МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГАОУ ВО «КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ МАТЕМАТИКИ И МЕХАНИКИ
им. Н.И. ЛОБАЧЕВСКОГО

Ф.Г. АВХАДИЕВ

ЧИСЛЕННЫЕ МЕТОДЫ АЛГЕБРЫ И АНАЛИЗА

УЧЕБНОЕ ПОСОБИЕ

КАЗАНЬ – 2019

УДК 517

Рекомендовано к опубликованию и размещению на сайте Казанского (Приволжского) федерального университета Учебно-методической комиссией Института математики и механики им. Н.И. Лобачевского (протокол № 6 от 21.05.19 г.)

Рецензенты:
к.ф.-м.н., доцент **Агачев Ю.Р.**
к.пед.н., доцент **Маклецов С.В.**

Научный редактор:
к.ф.-м.н., доцент **Насибуллин Р.Г.**

Авхадиев Ф.Г.

Численные методы алгебры и анализа / Ф.Г. Авхадиев. — Казань: Изд-во Казан. ун-та, 2019. — 200 с.

Учебное пособие представляет собой первую и вторую части обработанного курса лекций по численным методам, читаемого автором студентам Казанского федерального университета в Институте математики и механики им. Н. И. Лобачевского. Первая часть включает численные алгоритмы решения алгебраических уравнений и систем. Во второй части изложены основы теорий интерполяции, квадратурных формул и наилучших приближений в функциональных пространствах.

Книга предназначена для студентов-бакалавров, изучающих численные методы.

УДК 517

Содержание

1 Системы линейных алгебраических уравнений (СЛАУ)

1.1 Предисловие

В этой 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).

Будем рассматривать также базовые методы приближенного решения нелинейных алгебраических уравнений (НАУ), и системы нелинейных алгебраических уравнений (СНАУ).

Кроме того, будем знакомиться с методами нахождения собственных чисел и собственных векторов конечномерных линейных операторов, заданных матрицами. В этом случае для квадратной матрицы A порядка n рассматривается уравнение $Ax = \lambda x$, ищутся собственные числа λ такие, для которых указанная система линейных алгебраических уравнений имеет решение $x_\lambda \neq 0$. Такой вектор x_λ называется собственным вектором для матрицы (оператора) A, соответствующим собственному числу λ.

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.

1.2 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.

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 умножений 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 $n+1$

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

2 Метод Гаусса, его модификации и применения

Речь идет об известном с древнейших времен методе последовательного исключения переменных. В настоящее время этот подход принято называть методом Гаусса в честь величайшего математика всех времен Карла Фридрихса Гаусса (1777-1855), систематически использовавшего метод последовательного исключения переменных на практике, при обработке результатов проводимых им геодезических измерений в Нижней Саксонии.

2.1 Основные алгоритмы метода Гаусса

Рассмотрим систему линейных алгебраических уравнений общего вида

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

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}, $(k = 2, ..., n)$, и вычитаем из k-го уравнения. Получаем новые уравнения вида

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

Фактически мы имеем новую систему линейных алгебраических уравнений порядка $(n-1)$ с неизвестными $x_2, ..., x_n$. 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 (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 $a_{k2}^{(1)}$ $(k = 3, ..., n)$ и вычитаем из k-го уравнения. Уравнения с номерами $k = 3, ..., n$ преобразуются к следующему виду

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

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 (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

$$\begin{cases} x_1 + a_{12}^{(1)} x_2 + \ldots + a_{1n}^{(1)} x_n = b_1^{(1)} \\ x_2 + \ldots + a_{2n}^{(2)} x_n = b_2^{(2)} \\ \ldots\ldots\ldots \\ x_n = b_n^{(n)} \end{cases} . \qquad (4)$$

Переход от СЛАУ вида (1) к системе вида (4) называется **прямым ходом метода Гаусса**.

Обратный ход метода Гаусса — нахождение неизвестных $x_1, ..., x_n$ из системы (4) в порядке, обратном номеру неизвестной. Согласно (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 (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 $x_2, ..., x_n$.

Замечание 1. Если $\det A \neq 0$, то за счет перестановки строк в системе (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 переставить строки с номерами 1 и j, и проделать шаги 1.1 и 1.2 с новым $\tilde{a}_{11} = a_{1j} \neq 0$.

Если $\det A \neq 0$ и произведение $a_{11} a_{22}^{(1)} a_{33}^{(2)} \ldots a_{kk}^{(k-1)} \neq 0$, то детерминант системы

$$\begin{cases} a_{k+1\,k+1}^{(k)} x_{k+1} + a_{k+1\,k+2}^{(k)} x_{k+2} + \ldots + a_{k+1\,n}^{(2)} x_n = b_{k+1}^{(k)} \\ a_{k+2\,k+1}^{(k)} x_{k+1} + a_{k+2\,k+2}^{(k)} x_4 + \ldots + a_{k+2\,n}^{(k)} x_n = b_{k+2}^{(k)} \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n\,k+1}^{(k)} x_{k+1} + a_{n\,k+2}^{(k)} x_{k+2} + \ldots + a_{n\,n}^{(k)} x_n = b_n^{(k)} \end{cases} . \qquad (5)$$

отличен от нуля. Поэтому среди элементов первого столбца $a^{(k)}_{k+1\,k+1}, a^{(k)}_{k+2\,k+1}, ..., a^{(k)}_{n\,k+1}$ имеется хотя бы один элемент $a^{(k)}_{j\,k+1}$, отличный от нуля. Ясно, что перестановка строк с номерами $k+1$ и j позволяет продолжить прямой ход алгоритма Гаусса с ведущим элементом $\tilde{a}^{(k)}_{k+1\,k+1} = a^{(k)}_{j\,k+1} \neq 0$.

В заключение отметим, что 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).

2.2 Подсчет числа 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 $k = 2, 3, ..., n$, 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

$$N_2 = n^2 + (n-1)^2 + \ldots + 2^2 + 1^2 = \frac{n(n+1)(2n+1)}{6} = O(n^3).$$

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 $a_{kk}^{(k-1)}$, ..., $a_{nk}^{(k-1)}$. 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 $a_{kk}^{(k-1)}$, ..., $a_{kn}^{(k-1)}$. 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

$$\begin{cases} x_1 + a_{13}^{(2)} x_3 + \ldots + a_{1n}^{(2)} x_n = b_1^{(2)} \\ x_2 + a_{23}^{(2)} x_3 + \ldots + a_{2n}^{(2)} x_n = b_2^{(2)} \end{cases} . \qquad (6)$$

Шаг 2. Пользуясь (6), 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.

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

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

2.3 Вычисление определителей и обратных матриц

Речь идет о фактах, с которыми мы знакомы по курсу линейной алгебры. Рассмотрим сначала вычисление определителя

квадратной матрицы. Преобразования первого шага основного алгоритма Гаусса при условии $a_{11} \neq 0$ приводят к формуле

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

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

Продолжая процесс, получаем формулу

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

т. е. определитель матрицы равен произведению ведущих элементов.

Рассмотрим теперь задачу вычисления 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 необходимо 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, т. е. для определения чисел $x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)}$ имеем следующую систему 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 исключением 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, каждое преобразование строк типа (α), (β) и (γ) равносильно умножению слева преобразуемой матрицы 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 $BE = B = X = A^{-1}$.

В заключение приведем примеры, показывающие равносильность преобразований строк типа (α), (β) и (γ)

умножению слева преобразуемой матрицы 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) Умножение строки 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:

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

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

Итерационное уточнение обратной матрицы, вычисленной приближенно

Пусть X_0 — решение матричного уравнения $AX = E$. Если $\|E - AX_0\| = 0$, то решение определено точно. Вообще говоря, в результате вычислений обратная матрица определяется

приближенно из-за округлений и других погрешностей вычислений, и мы имеем лишь приближенное равенство $X_0 \approx A^{-1}$.

Предположим, что ε — заданная 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

$$X_k = X_{k-1}(2E - AX_{k-1}), \qquad k = 1, 2, \ldots \tag{7}$$

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

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

и, следовательно, итерационная последовательность X_k из (7) сходится к A^{-1}, т. е.

$$\lim_{k\to\infty} \|A^{-1} - X_k\| = 0.$$

Доказательство. Подставляя вместо X_k ее выражение из формулы (7), 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

$$\|A^{-1} - X_k\| \le \|A^{-1}\| \cdot \|E - AX_k\| \le \|A^{-1}\| \cdot \|E - AX_0\|^{2^k} =$$

$$= \|A^{-1}\| \cdot q^{2^k} \to 0 \quad \text{при} \quad k \to \infty.$$

2.4 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

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

когда матрица имеет вид

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

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

Излагаемый ниже метод называется методом прогонки и используется при численном решении краевой задачи для линейного дифференциального уравнения второго порядка.

Соответствующая система линейных алгебраических уравнений имеет вид

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

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}, ..., 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 отличность от нуля знаменателей в формулах для прогоночных коэффициентов.

Определение 2.1 *Говорят, что матрица имеет диагональное преобладание, если для любого номера* $i = 1, 2, ..., n$

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

В частности, для нашей трехдиагональной матрицы условие диагонального преобладания сводится к следующим требованиям

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

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 неравенство 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 2.0.1 *Пусть 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.

