $y_n(x)$ непрерывна на отрезке $[x_0, b]$. Значит, по теореме Вейерштрасса и предельная функция $y(x)$ непрерывна на $[x_0, b]$. Имеем

$$f(t, y_n(t)) \rightrightarrows f(t, y(t)) \text{ на } [x_0, b] \text{ при } n \to \infty.$$

Переходя к пределу в равенстве

$$y_n(x) = y_0 + \int_{x_0}^{x} f(t, y_{n-1}(t))dt$$

при $n \to \infty$, получаем, что предельная функция удовлетворяет интегральному уравнению (6.3). А значит, функция $y(x)$ является решением задачи Коши (6.1) – (6.2).

Замечание. Если $y_n \rightrightarrows y$, то понятно, что функцию y_n для достаточно большого n можно взять в качестве приближенного решения. Погрешность

$$\varepsilon_n = \max_{x \in [x_0, b]} |y(x) - y_n(x)| \to 0 \quad \text{при } n \to \infty.$$

Для явной оценки погрешности можно использовать следующее утверждение.

Теорема 6.2 *Пусть $f \in C(\Omega)$ и удовлетворяет условию Липшица по переменной y, а именно,*

$$|f(x,y) - f(x,z)| \le K|y - z|$$

для любых $y, z \in \mathbb{R}$, $K = const > 0$.

Если число $b > x_0$ выбрано так, что $K(b-x_0) < 1$, то справедлива следующая оценка погрешности

$$\varepsilon_n = \|y - y_n\|_{C[x_0,b]} \le \frac{\alpha^n}{1-\alpha} M(b - x_0),$$

где

$$\alpha = K(b - x_0) < 1, \qquad M = \max_{x \in [x_0,b]} |f(x, y_0)|.$$

Доказательство. В пространстве $C[x_0, b]$ определим оператор

$$A : C[x_0, b] \to C[x_0, b]$$

следующим равенством:

$$(Ay)(x) = y_0 + \int_{x_0}^{x} f(t, y(t))dt$$

для любой функции $y \in C[x_0, b]$. Легко показать, что A – сжимающий оператор, так как для любого $x \in [x_0, b]$

$$|(Ay)(x) - (Az)(x)| \le \int_{x_0}^{x} |f(t, y(t)) - f(t, z(t))|dt \le$$

$$\leq \int_{x_0}^{x} K|y(t) - z(t)|dt \leq$$

$$\leq K\|y - z\|_{C[x_0,b]} \int_{x_0}^{x} dt =$$

$$= K\|y - z\|_{C[x_0,b]}(x - x_0), \quad \forall x \in [x_0, b].$$

Следовательно,

$$\|Ay - Az\|_{C[x_0,b]} \leq K(b - x_0)\|y - z\|_{C[x_0,b]} =$$

$$= \alpha\|y - z\|_{C[x_0,b]}, \quad \alpha < 1.$$

По теореме Банаха о сжимающем операторе будем иметь

$$\|y - y_n\|_{C[x_0,b]} \leq \frac{\alpha^n}{1-\alpha}\|y_1 - y_0\|_{C[x_0,b]} \leq$$

$$\leq \frac{\alpha^n}{1-\alpha} M \left\| \int_{x_0}^{x} dt \right\|_{C[x_0,b]} =$$

$$= \frac{\alpha^n}{1-\alpha} M(b - x_0).$$

Этим и завершается доказательство.

Метод Пикара позволяет определить явно точное решение лишь в редких случаях, в частности, в тех случаях, когда функция $f(x, y)$ такова, что интегралы

$$\int_{x_0}^{x} f(t, y_{n-1}(t))dt$$

вычисляются в явном виде. Приведем иллюстративный

пример.

Рассмотрим следующую задачу Коши

$$y' = x^2 + 2xy, \quad y|_{x=0} = 0.$$

Поскольку уравнение $y' = x^2 + 2xy$ является линейным, стандартным методом из курса дифференциальных уравнений мы можем найти точное аналитическое решение. Оно имеет вид

$$y(x) = -\frac{x}{2} + \frac{e^{x^2}}{2} \int_0^x e^{-t^2}\, dt.$$

Вывод этой формулы пропускаю и оставляю для слушателей в качестве упражнения.

А теперь получим точное решение методом последовательных приближений Пикара. Полагая $y_0(x) \equiv 0$, непосредственными вычислениями получаем

$$y_1(x) = \frac{x^3}{3},$$

$$y_2(x) = \int_0^x \left(t^2 + \frac{2t^4}{3} \right) dt = \frac{x^3}{3} + \frac{2x^5}{3 \cdot 5},$$

$$y_3(x) = \int_0^x \left(t^2 + \frac{2t^4}{3} + \frac{4t^6}{3 \cdot 5} \right) dt = \frac{x^3}{3} + \frac{2x^5}{3 \cdot 5} + \frac{4x^7}{3 \cdot 5 \cdot 7}.$$

По индукции получаем

$$y_n(x) = \int_0^x (t^2 + 2ty_{n-1}(t))\,dt = \sum_{k=1}^{n} \frac{2^{k-1}x^{2k+1}}{(2k+1)!!}.$$

Разложив точное решение в ряд Тейлора и сравнив его с итерациями Пикара, мы легко убеждаемся в том, что итерационный процесс сходится и позволяет определить приближенное решение с любой степенью точности. В частности, для точного решения

$$y(x) = \lim_{n\to\infty} y_n(x)$$

мы нашли новую формулу. Сравнивая ее с точным решением, найденным выше аналитически, будем иметь нетривиальное тождество

$$\sum_{k=1}^{\infty} \frac{2^{k-1}x^{2k+1}}{(2k+1)!!} = -\frac{x}{2} + \frac{e^{x^2}}{2}\int_0^x e^{-t^2}\,dt.$$

6.1.2 Метод Коши

Предположим, что f — аналитическая функция переменных (x, y) в некоторой окрестности точки (x_0, y_0), т. е. эта функция представима там как сумма сходящегося степенного ряда.

Рассмотрим снова задачу (6.1) – (6.2):

$$y' = f(x, y), \quad x_0 \le x \le b,$$

$$y(x_0) = y_0.$$

В аналитической теории дифференциальных уравнений доказывается, что решение существует и единственно, причем в некоторой окрестности точки x_0 его можно представить в виде суммы сходящегося степенного ряда

$$y(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k.$$

Пусть этот ряд сходится на отрезке $[x_0, b]$.

Понятно, что на основании этого теоретического факта мы можем сделать простой вывод: в качестве приближенного решения задачи Коши можно взять частичную сумму ряда

$$y_n(x) = \sum_{k=0}^{n} a_k (x - x_0)^k$$

для достаточно большого n.

Коэффициенты a_k указанного степенного ряда зависят от y_0 и от частных производных функции f в точке (x_0, y_0). Действительно, в силу начального условия (6.2) имеем $a_0 = y(x_0) = y_0$, а уравнение (6.1) позволяет найти

$$a_1 = y'(x_0) = f(x_0, y(x_0)) = f(x_0, y_0).$$

Для определения a_k при $k \geq 2$ используем ряд

$$f(x, y(x)) = \sum_{k=0}^{\infty} b_k (x - x_0)^k,$$

где b_k также неизвестны, но их можно выразить через частные производные функции f в точке (x_0, y_0) и коэффициенты $a_1, \ldots, a_k$. Подставляя в уравнение (6.1) ряды для двух функций $y(x)$ и $f(x, y(x))$, получаем

$$\sum_{k=1}^{\infty} k\, a_k\, (x - x_0)^{k-1} = \sum_{k=0}^{\infty} b_k\, (x - x_0)^k,$$

отсюда следует равенство коэффициентов

$$n a_n = b_{n-1} \quad \forall n \geq 1.$$

Поскольку при $n = 1$

$$a_1 = b_0 = \sum_{k=0}^{\infty} b_k (x - x_0)^k \Big|_{x=x_0} = f(x_0, y(x_0)) = f(x_0, y_0),$$

то при $n = 2$ получаем

$$2a_2 = b_1 = \frac{df(x, y(x))}{dx}\Big|_{x=x_0} =$$

$$= \frac{\partial f}{\partial x}(x_0, y_0) + \frac{\partial f}{\partial y}(x_0, y_0) y'(x_0) =$$

$$= \frac{1}{2}\left[\frac{\partial f}{\partial x}(x_0, y_0) + \frac{\partial f}{\partial y}(x_0, y_0)a_1\right].$$

Для $n = 3$

$$3a_3 = b_2 = \frac{1}{2!}\left.\frac{d^2 f(x, y(x))}{dx^2}\right|_{x=x_0}.$$

Очевидно, в правую часть войдут частные производные f и найденные уже величины

$$y'(x_0) = a_1, \qquad y''(x_0) = 2a_2.$$

Таким образом a_3 выражается через a_1, a_2 и частные производные f в точке (x_0, y_0). Аналогично убеждаемся в том, что a_4 можно выразить через a_1, a_2, a_3 и значения частных производных f до порядка три в точке (x_0, y_0). Понятно, что по индукции мы можем определить коэффициенты a_k для любого натурального k.

В принципе, можно выписать и явные формулы для a_k, но в общем случае они будут весьма громоздкими. Но во многих специальных случаях неизвестные коэффициенты можно определить явно и выписать решение задачи. В качестве такого примера рассмотрим ту же задачу Коши

$$y' = x^2 + 2xy, \quad y|_{x=0} = 0,$$

с помощью которой мы иллюстрировали метод Пика-

ра. Подставляя в уравнение (6.1) ряды для $y(x)$ и ее производной, мы получаем тождество

$$y'(x) = \sum_{k=1}^{\infty} k\, a_k\, x^{k-1} = x^2 + 2\sum_{k=0}^{\infty} a_k\, x^{k+1},$$

справедливое в некоторой окрестности начала координат. Легко видеть, что $a_0 = y(x_0) = 0$, $a_1 = y'(0) = 0$, и для любого натурального числа n имеем

$$a_{2n} = 0,$$

а коэффициенты с нечетными индексами определяются последовательно из соотношения

$$(2n+1)a_{2n+1} = 2a_{2n-1}.$$

Очевидно, точное решение задачи в этом частном случае запишется в виде суммы знакомого нам ряда

$$y(x) = \sum_{k=1}^{\infty} \frac{2^{k-1}x^{2k+1}}{(2k+1)!!}.$$

Отметим, что радиус сходимости этого ряда равен бесконечности и поэтому полученное решение удовлетворяет дифференциальному уравнению на всей числовой оси.

6.1.3 Метод Эйлера

Снова рассматриваем задачу Коши (6.1) – (6.2). Предложенный Эйлером численный метод основан на приближении производной конечной разностью. Этот метод часто называют методом ломаных или явным методом Эйлера.

Идея метода проста: отрезок $[x_0, b]$ делим на N равных частей с шагом

$$h = \frac{b - x_0}{N}.$$

Берем сетку с равноотстоящими узлами

$$x_0,\, x_1,\, x_2, \ldots,\, x_N = b,$$

где

$$x_k = x_0 + kh, \quad k = 1, 2, \ldots, N.$$

Обозначим точное решение рассматриваемой задачи Коши как

$$y(x), \quad x_0 \le x \le b.$$

Приближенное решение ищем в виде ломаной с вершинами $(x_k,\, y_k)$, где $y_k \approx y(x_k)$.

Пусть $f \in C^2[x_0, b]$. Тогда по формуле Тейлора

$$y(x + h) - y(x) = hy'(x) + O(h^2),$$

и дифференциальное уравнение (6.1) можно записать

в виде равенства

$$y(x+h) - y(x) = hf(x, y(x)) + O(h^2).$$

Следуя Эйлеру, заменим в этом уравнении x на x_n и отбросим $O(h^2)$. Получим n приближенных равенств

$$y(x_{n+1}) - y(x_n) \approx hf(x_n, y(x_n)), \quad n = 0, \ldots, N-1.$$

Приближенное значение $y_n \approx y(x_n)$ естественно находить из соотношений

$$y_{n+1} - y_n = hf(x_n, y_n), \quad n = 0, \ldots, N-1.$$

Таким образом, мы приходим к следующему алгоритму Эйлера:

$$\begin{aligned}
&y_1 = y_0 + hf(x_0, y_0),\\
&y_2 = y_1 + hf(x_1, y_1),\\
&\ldots\ldots\ldots\\
&y_N = y_{N-1} + hf(x_{N-1}, y_{N-1}).
\end{aligned}$$

Очевидно, $y_1 = y(x_1) + O(h^2)$. При последовательных вычислениях следующих величин

$$\begin{aligned}
&y_2 \approx y(x_2),\\
&\ldots\ldots\ldots\\
&y_n \approx y(x_n),\\
&\ldots\ldots\ldots\\
&y_N \approx y(x_N)
\end{aligned}$$

теоретическая погрешность определения y_n на каждом шаге также равна $O(h^2)$ (подчеркнем, что мы говорим лишь о погрешности метода, которая определяется без учета погрешности нахождения $f(x_{n-1}, y_{n-1})$ и без учета округлений при вычислениях). Но погрешность может накапливаться, и, например, мы можем предположить лишь, что $y_N = y(x_N) + O(h)$, так как $N = (b - x_0)/h$ — величина порядка $1/h$. Таким образом, эти эвристические рассуждения дают, что погрешность метода Эйлера на конечном отрезке имеет порядок $O(h)$ (без учета погрешности вычислений).

6.2 Обобщения метода Эйлера

Рассмотрим обобщения метода Эйлера, основанные на применении квадратурных формул. Как мы убедились ранее, задача Коши (6.1) — (6.2) эквивалентна интегральному уравнению (6.3), т. е. уравнению

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))dt.$$

Отсюда следует

$$y(x + h) - y(x) = \int_{x}^{x+h} f(t, y(t))dt.$$

Основная идея такова: интеграл правой части вычисляем приближенно с помощью квадратурных формул,

используя, например, малые квадратурные формулы прямоугольников, трапеций и т. п. на каждом частичном отрезке $[x_{n-1},\, x_n]$, определяемом узлами

$$x_n = x_0 + nh, \quad n = 0, 1, \ldots, N \quad (h = (b - x_0)/N).$$

Для краткости мы будем использовать при этом обозначение

$$\varphi(t) = f(t, y(t)).$$

6.2.1 Метод квадратур

1) *Применим сначала малую формулу левых прямоугольников.*

Можем записать

$$\int_x^{x+h} \varphi(t)dt = \varphi(x)h + R_1.$$

Оценим погрешность, считая, что функция φ достаточное число раз дифференцируема. Так как

$$\int_x^{x+h} \varphi(x)dt = \varphi(x) \int_x^{x+h} dt = \varphi(x)h,$$

то будем иметь

$$R_1 = \int_x^{x+h} [\varphi(t) - \varphi(x)]\, dt = \int_x^{x+h} dt \int_x^t \varphi'(\tau)d\tau =$$

$$= \int_x^{x+h} \varphi'(\tau)d\tau \int_\tau^{x+h} dt = \int_x^{x+h} \varphi'(\tau)(x+h-\tau)d\tau.$$

По теореме о среднем для интегралов существует точка

$$\xi \in [x, x+h],$$

такая, что

$$R_1 = \varphi'(\xi)\left(x\tau + h\tau - \frac{\tau^2}{2}\right)\Bigg|_x^{x+h} = \varphi'(\xi)\frac{h^2}{2} = O(h^2).$$

Таким образом,

$$\int_x^{x+h} f(t, y(t))dt = hf(x, y(x)) + O(h^2),$$

и поэтому можем записать дифференциальное уравнение (6.1) в виде равенства

$$y(x+h) - y(x) = hf(x, y(x)) + O(h^2).$$

Полагая $x = x_n$ и $y_n \approx y(x_n)$, мы приходим к следующему алгоритму для определения y_{n+1}:

$$y_{n+1} = y_n + h\, f(x_n,\, y_n),$$

т. е. мы получаем по-другому алгоритм Эйлера.

2) *Применим теперь малую формулу правых прямоугольников.* По аналогии с предыдущим случаем по-

лучаем формулу

$$y(x+h) - y(x) = hf(x+h,\, y(x+h)) + O(h^2).$$

Полагая в ней $x = x_0, \ldots, x_n = x_0 + nh, \ldots$, приходим к следующему алгоритму

$$\begin{cases} y_1 = y_0 + hf(x_1, y_1) \\ y_2 = y_1 + hf(x_2, y_2) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}) \end{cases}$$

На каждом шаге y_{n+1} определяется как решение уравнения (в общем случае нелинейного)

$$y_{n+1} = y_n + h\, f(x_{n+1}, y_{n+1}).$$

Полученный алгоритм называют неявным методом Эйлера. Ясно, что погрешность метода на каждом шаге ($\equiv$ невязка) равна $O(h^2)$, а погрешность метода на отрезке имеет порядок малости $O(h)$.

Неявный метод Эйлера можно превратить в явный, если y_{n+1} в правой части равенства

$$y_{n+1} = y_n + h\, f(x_{n+1}, y_{n+1})$$

заменим его выражением

$$y_{n+1} = y_n + h\, f(x_n,\, y_n)$$

из явного метода Эйлера. Получим новый явный алгоритм вида

$$y_{n+1} = y_n + hf(x_{n+1},\, y_n + hf(x_n,\, y_n)).$$

3) Применим теперь малую формулу трапеций. С точностью до $O(h^3)$ имеем

$$y(x+h) - y(x) \approx h\frac{f(x, y(x)) + f(x+h, y(x+h))}{2},$$

отсюда для определения $y_{n+1} \approx y(x_{n+1})$ получаем следующий численный алгоритм:

$$y_{n+1} = y_n + \frac{h}{2}\left[f(x_n,\, y_n) + f(x_{n+1},\, y_{n+1})\right].$$

Понятно, что возник новый неявный метод, обобщающий метод Эйлера, причем погрешность нового метода равна $O(h^2)$ на отрезке, так как невязка имеет порядок малости $O(h^3)$.

Комбинируя этот неявный метод с явным методом Эйлера, мы получаем еще один явный метод, задаваемый формулой

$$y_{n+1} = y_n + \frac{h}{2}\, f(x_n,\, y_n) \,+ \frac{h}{2}\; f(x_{n+1},\, y_n \,+\, h\, f(x_n,\, y_n)).$$

Последняя формула представляет собой один из простейших алгоритмов, получаемых методом Рунге-Кутта. Погрешность этой формулы на каждом шаге

оказывается равной $O(h^3)$, хотя вывод этой формулы гарантирует лишь порядок $O(h^2)$ на каждом шаге из-за того, что мы использовали формулу из явного метода Эйлера с этой погрешностью.

6.2.2 Метод Рунге-Кутта (общая схема)

Все три формулы предыдущего пункта для приближенного интегрирования можно записать в едином виде. А именно, имеем

$$\int_x^{x+h} \varphi(t)dt \approx A_0\,\varphi(x) + A_1\,\varphi(x+h),$$

где

$$\varphi(t) = f(t, y(t)),$$

причем $A_0 = 1, A_1 = 0$ для формулы левых прямоугольников, $A_0 = 0, A_1 = 1$ для формулы правых прямоугольников и

$$A_0 = A_1 = \frac{1}{2}$$

для формулы трапеций. Отметим, что во всех случаях $A_0 + A_1 = 1$.

Понятно, что число подобных формул можно увеличить, применяя, например, формулы средних прямоугольников или малую формулу Симпсона.

Идея метода Рунге-Кутта состоит в применении но-

вых квадратурных формул вида

$$\int_x^{x+h} \varphi(t)dt \approx \sum_{i=0}^{q} A_i\,\varphi_i,$$

где q — заданное натуральное число, значения X_i и Y_i в формулах $\varphi_i = hf(X_i, Y_i)$ подбираются специальным образом.

Общая схема такова. Рассматриваются три группы констант:

(A) $\quad A_0, A_1, \ldots, A_q \qquad (\sum_{i=0}^{q} A_i = 1);$

(α) $\quad \alpha_1, \alpha_2, \ldots, \alpha_q;$

(β)

$$\begin{cases} \beta_{10} \\ \beta_{20}, \beta_{21} \\ \ldots\ldots\ldots \\ \beta_{q0}, \beta_{q1}, \ldots, \beta_{q(q-1)}. \end{cases}$$

Функции φ_i определяются следующим образом:

$$\begin{cases} \varphi_0 = h\,f(x, y(x)), \\ \varphi_1 = h\,f(x + \alpha_1 h,\ y(x) + \beta_{10}\varphi_0), \\ \varphi_2 = h\,f(x + \alpha_2 h,\ y(x) + \beta_{20}\varphi_0 + \beta_{21}\varphi_1), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \varphi_q = h\,f(x + \alpha_q h,\ y(x) + \sum_{i=0}^{q-1} \beta_{qi}\varphi_i). \end{cases}$$

Как и ранее, мы заменяем точное уравнение на прибли-

женное

$$y(x+h) - y(x) \approx \sum_{i=0}^{q} A_i \varphi_i,$$

тогда погрешность метода на каждом шаге определится формулой вида

$$r_q(h) = y(x+h) - y(x) - \sum_{i=0}^{q} A_i \varphi_i. \qquad (6.4)$$

Вопрос о том, каким образом определяются числовые параметры является очень важным, и мы займемся этой задачей чуть ниже. Предположим пока, что постоянные из групп $(A), (\alpha), (\beta)$ заданы. Тогда для приближенного решения задачи Коши (6.1) – (6.2) возникает следующий алгоритм Рунге-Кутта:

$$y_{n+1} = y_n + \sum_{i=0}^{q} A_i \varphi_i, \qquad n = 0, 1, \ldots, N-1,$$

где φ_i определяются выписанными формулами, в которых вместо x берем x_n, а вместо $y(x)$ подставляем $y_n \approx y(x_n)$, т. е.

$$\begin{cases} \varphi_0 = h\, f(x_n, y_n), \\ \varphi_1 = h\, f(x_n + \alpha_1 h,\ y_n + \beta_{10}\varphi_0), \\ \varphi_2 = h\, f(x_n + \alpha_2 h,\ y_n + \beta_{20}\varphi_0 + \beta_{21}\varphi_1), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \varphi_q = h\, f(x_n + \alpha_q h,\ y_n + \sum_{i=0}^{q-1} \beta_{q_i}\varphi_i). \end{cases}$$

Так формально выглядит алгоритм численного метода Рунге-Кутта в общем случае, если постоянные из групп $(A), (\alpha), (\beta)$ заданы. Но эти константы выбираются специальным образом, точнее, их определяют как решения систем уравнений, получаемых из некоторых систем тождеств.

Рассмотрим этот вопрос подробнее.

Основной становится следующая цель: минимизация погрешности метода (невязки) $r_q(h)$ за счет выбора констант из групп $(A), (\alpha), (\beta)$. Естественной является следующая

Задача. При заданном q нужно подобрать коэффициенты $(A), (\alpha), (\beta)$ так, чтобы $r_q(h) = O(h^{k+1})$, т. е. так, чтобы погрешность метода на отрезке $[x_0, b]$ имела порядок $O(h^k)$, $k \in N$, с максимально возможным числом k.

Понятно, что для этого нужно добиться, чтобы погрешность на каждом шаге равнялась $r_q(h) = O(h^{k+1})$ для любых достаточно гладких функций $f(x, y)$. Это максимальное число k принято называть *порядком метода Рунге-Кутта.*

Очевидно, $r_q(0) = 0$. Далее мы предполагаем, что функция f достаточное число раз дифференцируема. Тогда по формуле Тейлора

$$r_q(h) = r'_q(0)h + \frac{r''_q(0)}{2}h^2 + \ldots + \frac{r_q^{(k)}(0)}{k!}h^k + \frac{r_q^{(k+1)}(\xi)}{(k+1)!}h^{k+1}.$$

Потребуем, чтобы

$$r_q(h) = O(h^{k+1}),$$

т. е. потребуем выполнения равенств

$$r_q'(0) = 0, \ldots, r_q^{(k)}(0) = 0.$$

Таким образом, имеем k равенств, и постараемся выбрать неизвестные постоянные таким образом, чтобы эти равенства превратились в тождества.

По формуле Тейлора

$$y(x+h) - y(x) =$$

$$= y'(x)h + \frac{y''(x)}{2!}h^2 + \frac{y'''(x)}{3!}h^3 + \ldots + \frac{y^{(k)}(x)}{k!}h^k + O(h^{k+1}).$$

и

$$\sum_{i=0}^{q} A_i\varphi_i = c_1\,h + c_2\,h^2 + \ldots + c_k\,h^k \;+\; O(h^{k+1}),$$

поэтому равенства $r_q^{(i)}(0) = 0$ для $i = 1, \ldots, k$ запишутся в виде:

$$\begin{cases} c_1 = y'(x) \\ c_2 = \frac{y''(x)}{2} \\ \ldots\ldots\ldots\ldots \\ c_k = \frac{y^{(k)}(x)}{k!} \end{cases}$$

6.2.3 Алгоритмы Рунге-Кутта порядка 4, а также 2 и 3

Снова рассматриваем задачу Коши (6.1) — (6.2) на отрезке $[x_0, b]$. Выбираем шаг $h = \frac{b-x_0}{N}$ и узлы

$$x_n = x_0 + nh, \qquad n = 0, \ldots, N, \qquad x_n = x_0 + Nh = b.$$

Самым употребительным для решения задачи Коши (6.1) — (6.2) является метод Рунге-Кутта со следующим алгоритмом:

$$y_{n+1} = y_n + \frac{1}{6}(\varphi_0 + 2\varphi_1 + 2\varphi_2 + \varphi_3), \quad n = 0, \ldots, N-1,$$

где

$$\begin{cases} \varphi_0 = hf(x_n, y_n) \\ \varphi_1 = hf(x_n + \frac{h}{2}, y_n + \frac{\varphi_0}{2}) \\ \varphi_2 = hf(x_n + \frac{h}{2}, y_n + \frac{\varphi_1}{2}) \\ \varphi_3 = hf(x_n + h, y_n + \varphi_2) \end{cases}$$

Очевидно, этот алгоритм соответствует общему алгоритму, когда константы $(A), (\alpha), (\beta)$ даны равенствами

$$\begin{aligned} &A_0 = A_3 = \tfrac{1}{6}, \\ &A_1 = A_2 = \tfrac{1}{3}, \\ &\alpha_1 = \alpha_2 = \tfrac{1}{2}, \\ &\alpha_3 = 1, \\ &\beta_{10} = \beta_{21} = \tfrac{1}{2}, \\ &\beta_{32} = 1, \end{aligned}$$

а остальные константы $\beta_{ij} = 0$.

Таким образом, приведенный алгоритм соответствует случаю $q = 3$. Прямыми громоздкими вычислениями можно убедиться, что порядок этого алгоритма Рунге-Кутта равен четырем. Именно этот алгоритм Рунге-Кутта приводится в справочниках и широко используется на практике.

Последовательно рассмотрим случаи $q = 0$, $q = 1$ и $q = 2$, и найдем для этих случаев порядок $k = k(q)$ метода Рунге-Кутта. Как мы увидим, если $q \geq 1$, то система уравнений для определения числовых параметров является нелинейной и имеет, как правило, множество решений.

I. Метод Рунге-Кутта при $q = 0$. Имеем

$$y_{n+1} = y_n + A_0 h f(x_n, y_n),$$

$$r_0(h) = y(x + h) - y(x) - A_0 h f(x, y),$$

$r_0(0) = 0$, $x_n = x$, $y_n = y$.

Выпишем производные по h и приравняем к нулю их значения в точке $h = 0$. Имеем

$$r_0'(h) = y'(x + h) - A_0 f(x, y),$$

и

$$r_0''(h) = y''(x + h).$$

Очевидно

$$r_0^{(j)}(h) = y^{(j)}(x+h), \qquad r_0'(0) = 0,$$

отсюда $y'(x) = A_0 f(x,y)$, следовательно, $A_0 = 1$.

Таким образом, метод Рунге-Кутта при $q = 0$ эквивалентен методу Эйлера

$$y_{n+1} = y_n + hf(x_n, y_n).$$

Возникает вопрос: можно ли получить $k = 2$?

$$r_0''(0) = 0, \text{ отсюда } 0 = y''(x) = f_x' + f_y' y' = f_x' + f_y' f.$$

В общем случае $f_x' + f_y' f$ не эквивалентен нулю, таким образом, порядок $k = 2$ невозможен в общем случае. Но можно отметить, что для уравнений специального вида порядок $k = 2$ не исключен, например, тождество $f_x' + f_y' f \equiv 0$ верно для функции

$$f(x,y) = \frac{y}{x}, \qquad f_x' + f_y' f = -\frac{y}{x^2} + \frac{1}{x} \cdot \frac{y}{x}.$$

II. Метод Рунге-Кутта при $q = 1$.

Надо определить 4 постоянных A_0, A_1, α_1 и β_{10}. Проведем необходимые вычисления. Имеем:

$$r_1(h) = y(x+h) - y(x) -$$

$$-A_0 h f(x,y) - A_1 h f(x + \alpha_1 h, y + \beta_{10} h f(x,y)).$$

Поскольку

$$y(x+h) - y(x) = y'(x)h + y''(x)\frac{h^2}{2} + O(h^3) =$$

$$= f(x,y)h + [f'_x + f'_y f]\frac{h^2}{2!} + O(h^3),$$

$$A_0\varphi_0 + A_1\varphi_1 = A_0 hf + A_1 h[f + (f'_x\alpha_1 + f'_y\beta_{10}f)h + O(h^2)] =$$

$$= (A_0 + A_1)hf + A_1\alpha_1 f'_x h^2 + A_1\beta_{10}f'_y f h^2 + O(h^3),$$

то условие $r'_1(0) = 0$ приводит к равенству

$$(A_0 + A_1)f(x,y) = f(x,y)$$

для любых f, следовательно, $A_0 + A_1 = 1$.

Равенство $r''_1(0) = 0$ равносильно соотношению

$$A_1\alpha_1 f'_x + A_1\beta_{10}f'_y f = \frac{f'_x + f'_y f}{2}$$

(здесь и далее аргументы у функций f, f'_x, f'_y, т. е. x и y, для краткости записи опускаем).

В общем случае f, f'_x, f'_y являются линейно независимыми функциями. Поэтому равенство $r''_1(0) = 0$ (рассматриваемое как тождество, которое должно выполняться для всех допустимых значений x и y) приводит к двум уравнениям, получаемым путем приравни-

вания коэффициентов при функциях f'_x и $f'_y f$. Имеем

$$\begin{cases} A_1\alpha_1 = \frac{1}{2}, \\ A_1\beta_{10} = \frac{1}{2}, \\ A_0 + A_1 = 1. \end{cases}$$

Получили систему из 3-х уравнений с 4-мя неизвестными. Эта система имеет однопараметрическое множество решений. Можно взять, например, $A_1 \in (0,1)$ как параметр. Тогда

$$A_0 = 1 - A_1,$$
$$\alpha_1 = \beta_{10} = \frac{1}{2A_1}.$$

Укажем наиболее простые частные случаи, которым соответствуют различные алгоритмы Рунге-Кутта порядка $k = 2$, получаемые при $q = 1$.

а) Числовые значения

$$A_0 = A_1 = \frac{1}{2}, \quad \alpha_1 = \beta_{10} = 1,$$

приводят к следующему алгоритму Рунге-Кутта порядка $k = 2$:

$$y_{n+1} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_n + h, y_n + hf(x_n, y_n))].$$

б) Значениям

$$A_0 = 0, \quad A_1 = 1, \quad \alpha_1 = \beta_{10} = \frac{1}{2}$$

соответствует такой алгоритм Рунге-Кутта:

$$y_{n+1} = y_n + h\, f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2} f(x_n, y_n)\right),$$

получаемый в случае $q = 1$ с гарантированной оценкой погрешности $r_1(h) = O(h^3)$, т. е. порядок этого метода $k = 2$.

Отметим, что A_1 можно выбирать также из следующих соображений: представив погрешность в виде

$$r_1(h) = kh^3 + O(h^4),$$

минимизируем k за счет выбора A_1.

III. Методы Рунге-Кутта при $q = 2$.

Имеем

$$y_{n+1} = y_n + (A_0\varphi_0 + A_1\varphi_1 + A_2\varphi_2),$$

где

$$\begin{cases} \varphi_0 = hf(x, y) \\ \varphi_1 = hf(x + \alpha_1 h, y + \beta_{10}\varphi_0) \\ \varphi_2 = hf(x + \alpha_2 h, y + \beta_{20}\varphi_0 + \beta_{21}\varphi_1) \end{cases}$$

(предполагаем, что в этих формулах $x_n = x$, $y_n = y$). Имеем 8 неизвестных постоянных: A_0, A_1, A_2, α_1, α_2, $\beta_{10}, \beta_{20}, \beta_{21}$. Можно показать, что условие

$$r_2'(0) = r_2''(0) = r_2'''(0) = 0$$

будет выполнено при следующем выборе постоянных:

$$A_0 = A_1 = \frac{1}{6}, \quad A_2 = \frac{2}{3},$$

$$\alpha_1 = \beta_{10} = \frac{1}{2}, \quad \alpha_2 = 1, \quad \beta_{20} = -1, \quad \beta_{21} = 2.$$

Тогда $r_2(h) = O(h^4)$, следовательно, порядок метода $k = 3$.

Проведем вычисления для получения общего решения соответствующей системы тождеств. Имеем

$$r_2(h) = y(x+h) - y(x) - \sum_{i=0}^{2} A_i \varphi_i,$$

$$y(x+h) - y(x) = y'(x)h + y''(x)\frac{h^2}{2} + y'''(x)\frac{h^3}{6} + O(h^4) =$$

$$= f(x,y)h + [f'_x(x,y) + f'_y(x,y)f(x,y)]\frac{h^2}{2} +$$

$$+[f''_{xx}(x,y) + 2f''_{xy}(x,y)f(x,y) + f''_{yy}(x,y)f^2(x,y) +$$

$$+f'_y(x,y)f'_x(x,y) + f'^2_y(x,y)f(x,y)]\frac{h^3}{6} + O(h^4),$$

$$\frac{d\varphi_0}{dh} = f(x,y) = \varphi'_0, \quad \varphi''_0 = \varphi'''_0 = \ldots = 0,$$

$$\varphi'_1 = f(x+\alpha_1 h, y+\beta_{10}\varphi_0) + h[f'_x(x+\alpha_1 h, y+\beta_{10}\varphi_0)\alpha_1 +$$

$$+f'_y(x+\alpha_1 h, y+\beta_{10}\varphi_0)\beta_{10} f(x,y)].$$

Следовательно,

$$\varphi_1'(0) = f(x,y), \quad \varphi_2'(0) = f(x,y).$$

Тогда первое тождество имеет вид

$$f = A_0 f + A_1 f + A_2 f,$$

откуда следует, что

$$A_0 + A_1 + A_2 = 1.$$

Напомним, что мы рассматриваем случай $q = 2$. Цель заключается в том, чтобы порядок метода получился равным $k = 3$. Нужно определить подходящие постоянные $A_0, A_1, A_2, \alpha_1, \alpha_2, \beta_{10}, \beta_{20}, \beta_{21}$. Имеем $\varphi_0 = hf$, где $f = f(x,y)$, и далее

$$\varphi_1(h) = hf + (\alpha_1 f_x' + \beta_{10} f_y' f)h^2 +$$

$$(\alpha_2^2 f_{xx}'' + 2\alpha_1\beta_{10} f f_{xy}'' + \beta_{10}^2 f_{yy}'' f^2)\frac{h^3}{2} + O(h^4),$$

$$\varphi_2(h) = hf + (\alpha_2 f_x' + \beta_{20} f_y' f + \beta_{21} f_y' f)h^2 +$$

$$+[\alpha_2^2 f_{xx}'' + 2\alpha_2(\beta_{20} + \beta_{21}) f f_{xy}' + (\beta_{20} + \beta_{21})^2 f f_{yy}' +$$

$$+2\beta_{21} f_y'(\alpha_1 f_x' + \beta_{10} f_y' f)]\frac{h^3}{2} + O(h^4).$$

Приравниваем подобные члены в разложениях по сте-

пеням h для функций

$$y(x+h)-y(x) \quad \text{и} \quad \sum_{i=0}^{2} A_i\varphi_i,$$

и получаем следующие уравнения для определения 8-и постоянных $A_0, A_1, A_2, \alpha_1, \alpha_2, \beta_{10}, \beta_{20}, \beta_{21}$:

$$\begin{cases} A_0 + A_1 + A_2 = 1 & (1) \\ \alpha_1 A_1 + \alpha_2 A_2 = \frac{1}{2} & (2) \\ \alpha_1^2 A_1 + \alpha_2^2 A_2 = \frac{1}{3} & (3) \end{cases}$$

$$\begin{cases} \beta_{10} A_1 + (\beta_{20} + \beta_{21}) A_2 = \frac{1}{2} & (4) \\ \alpha_1 \beta_{10} A_1 + \alpha_2 (\beta_{20} + \beta_{21}) A_2 = \frac{1}{3} & (5) \\ \beta_{10}^2 A_1 + (\beta_{20} + \beta_{21})^2 A_2 = \frac{1}{3} & (6) \end{cases}$$

$$\begin{cases} \alpha_1 \beta_{21} A_2 = \frac{1}{6} & (7) \\ \beta_{10} \beta_{21}) A_2 = \frac{1}{6} & (8) \end{cases}$$

Из (7) и (8) следует, что $\alpha_1 \neq 0$, $\beta_{21} \neq 0$, $A_2 \neq 0$, $\beta_{10} \neq 0$.

Поделим (7) на (8), получим

$$\alpha_1 = \beta_{10}, \qquad (4')$$

$$\beta_{21} = \frac{1}{6\alpha_1 A_2}. \qquad (5')$$

Поскольку $\alpha_1 = \beta_{10}$, то из (2) и (4) следует, что

$$\beta_{20} + \beta_{21} = \alpha_2 \qquad (6').$$

Ясно, что уравнения (4), (5), (6) являются теперь следствиями уравнений (2) и (3).

Остается система из шести уравнений. А именно, к $(1), (2), (3)$ добавляются уравнения $(4'), (5'), (6')$, где

$$\begin{cases} \beta_{10} = \alpha_1 & (4') \\ \beta_{21} = \frac{1}{6\alpha_1 A_2} & (5') \\ \beta_{20} = \alpha_2 - \beta_{21} & (6') \end{cases}$$

Рассмотрим уравнения (1) — (3), не содержащих величин $\beta_{10}, \beta_{20}, \beta_{21}$. В системе (1) — (3) имеем пять неизвестных, а именно, $A_0, A_1, A_2, \alpha_1, \alpha_2$, а число уравнений равно 3. Ясно, что решение такой системы не является единственным, на самом деле существует бесконечное множество решений. А именно, можно взять за свободные параметры величины A_1, α_1 и найти все остальные. Таким образом, приходим к выводу о том, что полученная система будет иметь 2-х параметрическое семейство решений.

Одно из решений было указано выше:

$$A_0 = A_2 = \frac{1}{6}, \quad A_1 = \frac{2}{3},$$

$$\alpha_1 = \beta_{10} = \frac{1}{2}, \quad \alpha_2 = 1, \quad \beta_{20} = -1, \quad \beta_{21} = 2.$$

Укажем еще одно решение решение:

$$A_0 = \frac{1}{4}, \quad A_1 = 0, \quad A_2 = \frac{3}{4},$$

$$\alpha_1 = \beta_{10} = \frac{1}{3}, \quad \alpha_2 = \beta_{21} = \frac{2}{3}, \quad \beta_{20} = 0.$$

Соответствующий этому случаю алгоритм Рунге-Кутта имеет вид:

$$y_{n+1} = y_n + \frac{h}{4} f(x_n, y_n) +$$
$$+ \frac{3h}{4} f\left(x_n + \frac{2}{3}h, y_n + \frac{2}{3}hf\left(x_n + \frac{h}{3}, y_n + \frac{h}{3}f(x_n, y_n)\right)\right).$$

6.3 Оценки погрешности

Мы снова рассматриваем задачу Коши (6.1) — (6.2) о нахождении решения дифференциального уравнения

$$y' = f(x, y)$$

на отрезке $[x_0, b]$, удовлетворяющего условию

$$y|_{x=x_0} = y_0.$$

Напомню, что численные алгоритмы связаны с делением отрезка $[x_0, b]$ на N равных частей, с шагом

$$h = \frac{b - x_0}{N}$$

и с узлами

$$x_n = x_0 + nh, \qquad n = 0, \ldots, N \qquad (x_N = x_0 + Nh = b),$$

используемых при вычислении приближенных значений решения $y_n \approx y(x_n)$.

Метод численного решения задачи (6.1) — (6.2) называется одношаговым, если y_{n+1} определяется с использованием y_n формулой вида

$$y_{n+1} = y_n + h\Phi_f(h, x_n, y_n) \tag{6.5}$$

— *явная схема (явный метод)*, где Φ_f определяется методом. Также возможны *одношаговые неявные алгоритмы* вида

$$y_{n+1} = y_n + hF_f(h, x_n, y_n, y_{n+1}, x_{n+1}),$$

при этом y_{n+1} определяется как корень этого уравнения.