3 Другие точные методы решения СЛАУ

3.1 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 систему линейных алгебраических уравнений

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

Предположим, что $\det A \neq 0$. Тогда существует единственное решение $x^* = (x_1^*, x_2^*, ..., x_n^*)$ 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 $y = (x_1, \ldots, x_n, 1)$:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n - b_1 \cdot 1 = 0 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n - b_2 \cdot 1 = 0 \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n - b_n \cdot 1 = 0 \end{cases} . \tag{8}$$

Рассмотрим $(n+1)$-мерные векторы

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

и скалярные произведения

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

Тогда система (8) запишется в виде системы уравнений

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

Таким образом, решение СЛАУ свелось к следующей задаче:

нужно найти $(n+1)$-мерный вектор $y = (x_1, \ldots, x_n, 1)$ с последней координатой, равной единице, и ортогональный заданным $(n+1)$-мерным 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 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

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

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

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

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

$$v^{(1)} = \frac{a^{(1)}}{\|a^{(1)}\|}, \quad \ldots,$$

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

По построению имеем: вектор $v^{(n+1)} = (v_1^{(n+1)}, v_2^{(n+1)}, ..., v_{n+1}^{(n+1)})$ ортогонален векторам $v^{(1)}, v^{(2)}, \ldots, v^{(n)}$ и $a^{(1)}, a^{(2)}, \ldots, a^{(n)}$, 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 $y^* = (x_1^*, x_2^*, \ldots, x_n^*, 1)$.

Имеем: $(n+1)$-мерный вектор

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

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

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

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

3.2 Точные методы решения СЛАУ, основанные на факторизации матриц

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

Пусть A — квадратная матрица порядка n, представимая в виде

$$A = BC,$$

где C − верхнетреугольная матрица

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

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

Рассмотрим систему уравнений $Ax = b$, т. е. $BCx = b$. Решение этой системы может быть получено последовательным 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 $y_1, y_2, \dots, y_n$ по формулам

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

Зная $y_1, y_2, \dots, y_n$, последовательно определяем $x_n, x_{n-1}, \dots, x_1$ по формулам

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

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 алгебраических уравнений

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

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

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

Отметим, что 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

$$A = S^* D S, \tag{9}$$

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 (9) было возможным, необходимо еще отличие от нуля некоторых коэффициентов, возникающих в ходе преобразований (см. ниже примеры).

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

$$S^* D S x = b$$

сводится к последовательному решению двух простых систем

$$\begin{cases} S^* D y = b \\ S x = y \end{cases},$$

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 (9).

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 положить, что d_{11} равен плюс или минус

единице, точнее, полагаем $d_{11} = \operatorname{sign} a_{11}$. Тогда

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

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, позволяющие найти другие коэффициенты c_{ij}, b_{ij} (см., например, стр. 26-32 книги Д.К. Фаддеева и В.Н. Фаддеевой “Вычислительные методы линейной алгебры”, Физматгиз, М.-Л., 1963).

4 Новые числовые характеристики матриц

Для матриц определяют ряд числовых характеристик. Некоторые из них, в частности, детерминант, собственные числа и след матрицы, известны вам из стандартного курса линейной алгебры. Здесь мы рассмотрим несколько новых характеристик, которые нужны при изучении численных методов.

4.1 p - нормы векторов в евклидовом пространстве

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

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

согласованная со скалярным произведением $(x, y) = x_1y_1 + x_2y_2 + ... + x_ny_n$, $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$, $y = (y_1, y_2, ..., y_n) \in \mathbb{R}^n$ (или в $\mathbb{C}^n$, где скалярное произведение определено формулой $(x, y) = x_1\overline{y_1} + x_2\overline{y_2} + ... + x_n\overline{y_n}$). 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

$$c_1\|x\|' \leq \|x\|'' \leq c_2\|x\|'$$

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

Но 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 ($1 \leq p \leq +\infty$):

$$\|x\|_p := \left(|x_1|^p + |x_2|^p + ... + |x_n|^p\right)^{1/p}, \quad 1 \leq p < +\infty,$$

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

где $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$ (или $\mathbb{C}^n$). Очевидно, если все координаты вектора $x = (x_1, x_2, ..., x_n)$ 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 4.1 *Для любого вектора $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$ (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

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

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

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 $y(p)$ — невозрастающая функция. А этот факт проверяется простыми вычислениями, так как

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

с учетом неравенств $\ln \alpha_j \le 0$, $\ln\left(1 + \sum_{j=2}^{m} \alpha_j^p\right) > 0$.

Одновременно мы показали, что норма $\|x\|_p$ является строго убывающей функцией от параметра $p \in [1, +\infty]$, когда количество ненулевых координат вектора не меньше, чем 2. Этим и завершается доказательство.

4.2 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 мы будем называть норму этого линейного оператора A, определяемую 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 $c \ge 0$ является нормой матрицы A тогда и только тогда, когда выполняются два следующих свойства:

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

2) существует вектор $x \neq \theta$ такой, что $\|Ax\| = c\,\|x\|$.

В силу линейности оператора при определении нормы и проверке приведенных свойств 1) и 2) можно ограничиться 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 4.2 *Для квадратной матрицы A порядка n с элементами a_{kj} из $\mathbb{R}$ или $\mathbb{C}$ имеют место следующие формулы для норм:*

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

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

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

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

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

Доказательство. Обозначим $y = Ax$, где $x = (x_1, x_2, ..., x_n)$ и $y = (y_1, y_2, ..., y_n)$. Имеем

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

1) Очевидно,

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

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

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

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 $x^0 = (x_1^0, x_2^0, ..., x_n^0)$, у которой $x_{j_0}^0 = 1$, а все остальные координаты равны нулю, и вектор $Ax^0 = y^0 = (y_1^0, y_2^0, ..., y_n^0)$. 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 $x^0 = (x_1^0, x_2^0, ..., x_n^0)$ 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 $x^0 = (x_1^0, x_2^0, ..., x_n^0)$ 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 $\|x^0\|_\infty = 1$ и для вектора $Ax^0 = y^0 = (y_1^0, y_2^0, ..., y_n^0)$ 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 $D = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$ — 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, 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 $y^0 = (y_1^0, y_2^0, ..., y_n^0)$ такой, что $y_{k_0}^0 = 1$, а все остальные координаты этого вектора равны нулю.

Теорема доказана.

Следует отметить, что в кольце матриц порядка n в линейной алгебре определяется матричная норма как неотрицательный функционал, удовлетворяющий следующим условиям:

1) $\|A\| = 0 \Longleftrightarrow A = 0$;

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 всем этим требованиям.

Наряду с ними, требованиям 1)-4) удовлетворяет норма Фробениуса

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

которую иногда называют евклидовой нормой матрицы.

Заметим, что при любом определении нормы векторов операторная норма единичной матрицы $E = \mathrm{diag}(1, 1, ..., 1)$ равна единице, а $\|A\|_F = \sqrt{n}$. Следовательно, норма Фробениуса не является операторной нормой при $n \geq 2$.

Покажите, что при любом определении матричной нормы справедливы неравенства

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

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

Определение 4.1 *Пусть* $\lambda_i = \lambda_i(B)$ *— собственные значения матрицы* B*. Число* $\rho(B) = \max_i |\lambda_i(B)|$ *называется спектральным радиусом матрицы* B*.*

В терминах спектрального радиуса равенство

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

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

4.3 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

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

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

Имеем следующие числовые характеристики погрешностей:

$\|\delta\|$ — 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 4.3 *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 интересные 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.

Отметим попутно, что при любом определении нормы

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

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

5 Приближенные методы решения СЛАУ

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

$$Ax = b, \tag{10}$$

а также вида

$$x = Bx + c, \tag{11}$$

где $b = (b_1, \ldots, b_n)$ и $c = (c_1, \ldots, c_n)$ — заданные вектора из $\mathbb{R}^n$ или из $\mathbb{C}^n$. Свести (10) к эквивалентной системе вида (11) можно множеством разных способов. Опишем простейший прием.

Пусть α — фиксированное число, отличное от нуля, и пусть E — единичная матрица. Тогда СЛАУ вида (10) равносильна системе $0 = (b - Ax)\alpha$, следовательно, равносильна СЛАУ

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

Последняя система имеет вид (11) с матрицей $B = E - \alpha A$ и заданным вектором $c = \alpha b$.

5.1 Метод простой итерации

Этот метод применяется для нахождения решения $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ системы вида (11). Алгоритм метода простой итерации 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 (т.е. нулевого) приближения желательно назначить вектор, близкий к решению. Но если мы не имеем никакой информации о решении, то нулевое приближение берем "с потолка" , т.е. выбираем произвольно. Последующие приближения определяются по правилам:

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

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

$$\cdots\cdots\cdots$$

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

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 (11). Действительно, с учетом непрерывности операций вида Bx, $\|x\|$, при $k \to \infty$ из соотношения $x^k = Bx^{k-1} + c$ получаем, что

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

Справедливо следующее утверждение.

Теорема 5.1 *Метод простой итерации $x^k = Bx^{k-1} + c$ ($k = 1, 2, 3, \ldots$) для СЛАУ вида (11) сходится к единственному решению x^* при любом выборе нулевого приближения x^0 тогда и только тогда, когда спектральный радиус $\rho(B) < 1$ (т.е. все собственные числа по модулю < 1).*

Эту теорему мы возьмем без доказательства. Попробуйте доказать ее самостоятельно с учетом следующего указания. Так как $\rho(B) < 1$, то 1 не является собственным значением матрицы B. Поэтому $\det(E - B) \neq 0$, существует обратная матрица $(E - B)^{-1}$, и ее можно определить как сумму сходящегося ряда

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

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 простой итерации,

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 5.2 *Пусть* $\|B\| < 1$. *Тогда*

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

2) метод простой итерации $x^k = Bx^{k-1} + c$ $(k = 1, 2, 3, ...)$ *сходится при любом выборе нулевого приближения* x^0,

3) имеет место оценка

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

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 вектора $z^* = x^* - y^*$. Но тогда, пользуясь определением нормы оператора, получаем

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

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

Докажем теперь *существование* решения. Пусть $m \geq 1, \quad p \geq 1$ – 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

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

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 5.2 для $\|x^* - x^k\|$.