Отметим, что существуют многошаговые методы решения задачи (6.1) — (6.2). В многошаговых методах начало таблицы значений задано, т. е. известны

$$y_0, \quad y_1 \approx y(x_1), \ldots, y_{n-1} \approx y(x_{n-1}),\ y_n \approx y(x_n)$$

и явный k-шаговый метод задается алгоритмом вида

$$y_{n+1} = y_n + h\Psi_f(h, y_{n-k+1}, y_{n-k+2}, \ldots, y_n)$$

для $n = k-1, \ldots, N-1$.

Далее рассмотрим приближенное решение задачи (6.1) — (6.2) каким-либо явным одношаговым методом с алгоритмом вида

$$y_{n+1} = y_n + h\Phi_f(h, x_n, y_n), \quad n = 0, \ldots, N-1.$$

Определение 6.1 *Погрешностью метода при заданном шаге $h = (b - x_0)/N$ называется следующая величина*

$$\varepsilon(h) = \max_{1 \le n \le N-1} |\varepsilon_n|,$$

где $\varepsilon_n := y(x_n) - y_n$.

Нам потребуется также

Определение 6.2 *Пусть $y_n(x)$ — точное решение задачи Коши:*

$$y_n' = f(x, y_n), \quad x_n \le x \le b,$$

$$y_n(x_n) = y_n, \quad n = 0, 1, 2, \ldots, N-1.$$

В частности, при $n = 0$ имеем $y_0(x) = y(x)$ — решение (6.1) — (6.2). Невязкой называют величину

$$r_n := y_n(x_{n+1}) - y_{n+1}.$$

Задача. Оценить погрешность метода $\varepsilon(h)$ при условии, что известны свойства функции $f(x, y)$ и поведение невязки r_n для малых h. Стандартные требования

на f и r_n таковы:

$$f : [x_0, b] \times \mathbb{R} \to \mathbb{R}$$

— непрерывная функция, и

$$\left| \frac{\partial f(x,y)}{\partial y} \right| \leq L = \text{Const}$$

(следовательно, f удовлетворяет условию Липшица по второй переменной);

выполняется условие

$$\lim_{h \to 0} \frac{r(h)}{h} = 0$$

для функции $r(h) := \max_{1 \leq n \leq N-1} |r_n|$.

Напомню, что в известных нам методах оценки для невязки $r(h)$ определялись ранее. В частности, имеем

$$r(h) = O(h^2)$$

в методе Эйлера или

$$r(h) = O(h^{k+1}), \quad (k = 2, 3, 4)$$

в методе Рунге-Кутта порядка k.

Потребуются следующие предварительные утверждения.

Лемма 6.1 *Пусть* $[\alpha, \beta] \subset [x_0, b]$, *функция* f *диффе-*

ренцируема по y. *Пусть, далее, функции* Y *и* Z – *несовпадающие решения уравнения* $y' = f(x, y)$ *на отрезке* $[\alpha, \beta]$. *Тогда имеет место формула*

$$Y(\beta) - Z(\beta) = [Y(\alpha) - Z(\alpha)] \exp \int_\alpha^\beta \left(\frac{\partial f}{\partial y}\right)_{\Theta(x)} dx,$$

где использовано обозначение

$$\left(\frac{\partial f}{\partial y}\right)_{\Theta(x)} = f'_y(x, Y(x) + \Theta(x) \cdot (Z(x) - Y(z))),$$

с некоторой $\Theta(x) \in (0, 1)$.

Доказательство. Так как

$$Y'(x) = f(x, Y(x)), \quad Z'(x) = f(x, Z(x)),$$

то имеем

$$Y'(x) - Z'(x) = f(x, Y(x)) - f(x, Z(x)).$$

Применяя формулу Лагранжа о конечных приращениях к функции f по второй переменной, получаем

$$Y'(x) - Z'(x) = \left(\frac{\partial f}{\partial y}\right)_{\Theta(x)} [Y(x) - Z(x)],$$

что влечет равенство

$$\frac{d}{dx} \ln[Y - Z] = \left(\frac{\partial f}{\partial y}\right)_{\Theta(x)}.$$

Интегрируя от α до β и применяя экспоненту к обеим частям получаемого равенства, приходим к формуле

$$\ln \frac{Y(\beta) - Z(\beta)}{Y(\alpha) - Z(\alpha)} = \int_\alpha^\beta \left(\frac{\partial f}{\partial y}\right)_{\Theta(x)} dx.$$

Следовательно,

$$\frac{Y(\beta) - Z(\beta)}{Y(\alpha) - Z(\alpha)} = \exp \int_\alpha^\beta \left(\frac{\partial f}{\partial y}\right)_{\Theta(x)} dx,$$

что и требовалось доказать.

Лемма 6.2 *Для любого* $n = 1, 2, \ldots, N$ *справедливо равенство*

$$\varepsilon_n = \sum_{i=1}^{n} r_{i-1} e^{\int_{x_i}^{x_n} \left(\frac{\partial f}{\partial y}\right)_{\Theta_i} dx}, \quad \Theta_i(x) \in (0, 1).$$

Доказательство. Так как $y(x_n) = y_0(x_n)$, то

$$\varepsilon_n = y(x_n) - y_n = y_0(x_n) - y_n =$$

$$= y_0(x_n) - y_1(x_n) + y_1(x_n) - y_2(x_n) + y_2(x_n) -$$

$$-y_3(x_n) + \ldots + y_{i-1}(x_n) - y_i(x_n) + \ldots + y_{n-1}(x_n) - y_n.$$

Мы провели элементарные преобразования, а именно, вычитали и добавляли $y_{i-1}(x_n)$ при $i = \overline{2, n}$. Заметим также, что $y_i = y_i(x_i)$ по определению функции $y_i(x)$.

Согласно лемме 6.1 при выборе $\beta = x_n$, $\alpha = x_i$,

будем иметь

$$y_{i-1}(x_n) - y_i(x_n) = [y_{i-1}(x_i) - y_i(x_i)] \cdot e^{\int_{x_i}^{x_n} \left(\frac{\partial f}{\partial y}\right)_{\Theta_i} dx} =$$

$$= [y_{i-1}(x_i) - y_i] \cdot e^{\int_{x_i}^{x_n} \left(\frac{\partial f}{\partial y}\right)_{\Theta_i} dx} =$$

$$= r_{i-1} e^{\int_{x_i}^{x_n} \left(\frac{\partial f}{\partial y}\right)_{\Theta_i} dx}.$$

Суммируем полученные соотношения $i = \overline{1, n}$. Получаем искомую формулу. Тем самым, лемма доказана.

С использованием этих лемм докажем теперь основное утверждение этого пункта об оценке погрешности одношагового метода решения задачи Коши при условии, что невязка имеет вполне определенное поведение. Напомним, что

$$\varepsilon(h) = \max_{0 \le n \le N-1} |\varepsilon_n|, \qquad h = \frac{b - x_0}{N} - \text{шаг}.$$

Справедлива

Теорема 6.3 *Пусть y_n — приближенное значение $y(x_n)$ решения задачи (6.1) — (6.2) с использованием некоторого одношагового метода вида*

$$y_{n+1} = y_n + h\Phi_f(x_n, y_n, h).$$

Пусть, далее,

$$\left|\frac{\partial f}{\partial y}\right| \le L = Const \quad \forall x \in [x_0, b], y \in \mathbb{R},$$

существует мажоранта $r(h)$, такая, что $|r_i| \le r(h)$ для любого $i = 1, 2, \ldots, N-1$ и

$$\lim_{h\to 0} \frac{r(h)}{h} = 0.$$

Тогда имеет место оценка

$$\varepsilon(h) = O\left(\frac{r(h)}{h}\right).$$

Доказательство. В силу леммы 6.2

$$|\varepsilon_n| \le \sum_{i=1}^{n} |r_{i-1}| e^{\int_{x_i}^{x_n} \left(\frac{\partial f}{\partial y}\right)_{\Theta_i} dx}.$$

Поскольку

$$|r_{i-1}| \le r(h), \quad \left|\frac{\partial f}{\partial y}\right| \le L, \quad n \le N,$$

то получаем

$$|\varepsilon_n| \le r(h) \sum_{k=1}^{n} e^{L(x_n - x_k)} \le$$

$$\le r(h) \sum_{k=1}^{n} e^{L(b-x_0)} \le r(h) e^{L(b-x_0)} N.$$

Но $N = (b - x_0)/h$, поэтому

$$|\varepsilon_n| \le r(h) e^{L(b-x_0)} \frac{b - x_0}{h} = C\,\frac{r(h)}{h},$$

где

$$C = (b - x_0)e^{L(b-x_0)}.$$

Следовательно,

$$\varepsilon(h) \le C\frac{r(h)}{h},$$

поэтому можем записать

$$\varepsilon(h) = O\left(\frac{r(h)}{h}\right).$$

Имеем полезное

Следствие 6.3.1 *Пусть $k \ge 1$ — натуральное число. Если мажоранта $r(h) = O(h^{k+1})$, то погрешность метода $\varepsilon(h) = O(h^k)$.*

6.4 Правило Рунге для выбора шага

Предположим, что задача Коши (6.1) — (6.2) решается некоторым численным одношаговым методом, и мы теоретически знаем, что $r(h) = O(h^{k+1})$, $k \ge 1$. Точнее предполагаем известным следующий факт: для невязки справедлива формула

$$r_n = \varphi(x_n, y_n)h^{k+1} + O(h^{k+2})$$

при $h \to 0$, где φ — ограниченная функция. Из доказательства теоремы предыдущего пункта получаем, что

погрешность может быть представлена формулой вида

$$\varepsilon_n = \varrho(x_n)h^k + O(h^{k+1}),$$

где ϱ — ограниченная функция. Эта величина $\varrho(x_n)h^k$ называется главным членом погрешности.

В предположении, что φ, $\frac{\partial f}{\partial y}$ непрерывно дифференцируемые функции по обеим переменным и $k > 1$, можно показать, что

$$\varrho(x_n) = \int_{x_0}^{x_n} \varphi(t, y(t)) e^{\int_t^{x_n} f'_y(\tau, y(\tau))d\tau} dt.$$

Чтобы вычислить главный член погрешности, нужно знать явный вид $\varphi(x_n, y_n)$, а не просто существование функции φ и ее ограниченность.

Рунге предложил весьма практичный численный метод для оценки главного члена погрешности. Приведем эвристические рассуждения Рунге.

Предположим, что известен вид главного члена погрешности

$$\varepsilon_n = \varrho(x_n)h^k + O(h^{k+1}), \quad k \geq 1,$$

но сама функция $\varrho(x)$ неизвестна.

Рассматриваем какой-либо численный одношаговый метод. Пусть точка $\xi \in [x_0, b]$, $\xi > x_0$, является общей узловой точкой при разбиении отрезка $[x_0, b]$ с применением двух различных шагов $h_1 > 0$ и $h_2 > 0$, $h_1 \neq h_2$.

Вычисляем приближенное значение решения задачи (6.1) — (6.2) в узловой точке $\xi > x_0$ дважды по единому алгоритму с шагом h_1 и с шагом h_2. Предполагаем, что

$$\begin{cases} \xi = x_0 + n_1 h_1 \\ \xi = x_0 + n_2 h_2 \end{cases},$$

где $n_1, n_2 \in \mathbb{N}$, y_{h_1} — приближенное значение $y(\xi)$, вычисленное при шаге h_1, а y_{h_2} — приближенное значение $y(\xi)$, вычисленное для шага h_2. Тогда можем записать

$$\begin{cases} y(\xi) - y_{h_1} \approx \varrho(\xi) h_1^k \\ y(\xi) - y_{h_2} \approx \varrho(\xi) h_2^k \end{cases}$$

и, вычитая из первого соотношения второе, получаем

$$y_{h_2} - y_{h_1} \approx \varrho(\xi)(h_1^k - h_2^k).$$

Отсюда следует формула Рунге

$$\varrho(\xi) = \frac{y_{h_2} - y_{h_1}}{h_1^k - h_2^k},$$

получаемое заменой приближенного равенства точным равенством.

Предположим теперь, что нам нужно найти приближенное решение с точностью $\varepsilon > 0$. Если

$$|\varrho(\xi) h_1^k| \leq \varepsilon$$

или

$$|\varrho(\xi)h_2^k| \le \varepsilon,$$

то все в порядке, необходимая точность достигнута. В качестве шага берем h_1 или h_2, соответственно.

Если в обоих случаях главный член погрешности $> \varepsilon$, то возникает вопрос о выборе нового шага h_ε. А именно, требуется выбрать шаг h_ε так, чтобы погрешность $\le \varepsilon$. Рунге предложил учитывать при этом лишь следующую оценку для главного члена

$$|\varrho(\xi)|h_\varepsilon^k \le \varepsilon,$$

и определить оптимальный шаг h_ε из равенства

$$|\varrho(\xi)|h_\varepsilon^k = \frac{|y_{h_2} - y_{h_1}|}{|h_1^k - h_2^k|}h_\varepsilon^k = \varepsilon.$$

Отсюда получаем следующее **правило Рунге выбора оптимального шага**:

$$h_\varepsilon = \left(\varepsilon \left|\frac{h_1^k - h_2^k}{y_{h_2} - y_{h_1}}\right|\right)^{1/k}.$$

Рассмотрим употребительный частный случай, когда $h_1 = h$ и $h_2 = 2h$, где шаг $h > 0$. Тогда формула Рунге для $\varrho(\xi)$ приобретает вид

$$\varrho(\xi) = \frac{y_{2h} - y_h}{h^k - (2h)^k} = \frac{y_h - y_{2h}}{(2^k - 1)h^k}.$$

Полезной является также равенство

$$y(\xi) - y_h \approx \frac{y_h - y_{2h}}{(2^k - 1)},$$

из которой следует формула

$$y(\xi) \approx y_h + \frac{y_h - y_{2h}}{(2^k - 1)},$$

называемая двухшаговым правилом Рунге для увеличения точности вычисления $y(\xi)$.

6.5 Многошаговые методы Адамса

Снова рассматриваем задачу Коши (6.1) — (6.2)

$$y' = f(x, y)$$

на отрезке $[x_0, b]$, удовлетворяющее условию

$$y|_{x=x_0} = y_0.$$

Как и для одношаговых методов выбираем некоторый шаг

$$h = \frac{b - x_0}{N},$$

и через y_n обозначаем приближенное значение точного решения $y(x_n)$ в узловых точках

$$x_n = x_0 + nh, \qquad n = 0, \ldots, N \qquad (x_N = x_0 + Nh = b).$$

Ставится следующая задача: начало таблицы значений решения известно, а именно, известны значения

$$y_0, y_1, \ldots, y_n \quad n < N.$$

Требуется вычислить

$$y_{n+1}, y_{n+2}, \ldots, y_N.$$

Рассмотрим два классических многошаговых метода Адамса.

Начальные подходы для получения алгоритма для обоих методов одинаковы. А именно, пользуемся снова знакомым нам тождеством

$$y(x_{n+1}) = y(x_n) + \int_{n_n}^{x_{n+1}} y'(x)dx = y(x_n) + \int_{n_n}^{x_{n+1}} \varphi(x)dx,$$

где

$$\varphi(x) := y'(x) = f(x, y(x)).$$

Основная идея Адамса такова: для вычисления интеграла функцию $\varphi(x)$ приближаем интерполяционным полиномом Лагранжа, построенным по $(k+1)$ узлу. Можно заметить, что обсуждаемые ниже методы Адамса являются $(k+1)$ - шаговыми, где k — заданное вычислителем фиксированное число, $1 \leq k \leq n$.

Итак, берем формулу

$$\varphi(x) = \varphi_{k+1}(x) + r_{k+1}(x),$$

где φ_{k+1} — интерполяционный полином Лагранжа степени $\leq k$, $(k+1)$ — число узловых точек, r_{k+1} — остаточный член (т. е. погрешность) интерполяции.

Рассмотрим сначала **экстраполяционный метод Адамса.**

Считаем, что значения

$$\varphi(x_j) = f(x_j, y(x_j)) \approx f(x_j, y_j)$$

известны для $j = 0, 1, \ldots, n$. Полином $\varphi_{k+1}(x)$ строится как интерполяционный полином Лагранжа по узлам

$$x_n, x_{n-1}, x_{n-2}, \ldots, x_{n-k}.$$

На основании формулы

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} \varphi(x)dx$$

с заменой переменных $x = x_n + h\alpha$, $0 \leq \alpha \leq 1$, получаем

$$y(x_{n+1}) = y(x_n) + h\int_0^1 \varphi(x_n + \alpha h)d\alpha, \quad dx = h\,d\alpha.$$

Подставляем $\varphi(x) = \varphi_{k+1}(x) + r_{k+1}(x)$ в интеграл. Отбрасывая остаточный член, получаем формулу

$$y(x_{n+1}) \approx y(x_n) + h\int_0^1 \varphi_{k+1}(x_n + \alpha h)d\alpha,$$

что порождает следующий алгоритм для определения y_{n+1}:

$$y_{n+1} = y_n + h \int_0^1 \varphi_{k+1}(x_n + \alpha h) d\alpha.$$

Очевидно, экстраполяционный метод Адамса — явный метод, так как φ_{k+1} определяется явно известными значениями $y_n, y_{n-1}, \ldots, y_{n-k}$. При выводе формул используются значения интерполяционного полинома в точках $x \in (x_n, x_{n+1})$, лежащих вне отрезка $[x_{n-k}, x_n]$. Поэтому метод называется экстраполяционным.

Мы можем также записать формулу для погрешности

$$R_{k+1} = \int_{x_n}^{x_{n+1}} r_{k+1}(x) dx = h \int_0^1 r_{k+1}(x_n + \alpha h) d\alpha$$

и оценить $|R_{k+1}|$, пользуясь тем, что

$$r_{k+1}(x) = \frac{\varphi^{(k+1)}(\xi)}{(k+1)!} \omega_{k+1}(x), \quad \xi \in (x_{n-k}, x_{n+1}),$$

где

$$\omega_{k+1}(x) = \prod_{j=n-k}^{n} (x - x_j).$$

Введем обозначения $\varphi_j = \varphi(x_j)$ и запишем известную нам формулу для интерполяционного полинома $\varphi_{k+1}(x)$ в форме Ньютона с использованием конечных

разностей. Поскольку $x = x_n + \alpha h$, то имеем

$$\varphi_{k+1}(x_n+\alpha h) = \varphi_n+\Delta^1\varphi_{n-1}\frac{\alpha}{1!}+\Delta^2\varphi_{n-2}\frac{\alpha(\alpha+1)}{2!}+\ldots+$$

$$+\Delta^k\varphi_{n-k}\frac{\alpha(\alpha+1)\ldots(\alpha+k-1)}{k!}.$$

Следовательно, экстраполяционный алгоритм Адамса будет иметь вид

$$y_{n+1} = y_n + h\int_0^1 \varphi_{k+1}(x_n+\alpha h)d\alpha =$$

$$= y_n + h[\varphi_n + c_1\Delta^1\varphi_{n-1} + c_2\Delta^2\varphi_{n-2} + \ldots + c_k\Delta^k\varphi_{n-k}],$$

где коэффициенты выражаются формулами

$$c_1 = \int_0^1 \frac{\alpha}{1!}d\alpha = \frac{1}{2},$$

$$c_2 = \int_0^1 \frac{\alpha(\alpha+1)}{2!}d\alpha = \frac{5}{12},$$

$$\vdots$$

$$c_k = \int_0^1 \frac{\alpha(\alpha+1)\ldots(\alpha+k-1)}{k!}d\alpha.$$

Таким образом, в экстраполяционном методе Адамса y_{n+1} находится явно по формуле

$$y_{n+1} = y_n + h\sum_{j=0}^{n} c_j\Delta^j\varphi_{n-j}.$$

Очевидно,

$$\alpha \in [0,1], \quad \frac{\alpha + j}{j+1} < 1, \text{ а значит } c_1 > c_2 > c_3 > \dots .$$

Рассмотрим **интерполяционный метод Адамса.** Напомню, что мы ищем приближенное решение задачи Коши

$$\begin{cases} y' = f(x,y), \quad x_0 \le x \le b, \\ y(x_0) = y_0. \end{cases}$$

Задан шаг разбиения отрезка и соответствующие узлы:

$$h = \frac{b - x_0}{N}, \quad x_0, x_1 = x_0 + h, \dots$$

Рассуждая так же, как и при выводе алгоритма для экстраполяционного метода Адамса, получаем алгоритм вида

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} \varphi_{k+1}(x)dx. \tag{6.6}$$

Отличие интерполяционного метода от экстраполяционного состоит в следующем: берется иной интерполяционный полином Лагранжа $\varphi_{k+1}(x)$ для $\varphi(x)$. А именно, рассматривается интерполяционный полином, построенный по узлам

$$x_{n+1}, x_n, x_{n-1}, x_{n-2}, \dots, x_{n-k+1}.$$

Предполагаются известными значения

$$y_n, y_{n-1}, y_{n-2}, \ldots, y_{n-k+1},$$

поэтому будут известны

$$\varphi(x_j) = f(x_j, y_j), \quad j = n, n-1, n-2, \ldots, n-k+1.$$

Но при построении интерполяционного полинома по указанным узлам приходится пользоваться величиной

$$\varphi(x_{n+1}) = f(x_{n+1}, y_{n+1}),$$

где величина y_{n+1} заранее неизвестна.

Далее, в интеграле из (6.6) полагаем $x = x_{n+1} + \alpha h$, $-1 \le \alpha \le 0$, что приводит к формуле

$$y_{n+1} = y_n + \alpha \int_{-1}^{0} \varphi_{k+1}(x_{n+1} + \alpha h) d\alpha.$$

Интерполяционный полином φ_{k+1} можно записать в форме Ньютона через конечные разности. При обозначениях $\varphi_j = \varphi(x_j)$ будем иметь

$$\varphi_{k+1}(x) = \varphi_{n+1} + \Delta^1 \varphi_n \frac{\alpha}{1!} + \Delta^2 \varphi_{n-1} \frac{\alpha(\alpha+1)}{2!} + \ldots +$$

$$+ \Delta^k \varphi_{n-k+1} \frac{\alpha(\alpha+1)\ldots(\alpha+k-1)}{k!}.$$

Таким образом, алгоритм интерполяционного метода

Адамса можно записать в следующем виде:

$$y_{n+1} = y_n + h[\varphi_{n+1} + \widetilde{c}_1 \Delta^1 \varphi_n + \ldots + \widetilde{c}_k \Delta^k \varphi_{n-k+1}],$$

где

$$\widetilde{c}_j = \int_{-1}^{0} \frac{\alpha(\alpha+1)\ldots(\alpha+j-1)}{j!} d\alpha.$$

Нетрудно показать, что $\widetilde{c}_j = O\left(1/(j \ln j)\right)$.

Интерполяционный метод Адамса является неявным методом. Действительно, так как

$$\varphi_{n+1} = y'(x_{n+1}) = f(x_{n+1}, y_{n+1})$$

то правая часть в алгоритме содержит y_{n+1}. Поэтому y_{n+1} определяется как корень соответствующего нелинейного уравнения.

Уравнение для определения y_{n+1} в (6.6) имеет вид

$$y_{n+1} = y_n + \alpha \Psi(y_{n+1}).$$

Для его решения относительно неизвестной y_{n+1} можно воспользоваться методом простой итерации. Достаточное условие сходимости имеет вид

$$|\alpha \Psi'(y)| \leq q < 1.$$

В качестве начального приближения y_{n+1}^0 можно взять значение y_{n+1}, найденное по экстраполяционному методу Адамса.

Проведем оценку погрешности в предположении

$$\max |y^{(k+2)}(x)| \le M,$$

где M — заданная постоянная. Применяя известную оценку для погрешности интерполяции

$$r_{k+1}(x) = \frac{y^{(k+2)}(\xi)}{(k+1)!}\,\omega_{k+1}(x), \quad \xi \in (x_{n-k+1}, x_{n+1}),$$

где

$$\omega_{k+1}(x) = \prod_{j=n-k+1}^{n+1} (x - x_j),$$

легко получаем

$$|R_{k+1}| \le h^{k+1} M |\widetilde{c}_{k+1}|,$$

где

$$\widetilde{c}_{k+1} = \int_{-1}^{0} \frac{\alpha(\alpha+1)\dots(\alpha+k)}{(k+1)!} d\alpha.$$

6.6 Задача Коши для системы ОДУ

Изученные нами методы легко распространяются на более общие задачи Коши, в частности, на следующую **задачу Коши для системы ОДУ первого порядка**

$$Y' = F(x, Y), \quad x_0 \le x \le b,$$

$$Y(x_0) = Y_0,$$

где $Y_0 \in \mathbb{R}^m$ — заданный вектор, $Y : [x_0, b] \to \mathbb{R}^m$ — искомое непрерывно дифференцируемое решение (функция скалярного аргумента с векторными значениями), $F : [x_0, b] \times \mathbb{R}^m \to \mathbb{R}^m$ — заданная вектор-функция векторного аргумента.

Ясно, что $Y(x)$ — вектор с координатами

$$y_1(x), y_2(x), \ldots, y_m(x),$$

а $F(x, Y)$ — вектор с координатами

$$F_1(x, y_1, y_2, \ldots, y_m), \ldots, F_m(x, y_1, y_2, \ldots, y_m).$$

Поэтому мы фактически имеем систему дифференциальных уравнений первого порядка, состоящую из m уравнений вида

$$y_j' = F_j(x, y_1, y_2, \ldots, y_m), \quad j = 1, \ldots, m,$$

с начальными условиями (условиями Коши)

$$y_1(x_0) = y_{01}, \ \ldots, \ y_m(x_0) = y_{0m},$$

где y_{0j} — заданные числа, $j = 1, \ldots, m$.

Ясно, что описанная задача Коши для системы уравнений в векторной форме равносильно интегральному

уравнению

$$Y(x) = Y_0 + \int_{x_0}^{x} F(t, Y(t))dt, \quad x_0 \le x \le b,$$

где интеграл $\int_{x_0}^{x} F(t, Y(t))dt$ понимается как вектор с m координатами

$$\int_{x_0}^{x} F_1(t, Y(t))dt, \ \ \ldots, \ \ \int_{x_0}^{x} F_m(t, Y(t))dt.$$

Изученные методы приближенного решения задачи Коши для одного уравнения первого порядка мы можем распространить на случай задачи Коши для указанной выше системы ОДУ первого порядка. Если пользоваться векторной формой записи системы, то алгоритмы решения будут иметь тот же вид, что и алгоритмы решения для одного уравнения. Приведем три примера.

1) Метод последовательных приближений Пикара определится следующими формулами

$$Y_n(x) = Y_0 + \int_{x_0}^{x} F(t, Y_{n-1}(t))dt, \quad x_0 \le x \le b, \quad n \in \mathbb{N}.$$

2) Пусть $h = (b - x_0)/N$ и $x_n = x_0 + nh$ $(n = 0, 1, \ldots, N)$. Метод ломаных Эйлера для определения приближенных значений Y_n точного решения $Y(x)$ в

точке $x = x_n$ $(n = 1, \ldots, N)$ имеет вид

$$Y_{n+1} = Y_n + hF(x_n, Y_n), \quad n = 0, 1, \ldots, N-1.$$

3) Пусть снова $h = (b - x_0)/N$ и $x_n = x_0 + nh$ $(n = 0, 1, \ldots, N)$. Наиболее употребительный алгоритм Рунге-Кутта порядка $k = 4$ для решения задачи Коши в случае системы запишется в виде

$$Y_{n+1} = Y_n + \frac{1}{6}(\varphi_0 + 2\varphi_1 + 2\varphi_2 + \varphi_3), \quad n = 0, \ldots, N-1,$$

где $Y_n \approx Y(x_n)$, $\;Y_{n+1} \approx Y(x_{n+1})$,

$$\begin{cases} \varphi_0 = h\, F(x_n,\, Y_n) \\ \varphi_1 = h\, F(x_n + \frac{h}{2},\, Y_n + \frac{\varphi_0}{2}) \\ \varphi_2 = h\, F(x_n + \frac{h}{2},\, Y_n + \frac{\varphi_1}{2}) \\ \varphi_3 = h\, F(x_n + h,\, Y_n + \varphi_2) \end{cases}.$$

Обсудим теперь кратко **задачу Коши для дифференциального уравнения порядка** $m \geq 2$. Требуется найти решение $y \in C^m[x_0, b]$ дифференциального уравнения порядка $m \geq 2$

$$y^{(m)} = f\left(x,\, y,\, y',\ldots,\, y^{(m-1)}\right), \quad x_0 \leq x \leq b,$$

удовлетворяющее начальным условиям

$$y(x_0) = y_0, \quad y'(x_0) = y'_0, \quad \ldots, \quad y^{(m-1)}(x_0) = y_0^{(m-1)},$$

где $y_0^{(j)}$ — заданные числа. Здесь речь идет о числовой функции $y : [x_0, b] \to \mathbb{R}$ и заданной числовой функции $f : [x_0, b] \times \mathbb{R}^m \to \mathbb{R}$.

Как известно из курса дифференциальных уравнений, уравнение порядка $m \geq 2$ можно заменить равносильной системой ОДУ первого порядка, состоящей из $m \geq 2$ уравнений. Действительно, мы можем формально определить m-мерный вектор

$$Y(x) = \left(y(x),\, y'(x), \ldots,\, y^{(m-1)}(x)\right).$$

Тогда рассматриваемое дифференциальное уравнение порядка $m \geq 2$ равносильно системе ОДУ вида

$$Y' = F(x, Y), \quad x_0 \leq x \leq b,$$

где $F(x, Y(x))$ — m-мерный вектор с координатами

$$y'(x),\, y''(x), \ldots,\, y^{(m-1)}(x), f(x,\, Y(x)).$$

Очевидно, мы получаем задачу Коши для системы ОДУ первого порядка с начальным условием

$$Y(x_0) = Y_0 := (y_0,\, y_0', \ldots, y_0^{(m-1)}).$$

Например, задача Коши для уравнения

$$y'' = f\left(x,\, y,\, y'\right), \quad x_0 \leq x \leq b,$$

с начальными условиями

$$y(x_0) = y_0, \quad y'(x_0) = y'_0$$

сводится к решению задачи Коши для системы уравнений

$$y' = z, \quad z' = f\,(x,\, y,\, z)$$

с начальными условиями

$$y(x_0) = y_0, \quad z(x_0) = y'_0.$$

Следовательно, можем найти приближенное решение задачи Коши для дифференциального уравнения порядка $m \geq 2$ с использование любого из алгоритмов, приведенных выше для системы обыкновенных дифференциальных уравнение первого порядка.

6.7 Одна краевая задача для ОДУ

Рассмотрим 2-х точечную краевую задачу для линейных дифференциальных уравнений 2-го порядка.

Будем предполагать, что следующая задача однозначно разрешима. Ищется функция $y \in C^2[a,\, b]$, удовлетворяющая уравнению

$$y'' + p(x)y' + q(x)y = f(x), \quad a \leq x \leq b, \tag{6.7}$$

и двум следующим краевым условиям

$$\alpha_0 y(a) + \alpha_1 y'(a) = A, \tag{6.8}$$

$$\beta_0 y(b) + \beta_1 y'(b) = B. \tag{6.9}$$

Здесь p, q, f — вещественнозначные непрерывные функции, заданные на отрезке $[a, b]$, α_0, α_1, A — заданные вещественные постоянные, удовлетворяющие условию $\alpha_0^2 + \alpha_1^2 \neq 0$, а также β_0, β_1, B — заданные вещественные постоянные, такие, что $\beta_0^2 + \beta_1^2 \neq 0$.

Рассмотрим иллюстративный пример. Возьмем уравнение

$$y'' + y = 0$$

на отрезке $[0,\, 2\pi]$.

Общее решение этого уравнения известно и задается формулой

$$y = c_1 \cos x + c_2 \sin x, \quad c_1, c_2 - \text{Const}.$$

Рассмотрим различные краевые задачи.

а) Пусть ищется решение, удовлетворяющее краевым условиям $y(0) = 0$ и $y(2\pi) = 1$. В силу условия $y(0) = 0$ имеем $c_1 = 0$. Тогда $y = c_2 \sin x$ и условие $y(2\pi) = 1$ не может быть выполнено, так как будем иметь $y(2\pi) = 0 \neq 1$, Значит, поставленная краевая задача не имеет решения.

б) Пусть ищется решение, удовлетворяющее крае-

вым условиям $y(0) = 0$ и $y'(2\pi) = 1$. Тогда снова $c_1 = 0$ и $y = c_2 \sin x$, $y' = c_2 \cos x$. В силу условия $y(2\pi) = c_2$ имеем $c_2 = 1$. Решение задачи существует и единственно, причем решение определяется формулой $y = \sin x$.

в) Если задать краевые условия

$$\left.\begin{array}{l} y(0) = 0 \\ y(2\pi) = 0 \end{array}\right\},$$

то легко проверить, что функция $y = C \sin x$ является решением при любом C. Таким образом, в этом случае решение не единственно.

Для рассматриваемой задача 6.7 — 6.9 будем предполагать, что задача однозначно разрешима.

Кроме того, предполагаем, что решение $y \in C^4[a, b]$. Для этого достаточно предположить, что коэффициенты дифференциального уравнения $p, q, f \in C^2[a, b]$.

Докажем этот факт. Нам дано, что

$$p, q, f, y \in C^2[a, b].$$

Тогда из (6.7) вытекает

$$y'' = -py' - qy + f, \quad p \in C^2,$$

где правая часть содержит функции

$$y' \in C^1[a, b], \quad q, y, f \in C^2[a, b].$$

Отсюда следует, что левая часть $y'' \in C^1[a,b]$, т. е. $y \in C^3[a,b]$, следовательно, $y' \in C^2[a,b]$. Снова анализируем равенство

$$y'' = -py' - qy + f \in C^2[a,\, b].$$

Ясно, что левая часть $y'' \in C^2[a,b]$. Следовательно, $y \in C^4[a,\, b]$, что и требовалось показать.

Метод численного решения требует предварительной подготовки. Опишем необходимые сведения.

Отрезок $[a,b]$ делим на n частей с фиксированным $n \in \mathbb{N}$ и с шагом $h = (b-a)/n$. Имеем узлы

$$x_0 = a, \quad x_k = a+kh, \quad k = 0, \ldots, n \quad (x_n = a+nh = b).$$

Ищем числа $y_k \approx y(x_k)$, $k = 0, \ldots, n$. Для этого используем аппроксимацию производных. В итоге, для нахождения $y_0, \ldots, y_n$ получим систему линейных алгебраических уравнений. Опишем теперь подробнее этапы этого процесса.

Сначала выпишем разностную аппроксимацию. производных y', y''. Так как $y \in C^4[a;b]$, можем использовать формулу Тейлора. Пусть $x, x+h \in [a,\, b]$, тогда по формуле Тейлора

$$y(x+h) = y(x) + \frac{y'(x)}{1!}h + \frac{y''(x)}{2!}h^2 + \frac{y'''(x)}{3!}h^3 + O(h^4)$$

для малых $h = (b-a)/n$. Пусть $x, x-h \in [a,\, b]$, тогда

$$y(x-h) = y(x) - \frac{y'(x)}{1!}h + \frac{y''(x)}{2!}h^2 - \frac{y'''(x)}{3!}h^3 + O(h^4).$$

Таким образом, имеем формулы: $y(x+h) =$

$$= y(x) + y'(x)h + \frac{y''(x)}{2}h^2 + \frac{y'''(x)}{6}h^3 + O(h^4) \quad (6.10)$$

и $y(x-h) =$

$$= y(x) - y'(x)h + \frac{y''(x)}{2}h^2 - \frac{y'''(x)}{6}h^3 + O(h^4). \quad (6.11)$$

Из (6.10) следует разностная формула

$$y'(x) = \frac{y(x+h) - y(x)}{h} + O(h),$$

называемая аппроксимацией y' с шагом вперед.

Аналогично, из (6.11) следует

$$y'(x) = \frac{y(x) - y(x-h)}{h} + O(h)$$

— аппроксимация y' с шагом назад.

Далее, из (6.10) и (6.11) следует

$$y'(x) = \frac{y(x+h) - y(x-h)}{2h} + O(h^2)$$

— аппроксимация y' с использованием симметричной разности.

Можно получить различные конечно разностные представления для второй производной y''. Но нам потребуется лишь одна формула, которая получается путем сложения формул (6.10) и (6.11):

$$y''(x) = \frac{y(x+h) - 2y(x) + y(x-h)}{h^2} + O(h^2).$$

Далее, пользуемся соотношениями (6.7) — (6.9) во внутренних узлах $x_k \in (a, b)$. Подставляя эти представления производных в уравнение (6.7), получаем

$$\frac{y(x_k+h) - 2y(x_k) + y(x_k-h)}{h^2} +$$

$$+p(x_k)\frac{y(x_k+h) - y(x_k-h)}{2h} +$$

$$+q(x_k)y(x_k) + O(h^2) = f(x_k),$$

для $k = 1, 2, 3, \ldots, n-1$.

Так как $x_k + h = x_{k+1}$, а $x_k - h = x_{k-1}$, то предыдущие равенства можем записать в следующем виде

$$y(x_{k+1}) - 2y(x_k) + y(x_{k-1}) + \frac{h}{2}p(x_k)[y(x_{k+1}) - y(x_{k-1})] +$$

$$+h^2 q(x_k)y(x_k) = h^2 f(x_k) + O(h^4),$$

или, после элементарных преобразований, в виде ра-

венств

$$y(x_{k-1})\left[1-\frac{h}{2}p(x_k)\right]+y(x_k)[-2+h^2q(x_k)]+$$

$$+y(x_{k+1})\left[1+\frac{h}{2}p(x_k)\right]=h^2f(x_k)+O(h^4). \qquad (6.12)$$

Перейдем к аппроксимации граничных условий. Из (6.8) следует, что в силу равенства

$$\alpha_0y(x_0)+\alpha_1y'(x_0)=A$$

имеем

$$\alpha_0y(x_0)+\alpha_1\frac{y(x_1)-y(x_0)}{h}=A+O(h).$$

Отсюда

$$y(x_0)[\alpha_0h-\alpha_1]+\alpha_1y(x_1)=Ah+O(h^2). \qquad (6.13)$$

Граничное условие (6.9) обрабатывается аналогично. Имеем

$$\beta_0y(x_n)+\beta_1y'(x_n)=B,$$

отсюда

$$\beta_0y(x_n)+\beta_1\frac{y(x_n)-y(x_{n-1})}{h}=B+O(h),$$

поэтому получаем равенство

$$-\beta_1y(x_{n-1})+y(x_n)[\beta_0h+\beta_1]=hB+O(h^2). \qquad (6.14)$$

Обозначим

$$y_k \approx y(x_k), p_k = p(x_k), q_k = q(x_k), f_k = f(x_k).$$

Подставляем величины в уравнения (6.12)— (6.14) и отбрасываем малые величины (невязки r_k). В результате приходим к системе линейных алгебраических уравнений для неизвестных

$$y_0,\ y_1,\ \ldots,\ y_n.$$

А именно, имеем уравнения

$$y_0[\alpha_0 h - \alpha_1] + y_1\alpha_1 = hA;$$

$$y_{k-1}\left[1 - \frac{h}{2}p_k\right] + y_k[-2 + h^2 q_k] + y_{k+1}\left[1 + \frac{h}{2}p_k\right] = h^2 f_k,$$

для $k = 1, 2, \ldots, n-1$;

$$y_{n-1}(-\beta_1) + y_n(\beta_{0h} + \beta_1) = hB.$$

Таким образом, для определения неизвестных

$$y_0, y_1, \ldots, y_n$$

получили систему линейных алгебраических уравнений с 3-х диагональной матрицей.

Эту систему решаем методом прогонки. Напомню

кратко этот метод.

Из первого уравнения находим $y_0 = \xi_0 y_1 + \eta_1$, подставляем во второе. Получаем новое второе уравнение с двумя неизвестными y_1, y_2. Из нового второго уравнения находим $y_1 = \xi_1 y_2 + \eta_2$, подставляем в третье. Из нового третьего находим $y_2 = \xi_2 y_3 + \eta_3$ и т. д.

Предпоследний шаг: из нового предпоследнего уравнения находим $y_{n-1} = \xi_{n-1} y_n + \eta_{n-1}$, подставляем в последнее уравнение. Новое последнее уравнение имеет вид $\widetilde{a} y_n = b_n$, находим $y_n = b_n / \widetilde{a}$.

Мы провели прямой ход прогонки, который позволяет определить величину y_n. Обратная прогонка также стандартна: зная y_n, определяем y_{n-1}, зная y_{n-1}, определяем y_{n-2} и т. д., зная y_1, находим y_0.

Приведем достаточное условие для того, чтобы метод прогонки был корректен. Ясно, для этого достаточным является условие: матрица полученной СЛАУ обладает свойством диагонального преобладания. Для простоты рассмотрим случай, когда

$$\alpha_1 = \beta_1 = 0, \quad \alpha_0 = \beta_0 = 1.$$

Тогда первое уравнение имеет вид $y_0 = b_0$, а последнее $(n+1)$-е уравнение запишется в виде $y_n = b_n$. Для k-го уравнения условие диагонального преобладания озна-

чает, что

$$|-2+h^2q_k| > \left|1-\frac{h}{2}p_k\right| + \left|1+\frac{h}{2}p_k\right|. \qquad (6.15)$$

Для выполнения (6.15) достаточно потребовать следующие условия:

а) $q(x) < 0$ на отрезке $[a,b]$

б) $h\max|p(x)| \le 2, \quad x \in [a,b]$.