Замечание. В качественном плане теорема 5.2 является следствием теоремы 5.1, 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.

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

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

для любого $i = 1, 2, ..., n$.

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

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

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

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

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

$$c = \left(\frac{b_1}{a_{11}}, \frac{b_2}{a_{22}}, \ldots, \frac{b_n}{a_{nn}}\right), \quad B = (b_{ij}), \quad b_{ii} = 0, \quad b_{ij} = -\frac{a_{ij}}{a_{ii}} (i \neq j).$$

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 простой итерации и наших построений получаем следующее утверждение.

Теорема 5.3 *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 эквивалентна системе

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

Таким образом, 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. Наш первоначальный выбор $B = E - \alpha A$ — частный случай этого способа.

Известна такая теорема.

Теорема 5.4 *Пусть A — эрмитова (т.е. $A = A^*$) и положительно определенная (т.е. $(Ax, x) > 0, \forall x \neq 0$) матрица. Тогда для всех достаточно малых α спектральный радиус*

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

Способ 2. Матрицу A представляем в виде $A = C + D$, причем это разложение подбираем так, чтобы $\det C \neq 0$. 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

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

5.2 Итерационные 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 $x = (x_1, \ldots, x_n)$, $c = (c_1, \ldots, c_n)$. 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 простой итерации для любого номера i координаты последующей итерации определялись по формуле

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

Метод Зейделя представляет собой модификацию метода простой итерации, 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

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

Далее, 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 простой итерации 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 простой итерации к вопросу о сходимости метода Зейделя с заменой матрицы B на матрицу $\widetilde{B} = (E - H)^{-1}F$.

Теорема 5.5 *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. И справедлива оценка*

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

где $q = \|(E-H)^{-1}F\|$.

2) 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

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

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

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

$$C = \begin{pmatrix} 0 & a_{12} & \ldots & a_{1n} \\ a_{21} & 0 & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & 0 \end{pmatrix}.$$

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

$$Bx^k + Cx^{k-1} = b. \tag{12}$$

Поскольку

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

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 простой итерации

$$x^k = \widetilde{B}x^{k-1} + c, \tag{13}$$

где

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

Но тогда теоремы о сходимости метода простой итерации (13) позволяют сформулировать теоремы сходимости для изучаемого метода Зейделя.

Теорема 5.6 *1) 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

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

5.3 Методы градиентного спуска и их обобщения

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, т. е. $(Ax, x) > 0$ 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

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

Теорема 5.7 *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 $(x, Ay) = (Ay, x)$, $Ax^* = b$ 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

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

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

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

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

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

Сравнение последних формул наглядно показывает связь между решениями системы $Ax = b$ с действительной 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 $k = 1, 2, ..., n$ 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 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

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

Рассматриваем для определения x_2^1 новую вспомогательную 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 5.6 о сходимости.

Методы 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, достаточно малое, положительное число. Поскольку градиент функции направлен в сторону возрастания этой функции, то интуитивно понятно, что в пределе мы придем к точке минимума.

Геометрически понятно также, что если число τ велико, то итерационный процесс будет расходящимся. Поэтому возникает важный вопрос о выборе подходящего параметра τ. Более общий метод градиентного спуска, называемый нестационарным методом градиентного спуска, имеет вид

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

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

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

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

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 $0 < \tau < 1$.

Применим описанную идею спуска к поиску решения системы линейных алгебраических уравнений вида $Ax = b$, где $A = (a_{ij})_{i,j=1}^n$ — действительная симметричная 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 $t > 0$, 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 5.8 *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

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

$$= (E - t\,A)(E - t\,A)(x^* - x^{k-2}) = ... = (E - tA)^k(x^* - x^0).$$

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

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 $\lambda \in (0, \rho(A))$ — собственное значение самосопряженной, положительно определенной матрицы A. Так как $0 < t < 2/\rho(A)$, то получаем, что

$$q = \max_{\lambda \in (0,\rho(A))} |1 - t\,\lambda| < 1.$$

Теорема доказана.

Рассмотрим теперь один из нестационарных методов градиентного спуска, называемый методом наискорейшего спуска, который связан с параметром

$$t_k = \frac{(r^k, r^k)}{(r^k, Ar^k)}, \tag{14}$$

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 5.9 *Пусть A — действительная симметричная положительно определенная матрица. Тогда метод градиентного спуска с выбором (14), 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

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

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

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

отсюда следует, что

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

Применяя доводы, аналогичные тем, которые использовались при доказательстве предыдущей теоремы, 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 пример итерационного уточнения приближенного решения СЛАУ.

Рассмотрим систему $Ax = b$, где $A = (a_{ij})$ — квадратная матрица порядка n. Предполагаем, что $\det A \neq 0$. Теоретически мы можем тогда найти точное решение по формуле $x^* = A^{-1}b$. Часто на практике обратная матрица A^{-1} определяется приближенно. Но тогда $\widetilde{x}^* = \widetilde{A}^{-1}b$ является лишь приближенным решением. Возникает вопрос: как уточнить приближенное решение? Этого можно достичь с помощью итерационного уточнения найденного приближенного решения.

Рассуждаем так. Имеем $Ax = b$, отсюда $0 = \widetilde{A}^{-1}(b - Ax)$, 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 простой итерации $x^k = Bx^{k-1} + c$ сходится при любом выборе нулевого приближения $x^0 \in R^n$ и справедлива оценка

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

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

6 Методы решения нелинейных уравнений

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.

Универсальным является следующий элементарный метод нахождения корня x^* уравнения $f(x) = 0$.

6.1 Метод деления отрезка пополам

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: $x^* = x_1$ — искомый корень.

Случай 2: $f(x_1) \neq 0$ и $f(a)f(x_1) > 0$. Тогда $f(b)f(x_1) < 0$, 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 третьем шаге рассматриваем середину отрезка $[a_2, b_2]$:

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

Снова возможны 3 случая: либо $f(x_3) = 0$ и процесс завершается, либо существует 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 средних точек $x_{k+1} = (a_k + b_k)/2$ и счетная система отрезков $[a_{k+1}, b_{k+1}] \subset [a_k, b_k]$ $(k \in \mathbb{N})$ со 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.

6.2 Итерационные методы

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

Рассмотрим стандартный метод прямой итерации 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 6.1 *Пусть функция* $\varphi : \mathbb{R} \to \mathbb{R}$ *удовлетворяет условию Липшица*

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

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*

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

При удачном выборе x_0 теорему 6.1 можно распространить на функции, заданные на некотором отрезке.

Теорема 6.2 *Пусть точка* $x_0 \in \mathbb{R}$ *такова, что*

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

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

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

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

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

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

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

Замечание 1. В условиях теоремы 6.2

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

Поэтому имеет место следующая оценка скорости сходимости:

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

6.3 Порядок итерационного метода

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

Определение 6.1 *Число $m \in \mathbb{N}$ $(m \geq 2)$ называется порядком итерационного метода, если в точке $x^* = \varphi(x^*)$*

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

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:

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

Предположим, что существует такая постоянная $M_m > 0$, 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

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

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: удачный выбор нулевого приближения.

6.4 Метод Ньютона и его модификации

Метод Ньютона, называемый также методом касательных, является классическим итерационным методом для решения уравнения $f(x) = 0$. 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. Пусть $f'(x) \neq 0$, тогда уравнение $f(x) = 0$ равносильно уравнению

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

Ясно, что метод простой итерации $x_k = \varphi(x_{k-1})$ 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

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

6.5 Проблема собственных значений 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

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

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

Таким образом, мы можем сказать, что спектр матрицы A

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

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 $k = 1, 2, ..., n$.

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 6.3 *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

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

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. Все они основаны на приведении матрицы преобразованиями подобия к матрице 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 $k = 1, 2, ..., n$. Проблема локализации собственных значений заданной матрицы частично решается следующей теоремой.

Теорема 6.4 *(Первая теорема Гершгорина.) Любое собственное значение λ матрицы $A = (a_{kj})$ лежит в одном из кругов $D_k(A)$.*

Доказательство. Пусть $x = (x_1, x_2, ..., x_n) \neq \theta$ — 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 кратко 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 $x_1, x_2, ..., x_n, x_{n+1} \in \mathbb{R}$ и вычислим

$$y_1 = \det(A - x_1 E), y_2 = \det(A - x_2 E), ..., y_{n+1} = \det(A - x_{n+1} E).$$

Обозначим $f(\lambda) = P_n(A; \lambda) \equiv \det(A - \lambda E)$. Зная $y_j = f(x_j)$, мы можем построить интерполяционный полином Лагранжа

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

Поскольку степень полинома f меньше числа узлов, то $L_{n+1}(f; x) \equiv f(x)$. Следовательно, характеристический полином явно определяется 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 собственные значения, решая характеристическое уравнение $P_n(A; \lambda) = 0$ с помощью изученных нами методов решения нелинейных уравнений.

7 Решение систем нелинейных уравнений

7.1 Метод Ньютона для систем уравнений

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 $x^* = (x_1^*, x_2^*, x_3^*, ..., x_n^*) \in \mathbb{R}^n$ 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

$$x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n, \quad F(x) = (f_1(x), f_2(x), ..., f_n(x)) \in \mathbb{R}^n.$$

Очевидно, рассматриваемая нелинейная система уравнений может быть записана как одно уравнение

$$F(x) = \theta,$$

где θ – нулевой вектор.

Рассмотрим матрицу Якоби отображения $F : \mathbb{R}^n \to \mathbb{R}^n$:

$$F'(x) := \begin{pmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 & \dots & \partial f_1/\partial x_n \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 & \dots & \partial f_2/\partial x_n \\ \dots & \dots & \dots & \dots \\ \partial f_n/\partial x_1 & \partial f_n/\partial x_2 & \dots & \partial f_n/\partial x_n \end{pmatrix}.$$

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. Выбираем нулевое приближение $x^0 = (x_1^0, x_2^0, ..., x_n^0)$. Итерации $x^k = (x_1^k, x_2^k, ..., x_n^k)$ 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

$$c = a_1 a_2, \quad 0 < b < \min\{a, 1/c\}, \quad B_b(x^*) = \{x \in \mathbb{R}^n : \|x - x^*\| \leq b\}.$$

Теорема 7.1 *Пусть* $x^* = (x_1^*, x_2^*, ..., x_n^*) \in \mathbb{R}^n$ *— решение уравнения* $F(x) = \theta$. *Предположим, что для любых точек* $x \in B_a(x^*)$ *и* $y \in B_a(x^*)$ *выполнены условия:*

$$\left\| [F'(x)]^{-1} \right\| \leq a_1, \quad \|F(x) - F(y) - F'(y)(x - y)\| \leq a_2 \|x - y\|^2.$$

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 и

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

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

$$\left\|[F'(x^k)]^{-1}F(x^k) + (x^* - x^k)\right\| \le$$

$$\le \|[F'(x^k)]^{-1}\| \, \|F'(x^k)\{[F'(x^k)]^{-1}F(x^k) + (x^* - x^k)\}\| =$$

$$= \|[F'(x^k)]^{-1}\| \, \|F(x^k) + F'(x^k)(x^* - x^k)\|,$$

с учетом неравенства $\left\|[F'(x)]^{-1}\right\| \le a_1$, имеем

$$\left\|[F'(x^k)]^{-1}F(x^k) + (x^* - x^k)\right\| \le a_1 a_2 \|x^* - x^k\|^2 = c\|x^* - x^k\|^2.$$

Поскольку

$$\|x^* - x^{k+1}\| = \|x^* - x^k + \left[F'(x^k)\right]^{-1} F(x^k)\|,$$

то по индукции получаем

$$\|x^* - x^{k+1}\| \le c\|x^* - x^k\|^2 \le \ldots \le c^{2^k-1}\|x^* - x^0\|^{2^k}.$$

Индукцией также легко получаем, что условие $x^0 \in B_b(x^*)$ влечет $x^k \in B_b(x^*)$ для любой итерации, так как

$$\|x^* - x^{k+1}\| \le c\|x^* - x^k\|^2 \le c\,b^2 < b.$$

Далее, имеем: число $q = c\,b < 1$ в силу выбора b. Следовательно, предыдущая оценка запишется в виде неравенства

$$\|x^* - x^{k+1}\| \le \frac{1}{c}\, q^{2^k},$$

что влечет сходимость итераций к точному решению.

Этим и завершается доказательство теоремы.

Нахождение обратной матрицы представляет собой трудоемкую задачу. Поэтому имеет смысл пользоваться упрощенной версией метода Ньютона, задавая итерации формулой

$$x^k = x^{k-1} - \left[F'(x^0)\right]^{-1} F(x^{k-1}), \quad \forall k \in \mathbb{N}.$$

7.2 Другие итерационные методы

Рассмотрим теперь систему нелинейных уравнений следующего вида

$$\begin{cases} x_1 = \varphi_1(x_1, x_2, x_3, \ldots, x_n) \\ x_2 = \varphi_2(x_1, x_2, x_3, \ldots, x_n) \\ \quad \ldots\ldots\ldots\ldots \\ x_n = \varphi_n(x_1, x_2, x_3, \ldots, x_n) \end{cases}.$$

Предположим, что функции $\varphi_j : \mathbb{R}^n \to \mathbb{R}$ являются непрерывными. Обозначим решение этой системы уравнений как $x^* = (x_1^*, x_2^*, x_3^*, ..., x_n^*) \in \mathbb{R}^n$.

Указанную систему формально можно записать в виде одного уравнения для отображения $\Phi : \mathbb{R}^n \to \mathbb{R}^n$, определяя вектор $\Phi(x)$ равенствами

$$x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n, \quad \Phi(x) = (\varphi_1(x), \varphi_2(x), ..., \varphi_n(x)) \in \mathbb{R}^n.$$

Тогда наша нелинейная система уравнений может быть записана как одно уравнение

$$x = \Phi(x).$$

Рассмотрим метод прямой итерации. А именно, задаем нулевое приближение $x^0 = (x_1^0, x_2^0, ..., x_n^0)$. Итерации $x^k = (x_1^k, x_2^k, ..., x_n^k)$ 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 7.2 *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*

$$\|x^* - x^k\| \leq \frac{\alpha^k}{1 - \alpha}\|x^1 - x^0\|.$$

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

$$\begin{cases} x_1^k = \varphi_1(x_1^{k-1}, x_2^{k-1}, x_3^{k-1}, \ldots, x_n^{k-1}) \\ x_2^k = \varphi_2(x_1^k, x_2^{k-1}, x_3^{k-1}, \ldots, x_n^{k-1}) \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^k = \varphi_n(x_1^k, x_2^k, x_3^k, \ldots, x_{n-1}^k, x_n^{k-1}) \end{cases}.$$

8 Задачи и упражнения

Задачи 1-7 взяты из задачника Дробышева, Дымникова и Ривина, в этом задачнике приведены и решения.

1. Пусть A — невырожденная матрица. Докажите, 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. Докажите, что для произвольных матриц A, 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

$$\left\|(A+\alpha E)^{-1}\right\|_2 \ \|A+\alpha E)\|_2$$

является монотонно убывающей функцией от α при $\alpha > 0$.

6. Найдите миллионный член последовательности чисел Фибоначчи

$$0,\ 1,\ 1,\ 2,\ 3,\ 5,\ 8,\ 13,\ 21,\ 34,\ 55, \ldots.$$

7. Докажите вторую теорему Гершгорина: если объединение кругов $D_k(A)$ распадается на несколько связных частей, то каждая такая связная часть содержит столько собственных значений, сколько кругов ее составляют.

8. Найдите характеристический полином следующей матрицы Фробениуса

$$F_n = \begin{pmatrix} p_{n-1} & p_{n-2} & \ldots & p_2 & p_1 & p_0 \\ 1 & 0 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 & 0 \\ 0 & 0 & \ldots & 0 & 1 & 0 \end{pmatrix}.$$

9. Вычислите решение системы $10^{-3}x_1 + x_2 = 5, x_1 - x_2 = 6$ 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 из учебника В.С. Рябенького "Введение в вычислительную математику" Москва, Физматлит, 2008).

10. Пользуясь методом Ньютона, постройте алгоритм для вычисления числа $\sqrt{7}$, рассматриваемая это число как корень уравнения $x^2 = 7$.

11. 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 из учебника В.С. Рябенького “Введение в вычислительную математику”, Москва, Физматлит, 2008).

12. 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.

13. Докажите формулу

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

для определителя Вандермонда.

Указание: легко найти факторизацию полинома $\Delta_n(x)$, получаемого из определителя Вандермонда заменой последней

строки элементами $1, x, x^2, ..., x^{n-1}$, а именно, полинома

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

Далее применяем метод математической индукции.

14. Пусть A — квадратная матрица порядка n, норма $\|A\|$ определена как норма линейного оператора. Мы знаем простую оценку $\rho(A) \leq \|A\|$ для спектрального радиуса. Известен более тонкий и глубокий факт, а именно, известна формула

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

Докажите эту формулу.

9 Приближение функций полиномами

Математические модели многих проблем естествознания используют 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 алгебраическими 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 называется интерполяционным. Как мы убедимся, замена функции ее интерполяционным полиномом позволяет найти легко приближенную формулу при интегрировании. Получаемые формулы будут зависеть лишь от конечного числа значений функции, использованных при построении интерполяционного полинома.

Наиболее известной и употребительной является интерполяционная формула, открытая Лагранжем (1795), 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 (С.Н. Бернштейн, Джексон и ряд других математиков). При этом существенно использовались результаты Вейерштрасса, П.Л. Чебышева и Лебега.

Отметим также, что интерполяция представляет собой лишь

один из разделов обширной теории приближения функций.

9.1 Интерполяционный полином Лагранжа

Пусть на отрезке $[a,b]$ заданы точки $x_1, x_2, \ldots, x_n \in [a,b]$. 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.

Теорема 9.1 *Для любой функции $f \in C[a,b]$ и заданных узлов $x_1, x_2, \ldots, x_n$ интерполяционный полином $L_n(f;x)$ степени не выше $n-1$ существует и определяется единственным образом.*

Доказательство. 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:

$$L_n(f;x_j) = f(x_j)\, j = 1,2,...,n \Leftrightarrow$$

$$\begin{cases} a_1 + a_2 x_1 + \ldots + a_n x_1^{n-1} = f(x_1) \\ a_1 + a_2 x_2 + \ldots + a_n x_2^{n-1} = f(x_2) \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots \\ a_1 + a_2 x_n + \ldots + a_n x_n^{n-1} = f(x_n) \end{cases}.$$

Для определения 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

$$\begin{matrix} f(x_1) \\ f(x_2) \\ \dots \\ f(x_n) \end{matrix}.$$

Поэтому интерполяционный полином Лагранжа запишется в виде:

$$L_n(f;x) = \sum_{k=1}^{n} \frac{\Delta_{n,k}}{\Delta_n} \cdot x^{k-1}.$$

Заметим, что $L_n(f;x)$ — полином степени $\leq n-1$. По построению $f(x) \approx L_n(f;x)$ на $[a,b]$.