Действительно, при выполнении этих условий будем иметь

$$|-2+h^2q_k| = 2-h^2q_k > 2,$$

$$\left|1-\frac{h}{2}p_k\right| + \left|1+\frac{h}{2}p_k\right| = 1-\frac{h}{2}p_k+1+\frac{h}{2}p_k = 2.$$

Таким образом, условия а), б) влекут (6.15). Следовательно, СЛАУ для определения неизвестных

$$y_0, y_1, \ldots, y_n$$

однозначно разрешима и ее решение можно найти методом прогонки.

Нетрудно оценить и погрешность метода. Действительно, пусть $\varepsilon_k = y(x_k) - y_k$ — погрешность определения $y(x_k)$. Покажем, что $\max|\varepsilon_k| = O(h^2)$.

Удобно рассмотреть векторную погрешность

$$\varepsilon = (\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_n)$$

и векторную невязку

$$r = (r_0, r_1, \ldots, r_n), \quad r = r(h).$$

Заметим, что для $k = 1, 2, \ldots, n-1$

$$y_{k-1}\left[1 - \frac{h}{2}p_k\right] + y_k[-2 + h^2 q_k] + y_{k+1}\left[1 + \frac{h}{2}p_k\right] = h^2 f_k,$$

и

$$y(x_{k-1})\left[1 - \frac{h}{2}p_k\right] + y(x_k)[-2 + h^2 q_k] +$$
$$+ y(x_{k+1})\left[1 + \frac{h}{2}p_k\right] = h^2 f_k + O(h^4),$$

поэтому

$$[y(x_{k-1}) - y_{k-1}]\left[1 - \frac{h}{2}p_k\right] + [y(x_k) - y_k][-2 + h^2 q_k] +$$

$$+[y(x_{k+1}) - y_{k+1}]\left[1 + \frac{h}{2}p_k\right] = h^2 f_k + r_k(h), r_k(h) = O(h^4).$$

Очевидно, систему для определения векторной погрешности можно записать в следующем виде

$$\varepsilon_0 = 0,$$

$$\varepsilon_{k-1}\left[1 - \frac{h}{2}p_k\right] + \varepsilon_k[-2 + h^2 q_k] + \varepsilon_{k+1}\left[1 + \frac{h}{2}p_k\right] = r_k(h),$$

для $k = 1, 2, \ldots, n-1$,

$$\varepsilon_n = 0,$$

где $|r_k(h)| = O(h^4)$. Эту систему можно привести к системе вида $\varepsilon = B\varepsilon + c$, где B — двухдиагональная квадратная матрица порядка $n+1$. Более подробно, эту систему можно записать так

$$\varepsilon_0 = 0,$$

$$\varepsilon_k = \varepsilon_{k-1}\frac{1 - hp_k/2}{2 - h^2q_k} + \varepsilon_{k+1}\frac{1 + hp_k/2}{2 - h^2q_k} + \frac{r_k(h)}{[-2 + h^2q_k]},$$

для $k = 1, 2, \ldots, n-1$,

$$\varepsilon_n = 0.$$

Обозначим

$$q := \min_{0 \le k \le n} |q_k| > 0.$$

Имеем: $2 - h^2q_k \ge 2 + h^2q$,

$$\|B\|_\infty \le \frac{2}{2 + h^2q} < 1, \quad 1 - \|B\|_\infty \ge \frac{h^2q}{2 + h^2q},$$

и

$$\max_{0 \le k \le n} |\varepsilon_k| := \|\varepsilon\|_\infty \le \|B\|_\infty \|\varepsilon\|_\infty + \frac{\|r(h)\|_\infty}{2 + h^2q}.$$

Следовательно,

$$\max_{0 \le k \le n} |\varepsilon_k| := \|\varepsilon\|_\infty \le \frac{\|r(h)\|_\infty}{(2 + h^2q)(1 - \|B\|_\infty)} \le$$

$$\le \frac{\|r(h)\|_\infty}{h^2q} = O(h^2),$$

что и требовалось доказать.

Случай нелинейных ОДУ

Рассмотрим численный метод решения краевой задачи для нелинейного обыкновенного дифференциального уравнения 2-го порядка

$$y'' = f(x, y, y'), \quad a \le x \le b. \tag{6.16}$$

Пусть краевые условия имеют вид

$$\varphi(y(a), y'(a)) = A \tag{6.17}$$

и

$$\psi(y(b), y'(b)) = B, \tag{6.18}$$

где A, B — постоянные, f, φ, ψ — гладкие функции. Предполагаем, что $f \in C^2([a, b] \times \mathbb{R}^2)$, $\varphi \in C^1(\mathbb{R}^2)$, $\psi \in C^1(\mathbb{R}^2)$.

Предположим также, что поставленная краевая задача имеет единственное решение $y \in C^4[a, b]$. Тогда идея построения численного решения остается такой же, что и для линейного уравнения.

Отрезок $[a, b]$ делим на n равных частей с шагом $h = \frac{b-a}{n}$, возьмем узловые точки

$$x_0 = a, \quad x_k = a+kh, \quad k = 0, \ldots, n \quad (x_n = a+nh = b).$$

Производим аппроксимацию производных по тем же формулам, что и ранее, Например, для второй произ-

водной берем формулу

$$y''(x) = \frac{y(x+h) - 2y(x) + y(x-h)}{h^2} + O(h^2).$$

Обозначаем $y_k \approx y(x_k)$, для определения $y_0, y_1, \ldots, y_n$ получаем систему уравнений:

$$\begin{cases} \varphi(y_0, \frac{y_1 - y_0}{h}) = A \\ y_{k+1} - 2y_k + y_{k-1} = h^2 f(x_k, y_k, \frac{y_{k+1} - y_{k-1}}{2h}), \\ (k = 1, 2, \ldots, n-1), \\ \psi(y_n, \frac{y_n - y_{n-1}}{h}) = B \end{cases}$$

Получили нелинейную систему уравнений, содержащую $n+1$ уравнение. Решая эту систему уравнений, например, методом итераций, получаем приближенные значения $y_0, y_1, \ldots, y_n$ искомого решения в узловых точках.

В общем случае оценка погрешности метода для нелинейных уравнений является сложной проблемой, изучаемой в специальной литературе по численным методам.

6.8 Краевые задачи для УЧП

Математические модели огромного числа прикладных проблем основаны на краевых или начально-краевых задачах для дифференциальных уравнений в частных

производных (УЧП). Естественно, разработаны различные методы численного решения подобных задач.

Мы ограничимся изложением метода конечных разностей (МКР) решения базовых краевых задач для дифференциальных уравнений математической физики 2-го порядка в случае 2-х независимых переменных.

Пусть D — область (непустое открытое связное множество), $D \subset \mathbb{R}^2$, $u = u(x, y)$ — функция, определяемая в этой области. В дальнейшем через Γ будем обозначать границу этой области D и предполагать, что Γ — замкнутая кусочно-гладкая кривая.

Будем рассматривать лишь линейные уравнения. Приведем сначала формулировки нескольких стандартных краевых задач для уравнения Лапласа

$$\Delta u := \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad (x,\, y) \in D.$$

Основные краевые задачи для уравнения Лапласа, часто встречающиеся в прикладных вопросах, таковы.

1) Задача Дирихле для уравнения Лапласа. Требуется определить функцию $u \in C^2(D) \cap C(\overline{D})$, удовлетворяющую условиям

$$\begin{cases} \Delta u = 0 \text{ в } D, \\ u|_\Gamma = \varphi(x, y), \quad (x, y) \in \Gamma. \end{cases}$$

Здесь $\Gamma = \partial D$ — жорданова кривая, $\varphi : \Gamma \to \mathbb{R}$ —

непрерывная функция.

Кратко сформулируем еще две краевых задачи для того же уравнения Лапласа, отличающиеся от задачи Дирихле видом краевых условий.

2) Задача Неймана: Γ — кусочно гладкая кривая,

$$\begin{cases} \Delta u = 0 \text{ в } D, \\ \frac{\partial u}{\partial n} = \psi(x,y), \quad (x,y) \in \Gamma. \end{cases}$$

Здесь $\partial u/\partial n$ — производная по внешней нормали к граничной кривой.

3) Смешанная задача:

$$\begin{cases} \alpha\, u + \beta\, \frac{\partial u}{\partial n} = \gamma \text{ на } \Gamma, \\ \Delta u = 0 \text{ в } D. \end{cases}$$

Здесь

$$\alpha : \Gamma \to \mathbb{R},\ \beta : \Gamma \to \mathbb{R},\ \gamma : \Gamma \to \mathbb{R}$$

— заданные функции, причем $\alpha^2 + \beta^2 \neq 0$.

Часто встречается следующий специальный случай смешанной задачи. Граничная кривая состоит из двух дуг: $\Gamma = \Gamma_1 \cup \Gamma_2$. Функция, удовлетворяющая в области уравнению Лапласа и непрерывно продолжимая на границу, должна удовлетворять и краевым условиям вида

$$\begin{cases} u|_{\Gamma_1} = \varphi, \\ \frac{\partial u}{\partial n}|_{\Gamma_2} = \psi. \end{cases}$$

Отметим, что задачи вида 1) — 3) рассматриваются не только для уравнения Лапласа, но и для более общих дифференциальных уравнений эллиптического типа.

А теперь вспомним классификацию наиболее употребительных УЧП математической физики. Для простоты рассмотрим лишь общие линейные уравнения 2-го порядка в случае 2-х независимых переменных. Пусть $D \subset \mathbb{R}^2$ — область. Рассмотрим в этой области дифференциальное уравнение

$$A\frac{\partial^2 u}{\partial x^2}+B\frac{\partial^2 u}{\partial x \partial y}+C\frac{\partial^2 u}{\partial y^2}+D\frac{\partial u}{\partial x}+E\frac{\partial u}{\partial y}+Fu=f, \quad (6.19)$$

где $A = A(x, y)$, $\ldots$, $F = F(x, y)$, $f = f(x, y)$ — заданные вещественнозначные непрерывные функции, определенные в D.

Рассмотрим дискриминант

$$\alpha(x, y) := 4A(x, y)C(x, y) - B^2(x, y).$$

1) Если $\alpha(x, y) > 0$ во всех точках области D, то говорят, что уравнение (6.19) является уравнением эллиптического типа.

Примеры уравнений эллиптического типа: уравнение Лапласа $\Delta u = 0$, уравнение Пуассона $\Delta u = f$.

2) Если $\alpha(x, y) < 0$ во всех точках области D, то говорят, что уравнение (6.19) является уравнением гиперболического типа.

Пример:

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y),$$

для этого уравнения $A = 1$, $C = -1$, $B = 0$.

3) Если $\alpha(x, y) \equiv 0$ во всех точках области D, то говорят, что уравнение (6.19) является уравнением параболического типа.

Пример: заменим y на t, т. е. $(x, t) \in D$, и рассмотрим уравнение

$$\frac{\partial^2 u}{\partial x^2} + \varphi(x, t) = \frac{\partial u}{\partial t},$$

для этого уравнения $A = 1$, $C, B = 0$.

Из курса УЧП известно, что для каждого типа уравнений существует хорошо разработанная теория о свойствах решений, о корректных постановках краевых и начально краевых задач и методах их решения. Эти теории существенно зависят от типа уравнений и отличаются друг от друга.

Для УЧП, которые имеют применения на практике, хорошо разработаны методы аналитического представления решений и разнообразные численные методы. Ниже мы рассмотрим лишь типичные задачи для нескольких уравнений математической физики, когда численное решение может быть найдено методом конечных разностей (МКР).

6.8.1 Задача Дирихле для уравнения Пуассона

Будем рассматривать модельную задачу в прямоугольнике

$$\overline{D} = [a, b] \times [c, d].$$

А именно, рассмотрим задачу Дирихле для уравнения Пуассона

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y) \text{ в } D \tag{6.20}$$

с краевыми условиями

$$u|_\Gamma = \varphi(x, y) \tag{6.21}$$

на границе прямоугольника. Краевые условия можно записать так:

$$\begin{aligned} u(x, c) &= \varphi_1(x), \quad a \le x \le b, \\ u(x, d) &= \varphi_2(x), \quad a \le x \le b, \\ u(a, y) &= \varphi_3(y), \quad c \le y \le d, \\ u(b, y) &= \varphi_4(y), \quad c \le y \le d, \end{aligned}$$

где $\varphi_1, \varphi_2, \varphi_3, \varphi_4$ — заданные дважды непрерывно дифференцируемые функции.

Будем предполагать, что $f \in C^2(\overline{D})$.

Предполагая, что существует единственное решение $u \in C^4$ поставленной краевой задачи, применим МКР для численного решения.

Этап 1. Строим сетку, покрывающую прямоуголь-

ник D следующим образом, Пусть $n \in \mathbb{N}$, возьмем шаг $h = \frac{b-a}{n}$ для переменной x и узлы

$$x_k = a + kh, \quad k = 0, 1, \ldots, n.$$

Далее, пусть $m \in \mathbb{N}$, шаг $l = \frac{d-c}{m}$ для переменной y. Выбираем узлы

$$y_j = c + jl, \quad j = 0, 1, \ldots, m$$

для переменной y. Наша цель состоит в определении приближенных значений u_{kj} сеточной функции

$$u(x_k, y_j), \quad k = 0, 1, \ldots, n, \quad j = 0, 1, \ldots, m.$$

Этап 2 — замена производных разностными отношениями.

Так же, как и для функции одной переменной, мы можем получить конечно разностные представления для производных второго порядка. А именно, предположим, что

$$(x, y), (x \pm h, y) \in D \quad \text{и} \quad u \in C^4 \quad \text{по} \quad x \in [a, b].$$

Тогда

$$\frac{\partial^2 u}{\partial x^2} = \frac{u(x+h, y) - 2u(x, y) + u(x-h, y)}{h^2} + O(h^2).$$

Аналогично, предполагая

$$(x, y), (x, y \pm l) \in D \quad \text{и} \quad u \in C^4 \quad \text{по} \quad y,$$

получаем формулу

$$\frac{\partial^2 u}{\partial y^2} = \frac{u(x, y + l) - 2u(x, y) + u(x, y - l)}{l^2} + O(l^2).$$

Через $u_{k,j}$ обозначим приближенное значение решения в точке (x_k, y_j), т. е.

$$u_{k,j} \approx u(x_k, y_j).$$

Обозначим также

$$f_{k,j} = f(x_k, y_j).$$

Далее, подставляем аппроксимации для производных

$$\frac{\partial^2 u}{\partial x^2}, \quad \frac{\partial^2 u}{\partial y^2}$$

в узловых точках (x_k, y_j) в уравнение Пуассона, отбрасываем малые величины $O(h^2)$, $O(l^2)$. В результате получаем следующую систему уравнений для u_{kj}:

$$\frac{u_{k+1,j} - 2u_{k,j} + u_{k-1,j}}{h^2} + \frac{u_{k,j+1} - 2u_{k,j} + u_{k,j-1}}{l^2} = f_{k,j}.$$

Здесь: $k = 1, \ldots, n-1$, $j = 1, \ldots, m-1$. Отметим, что

величины

$$u_{0,j}, u_{n,j}, u_{k,0}, u_{k,m}$$

известны в силу краевых условий.

Этап 3 — определение приближенных значений $u_{k,j}$ сеточной функции. Для этого пользуемся выписанной выше СЛАУ для неизвестных $u_{k,j}$ при $k = 1, \ldots, n-1$, $j = 1, \ldots, m-1$, дополняя эту систему граничными условиями.

Легко проверить, что в полученной СЛАУ число неизвестных равно числу уравнений. Эта система решается методом матричной прогонки (для разреженной системы). Можно показать, что погрешность метода

$$O(h^2 + l^2) = O(h^2)$$

при условии, что h и l — величины одного порядка.

Опишем кратко упомянутый *метод матричной прогонки.*

Полагаем $h^2 = \alpha l^2$, $\alpha = \text{const} \in (0, 1)$. Тогда СЛАУ из предыдущего пункта запишется в виде

$$(u_{k+1,j} - 2u_{k,j} + u_{k-1,j}) + \alpha(u_{k,j+1} - 2u_{k,j} + u_{k,j-1}) = f_{k,j}h^2,$$

где $k = 1, \ldots, n-1$, $j = 1, \ldots, m-1$. Перегруппируя слагаемые, будем иметь систему

$$u_{k+1,j} + u_{k-1,j} +$$

$$+[\alpha u_{k,j+1} - (2+2\alpha)u_{k,j} + 2u_{k,j-1}] = h^2 f_{k,j}, \qquad (6.22)$$

где $k = 1, \ldots, n-1$, $j = 1, \ldots, m-1$.

Запишем новую СЛАУ, эквивалентную (6.22), относительно векторов

$$\overline{u}_k = (u_{k,1}, u_{k,2}, \ldots, u_{k,m-1}).$$

Получаем новую систему уравнений

$$\overline{u}_{k+1} + B\overline{u}_k + \overline{u}_{k-1} = \overline{g}_k \qquad (6.23)$$

для векторов. Здесь B — следующая квадратная матрица порядка $(m-1)$:

$$B = \begin{pmatrix} -2-2\alpha & \alpha & 0 & \ldots & 0 & 0 \\ \alpha & -2-2\alpha & \alpha & \ldots & 0 & 0 \\ 0 & \alpha & -2-2\alpha & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \alpha & -2-2\alpha \end{pmatrix}$$

где $k = 1, \ldots, m-1$, B — трехдиагональная матрица. Кроме того, отметим, правые части уравнений определены следующим образом. Для $2 \le k \le m-2$ имеем

$$\overline{g}_k = h^2 \overline{f}_k = h^2 \cdot (f_{k,1}, f_{k,2}, \ldots, f_{k,m-1}).$$

Система (6.23) решается методом матричной прогонки.

При $k = 1$ имеем из первого уравнения системы:

$$\overline{u}_2 + B\overline{u}_1 = \overline{g}_1 - \overline{u}_0,$$

где $\overline{u}_0$ — известный вектор, определяемый краевыми условиями.

Применяя обратную матрицу B^{-1}, получаем отсюда

$$\overline{u}_1 = -B^{-1}\overline{u}_2 + B^{-1}(\overline{g}_1 - \overline{u}_0),$$

т. е. получаем равенство вида

$$\overline{u}_1 = X_1\overline{u}_2 + \overline{z}_1.$$

Здесь и далее заглавными буквами обозначены матрицы.

Во втором уравнении $\overline{u}_1$ заменяем ее представлением $X_1\overline{u}_2 + \overline{z}_1$. Тогда уравнение приобретает вид

$$\overline{u}_3 + B\overline{u}_2 + X_1\overline{u}_2 + \overline{z}_1 = \overline{g}_2$$

или, что то же самое,

$$\overline{u}_3 + (B + X_1)\overline{u}_2 = \overline{g}_2 - \overline{z}_1.$$

Отсюда определяем $\overline{u}_2$:

$$\overline{u}_2 = -(B + X_1)^{-1}\overline{u}_3 + \overline{z}_2,$$

т. е. будем иметь формулу вида

$$\overline{u}_2 = X_2\overline{u}_3 + \overline{z}_2$$

и т. д. Такая процедура после подстановки в последнее уравнение дает одно уравнение с одним неизвестным. Из предпоследнего уравнения получаем:

$$\overline{u}_{m-2} = X_{m-2}\overline{u}_{m-1} + \overline{z}_{m-2}.$$

Подставляем это выражение для $\overline{u}_{m-2}$ в последнее уравнение. В итоге получаем:

$$\overline{u}_{m-1} = \overline{z}_{m-1},$$

где правая часть представляет собой известный вектор.

Затем идет обратный ход прогонки по известным формулам, полученным выше. Последовательно определяем вектора

$$\overline{u}_{m-1} \to \overline{u}_{m-2} \to \ldots \to \overline{u}_1.$$

Таким образом, метод матричной прогонки выглядит так же, как и метод прогонки для обычной СЛАУ, но вместо прогоночных коэффициентов имеем прогоночные матрицы, так как при выводе прогоночных формул мы пользовались обратными матрицами. Понятно, что указанный алгоритм включает в себя знакомую нам задачу обращения матриц.

Замечание 1. Ожидаемая погрешность решения имеет порядок $O(h^2)$. Обоснование этого факта можно найти в специальной литературе.

Замечание 2. Рассмотренный подход годится для решения задачи Дирихле, Неймана и смешанной краевой задачи для общего линейного уравнения эллиптического типа в прямоугольнике.

Замечание 3. Метод можно приспособить и для решения задач, когда уравнение рассматривается в области произвольного вида, а не в прямоугольнике. Но здесь возникают трудности следующего характера:

1) область нужно аппроксимировать элементарными областями, составленными из прямоугольников (так же, как и при определении внутренней меры Жордана);

2) естественно, требуется аппроксимация граничных условий.

6.8.2 Уравнение параболического типа

Рассмотрим уравнение

$$\frac{\partial^2 u}{\partial x^2} + \varphi(x,t) = \frac{\partial u}{\partial t} \tag{6.24}$$

в полуплоскости: $t \geq 0$, $-\infty < x < \infty$. На границе области, т. е. на оси абсцисс задается начальное условие

$$u(x,0) = \psi(x), \quad -\infty < x < \infty. \tag{6.25}$$

Предполагаем, что существует единственное решение $u(x,t)$, причем функция ψ непрерывна, а функции u и φ непрерывны в полуплоскости $t \geq 0$, $-\infty < x < \infty$ и удовлетворяют некоторым дополнительным условиям, которые будут указаны ниже.

Разностный метод решения получается с помощью аппроксимации производных решения.

Этап 1. Строим сеточную область. Пусть $h > 0$ — шаг по переменной x, тогда узлы определяются формулами $x_k = kh$, $k = 0, \pm 1, \pm 2, \ldots$, т. е. $k \in \mathbb{Z}$.

Пусть τ — шаг по переменной t, т.е узлы определяются формулами $t_j = \tau j$, $j = 0, 1, 2, \ldots$, т. е. индекс $j \in \mathbb{N} \cup \{0\}$. Так называемая сеточная область состоит из узлов $\{(x_k, t_j)\}$.

Рекомендуемая связь между шагами по разным переменным такова: $\tau = \tau(h) = O(h^2)$.

Этап 2. Аппроксимируем производные, считая, что решение достаточное число раз дифференцируемо. А именно, считаем справедливыми стандартные формулы

$$\frac{\partial u}{\partial t} = \frac{u(x, t+\tau) - u(x,t)}{\tau} + O(\tau),$$

$$\frac{\partial^2 u}{\partial x^2} = \frac{u(x+h,t) - 2u(x,t) + u(x-h,t)}{h^2} + O(h^2).$$

Подставляем эти выражения в дифференциальное урав-

нение. Считая $\tau = O(h^2)$, получаем

$$\frac{u(x,t+\tau)-u(x,t)}{\tau} -$$

$$-\frac{u(x+h,t)-2u(x,t)+u(x-h,t)}{h^2} =$$

$$= \varphi(x,t) + O(h^2).$$

Как и ранее, обозначим $u_{k,j} \approx u(x_k, t_j)$. Далее, в предыдущем уравнении заменим x на x_k, t на t_j. Отбрасываем $O(h^2)$, тогда получается СЛАУ для определения $u_{k,j}$.

Начальное условие влечет равенства

$$u_{k,0} = \psi_k := \psi(x_k), \quad k \in \mathbb{Z}.$$

Из уравнения (6.24) следует, что

$$\frac{u_{k,j+1}-u_{k,j}}{\tau} - \frac{u_{k+1,j}-2u_{k,j}+u_{k-1,j}}{h^2} = \varphi_{k,j}.$$

По определению, полагаем

$$\varphi_{k,j} = \varphi(x_k, t_j).$$

Здесь

$$k \in \mathbb{Z}, j = 0, 1, \ldots.$$

Обозначим $\tau = rh^2$ и запишем систему следующим

образом

$$u_{k,j+1} = u_{k,j} + r[u_{k+1,j} - 2u_{k,j} + u_{k-1,j}] + \ldots + r\varphi_{k,j}, \quad (6.26)$$

$k \in \mathbb{Z}$, $j = 0, 1, \ldots$.

Отметим, что в силу начального условия известны величины $u_{k0} = \psi_k$, $\forall k \in \mathbb{Z}$.

Получили СЛАУ бесконечного порядка. Но эта система устроена просто, поэтому ее разрешимость доказывается элементарно. Действительно, на нулевом слое по t, т. е. при $t = 0$ имеем:

бесконечномерный вектор

$$\overline{u}_0 = (\ldots, u_{-2,0}, u_{-1,0}, u_{0,0}, u_{1,0}, u_{2,0}, \ldots, u_{k,0}, \ldots)$$

известен. А именно, согласно начальным условиям

$$\overline{u}_0 = \overline{\psi} = (\ldots, \psi_{-2}, \psi_{-1}, \psi_0, \psi_1, \psi_2, \ldots, \psi_k, \ldots).$$

Система (6.26) при $j = 0$ дает:

для любого k

$$u_{k,1} = u_{k,0} + r[u_{k+1,0} - 2u_{k,0} + u_{k-1,0}] + \tau\varphi_{k,0}.$$

Пользуясь этими формулами мы можем вычислить приближенные значения решения на первом слое, т. е. определить явно все координаты бесконечномерного вектора $\overline{u}_1 = \{u_{k,1}\}$ при $k \in \mathbb{Z}$.

Итак, мы находим все координаты вектора

$$\overline{u}_1 = (\ldots, u_{-2,1}, u_{-1,1}, u_{0,1}, u_{1,1}, u_{2,1}, \ldots, u_{k,1}, \ldots).$$

Зная сеточную функцию на 1-м слое и пользуясь формулами

$$u_{k,2} = u_{k,1} + r[u_{k+1,1} - 2u_{k,1} + u_{k-1,1}] + \tau\varphi_{k,1},$$

определяем сеточную функцию на 2-м слое, и так далее. Условно можем записать процесс последовательного определения сеточной функции так:

$$\overline{u}_2 \to \overline{u}_3 \to \ldots.$$

Рассмотрим вопрос об устойчивости процесса за конечное число шагов. Пусть N — число шагов (слоев), и пусть

$$0 \le t \le T = \tau N, \;\; 0 < \tau \le 1/2.$$

Постановку задач дадим в виде некоторых замечаний и определений.

Замечание 6.8.1 *Задачу (6.24) – (6.25) мы можем записать формально следующим образом: найти функцию u как решение операторного уравнения $Lu = f$, где оператор определен формулами*

$$Lu = \begin{cases} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} & \text{в области } D \\ u & \text{на } \Gamma \end{cases},$$

$$f = \begin{cases} \varphi(x,t) & \text{в области } D \\ \psi(x) & \text{на } \Gamma \end{cases}.$$

Далее, сеточные функции обозначим u^h, f^h, речь идет о векторах с координатами $u_{k,j}, f_{k,j}$.

Замечание 6.8.2 *Численный метод приводит к решению задачи: найти сеточную функцию u^h по условию*

$$L_h u^h = f^h,$$

где L_h — линейный оператор, определяемый выписанной выше системой уравнений (6.26).

Нормы бесконечномерных векторов u^h, f^h задаем формулами

$$\|u^h\| = \sup_{k,j} |u_{k,j}|,$$

$$\|f^h\| = \sup_{k,j} |f_{k,j}|.$$

Определение 6.3 (устойчивости) *Существует*

$$M > 0$$

такое, что

$$\|u^h\| \le M\|f^h\|, \quad \forall f^h, \|f^h\| < \infty.$$

Это означает, что ограничена норма обратного оператора, точнее, выполнено неравенство $\|L_h^{-1}\| \le M$.

Напомню, что мы имеем формулы

$$u_{k0} = \psi_k, \quad k \in \mathbb{Z}, \tag{6.27}$$

$$u_{k,j+1} = u_{k,j} + r(u_{k+1,j} - 2u_{k,} + u_{k-1,j}) + \tau\varphi_{k,j}, \tag{6.28}$$

где $k \in \mathbb{Z}$, $j = 0, 1, \ldots$. Из (6.28) следует, что

$$u_{k,j+1} = r\left[u_{k+1,j} + u_{k-1,j}\right] + (1 - 2r)\, u_{k,j} + \tau\, \varphi_{k,j}.$$

Задачу рассматриваем для конечного времени

$$0 \le t \le T = \tau N$$

(можно взять более общий случай, когда

$$\tau N \le T < \tau(N + 1)).$$

Итак, рассматриваем операторное уравнение

$$L_h u^h = f^h,$$

где $u^h = \{u_{k,j}\}$ — сеточная функция, причем нормы векторов заданы формулами

$$\|u^h\| = \sup_{k\in\mathbb{Z}, 0\le j\le N} |u_{k,j}|,$$

$$\|f^h\| = \max\{ \sup_{k\in\mathbb{Z}} |\psi_k|, \quad \sup_{k\in\mathbb{Z}, 0\le j\le N} |\varphi_{k,j}|\}.$$

Теорема 6.4 *Пусть* $\tau = rh^2$, *причем*

$$0 < r \le 1/2, \quad 0 \le t \le T = \tau N.$$

Тогда задача $L_h u^h = f^h$ *устойчива, т. е. существует постоянная* $M > 0$, *такая, что*

$$\|u^h\| \le M\|f^h\|.$$

Доказательство. Имеем

$$\sup_{k\in\mathbb{Z}} |u_{k,0}| = \sup_{k\in\mathbb{Z}} \psi_k.$$

Далее, величину $\sup_{k\in\mathbb{Z}} |u_{k1}|$ оценим, пользуясь (6.26) при $j = 0$ и соотношениями

$$\sup|A + B| \le \sup|A| + \sup|B|.$$

Будем иметь

$$\sup_{k\in\mathbb{Z}} |u_{k,1}| \le r(\sup_k |u_{k+1,0}| + \sup_k |u_{k-1,0}|)+$$

$$+(1-2r)\sup_{k\in\mathbb{Z}} |u_{k,0}| + \tau \sup_{k\in\mathbb{Z}} |\varphi_{k,0}| =$$

$$= 2r\sup_{k\in\mathbb{Z}} |u_{k,0}| + (1-2r)\sup_{k\in\mathbb{Z}} |u_{k,0}| + \sup_{k\in\mathbb{Z}} |\varphi_{k,0}| =$$

$$= \sup_{k\in\mathbb{Z}} |u_{k,0}| + \tau \sup_{k\in\mathbb{Z}} |\varphi_{k,0}|.$$

Итак, полагая $j = 1$, имеем

$$\sup_{k\in\mathbb{Z}} |u_{k,2}| \le \sup_{k\in\mathbb{Z}} |u_{k,1}| + \tau \sup_{k\in\mathbb{Z}} |\varphi_{k,1}|.$$

Далее, продолжаем процесс. В итоге получаем неравенства

$$\begin{cases} \sup_k |u_{k,0}| \le \sup_k |\varphi_k| \\ \sup_{k\in\mathbb{Z}} |u_{k,1}| \le \sup_{k\in\mathbb{Z}} |u_{k,0}| + \tau \sup_{k\in\mathbb{Z}} |\varphi_{k,0}| \\ \sup_{k\in\mathbb{Z}} |u_{k,2}| \le \sup_{k\in\mathbb{Z}} |u_{k,1}| + \tau \sup_{k\in\mathbb{Z}} |\varphi_{k,1}| \\ \dots\dots\dots\dots\dots\dots \\ \sup_{k\in\mathbb{Z}} |u_{k,m}| \le \sup_{k\in\mathbb{Z}} |u_{k,m-1}| + \tau \sup_{k\in\mathbb{Z}} |\varphi_{k,m-1}| \\ \dots\dots\dots\dots\dots\dots \\ \sup_{k\in\mathbb{Z}} |u_{k,N}| \le \sup_{k\in\mathbb{Z}} |u_{k,N-1}| + \tau \sup_{k\in\mathbb{Z}} |\varphi_{k,N-1}|. \end{cases}$$

Складываем эти неравенства, и для любого

$$m = 1, 2, \dots, N$$

будем иметь неравенство

$$\sup_{k\in\mathbb{Z}} |u_{k,m}| \le \sup_{k\in\mathbb{Z}} |\psi_k| + \tau m \sup_{k\in\mathbb{Z}, 1\le j\le N} |\varphi_{k,j}| \le \|f^h\| + T\|f^h\|.$$

Отсюда следует, что

$$\|u^h\| \le M\|f^h\|,$$

где $M = T + 1$. Таким образом, теорема доказана.

Обратим внимание на то, что в доказательстве су-

щественно используется малость величины $\tau \in (0, 1/2]$.

Общие замечания.

1) Ясно, что рассмотренный метод можно применить и для параболических линейных уравнений общего вида с гладкими коэффициентами в верхней полуплоскости.

2) Существуют и иные подходы численного решения уравнений параболического типа, в частности, численно-аналитические. Начальная идея такова: аппроксимируется лишь производная $\partial u/\partial t$ по формуле

$$\frac{\partial u}{\partial t} \approx \frac{u(x, t-\tau) - u(x,t)}{-\tau}.$$

В итоге на каждом слое имеем ОДУ второго порядка относительно переменной x.

6.8.3 МКР для волнового уравнения

В верхней полуплоскости

$$D = \{(x, y) : \ -\infty < x < +\infty, y > 0\}$$

рассмотрим волновое уравнение

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y), \quad (x, y) \in D, \tag{6.29}$$

с начальными условиями

$$u(x,0)=\varphi(x),\quad -\infty<x<+\infty, \tag{6.30}$$

$$\left.\frac{\partial u}{\partial y}\right|_{y=0}=\psi(x),\quad -\infty<x<+\infty. \tag{6.31}$$

Снова предполагаем существование, единственность и достаточную гладкость решения поставленной задачи. Схематично опишем метод конечных разностей для численного решения.

Этап 1. Строим сеточную область, состоящую из точек (x_k, y_j), беря h в качестве шага по переменной x, $x_k = kh$, $k \in \mathbb{Z}$, и l в качестве шага по y, $y_j = jl$, $j \in \mathbb{N} \cup \{0\}$.

Этап 2. Аппроксимируем производные $\frac{\partial^2 u}{\partial x^2}$, $\frac{\partial^2 u}{\partial y^2}$, считая решение u и его производную $\frac{\partial u}{\partial y}$ в начальных условиях достаточно гладкими функциями. Имеем

$$\frac{\partial^2 u(x,y)}{\partial x^2}=\frac{u(x+h,y)-2u(x,y)+u(x-h,y)}{h^2}+O(h^2),$$

$$\frac{\partial^2 u(x,y)}{\partial y^2}=\frac{u(x,y+l)-2u(x,y)+u(x,y-l)}{l^2}+O(l^2)$$

и

$$\frac{\partial u(x,0)}{\partial y}=\frac{u(x,l)-u(x,0)}{l}+O(l).$$

Этап 3. Подставляем указанные выражения для производных в уравнение и начальные условия, полагая $x = x_k$, $y = y_j$, $x_k \pm h = x_{k\pm1}$, $y_j \pm l = y_{j\pm1}$.

Отбрасываем малые величины $O(h^2)$, $O(l^2)$ в уравнении и $O(l)$ в начальных условиях. Получаем систему для определения $u_{k,j} \approx u(x_k, y_j)$. Имеем

$$\frac{u_{k+1,j} - 2u_{k,j} + u_{k-1,j}}{h^2} + \\ + \frac{u_{k,j+1} - 2u_{k,j} + u_{k,j-1}}{l^2} = f_{k,j}, \qquad (6.32)$$

где $f_{k,j} := f(x_k, y_j)$,

$$k = 0, \pm 1, \pm 2, \ldots, \quad j = 0, 1, 2, \ldots.$$

Из (6.30) следует $u_{k0} = \varphi_k = \varphi(x_k)$, $k \in \mathbb{Z}$.

Из (6.31) следует $u_{k,1} = u_{k,0} + l\psi_k$, $\psi_k = \psi(x_k)$, $k \in \mathbb{Z}$.

Полученная система относительно $u_{k,j}$ однозначно разрешима. Действительно, значение сеточной функции известно на нулевом слое, так как $u_{k0} = \varphi_{k1}$, и на первом слое, так как $u_{k1} = \varphi_k + l\psi_k$. Но тогда, зная $u_{k,j}$, $u_{k,j-1}$, из (6.32) мы определяем $u_{k,j+1}$ для любого k. Зная $\{u_{k,0}, u_{k,1}\}$, находим $u_{k,2}$, зная $\{u_{k,1}, u_{k,2}\}$, определим $u_{k,3}$. Процесс можно продолжить, так как ясно, что знание сеточной функции на двух соседних слоях позволяет найти значения сеточной функции на следующем слое.

Отметим, что при выборе шагов полагают $l = \alpha h$, $0 < \alpha \leq 1$. Тогда $l = O(h)$.

6.9 Задачи и упражнения

1. Найдите решение $y \in C^1[0,1]$ интегрального уравнения

$$y(x) - 2\int_0^x ty(t)\,dt = -\frac{x^3}{3}.$$

2. Для задачи Коши

$$y' = f(x,y), \quad y|_{x=0} = 0,$$

где

$$f(x,y) := x^2 + 2xy, \quad 0 \le x \le 1,$$

запишите алгоритм численного решения порядка 4, пользуясь следующим алгоритмом Рунге — Кутта

$$y_{n+1} = y_n + \frac{1}{6}(\varphi_0 + 2\varphi_1 + 2\varphi_2 + \varphi_3), \quad n = 0, \ldots, N-1,$$

взяв $h = 1/10$, $x_n = hn$ $(n = 0, 1, 2, \ldots, 10)$,

$$\begin{cases} \varphi_0 = hf(x_n, y_n) \\ \varphi_1 = hf(x_n + \frac{h}{2}, y_n + \frac{\varphi_0}{2}) \\ \varphi_2 = hf(x_n + \frac{h}{2}, y_n + \frac{\varphi_1}{2}) \\ \varphi_3 = hf(x_n + h, y_n + \varphi_2) \end{cases}.$$

3. Пользуясь методом Рунге — Кутта, попробуйте получить свой алгоритм четвертого порядка для реше-

ния задачи Коши

$$y' = f(x, y), \quad 0 \le x \le 1, \quad y|_{x=0} = y_0.$$

4. Рассмотрите задачу Коши на отрезке $[x_0, b]$ для системы из двух уравнений

$$u'(x) = f(x, u(x), v(x)), \quad v'(x) = g(x, u(x), v(x))$$

с начальными условиями

$$u(x_0) = u_0, \quad v(x_0) = v_0.$$

Напишите для этой задачи алгоритмы Рунге-Кутта второго порядка точности, аналогичные следующим двум алгоритмам:

$$y_{n+1} = y_n + hf(x_n + h/2, y_n + (h/2)f(x_n, y_n))$$

и

$$y_{n+1} = y_n + (h/2)f(x_n, y_n) + \\ +(h/2)f(x_n + h, y_n + hf(x_n, y_n)).$$

5. На лекциях был указан способ получения обобщений метода ломаных Эйлера с применением малых квадратурных формул прямоугольников и трапеций. Пользуясь этим способом с привлечением малой квадратурной формулы Симпсона, получите новый алго-

ритм для решения задачи Коши

$$y' = f(x, y), \quad 0 \le x \le 1, \quad y|_{x=0} = y_0.$$

6. Напомню, что алгоритм интерполяционного метода Адамса можно записать в следующем виде

$$y_{n+1} = y_n + h[\varphi_{n+1} + \widetilde{c}_1 \Delta^1 \varphi_n + \ldots + \widetilde{c}_k \Delta^k \varphi_{n-k+1}],$$

где

$$\widetilde{c}_j = \int_{-1}^{0} \frac{\alpha(\alpha+1)\ldots(\alpha+j-1)}{j!} d\alpha.$$

Докажите, что

$$\widetilde{c}_j = O\left(1/(j \ln j)\right).$$

Глава 7

Операторные уравнения

Пусть Y и F — бесконечномерные линейные нормированные пространства над полем вещественных чисел, и пусть $A : Y \to F$ — заданный линейный непрерывный оператор, имеющий ограниченный обратный оператор A^{-1}. Рассмотрим уравнение

$$Ay = f, \tag{7.1}$$

где $f \in F$ — заданный элемент, требуется определить $y \in Y$. Если обратный оператор A^{-1} известен, то точное решение уравнения (7.1) определяется формулой $y^* = A^{-1}f$.

Задача: Найти приближенное решение y_n^* уравнения (7.1) в ситуации, когда известно лишь существование обратного оператора A^{-1}, но обратный оператор A^{-1} неизвестен.

С операторным уравнением вида (7.1) мы уже встречались при обосновании сходимости метода последо-

вательных приближений Пикара (см. методы решения задачи Коши для ОДУ первого порядка). Теперь же нашей основной целью является изучение нескольких методов приближенного решения уравнения (7.1), применяемых при численном решении интегральных уравнений.

Предположение о бесконечномерности пространств Y и F не является случайным.

Если Y и F являются конечномерными, то, очевидно, уравнение вида (7.1) сводится к обычной СЛАУ, с методами решения которой мы уже знакомы.