Приведем второе доказательство единственности, показывающее, в частности, что $L_n(f;x) = f(x)$ 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 $\leq n-1$:

$$Q(x) = \sum_{k=1}^{n} b_k x^{k-1}, \quad Q(x_j) = f(x_j), \quad j = 1,2,\dots,n.$$

Рассмотрим разность

$$p(x) = L_n(f;x) - Q(x) = \sum_{k=1}^{n} (a_k - b_k) x^{k-1}$$

— полином степени $\leq n-1$. Имеем для любого $j = 1,\dots,n$

$$p(x_j) = L_n(f;x_j) - Q(x_j) = f(x_j) - f(x_j) = 0.$$

Таким образом, получаем, что полином $p(x)$ степени не выше $n-1$ имеет n различных корней $x_1, x_2, \dots, x_n$.

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 9.1.1 *Если $Q(x)$ — алгебраический полином степени $\leq n-1$, то*

$$L_n(Q;x) \equiv Q(x).$$

Представление Лагранжа для интерполяционного полинома

Приведем теперь второе доказательство существования интерполяционного полинома. Одновременно мы дадим основное представление для полинома Лагранжа в виде явной формулы, включающей узлы интерполяции $x_1, x_2, \ldots, x_n$ и значения интерполируемой функции в этих точках.

Нам потребуются следующие полиномы степени $n-1$, которые называются **фундаментальными полиномами Лагранжа**.

$$l_k(x) = \prod_{j=1, j\neq k}^{n} (x - x_j) / \prod_{j=1, j\neq k}^{n} (x_k - x_j) =$$

$$= \frac{(x-x_1)\ldots(x-x_{k-1})(x-x_{k+1})\ldots(x-x_n)}{(x_k-x_1)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_n)}.$$

В узлах интерполяции получаем

$$l_k(x_j) = \delta_{k_j} = \begin{cases} 1, \text{ если } k = j; \\ 0, \text{ если } k \neq j. \end{cases}$$

Рассмотрим полином

$$Q(x) = \sum_{k=1}^{n} f(x_k) l_k(x).$$

Имеем: степень $Q \le n-1$, кроме того,

$$Q(x_j) = \sum_{k=1}^{n} f(x_k) l_k(x_j) = \sum_{k=1}^{n} f(x_k)\delta_{k_j} = f(x_j)$$

для любого $j = 1, \ldots, n$.

В силу единственности интерполяционного полинома получаем

$$Q(x) \equiv L_n(f; x) = \sum_{k=1}^{n} f(x_k) l_k(x).$$

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

$$l_k(x) = \frac{A}{B},$$

где

$$A = \frac{\omega_n(x)}{x - x_k}, \qquad B = \omega_n'(x_k) = \prod_{j=1, j\neq k}^{n} (x_k - x_j),$$

так как

$$\omega_n'(x) = (x - x_2)\ldots(x - x_n) + (x - x_1)(x - x_3)\ldots(x - x_n) +$$

$$\ldots + (x - x_1)(x - x_2)\ldots(x - x_{n-1}).$$

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 9.2 *Для любой функции $f \in C[a,b]$ и заданных узлов $x_1, x_2, \ldots, x_n$ 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, а также 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.

9.2 Оценки погрешности для гладких функций

Будем рассматривать снова n узлов $x_1, x_2, \ldots, x_n \in [a,b]$.

Нас будет интересовать остаточный член интерполяции

$$r_n(x) = f(x) - L_n(f;x),$$

называемый также погрешностью интерполяции.

Теорема 9.3 *Пусть $f \in C^{(n-1)}[a,b]$ и в интервале (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 корнями 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 9.3.1 *Если* $|f^{(n)}(x)| \leq M_n = \text{const}$ *для всех* $x \in [a,b]$, *то*

$$|r_n(x)| = |f(x) - L_n(f;x)| \leq \frac{M_n}{n!}(b-a)^n$$

для любого $x \in [a,b]$.

Доказательство. Действительно, имеем

$$|r_n| \leq \frac{M_n}{n!}\,|\omega_n(x)|,$$

кроме того,

$$|\omega_n(x)| = \left| \prod_{k=1}^{n} (x - x_k) \right| \leq (b-a)^n.$$

Следствие 9.3.2 *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 $n \to \infty$. Из условия

$$\varepsilon_n = \frac{\sqrt[n]{M_n}}{n} \to 0$$

следует $\varepsilon_n\, e\,(b-a) \to 0$, а значит и $\{\varepsilon_n\, e\,(b-a)\}^n \to 0$. Таким образом,

$$\max_{x\in[a,b]} |r_n(x)| := \|r_n\|_{C[a,b]} \to 0$$

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

$$|r_n(x)| \le \frac{M_n}{n!}(1-0)^n = \frac{1}{n!}$$

показывают, что неравенство

$$|r_n(x)| < 0,01$$

будет выполняться наверняка при $n \ge 5$.

9.3 Полиномы Чебышева и оптимальный выбор узлов

Рассмотрим функции, определяемые формулами:

$$T_0(t) = 1,\ T_1(t) = t,\ T_n(t) = \cos(n \arccos t),\ n \geq 2.$$

Как показывают следующие лемма и теорема 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 9.1 *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.

$$T_n(t) = 2^{n-1} \prod_{k=1}^{n} \left(t - \cos\frac{2k-1}{2n}\pi\right).$$

Кроме того, максимум и минимум $T_n(t)$ достигаются в точках $t_k^* = \cos\frac{k\pi}{n}$, причем $T_n(t_k^*) = (-1)^k$.

Доказательство. Рассмотрим сначала случаи $n = 1$ и $n = 2$. 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 $k = 1, \ldots, n$. Максимальное и минимальное значения $T_n(t) = \cos(n \arccos t)$ равен ± 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 $(n+1)$ точке t_k^* $k = 0, \ldots, n$. Лемма доказана полностью.

Теорема 9.4 *(Теорема Чебышева) Для любого $n \in N$ 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

$$q(t) = \frac{T_n(t)}{2^{n-1}} - Q_n(t) = \prod_{k=1}^{n}(t - t_k^0) - \prod_{k=1}^{n}(t - t_k) =$$

$$= t^n + a_{n-1}t^{n-1} + \ldots + a_0 - (t^n + b_{n-1}t^{n-1} + \ldots + b_0).$$

Видно, что $q(t)$ — полином степени $\leq (n-1)$. С другой стороны, в точках экстремума полинома Чебышева получаем

$$q(t_k^*) = \frac{(-1)^k}{2^{n-1}} - Q(t_k^*), \quad |Q_n(t_k^*)| < \frac{1}{2^{n-1}},$$

$$q(t_0^*) = \frac{1}{2^{n-1}} - Q(t_0^*) > 0,$$

$$q(t_1^*) = \frac{-1}{2^{n-1}} - Q(t_1^*) < 0,$$

$$q(t_2^*) > 0 \ \ldots.$$

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.

Общая задача. Дано некоторое семейство $F \subset C[a, b]$. Нужно найти величину

$$V_n(F) = \inf_{x_1, x_2, \ldots, x_n \in [a,b]} \sup_{f \in F} \max_{a \leq x \leq b} |r_n(x)|.$$

Иными словами, необходимо подобрать узлы $x_1, x_2, \ldots, x_n$ на отрезке $[a, b]$ так, чтобы полученная сетка узлов была бы оптимальной для выбранного семейства F.

Рассмотрим эту задачу для следующего семейства функций

$$W^n M = \{f \in C[a,b] \,:\, \exists f^{(m)}(x)\,(x \in [a,b],$$

$$m = 1, ..., n), |f^{(n)}(x)| \le M\},$$

где M — некоторая положительная постоянная.

Оказывается, что можно найти $V_n(W^n M)$ с применением теоремы Чебышева.

Теорема 9.5 *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. Ранее было доказано, что из условия $|f^{(n)}(x)| \le M$ следует оценка

$$|r_n(x)| \le \frac{M}{n!}|\omega_n(x)|$$

для любого $x \in [a,b]$, где $\omega_n(x) = (x-x_1)(x-x_2)\ldots(x-x_n)$.

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

$$\sup_{f\in W^nM}\max_{x\in[a,b]}|r_n(x)|=\frac{M}{n!}\max_{x\in[a,b]}|\omega_n(x)|,$$

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

$$V_n(W^nM)=\frac{M}{n!}\frac{(b-a)^n}{2^n}\inf_{t_1,t_2,\ldots,t_n\in[-1,1]}\prod_{k=1}^{n}|t-t_k|.$$

По теореме П. Л. 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

$$V_n(W_nM)=\frac{M}{n!}\frac{(b-a)^n}{2^{2n-1}}.$$

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.

9.4 Лебеговы оценки погрешности интерполяции

Оценки Лебега для остаточного члена зависят от двух констант: от наименьшего равномерного приближения $E_n f$ и константы Лебега Λ_n.

Величина $E_n f$, называемая наименьшим равномерным приближением $f \in C[a,b]$ алгебраическими полиномами степени $\leq n-1$, определяется следующим образом

$$E_n f = \inf \left\{ \left\| f(x) - \sum_{k=1}^{n} a_k x^{k-1} \right\|_{C[a,b]} : \; a_1, a_2, \ldots, a_n \in \mathbb{R} \right\}.$$

Известно, что для любой функции $f \in C[a,b]$ величина $E_n f \to 0$ при $n \to \infty$. Этот факт является простым следствием следующей теоремы Вейерштрасса.