В качестве бесконечномерных пространств Y и F при решении интегральных уравнений на практике используются функциональные пространства, в частности, банахово пространство $C[a, b]$ и гильбертово пространство $L^2[a, b]$.

7.1 Методы моментов (общая схема)

Дополнительно предполагаем, что F — гильбертово пространство. Следовательно, определено скалярное произведение $(f,\, g)$ для любых двух элементов $f \in F$ и $g \in F$.

Пусть заданы две системы линейно-независимых элементов:

система элементов $l_1, l_2, \ldots, l_n$ в пространстве Y и

система элементов $g_1, g_2, \ldots, g_n$ в пространстве F.

Приближенное решение уравнения (7.1) ищем в виде следующей линейной комбинации

$$y_n = \sum_{k=1}^{n} \alpha_k l_k,$$

где α_k — числовые коэффициенты из $\mathbb{R}$. Числа α_k заранее неизвестны, для нахождения этих коэффициентов нужны дополнительные условия. Выбор дополнительных условий представляет собой важную проблему.

В методе моментов для определения постоянных α_k накладываются следующие дополнительные условия.

Рассматривается невязка $r_n = f - Ay_n$ и требуется выполнение условий: вектор r_n ортогонален вектору g_j для всех $j = 1, 2, \ldots, n$. Условие ортогональности означает, что

$$(r_n, g_j) = 0, \quad j = 1, 2, \ldots, n.$$

Отсюда следует

$$(f, g_j) - (Ay_n, g_j) = 0,$$

поэтому

$$\sum_{k=1}^{n} \alpha_k (Al_k, g_j) = (f, g_j) \quad (j = 1, 2, \ldots, n). \tag{7.2}$$

Предположим, что СЛАУ (7.2) однозначно разрешима, обозначим ее решение как $\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*$. Тогда искомое приближенное решение определяется формулой

$$y_n^* = \sum_{k=1}^{n} \alpha_k^* l_k.$$

Такой метод называется методом моментов.

Замечание. Если F — n-мерное пространство, то совокупность векторов $\{g_j\}$, $j = 1, 2, \ldots, n$, является базисом. Следовательно, условия

$$(r_n, g_j) = 0, \quad j = 1, 2, \ldots, n,$$

влекут, что $r_n = 0$, т. е. y_n^* будет точным решением.

В общем случае, можно лишь надеяться на то, что $y_n^* \approx y^*$, где $y^* = A^{-1} f$ — точное решение уравнения (7.1).

Возникает важный вопрос, как оценить норму

$$\|y^* - y_n^*\|.$$

Понятно, что нужны дополнительные условия для того, чтобы гарантировать малость нормы $\|y^* - y_n^*\|$.

Отметим, что ставший уже классическим, глубоко разработанный метод Галеркина представляет собой специальный случай метода моментов, когда $Y = F$, $g_k = l_k$ при $k = 1, 2, \ldots, n$. Система для определения

α_k в этом случае имеет вид

$$\sum_{k=1}^{n} \alpha_k \, (Al_k, l_j) = (f, l_j), \quad j = \overline{1, n}. \tag{7.3}$$

7.2 Метод наименьших квадратов

Так называют другой специальный случай метода моментов, когда заданные линейно-независимые системы взаимосвязаны, а именно, имеют место равенства

$$g_j = Al_j, \;\; j = 1, 2, \ldots, n.$$

Очевидно, этом случае система для определения α_k имеет вид

$$\sum_{k=1}^{n} \alpha_k (Al_k, Al_j) = (f, Al_j), \quad j = \overline{1, n}. \tag{7.4}$$

Название "метод наименьших квадратов" оправдано тем, что коэффициенты α_k можно найти из требования минимальности некоторой суммы квадратов. Поясним этот факт подробнее. Пусть

$$y_n = \sum_{k=1}^{n} \alpha_k l_k,$$

рассмотрим задачу минимизации функции Φ, определенную равенством

$$\Phi(\alpha_1, \alpha_2, \ldots, \alpha_n) = \|f - Ay_n\|^2 \to \min.$$

Функция Φ зависит от n переменных. Имеем

$$0 \leq \Phi(\alpha_1, \alpha_2, \ldots, \alpha_n),$$

кроме того,

$$\Phi(\alpha_1, \alpha_2, \ldots, \alpha_n) = \left(f - \sum_{k=1}^{n} \alpha_k Al_k, f - \sum_{k=1}^{n} \alpha_k Al_k \right) =$$

$$= \|f\|^2 - 2\sum_{k=1}^{n} \alpha_k (f, Al_k) + \sum_{k=1}^{n} \sum_{j=1}^{n} \alpha_k \alpha_j (Al_k, Al_j)$$

— квадратичная функция от переменных $\alpha_1, \alpha_2, \ldots, \alpha_n$.

Ясно, что эта функция является непрерывно дифференцируемой, минимум существует и точка минимума находится как решение системы уравнений

$$\frac{\partial \Phi}{\partial \alpha_j} = 0, \quad j = \overline{1, n}.$$

Вычисляя указанные производные, мы легко убеждаемся в том, что получаемая таким образом система уравнений эквивалентна системе (7.4) из метода моментов.

Рассмотрим теперь вопрос оценки погрешности для метода наименьших квадратов при следующих услови-

ях (требованиях):

I) Y — банахово пространство, $\{l_k\}_{k=1}^{\infty}$ — полная система в пространстве Y;

II) линейный непрерывный оператор $A : Y \to F$ имеет обратный ограниченный оператор A^{-1}, и, следовательно, существует единственное решение уравнения (7.1), определяемое формулой $y^* = A^{-1}f$;

III) будем рассматривать метод наименьших квадратов, т. е. $g_j = Al_j$, $(j = 1, 2, \ldots, n)$.

Теорема 7.1 *При сделанных предположениях I), II), III) существует единственное решение $\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*$ системы*

$$\sum_{k=1}^{n} \alpha_k (Al_k, Al_j) = (f, Al_j), \quad j = \overline{1, n},$$

определяющее приближенное решение

$$y_n^* = \sum_{k=1}^{n} \alpha_k^* l_k$$

на основе метода наименьших квадратов, причем

$$\|y^* - y_n^*\| \to 0 \quad \textit{при } n \to \infty.$$

Более точно, имеем оценки

$$E_n(y^*) \le \|y^* - y_n^*\| \le \eta(A) E_n(y^*),$$

где

$\eta(A) = \|A^{-1}\| \cdot \|A\|$ — *число обусловленности оператора* A,

$E_n(y^*)$ — *наилучшее приближение элемента*

$$y^* \in Y$$

элементами вида

$$\widetilde{y}_n = \sum_{k=1}^{n} \widetilde{\alpha}_k l_k.$$

Доказательство проведем в несколько шагов.

1) Как мы знаем, в нормированном пространстве существует элемент y_n^{**} наилучшего приближения, т. е. существует элемент вида

$$y_n^{**} = \sum_{k=1}^{n} \alpha_k^{**} l_k,$$

обладающий свойством:

$$E_n(y^*) = \|y^* - y_n^{**}\| \leq \|y^* - y_n\|$$

для любого элемента y_n вида

$$y_n = \sum_{k=1}^{n} \alpha_k l_k.$$

2) Докажем теперь существование приближенно-

го решения y_n^*, определяемого по методу наименьших квадратов. Дано, что

$$g_1 = Al_1, \ \ldots, \ g_n = Al_n.$$

Покажем, что система $\{g_j\}_{j=1}^n$ линейно-независима.

Предположим обратное: пусть существуют коэффициенты $\beta_1, \ldots, \beta_n$, такие, что

$$\beta_1^2 + \ldots + \beta_n^2 \neq 0 \quad \text{и} \quad \sum_{j=1}^{n} \beta_j g_j = 0.$$

Следовательно,

$$\sum_{j=1}^{n} \beta_j A l_j = 0.$$

Применяя к обеим частям этого равенства обратный оператор A^{-1}, получаем

$$\sum_{j=1}^{n} \beta_j \, l_j = 0.$$

Пришли к противоречию с предположением о линейной независимости системы $\{l_j\}_{j=1}^n$.

Тогда из того, что система $Al_1, Al_2, \ldots, Al_n$ линейно-независима, следует

$$\det\{(Al_k, Al_j)\}_{k,j=1}^n \neq 0.$$

Значит система (7.4) однозначно разрешима и позволя-

ет определить приближенное решение, имеющее вид

$$y_n^* = \sum_{j=1}^{n} \alpha_k^* l_k.$$

Как мы показали выше при объяснении названия метода наименьших квадратов, этот элемент минимизирует $\|f - Ay_n\|$. Поэтому имеем

$$\|f - Ay_n^*\| \leq \|f - Ay_n\|$$

для любого элемента y_n вида

$$y_n = \sum_{j=1}^{n} \alpha_k l_k.$$

В частности, будем иметь

$$\|f - Ay_n^*\| \leq \|f - Ay_n^{**}\|.$$

3) Теперь легко получается оценка погрешности, а именно, оценка для величины

$$\|y^* - y_n^*\|.$$

Имеем

$$E_n(y^*) \leq \|y^* - y_n^*\| =$$
$$= \|A^{-1} \cdot A \cdot (y^* - y_n^*)\| \leq$$
$$\leq \|A^{-1}\| \cdot \|Ay^* - Ay_n^*\| =$$

$$= \|A^{-1}\| \cdot \|f - Ay_n^*\| \leq$$
$$\leq \|A^{-1}\| \cdot \|f - Ay_n^{**}\| =$$
$$= \|A^{-1}\| \cdot \|Ay^* - Ay_n^{**}\| \leq$$
$$\leq \|A^{-1}\| \cdot \|A\| \cdot \|y^* - y_n^{**}\| = \eta(A)E_n(y^*),$$

где

$$\eta(A) = \|A^{-1}\| \cdot \|A\|.$$

Таким образом, доказаны неравенства

$$E_n(y^*) \leq \|y^* - y_n^*\| \leq \eta(A)E_n(y^*).$$

В силу того, что система $\{l_k\}_{k=1}^{\infty}$ является полной в банаховом пространстве Y, будем иметь, что $E_n(y^*) \to 0$ при $n \to \infty$. Следовательно,

$$\|y^* - y_n^*\| \to 0 \quad \text{при } n \to \infty.$$

Этим и завершается доказательство теоремы.

7.3 Прямые методы, основанные на аппроксимации операторов

Следуя лауреату нобелевской премии Л.В. Канторовичу, прямыми называют методы, сводящие приближенное решение операторных уравнений к решению конечномерных СЛАУ. Очевидно, изученные нами методы

моментов относятся к прямым методам.

В этом пункте рассмотрим новое семейство прямых методов.

Пусть F, Y — линейные нормированные пространства, и пусть $f \in F$ — заданный элемент. Рассмотрим снова уравнение

$$Ay = f \tag{7.5}$$

в предположении, что $A : Y \to F$ — линейный обратимый оператор. Ищется решение $y \in Y$.

Пусть заданы конечномерные подпространства

$$Y_n \subset Y, \quad F_n \subset F$$

размерности

$$\dim Y_n = \dim F_n = n \in \mathbb{N}.$$

И пусть для каждого $n \in \mathbb{N}$ задан также элемент

$$f_n \in F_n,$$

и линейный конечномерный оператор

$$A_n \,:\, Y_n \to F_n.$$

Если

$$f_n \approx f, \quad A_n \approx A,$$

в некотором смысле, то естественно искать приближен-

ное решение $y_n \in Y_n$ уравнения (7.5) как решение уравнения

$$A_n\, y_n = f_n.$$

Это уравнение эквивалентно некоторой системе линейных алгебраических уравнений. В этом легко убедиться следующим образом.

В подпространстве Y_n существует базис $l_1, \ldots, l_n$, а в подпространстве F_n существует базис $g_1, \ldots, g_n$. Тогда справедливы следующие разложения соответствующих элементов по базисам:

$$f_n = \sum_{j=1}^{n} \beta_j g_j, \qquad A_n l_k = \sum_{j=1}^{n} l_{kj} g_j,$$

где $k = 1, \ldots, n$.

Возьмем элемент

$$y_n = \sum_{k=1}^{n} \alpha_k l_k \in Y_n$$

и подставим в уравнение $A_n y_n = f_n$. Получим

$$\sum_{k=1}^{n} \alpha_k \sum_{j=1}^{n} l_{kj} g_j = \sum_{k=1}^{n} \beta_j g_j.$$

Приравнивая коэффициенты при g_j, получаем следующую систему линейных алгебраических уравнений для

определения α_k:

$$\sum_{k=1}^{n} \alpha_k l_{kj} = \beta_j, \quad j = \overline{1, n}.$$

Как мы увидим ниже, эта система линейных алгебраических уравнений, а значит и уравнение $A_n\, y_n = f_n$, однозначно разрешима при некоторых дополнительных предположениях.

Теперь рассмотрим подробнее вопросы существования приближенного решения (т. е. решения уравнения $A_n\, y_n = f_n$) и его близости к точному решению операторного уравнения (7.5).

Точнее, рассмотрим обоснование сходимости описанного метода при следующих предположениях:

I) оператор $A : Y \to F$ является линейным и непрерывным, и существует непрерывный обратный оператор

$$A^{-1} : F \to Y;$$

II) для любого $n \in \mathbb{N}$ выбраны подпространства

$$Y_n \subset Y, \quad F_n \subset F$$

такие, что

$$\dim Y_n = \dim F_n = n;$$

III) элементы $f_n \in F_n$ выбраны так, что

$$\delta_n := \|f - f_n\| \to 0 \quad \text{при } n \to \infty;$$

IV) конечномерные линейные операторы

$$A_n : Y_n \to F_n$$

выбраны так, что

$$\varepsilon_n := \|A - A_n\|_{Y_n \to F} \to 0 \quad \text{при } n \to \infty.$$

Теорема 7.2 *При указанных предположениях I—IV справедливы следующие утверждения:*

для достаточно больших n уравнение $A_n y_n = f_n$ однозначно разрешимо и

$$\|y^* - y_n^*\| = O(\varepsilon_n + \delta_n), \quad \varepsilon_n + \delta_n \to 0,$$

где

$$y^* = A^{-1} f, \qquad y_n^* = A_n^{-1} f_n.$$

Доказательство. Выберем n настолько большим, чтобы

$$q_n := \varepsilon_n \|A^{-1}\| < 1.$$

Далее считаем, что $q_n < 1$. Рассмотрим произвольный элемент $y_n \in Y_n$. Имеем

$$\|y_n\| = \|A^{-1} A y_n\| \leq$$

$$\leq \|A^{-1}\| \cdot \|A y_n\| \leq$$

$$\leq \|A^{-1}\| \cdot \|A y_n - A_n y_n + A_n y_n\| \leq$$

$$\leq \|A^{-1}\| \cdot \|(A - A_n)y_n\| + \|A^{-1}\| \cdot \|A_n y_n\| \leq$$
$$\leq \|A^{-1}\|\|A - A_n\| \cdot \|y_n\| + \|A^{-1}\| \cdot \|A_n y_n\| =$$
$$= q_n\|y_n\| + \|A^{-1}\| \cdot \|A_n y_n\|.$$

Отсюда следует, что

$$(1 - q_n)\|y_n\| \leq \|A^{-1}\| \cdot \|A_n y_n\|,$$

поэтому

$$\|A_n y_n\| \geq m_n \cdot \|y_n\|, \tag{7.6}$$

где

$$m_n = \frac{1 - q_n}{\|A^{-1}\|} > 0.$$

Из неравенств $q_n < 1$ и (7.6) следует, что уравнение $A_n y_n = f_n$ однозначно разрешимо, причем существует ограниченный обратный оператор A_n^{-1}, такой, что

$$\|A_n^{-1}\| \leq \frac{1}{m_n}.$$

Таким образом, доказано, что существует искомое приближенное решение $y_n^* = A_n^{-1} f_n$ при достаточно больших n, для которых $q_n := \varepsilon_n\|A^{-1}\| < 1$.

Напомним, что $y^* = A^{-1}f$ — точное решение уравнения $Ay = f$. Оценим погрешность, т. е. норму разности $y^* - y_n^*$. Имеем

$$\|y^* - y_n^*\| = \|A^{-1}A(y^* - y_n^*)\| \leq$$

$$\leq \|A^{-1}\| \cdot \|Ay^* - Ay_n^*\|.$$

Так как $Ay^* = f$ и $A_n y_n^* = f_n$, то

$$\|Ay^* - Ay_n^*\| =$$

$$= \|f - Ay_n^*\| =$$

$$= \|f - f_n + f_n - Ay_n^*\| =$$

$$= \|f - f_n + (A_n - A)y_n^*\| \leq$$

$$\leq \|f - f_n\| + \|A_n - A\| \cdot \|y_n^*\| =$$

$$= \delta_n + \varepsilon_n \|y_n^*\|.$$

Следовательно,

$$\|y^* - y_n^*\| \leq \|A^{-1}\| \cdot (\delta_n + \varepsilon_n \|y_n^*\|).$$

С учетом неравенства

$$\|y_n^*\| \leq \|y_n^* - y^*\| + \|y^*\|,$$

получаем

$$\|y^* - y_n^*\| \leq \|A^{-1}\|\, \delta_n + \varepsilon_n \|A^{-1}\| \|y^* - y_n^*\| + \varepsilon_n \|A^{-1}\| \|y^*\|.$$

Так как $q_n = \varepsilon_n \|A^{-1}\|$, то это неравенство равносильно неравенству

$$(1 - q_n)\|y^* - y_n^*\| \leq \|A^{-1}\|(\delta_n + \varepsilon_n \|y^*\|).$$

Отсюда следует неравенство

$$\|y^* - y_n^*\| \le \frac{\|A^{-1}\|(1 + \|y^*\|)}{1 - q_n}(\delta_n + \varepsilon_n)$$

при $q_n = \varepsilon_n\|A^{-1}\| < 1$, что в свою очередь влечет соотношение

$$\|y^* - y_n^*\| = O(\varepsilon_n + \delta_n)$$

при $n \to \infty$, так как $\lim_{n\to\infty} q_n = 0$. Этим и завершается доказательство теоремы.

Замечание. *Обычно, подпространства F_n, Y_n и оператор A_n выбирают согласованно. Типичная ситуация такова. Предположим, что существует линейный оператор проектирования $P_n : F \to F_n$, обладающий характеристическим свойством $P_n^2 = P_n$. Тогда полагают $A_n := P_nA$, и приближенное решение определяют из уравнения*

$$P_nAy_n = P_nf \quad (y_n \in Y_n).$$

Область определения оператора P_n может быть шире, чем пространство F. Если пространство Y также лежит а области определения оператора P_n, то можно взять $Y_n = P_nY := \{P_ny : y \in Y\}$. Получаемые таким образом приближенные методы решения операторных уравнений называют проекционными.

Глава 8

Интегральные уравнения

Рассмотрим интегральное уравнение Фредгольма II рода

$$y(x) + \lambda \int_a^b h(x,t)\, y(t) dt = f(x), \quad a \le x \le b, \quad (8.1)$$

где ядро h и правая часть f — заданные непрерывные функции, λ — постоянная. В общем случае решение уравнения (8.1) не удается определить в явном виде. Поэтому возникает задача численного решения.

Очевидно, уравнение (8.1) можно рассматривать как уравнение $Ay = f$, где линейный оператор

$$A : C[a,b] \to C[a,b]$$

определен формулой

$$(Ay)(x) := y(x) + \lambda \int_a^b h(x,t)\, y(t) dt, \quad a \le x \le b.$$

Для малых λ решение может быть найдено с примене-

нием сходящегося итерационного процесса, заданного формулами

$$y_0(x) = 0, \;\; y_k(x) + \lambda \int_a^b h(x,t)\, y_{k-1}(t) dt = f(x), \;\; k \in \mathbb{N}.$$

Далее мы рассмотрим иные методы в предположении однозначной разрешимости уравнения (8.1) при заданных h, f и λ. Точное решение (8.1) будем обозначать символом y^*.

8.1 Полиномиальная коллокация

Метод полиномиальной коллокации для приближенного решения уравнения (8.1) заключается в следующем.

Задаем узлы $x_j \in [a,b]$, $j = 1, 2, \ldots, n$. Приближенное решение уравнения (8.1) ищем в виде полинома степени $n-1$, полагая

$$y_n(x) = \sum_{k=1}^{n} \alpha_k x^{k-1}.$$

Коэффициенты $\alpha_1, \alpha_2, \ldots, \alpha_n$ нужно определить таким образом, чтобы функция $y_n(x)$ представляла собой приближенное решение уравнения (8.1).

Требуем, чтобы $y_n(x)$ удовлетворяла уравнению (8.1) в узлах, т. е. требуем, чтобы

$$(Ay_n)(x_j) = f(x_j), \quad j = 1, 2, \ldots, n.$$

Легко видеть, что эта система равносильна следующей системе линейных алгебраических уравнений

$$\sum_{k=1}^{n} a_{jk}\alpha_k = f(x_j), \quad j = 1, 2, \ldots, n, \tag{8.2}$$

где

$$a_{jk} = x_j^{k-1} + \lambda \int_a^b h(x_j, t)t^{k-1}dt.$$

Предположим, что система (8.2) имеет решение

$$\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*.$$

Тогда в качестве приближенного решения (8.1) берем полином

$$y_n^* = \sum_{k=1}^{n} \alpha_k^* x^{k-1}.$$

Близость найденного таким образом приближенного решения y_n^* к точному решению $y^* = A^{-1}f$ уравнения (8.1) удается подтвердить лишь при некоторых дополнительных условиях.

Приведем теорему Л.В. Канторовича.

Теорема 8.1 *Пусть выполняются условия:*

1) существуют числа $r \geq 1$ и $M > 0$, такие, что функции $f(.)$ и $h(.,t)$ при любом $t \in [a,b]$ принадлежат пространству $W^rM :=$

$$:= \{g \in C[a,b],\ \exists g^{(r)},\ |g^{(r)}(x)| \leq M\ \ \forall x \in [a,b]\};$$

2) ядро h и параметр λ таковы, что уравнение (8.1) однозначно разрешимо при любом $f \in C[a,b]$;

3) $x_j\,(j = 1, 2, \ldots, n)$ — узлы Чебышева.

Тогда СЛАУ (8.2) однозначно разрешима для достаточно больших n и

$$\|y^* - y_n^*\|_{C[a,b]} = O\left(\frac{\ln n}{n^r}\right).$$

Схема доказательства. Доказательство основано на нескольких фактах.

а) Следуя Канторовичу, интерпретируем систему (8.2) как уравнение вида

$$A_n y_n = f_n,$$

где f_n — интерполяционный полином Лагранжа по узлам x_j для f, y_n — полином степени $\leq n-1$. А именно,

$$f_n(x) = \sum_{j=1}^{n} f(x_j) l_j(x),$$

где $l_j(x)$ — фундаментальные полиномы Лагранжа,

$$y_n(x) = \sum_{k=1}^{n} \alpha_k x^{k-1}.$$

Можем записать равенства $f_n = P_n f$, $y_n = P_n y_n$, где P_n — линейный оператор, сопоставляющий функции ее интерполяционный полином Лагранжа, постро-

енный по узлам x_j. В частности, имеем

$$f_n(x) = (P_n f)(x) := L_n(f; x) = \sum_{j=1}^{n} f(x_j) l_j(x).$$

Оператор P_n является проекционным оператором. Тогда уравнение

$$A_n y_n = f_n$$

запишется в виде уравнения

$$P_n A y_n = P_n f.$$

Поскольку

$$(Ay_n)(x) := y_n(x) + \lambda \int_a^b h(x,t)\, y_n(t) dt$$

и

$$(A_n y_n)(x) := y_n(x) + \lambda P_n \left(\int_a^b h(.,\, t)\, y_n(t) dt \right)(x) =$$

$$= y_n(x) + \lambda \int_a^b P_n \left(h(.,\, t) \right)(x)\, y_n(t) dt,$$

то справедливо равенство $((A - A_n) y_n)(x) =$

$$= \lambda \int_a^b y_n(t) [h(x,t) - (P_n h(.,\, t))(x)] dt$$

для всех точек $x \in [a, b]$. Поэтому

$$||(A - A_n)y_n|| \leq const\frac{M \ln n}{n^r}||y_n||,$$

что влечет оценку

$$\varepsilon_n := ||A - A_n|| = O\left(\frac{\ln n}{n^r}\right), \ \ (n \to \infty).$$

б) Далее Л.В. Канторович применяет теорему 7.2, полагая $A_n := P_nA$, $f_n = P_nf$, пользуясь равенством $y_n = P_ny_n$ для полинома y_n степени $\leq n-1$ и оценками

$$\|P_n\| = O(\ln n), \quad \varepsilon_n = O\left(\frac{\ln n}{n^r}\right), \quad \delta_n = O\left(\frac{\ln n}{n^r}\right).$$

С этими оценками мы знакомились ранее при изучении численных методов анализа. Как мы знаем, оценка $\|P_n\| = O(\ln n)$ доказана С.Н. Бернштейном в предположении, что узлы интерполяции являются узлами Чебышева. Оценки

$$\varepsilon_n = O\left(\frac{\ln n}{n^r}\right), \quad \delta_n = O\left(\frac{\ln n}{n^r}\right)$$

вытекают из теоремы Лебега об оценке погрешности интерполяции в сочетании с упомянутой теоремой Бернштейна и теоремой типа Джексона-Стечкина о скорости аппроксимации полиномами степени $n-1$ функции, принадлежащей семейству W^rM.

8.2 Метод сплайн-коллокации

Применяется та же идея, что и в методе полиномиальной коллокации, а именно, сведение к теореме 7.2. Но явные представления будут другими, так как рассматриваются сплайн-интерполяции.

Обсудим подробнее. На отрезке $[a, b]$ рассматриваем равноотстоящие узлы с шагом $h = (b - a)/n$, т. е. узлы

$$x_j = a + hj, \quad j = 0, 1, \ldots, n.$$

Функцию $y_n(x)$ ищем как сплайн 1-степени, а функции f сопоставляем ее сплайн 1-ой степени, т. е. полагаем

$$y_n(x) = \sum_{k=0}^{n} \alpha_k s_k(x),$$

$$f_n(x) = \sum_{k=0}^{n} f(x_k) s_k(x),$$

где s_k — фундаментальный сплайн 1-ой степени, удовлетворяющий условиям $s_k(x_j) = \delta_{kj}$ $(k, j = 0, 1, \ldots, n)$.

Очевидно, СЛАУ для определения α_k легко выписывается. Система имеет вид

$$\alpha_j + \sum_{k=0}^{n} b_{jk} \alpha_k = f(x_j), \quad j = 0, 1, 2, \ldots, n, \qquad (8.3)$$

где

$$b_{jk} = \lambda \int_a^b h(x_j, t) s_k(t) dt.$$

Предположим, что система (8.3) имеет решение

$$\alpha_0^*, \alpha_1^*, \ldots, \alpha_n^*.$$

Тогда в качестве приближенного решения уравнения (8.1) берем сплайн

$$y_n^* = \sum_{k=0}^{n} \alpha_k^* s_k(x).$$

Теорема 8.2 *Пусть интегральное уравнение (8.1) имеет решение $y^* = A^{-1} f \in C[a, b]$ при любой правой части $f \in C[a, b]$. Тогда для достаточно больших n приближенное решение y_n^*, определяемое описанным выше методом сплайн-коллокации, существует и единственно, причем*

$$\|y^* - y_n^*\|_{C[a,b]} = O(\varepsilon_n + \delta_n).$$

Здесь

$$\varepsilon_n = |\lambda| \int_a^b \omega\left(h(., t), \frac{b-a}{n}\right) dt, \quad \delta_n = \omega\left(f, \frac{b-a}{n}\right),$$

где

$$\omega\left(f, \frac{b-a}{n}\right), \quad \omega\left(h(., t), \frac{b-a}{n}\right)$$

— модули непрерывности соответствующих функций по переменной $x \in [a, b]$.

Схема доказательства. Пользуемся тем, что нам известна оценка погрешности при аппроксимации непрерывной функции ее сплайном 1-ой степени. При этом будут использованы соответствующие модули непрерывности. Имеем

$$\delta_n = \|f - f_n\|_{C[a,b]} = \max_x |f(x) - S_n^1(f, x)| \le \omega\left(f, \frac{b-a}{n}\right).$$

Далее, для любого y_n вида

$$y_n(x) = \sum_{k=0}^{n} \alpha_k s_k(x)$$

имеем

$$\|(A - A_n)y_n\|_{C[a,b]} =$$

$$= \left\| \lambda \int_a^b y_n(t)[h(x,t) - h_n(x,t)]dt \right\|_{C[a,b]} \le$$

$$\le |\lambda| \int_a^b \omega\left(h(.,t), \frac{b-a}{n}\right) dt \cdot \|y_n\|_{C[a,b]}.$$

Отсюда получаем

$$\varepsilon_n = \|A - A_n\| \le |\lambda| \int_a^b \omega\left(h(.,t), \frac{b-a}{n}\right) dt.$$

Поскольку h и f — заданные непрерывные функции, то $\varepsilon_n \to 0$ и $\delta_n \to 0$ при $n \to \infty$. Поэтому мы можем

применить теорему 7.2 и получить искомую оценку

$$\|y^* - y_n^*\|_{C[a,b]} = O(\varepsilon_n + \delta_n).$$

Замечание. *Пусть Y и F — линейные нормированные пространства функций, определенных на отрезке $[a,b]$. В этом случае методы коллокации можно использовать и для численного решения операторных уравнений вида $Ay = f$, где $A : Y \to F$ — заданный линейный непрерывный оператор.*

А именно, как и в методе моментов ищем приближенное решение вида $y_n(x) = \sum_{k=1}^n \alpha_k l_k(x)$, $x \in [a,b]$, определяем невязку $r_n(x) = f(x) - (Ay_n)(x)$.

Далее, задаем узлы $x_j \in [a,b]$, $j = 1, 2, \ldots, n$, и числа α_k определяем как решение системы уравнений $r_n(x_j) = 0$, $j = 1, 2, \ldots, n$.

8.3 Метод механических квадратур

Рассмотрим следующее уравнение Фредгольма II рода

$$(Ay)(x) := y(x) - \lambda \int_a^b \rho(t) h(x,t) y(t) dt = f(t).$$

Считаем, что заданные функции f, h и ρ являются непрерывными, а также $\rho(x) > 0$ при $x \in [a,b]$.

Предполагаем, что при любой непрерывной правой части существует единственное решение $y^* \in C[a,b]$.

Опишем кратко метод механических квадратур.

Обозначим $\psi(t) = h(x,t)y(t)$. Берем квадратурную формулу вида

$$\int_a^b \rho(t)\psi(t)dt = \sum_{k=1}^{n} A_k\psi(x_k) + R_n(\psi),$$

$x_1, \ldots, x_n$ — заданные узлы, $\psi(x_k) = h(x,x_k)y(x_k)$, приближенные значения $y_n(x) \approx y(x)$ в узлах x_j ищем как решение СЛАУ:

$$y_n(x_j) - \lambda \sum_{k=1}^{n} A_k h(x_j, x_k) y(x_k) = f(x_j), \quad j = 1, 2, \ldots, n.$$

т. е. $c_j = y_n(x_j)$ — неизвестные, подлежащие определению числа, удовлетворяющие СЛАУ

$$c_j - \lambda \sum_{k=1}^{n} A_k h(x_j, x_k) c_k = f(x_j), \quad j = 1, 2, \ldots, n.$$

Таким образом, мы получаем ряд новых вычислительных схем, основанных на применении квадратурных формул, так как имеется множество квадратурных формул (в частности, квадратурные формулы прямоугольников, трапеций и Симпсона, а также квадратурные формулы Гаусса).

Многие задачи для ОДУ являются эквивалентными соответствующим интегральным уравнениям. Например, нетрудно показать, что интегральное уравне-

ние

$$y(x) + (1-x)\int_0^x tf(t, y(t))dt +$$

$$+x\int_x^1 (1-t)f(t, y(t))dt = Bx$$

эквивалентно следующей краевой задаче

$$y'' = f(x, y), \quad y(0) = 0,\ y(1) = B,$$

где $f : [0,1] \times \mathbb{R} \to \mathbb{R}$ предполагается непрерывной функцией. Поэтому естественно применять методы коллокации при решении краевых задач для ОДУ.

Рассмотрим дифференциальное уравнение

$$y'' + p(x)y' + q(x)y = f(x), \quad a \le x \le b, \tag{8.4}$$

где коэффициенты $p, q, f \in C[a,b]$. Ищем $y \in C^2[a,b]$, удовлетворяющее краевым условиям

$$y(a) = y(b) = 0. \tag{8.5}$$

В этом случае конструкция приближенного решения $y_n(x)$ методом полиномиальной коллокации такова. Приближенное решение ищем в виде следующего полинома

$$y_n(x) = (x-a)(b-x)\sum_{k=1}^{n} \alpha_k x^{k-1}. \tag{8.6}$$

Краевые условия $y_n(a) = y_n(b) = 0$ обеспечены автоматически. Неизвестные коэффициенты $\alpha_1, \ldots, \alpha_n$ опре-

деляем так:

* задаем узлы $x_1, x_2, \ldots, x_n \in (a, b)$,

** потребуем, чтобы $y_n(x)$ в виде (8.6) удовлетворяло (8.4) точно, но только в узлах, т. е. в точках $x = x_j$.

Обозначим

$$\psi_k(x) = (x - a)(b - x)x^{k-1},$$

следовательно,

$$y_n(x) = \sum_{k=1}^{n} \alpha_k \psi_k(x).$$

Подставляем $y_n(x)$ в (8.4) и требуем выполнение уравнения в узловых точках $x = x_j$. Получаем

$$\sum_{k=1}^{n} \alpha_k \{\psi_k''(x_j) + p(x_j)\psi_k'(x_j) + q(x_j)\psi_k(x_j)\} = f(x_j)$$

для любого $j = 1, 2, \ldots, n$. Если эта СЛАУ имеет решение $\alpha_1^*, \ldots, \alpha_n^*$, то получаем искомое приближенное решение в виде следующего полинома

$$y_n^*(x) = (x - a)(b - x) \sum_{k=1}^{n} \alpha_k^* x^{k-1}.$$

Л.В. Канторович доказал, что *указанное приближенное решение y_n^* сходится к точному решению при $n \to \infty$, если выполнены следующие условия:*

1) задача (8.4) – (8.5) имеет единственное решение $y^* \in C^2[a,b]$*;*

2) $p, q, f \in W^r M$*, т. е. эти функции имеют производные порядка* $r \geq 1$*, ограниченные по модулю некоторой константой* M*;*

3) $x_1, x_2, \ldots, x_n \in (a,b)$ *— узлы Чебышева.*

8.4 Задачи и упражнения

1. Убедитесь в том, что функция

$$y(x) = x^2 + 1$$

является решением интегрального уравнения

$$y(x) - \int_0^1 \frac{x^2}{t^2+1}\, y(t)\, dt = 1.$$

2. Подбором ядра $K(x,t)$ и правой части $f(x)$ постройте три разных интегральных уравнения вида

$$y(x) + \int_a^b K(x,t)\, y(t)\, dt = f(x), \quad a \leq x \leq b,$$

решением которых является одна и та же функция

$$y(x) = e^x.$$

3. Методом полиномиальной коллокации найдите

приближенное решение интегрального уравнения

$$y(x) - \int_0^1 \frac{x^2}{t^2+1}\, y(t)\, dt = 1, \quad 0 \le x \le 1.$$

Рассмотрите три случая с узлами Чебышева при $n = 3$, $n = 6$ и $n = 9$. Сравните с точным решением.

4. Методом сплайн-коллокации найдите приближенное решение интегрального уравнения

$$y(x) - \int_0^1 \frac{x^2}{t^2+1}\, y(t)\, dt = 1, \quad 0 \le x \le 1.$$

Рассмотрите три случая: $n = 3$, $n = 6$ и $n = 9$.

Сравните полученные решения с точным.

5. Пусть $a < b$, $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ — непрерывная функция. Докажите, что интегральное уравнение

$$y(x) + \frac{b-x}{b-a} \int_a^x (t-a) f(t, y(t)) dt +$$

$$+ \frac{x-a}{b-a} \int_x^b (b-t) f(t, y(t)) dt =$$

$$= \frac{(\beta - \alpha)x + (\alpha b - \beta a)}{b-a}$$

эквивалентно следующей краевой задаче

$$y'' = f(x, y), \quad y(a) = \alpha,\ y(b) = \beta$$

для ОДУ второго порядка.

6. Докажите следующее свойство проекционных операторов.

Пусть X — векторное пространство над полем K, $P : X \to X$ — линейный оператор, обладающий свойствами: $P^2 = P$, $P \neq I$, $P \neq 0$. Тогда $0 \in K$ и $1 \in K$ являются собственными значениями оператора P. Если $\lambda \in K \setminus \{0, 1\}$, то $P - \lambda I$ является биекцией, причем

$$(P - \lambda I)^{-1} = \frac{1}{\lambda(1-\lambda)} P - \frac{1}{\lambda} I.$$

Глава 9

Рекомендуемая литература

Большинство включенных в библиографию работ представляют собой учебные пособия. Каждая книга имеет свои особенности. Так, например, книга Н.Н. Калиткина рассчитана на читателя, который занимается применением численных методов к прикладным проблемам. *В этой книге вы найдете огромное количество практических рекомендаций вычислителю.*

Включены также работы, изучение которых будет весьма полезно, но требует серьезных усилий. К ним относятся книга И.П. Натансона по конструктивной теории функций, книга С.Б. Стечкина и Ю.Н. Субботина по сплайнам и обзорная статья В.М. Тихомирова по теории приближения функций.

Следует также указать, что научная и учебная литература по численным методам весьма обширна. Ниже приведена лишь часть известных книг, написанных специалистами по вычислительной математике.

Литература

[1] Бабенко К.И. *Основы численного анализа.* Москва: Наука. Гл. ред. физ.-мат. лит. 1986.

[2] Бадриев И.Б., Волошановская С.Н. *Численные методы. Приближение функций и численное интегрирование.* Учебное пособие. Под ред. Р.З. Даутова. Казань: изд-во Казанского ун-та. 1990.

[3] Бахвалов Н.С., Лапин А.В., Чижонков Е.В. *Численные методы в задачах и упражнениях.* Учебное пособие. Под ред. В.А. Садовничего. Москва: Высшая школа. 2000.

[4] Березин И.С., Жидков Н.П. *Методы вычислений.* Ч. 1, Москва: Наука, 1966. То же. Ч. 2. Физматгиз, 1962.

[5] Богачев К.Ю. *Практикум на ЭВМ. Методы приближения функций.* Москва: Изд-во ЦПИ при механико-математическом ф-те МГУ. 2002.

[6] Бут Э.Д. *Численные методы.* Москва: ГИФМЛ. 1959.

[7] Дробышевич В.И., Дымников В.П., Ривин Г.С. *Задачи по вычислительной математике.* Учебное пособие. Под ред. Г.И. Марчука. Москва: Наука. 1980.

[8] Калиткин Н.Н. *Численные методы.* Учебное пособие для вузов. Под ред. А.А. Самарского. Москва: Наука. Гл. ред. физ.-мат. лит. 1978.

[9] Канторович Л.В., Крылов В.И. *Приближенные методы высшего анализа.* Москва–Ленинград: ГИТТЛ, 1949.

[10] Крылов В.И., Бобков В.В., Монастырный П.И. *Вычислительные методы.* Т.1, Москва: Наука, 1976. То же. Т. 2. Москва: Наука, 1977.

[11] Натансон И.П. *Конструктивная теория функций.* Москва: Гостехиздат. 1949.

[12] Рябенький В.С. *Введение в вычислительную математику.* Серия "Физтеховский учебник". Москва: Физматлит. 2008.

[13] Самарский А.А., Гулин А.В. *Численные методы.* Учебное пособие для вузов. Москва: Наука. Гл. ред. физ.-мат. лит. 1989.

[14] Срочко В.А. *Численные методы. Курс лекций.* Учебное пособие для вузов. Санкт-Петербург: Изд-во ЛАНЬ. 2010.

[15] Стечкин С.Б., Субботин Ю.Н. *Сплайны в вычислительной математике*. Москва: Наука. Гл. ред. физ.-мат. лит. 1976.

[16] Тихомиров В.М. *Теория приближений*. В книге "Современные проблемы математики. Фундаментальные направления. Т. 14. Итоги науки и техники, ВИНИТИ АН СССР"Москва. 1987, с. 103-260.

[17] Фаддеев Д.К., Фаддеева В.Н. *Вычислительные методы линейной алгебры*. Москва–Ленинград: Физматгиз, 1963.

Учебное пособие

Авхадиев Фарит Габидинович

ОСНОВЫ
ЧИСЛЕННЫХ МЕТОДОВ

Компьютерная верстка
Ф.Г. Авхадиева

Дизайн обложки
Р.М. Абдрахмановой

Подписано в печать 5.07.2022.
Бумага офсетная. Печать цифровая.
Формат 60x84 1/16. Гарнитура " Book Antiqua".
Усл. печ. л. 25,8.
Тираж 50 экз. Заказ 80/6

Отпечатано в типографии
Издательства Казанского университета

420008, г. Казань, ул. Профессора Нужина, 1/37
тел. (843) 233-73-59, 233-73-28