Теорема 9.6 *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 самого Вейерштрасса (1885), известны и другие доказательства этой фундаментальной теоремы, данные А. Лебегом (1898), С.Н. Бернштейном (1912) и другими математиками. Мы приведем доказательство Лебега, рассуждения которого оказались полезными при построении обобщений, а именно, при доказательстве теоремы Стоуна-Вейерштрасса.

Лебег выводит утверждение теоремы Вейерштрасса из трех простых фактов.

Шаг 1. Согласно теореме Кантора, непрерывная на отрезке функция является равномерно непрерывной, поэтому она равномерно приближаема 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 $x_0 = a < x_1 < ... < x_{m-1} < x_m = b$ — абсциссы вершин ломаной. Это утверждение устанавливается элементарными рассуждениями, так как указанное представление задает непрерывную кусочно-линейную функцию при любом выборе $a_0, a_1, ..., a_m$, 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 $a_1, a_2, ..., a_m$ явно определяются из системы линейных уравнений

$$a_1 - \sum_{j=2}^{m} a_j = k_1,$$

$$\sum_{j=1}^{s} a_j - \sum_{j=s+1}^{m} a_j = k_s, \quad s = 2, ..., m-1,$$

$$\sum_{j=1}^{m} a_j = k_m.$$

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.

Вернемся теперь к интерполяционным полиномам. Получим сначала формулу для погрешности r_n без предположения дифференцируемости интерполируемой функции.

Теорема 9.7 *Для любого $f \in C[a,b]$ с n узлами интерполяции $x_1, \ldots, x_n$ ($n \ge 2$)*

$$r_n(x) = f(x) - L_n(f;x) = \sum_{k=1}^{n} [f(x) - f(x_k)]\, l_k(x),$$

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 интерполированием 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 9.1 *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 9.8 *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

$$r_n(\overline{x}) = f(\overline{x}) - L_n(f;\overline{x}) \to 0 \quad при \quad n \to \infty.$$

б) если $\Lambda_n E_n f \to 0$ при $n \to \infty$, то равномерно на отрезке $[a,b]$

$$r_n(x) \to 0 \quad при \quad n \to \infty.$$

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 $\le n-1$. Следовательно,

$$L_n(Q;x) \equiv Q(x).$$

Поэтому

$$|r_n(x)| \le |f(x) - Q(x)| + \left| \sum_{k=1}^{n} [f(x_k) - Q(x_k)] l_k(x) \right| \le$$

$$\le \|f(x) - Q(x)\|_{C[a,b]} + \|f(x) - Q(x)\|_{C[a,b]} \sum_{k=1}^{n} |l_k(x)| =$$

$$= \|f(x) - Q(x)\|_{C[a,b]} \, (1 + \Lambda_n(x)).$$

В силу произвольности $Q(x)$ отсюда следует

а)

$$|r_n(\overline{x})| \le (E_n f)[1 + \Lambda_n(\overline{x})] \le$$

$$\le 2\Lambda_n(\overline{x}) E_n f \to 0 \qquad \text{при} \qquad n \to \infty,$$

и, аналогично,

б)

$$\|r_n(x)\|_{C[a,b]} \le 2\Lambda_n \cdot E_n f \to 0 \qquad \text{при} \qquad n \to \infty.$$

Таким образом, теорема Лебега доказана.

Замечания. Понятно, что сходимость или расходимость интерполяционного процесса зависит как от выбора последовательности точек интерполяции, т. е. последовательности сеток

$$\Omega_n = \{x_{n1}, x_{n2}, ..., x_{nn}\},$$

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, построенных для функции $f(x) = |x|$ по равноотстоящим узлам на отрезке $[-1, 1]$, не сходится к функции $|x|$ ни в одной точке этого отрезка, кроме трех точек -1, 0, 1.

Пример 2 (пример Рунге (Runge C.)): последовательность интерполяционных полиномов, построенных для гладкой функции

$$f(x) = \frac{1}{25x^2 + 1}$$

по равноотстоящим узлам на отрезке $[-1, 1]$, не сходится равномерно к этой функции 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*

$$\Omega_n = \{x_{n1}, x_{n2}, ..., x_{nn}\} \subset [a, b]$$

существует непрерывная на этом отрезке функция $f(x)$ *такая, что последовательность интерполяционных полиномов Лагранжа* $L_n(f; x)$ *не сходится равномерно к этой функции на отрезке* $[a, b]$.

Известно также, что для каждой непрерывной функции существует своя оптимальная последовательность сеток. А именно, имеет место следующий положительный результат.

Теорема Марцинкевича: для любой функции $f(x)$, 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 $[a, b]$.

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. Так, например, в теории приближений доказан следующий факт:

для любой функции, абсолютно непрерывной на отрезке $[a, b]$ *(следовательно, для любой гладкой функции), интерполяционные полиномы равномерно сходятся к ней на* $[a, b]$ *для некоторых специальных последовательностей сеток, например, для последовательности сеток с узлами Чебышева.*

9.5 Свойства оператора интерполирования

Для фиксированной сетки

$$\Omega_n = \{x_{n1}, x_{n2}, ..., x_{nn}\} \subset [a, b]$$

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 9.9 *Норма линейного оператора $P_n : C[a,b] \to C[a,b]$ 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 непрерывной функции $\Lambda_n(x)$. Очевидно, существует непрерывная на отрезке $[a,b]$ функция f_0 такая, что $f_0(x_k) = sign\, l_k(x_0)$ для всех $k = 1, 2, ..., n$ 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 $x = a + (b - a)(t + 1)/2$ легко убедиться в том, что константа Лебега (норма линейного оператора $P_n : C[a, b] \to C[a, b]$) не зависит от длины отрезка интерполирования, а зависит только от относительного расположения узлов. Понятно, что зависимость константы Лебега от числа узлов имеет большое значение, так как через эту константу оценивается погрешность интерполяции.

Теорема 9.10 *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 $x_1 = 1, x_2 = 2, ..., x_n = n$. Тогда

$$\Lambda_n = \max_{1 \leq x \leq n} \sum_{k=1}^{n} \prod_{j \neq k} \left| \frac{x - j}{k - j} \right| =$$

$$= \max_{1 \leq x \leq n} \sum_{k=1}^{n} \frac{1}{(n - k)!(k - 1)!} \prod_{j \neq k} |x - j|.$$

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

$$x_k = \cos \frac{\pi(2k+1)}{2n}, \quad k = 0, 1, ..., n-1,$$

можем записать

$$\Lambda_n = \max_{-1 \leq x \leq 1} \sum_{k=0}^{n-1} \frac{|T_n(x)|}{|x - x_k||T_n'(x_k)|} =$$

$$= \max_{-1 \leq x \leq 1} \sum_{k=0}^{n-1} \frac{|\cos(n \arccos x)|\sqrt{1 - x_k}}{n|x - x_k|}.$$

Имеет место следующая теорема С.Н. Бернштейна.

Теорема 9.11 *Для узлов Чебышева при интерполяции алгебраическими полиномами 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. А именно, существует постоянная $c > 0$ такая, что

$$\Lambda_n \geq c \ln n$$

для любой сетки из n узлов.

10 Интерполяционный полином Ньютона

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

$$\omega_{n+1}(x) = (x-x_1)(x-x_2)\ldots(x-x_{n+1}).$$

Понятно, что при добавлении нового узла приходится пересчитывать все слагаемые в представлении Лагранжа.

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 Лагранжа 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 Лагранжа запишем в виде

$$L_n(f;x) = A_0 + A_1(x-x_1) + A_2(x-x_1)(x-x_2) +$$

$$+ \ldots + A_{n-1}(x-x_1)\ldots(x-x_{n-1}),$$

т. е. в виде

$$L_n(f;x) = \sum_{j=1}^{n} A_{j-1}\omega_{j-1}(x),$$

где $\omega_0(x) = 1$, $\omega_1(x) = (x - x_1)$, $\omega_k(x) = (x - x_1)\dots(x - x_k)$.

Неизвестные коэффициенты $A_0, A_1, A_2, \dots, A_{n-1}$ можно попытаться определить из n уравнений

$$L_n(f; x_1) = f(x_1), \ \dots, \ L_n(f; x_n) = f(x_n).$$

Легко показать, что коэффициенты A_k однозначно определяются этими уравнениями, зависят лишь от значений функции в точках $x_1, x_2, \dots, x_k$, следовательно, не меняются при добавлении нового узла x_{n+1}.

Для первых двух коэффициентов вычисления весьма просты: из первых двух уравнений имеем

$$f(x_1) = A_0,$$

$$f(x_2) = A_0 + A_1(x_2 - x_1),$$

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 $x_1, x_2, \dots, x_k$.

10.1 Разделенные разности

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 $f(x_1;x_2;...;x_k)$ порядка $k-1$ определяется так:

$$f(x_1;x_2;...;x_k)=\frac{f(x_2;x_3;...;x_k)-f(x_1;x_2;...;x_{k-1})}{x_k-x_1}.$$

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 10.1 *Справедлива следующая формула*

$$f(x_1;x_2;\ldots;x_k)=\sum_{j=1}^{k}\frac{f(x_j)}{\omega_k'(x_j)}= \tag{15}$$

$$=\sum_{j=1}^{k}f(x_j)\prod_{i=1,i\neq j}^{k}\frac{1}{x_j-x_i},$$

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

$$f(x_1; x_2; ...; x_{k+1}) = \frac{f(x_2; x_3; ...; x_{k+1}) - f(x_1; x_2; ...; x_k)}{x_{k+1} - x_1},$$

тогда по предположению индукции $f(x_1; x_2; ...; x_{k+1}) =$

$$= \frac{1}{x_{k+1} - x_1} \sum_{j=2}^{k+1} f(x_j) \prod_{i=2, i\neq j}^{k+1} \frac{1}{x_j - x_i} -$$

$$- \frac{1}{x_{k+1} - x_1} \sum_{j=1}^{k} f(x_j) \prod_{i=1, i\neq j}^{k} \frac{1}{x_j - x_i}.$$

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 $f(x_1, \ldots, x_n)$ является симметричной функцией своих аргументов т. е. инвариантна относительно перестановки аргументов (например, $f(x_1; x_2) = f(x_2; x_1)$).

10.2 Представление Ньютона

Как уже отмечалось выше, в формуле

$$L_n(f; x) = A_0 + A_1(x - x_1) + A_2(x - x_1)(x - x_2) + \ldots$$

$$\ldots + A_{n-1}(x - x_1)(x - x_2)\ldots(x - x_{n-1}) = \sum_{k=1}^{n} A_{k-1}\, \omega_{k-1}(x),$$

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, $1 \leq k \leq n$,

$$A_{k-1} = f(x_1; x_2; \ldots; x_k).$$

Теорема 10.2 *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 $x_1, x_2, ..., x_m$, $1 \le m \le n$. 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. А именно, 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 Q порядка $(n+1)$ и выше равны 0.

Действительно, пусть $m \geq n+1$, тогда $Q(x) \equiv L_m(Q; x)$ 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 $m \geq n+2$.

10.3 Переход от разделенных к конечным разностям

В этом пункте мы запишем формулу Ньютона для интерполяционного полинома с заменой разделенных разностей на конечные разности.

Рассмотрим узлы $x_1, \ldots, x_n \in [a, b]$ и функцию $f \in C[a, b]$, 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

$$a \le x_1, x_2 = x_1 + h, x_3 = x_1 + 2h,\ \ldots,\ x_n = x_1 + (n-1)h \le b,$$

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

$$L_n(f; x) = \sum_{k=1}^{n-1} \frac{\Delta^k y_1}{k!} t(t-1) \ldots (t-k+1)$$

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

$$x_k = x_1 + kh, \ (h > 0, \ \ k = 0, 1, ..., n-1)$$

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. Заинтересованный читатель найдет замечательные формулы Гаусса, Бесселя, Стирлинга и других классиков для интерполяционного полинома в ряде книг, например, в учебнике Березина и Жидкова (см. [4], том 1, стр.125-142).

11 Кратное интерполирование

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 интерполирования $x_1, x_2, \ldots, x_n \in [a, b]$ и непрерывно дифференцируемую функцию f на этом отрезке.

Интерполяционный полином $H_n(f; x)$ ищется как полином наименьшей степени, удовлетворяющий следующим условиям

$$f(x_1) = H_n(f; x_1),\ f(x_2) = H_n(f; x_2), \ldots, f(x_n) = H_n(f; x_n);$$

$$f'(x_1) = H'_n(f; x_1), f'(x_2) = H'_n(f; x_2), \ldots, f'(x_n) = H'_n(f; x_n).$$

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.

11.1 Интерполяционный полином 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

$$H^{(k)}(x_j) = f^{(k)}(x_j) \tag{16}$$

для всех

$$k = 0, 1, \ldots, a_j - 1.$$

Очевидно, для записи системы уравнений (16) достаточно, чтобы функция f была бы непрерывно дифференцируемой $(a_j - 1)$-раз в некоторой окрестности точки x_j, где $j = 1, 2, ..., n$.

Число уравнений для определения $H(x)$ равно

$$m = a_1 + a_2 + \ldots + a_n,$$

поэтому естественно искать полином $H(x)$ степени $\leq m - 1$.

Теорема 11.1 *Интерполяционный полином Эрмита степени* $\leq m - 1$ *существует и определяется единственным образом, причем его можно представить в следующей форме*

$$H(x) = P_1(x) + (x - x_1)^{a_1} P_2(x) + (x - x_1)^{a_1}(x - x_2)^{a_2} P_3(x) + \ldots \tag{17}$$

$$+ (x - x_1)^{a_1}(x - x_2)^{a_2} \ldots (x - x_{n-1})^{a_{n-1}} P_n(x),$$

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 (17) 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 (17).

Поэтому интерполяционный полином Эрмита можно искать в виде (17), при этом коэффициенты P_1, P_2, ..., P_n 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

$$P_1(x) = b_0 + \frac{b_1}{1!}(x - x_1) + \frac{b_2}{2!}(x - x_1)^2 + \ldots + \frac{b_{a_1-1}}{(a_1 - 1)!}(x - x_1)^{a_1-1}.$$

Зная P_1 и условия интерполяции в точке x_2, определяем $P_2(x)$. Из (17) 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

$$\left[\frac{H(x) - P_1(x)}{(x - x_1)^{a_1}} - P_2(x)\right]^{(k)} \Bigg|_{x=x_2} = 0$$

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 функций P_1 и $1/(x - x_1)$ и их производных в точке x_2 известны, а числа

$$H(x_2), H'(x_2), \ldots, H^{(a_2-1)}(x_2)$$

заданы условиями интерполяции

$$H(x_2) = f(x_2), \;\; H'(x_2) = f'(x_2), \;\; ..., H^{(a_2-1)}(x_2) = f^{(a_2-1)}(x_2).$$

Кроме того, $P_2(x)$ — полином степени $\leq a_2 - 1$, поэтому $P_2^{(k)}(x) \equiv 0$ для $k \geq a_2$. По формуле Тейлора можем найти P_2.

Продолжая процесс, по индукции находим все P_k ($k = 1, 2, ..., n$) по той же схеме, причем P_k определяется единственным образом условиями интерполяции в точке x_k.

Приведем прямое доказательство единственности $H(x)$ в форме (17). 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 $q(x)$ 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 11.2 *Если функция $f(x)$ является $m = a_1 + a_2 + \ldots + a_n$ раз непрерывно дифференцируемой, то существует точка ξ 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, ..., a_n, соответственно. По теореме Ролля $\varphi'(t) = 0$ в некоторых промежуточных точках $\xi_1, \xi_2, \ldots, \xi_n \in (a, b)$. 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.

11.2 Полином Эрмита-Фейера. Другие частные случаи

Рассмотрим 3 частных случая.

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 (17) имеют нулевую степень, т. е. являются константами. Поэтому формула (17) при $a_1 = a_2 = \ldots = a_n = 1$ сводится к формуле Ньютона для интерполяционного полинома Лагранжа с коэффициентами $P_k = A_{k-1} = f(x_1; x_2; ...; x_k)$.

2) Пусть $n = 1$, $m = a_1 \geq 2$. Тогда в представлении (17) 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) Пусть $n \geq 2$ и все узлы имеют одинаковую кратность 2, 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 11.3 *Справедлива следующая формула для полинома Эрмита-Фейера:*

$$H_n(f;x) = \sum_{k=1}^{n} y_k\, l_k^2(x)\,[1 - c_k(x - x_k)] + \sum_{k=1}^{n} y'_k\, l_k^2(x)\,(x - x_k), \quad (18)$$

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 (18), 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

$$f(x_1) = H_n(f;x_1),\ \ f(x_2) = H_n(f;x_2), \ldots, f(x_n) = H_n(f;x_n);$$

$$f'(x_1) = H'_n(f;x_1), f'(x_2) = H'_n(f;x_2), \ldots, f'(x_n) = H'_n(f;x_n).$$

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: $H'_n(f;x) =$

$$= \sum_{k=1}^{n} l_k^2(x)(y'_k - y_k c_k) + \sum_{k=1}^{n} 2l_k(x)l'_k(x)\{y_k + (y'_k - y_k c_k)\,(x - x_k)\},$$

отсюда следует

$$H'_n(f;x_j) = y'_j - y_j\, c_j + 2l'_j(x_j)\, y_j.$$

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.

12 Приближение периодических функций

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

$$T_n(f; x_0) = f(x_0),\ T_n(f; x_1) = f(x_1),\ \ldots,\ T_n(f; x_{2n}) = f(x_{2n}).$$

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

$$\begin{pmatrix} 1/2 & \cos x_0 & \sin x_0 & \ldots & \cos nx_0 & \sin nx_0 \\ 1/2 & \cos x_1 & \sin x_1 & \ldots & \cos nx_1 & \sin nx_1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1/2 & \cos x_{2n} & \sin x_{2n} & \ldots & \cos nx_{2n} & \sin nx_{2n} \end{pmatrix} \tag{19}$$

отличен от нуля, поэтому интерполяционный тригонометрический полином определится единственным образом.

Но мы избираем другой, более плодотворный путь доказательства. А именно, предъявим аналог формулы Лагранжа для $T_n(f; x)$ и проверим лишь выполнение условий интерполирования и единственности.

12.1 Тригонометрический интерполяционный полином

Нам потребуются фундаментальные тригонометрические полиномы Лагранжа $t_k(x)$, заданные формулами

$$t_k(x) = \frac{\prod_{j=0,j\neq k}^{2n} \sin\frac{x-x_j}{2}}{\prod_{j=0,j\neq k}^{2n} \sin\frac{x_k-x_j}{2}},\ k = 0, 1, ..., 2n.$$

Теорема 12.1 *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:

$$T_n(f; x_j) = \sum_{k=0}^{2n} f(x_k)\delta_{kj} = f(x_j)\delta_{jj} = f(x_j), \quad j = 0, 1, ..., 2n.$$

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

$$t_k(x) = C \prod_{j=0,j\neq k}^{2n} \sin\frac{x - x_j}{2}\ \ (k = 0, 1, ..., 2n),$$

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 $\le n$ и обращается в нуль в узлах сетки:

$$q(x_j) = T_n(f;x_j) - T_n(x_j) = f(x_j) - f(x_j) = 0, \quad j = 0,\ldots,2n.$$

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, $q(x) \equiv 0$.

12.2 Случай равноотстоящих узлов

Рассмотрим равноотстоящие узлы

$$x_0 = 0, \quad x_1 = h, \quad x_2 = 2h, \quad \ldots, \quad x_{2n} = 2nh = \frac{4n\pi}{2n+1}$$

с шагом $h = \frac{2\pi}{2n+1}$. В этом случае формулы для фундаментальных тригонометрических полиномов Лагранжа 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 12.2 *Для каждой непрерывной 2π-периодической функции f и равноотстоящих узлов $x_k = \frac{2\pi}{2n+1}k,\ k = 0, 1, \ldots, 2n,$*

$$T_n(f;x) = \frac{2}{2n+1}\sum_{k=0}^{2n} f(x_k)D_n(x - x_k),$$

$$m.\ e.\quad t_k(x) = \frac{2}{2n+1}D_n(x - x_k),$$

а коэффициенты Фурье $T_n(f;x)$ 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 интерполяционного полинома равенство $\tau_k(x) = t_k(x)$ будет верно для любого $k = 0, 1, 2, ..., 2n$, 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 $\frac{\sin\alpha}{\alpha} \to 1$ при $\alpha \to 0$, 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

$$T_n(f;x) = \frac{2}{2n+1}\sum_{k=0}^{2n} f(x_k)\left(1/2 + \sum_{m=1}^{n}\cos m(x-x_k)\right) =$$

$$= \frac{1}{2n+1}\sum_{k=0}^{2n} f(x_k) +$$

$$+\frac{2}{2n+1}\sum_{m=1}^{n}\sum_{k=0}^{2n} f(x_k)[\cos mx_k \cos mx + \sin mx_k \sin mx].$$

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, соответственно, 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 $n \to \infty$

$$a_m \to \frac{1}{\pi}\int_0^{2\pi} f(x)\cos mx\,dx, \quad b_m \to \frac{1}{\pi}\int_0^{2\pi} f(x)\sin mx\,dx.$$

13 Сплайн-интерполяция

Как мы видели выше, вопрос о равномерной сходимости интерполяционных полиномов к интерполируемой функции при неограниченном росте числа точек интерполяции является сложным. В общем случае успеха можно добиться лишь специальным подбором узлов. Вопросы сходимости сильно упрощаются, если в качестве приближающих функций используются кусочно-полиномиальные функции. Такие функции называются сплайнами.

Теория сплайн-интерполяции бурно развивается с сороковых годов 20-го столетия.

Можно отметить, что кусочно-полиномиальные функции (сплайны) возникли уже на заре математического анализа в работах Лейбница и особенно в трудах Эйлера при разработке прямых методов вариационного исчисления. Английское слово "сплайн" означает балка, рейка. Оно стало математическим термином по праву: американские инженеры и чертежники издавна использовали гибкие рейки для ручной интерполяции функций, заданных значениями на конечном числе точек.

Перейдем к точным определениям. Непрерывная функция

$$g : [a, b] \to \mathbb{R}$$

называется сплайном, если существует разбиение

$$a = x_0 < x_1 < x_2 < ... < x_n = b$$

такое, что на каждом частичном отрезке $[x_{k-1}, x_k]$ функция g является некоторым полиномом. Таким образом, ограничение $g\,|_{[x_{k-1},x_k]}$ является полиномом, для простоты мы обозначим его как

$$g_k : [x_{k-1}, x_k] \to \mathbb{R}.$$

Определение 13.1 *Пусть $f \in C[a, b]$, и пусть заданы узлы*

$$a = x_0 < x_1 < x_2 < ... < x_n = b, \qquad n \in \mathbb{N}.$$

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

$$|f(x'') - f(x')| \le M|x'' - x'|$$

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

$$|f(x'') - f(x')| = |(|x''| - |x'|)| \le |x'' - x'|.$$

13.1 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 $g \in C[a,b]$, $g(x_k) = f(x_k)$, $k = 0,1,...,n$, 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 $f \in C[a,b]$ ломаными, т. е. 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 степени ≤ 1, 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 *представление 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

$$s_0(x) = \begin{cases} \frac{x_1 - x}{x_1 - a} & \text{при } \ a \leq x \leq x_1, \\ 0 & \text{при } \ x_1 \leq x \leq b; \end{cases}$$

$$s_n(x) = \begin{cases} 0 & \text{при } \ a \le x \le x_{n-1}, \\ \frac{x-x_{n-1}}{b-x_{n-1}} & \text{при } \ x_{n-1} \le x \le b; \end{cases}$$

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 13.1 *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, имеет следующие свойства:*

$$1) \ \left\| f(x) - S_n^1(f;x) \right\|_{C[a,b]} \le \omega(f, \delta_n);$$

$$2) \ S_n^1(f;x) \rightrightarrows f(x) \quad \textit{при} \quad \delta_n \to 0.$$

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

$$|f(x) - f(x_{k-1})| \le \omega(f, \delta_n), \quad |f(x) - f(x_k)| \le \omega(f, \delta_n).$$

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 $f \in Lip\,\alpha$ $(0 < \alpha \le 1)$, 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 13.2 *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 13.1 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 13.2.1 *Если* $f' \in Lip\,\alpha \quad (0 < \alpha \le 1)$, *то*

$$\|f(x) - S_n^1(f; x)\|_{C[a,b]} = O(\delta_n^{1+\alpha}).$$

Следствие 13.2.2 *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 следующем: дальнейшее увеличение порядка гладкости интерполируемой функции, например, требование $f \in C^r[a,b]$, $r \geq 3$, 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

$$x_k = -1 + kh, h = 2/n, \;\; k = 0,1,...,n.$$

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 (экстремальное) свойство сплайнов первой степени. Нам потребуется пространство Соболева $W_2^1[a,b]$ абсолютно непрерывных функций $F : [a,b] \to \mathbb{R}$ с нормой

$$\|F\|_{W_2^1} = \|F\|_{C[a,b]} + \|F'\|_{L_2[a,b]}.$$

Напомню, 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 кусочно-гладкие функции, в частности, сплайны, определенные на отрезке $[a,b]$.

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 $F(x_k) = f(x_k)$, где $k = 0, \ldots, n$.

Очевидно, сплайн $g(x) = S_n^1(f;x)$ является одной из функций, удовлетворяющей обоим условиям.

Теорема 13.3 *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.

13.2 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 частичном 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 третьей степени;

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. Дополнительные 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 13.4 *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):

$$y_1 = g''(x_1),\ y_2 = g''(x_2), ..., y_{n-1} = g''(x_{n-1}).$$

Покажем, что по этим параметрам однозначно определяются $g_k(x)$, а сами числа y_k $(k = 1, 2, ..., n-1)$ 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 $g'_k(x_k) = g'_{k+1}(x_k)$ $(k = 1, 2, ..., n-1)$, т. е. условия непрерывной склейки первых производных, приводят к линейной системе для моментов y_k $(k = 1, 2, ..., n-1)$:

$$\frac{f(x_k) - f(x_{k-1})}{\Delta x_k} + \frac{\Delta x_k}{6} y_{k-1} + \frac{\Delta x_k}{3} y_k =$$

$$= \frac{f(x_{k+1}) - f(x_k)}{\Delta x_{k+1}} - \frac{\Delta x_{k+1}}{3} y_k - \frac{\Delta x_{k+1}}{6} y_{k+1}$$

или, что то же самое, к системе

$$\Delta x_k \; y_{k-1} + 2\left(\Delta x_k + \Delta x_{k+1}\right) \; y_k + \Delta x_{k+1} \; y_{k+1} = b_k,$$

где $k = 1, 2, ..., n-1$, $y_0 = y_n = 0$, 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 $z_k = g'(x_k)$ $(k = 0, 1, ..., n)$. При таком подходе получается формула (докажите!)

$$g_k(x) = (1-t)^2(1+2t)f(x_{k-1}) + t^2(3-2t)f(x_k) +$$

$$+ t(1-t)\Delta x_k\left[(1-t)\,z_{k-1} - t\,z_k\right],$$

и система для определения параметров z_k $(k = 0, 1, ..., n)$ 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) имеют место равенства $F(x_k) = f(x_k)$, где $k = 0, \ldots, n$.

Очевидно, кубический сплайн $g(x) = S_n^3(f;x)$ является одной из функций, удовлетворяющей обоим условиям.

Теорема 13.5 *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 13.3. 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.

14 Наилучшие приближения функций

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 f_n^0 необходимы дополнительные предположения о структуре пространства F. Вопрос 3) мы рассмотрим в двух случаях, когда пространство F является гильбертовым или F — банахово пространство $C[a,b]$.

14.1 Теоремы существования и единственности

Докажем сначала теорему существования.

Теорема 14.1 *Пусть F — линейное нормированное пространство над полем вещественных чисел. Тогда для любого $f \in F$ существует элемент наилучшего приближения f_n^0.*

Доказательство. Если $f \in F_n$, то, очевидно, $f_n^0 = f$ 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 подпоследовательность 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 14.1 *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 $\lambda \geq 0$ 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 14.2 *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

$$E_n f \le \|f - g\| = \left\| f - \frac{f_n^0 + f_n^1}{2} \right\| = \frac{\|f - f_n^0 + f - f_n^1\|}{2} \le$$

$$\le \frac{\|f - f_n^0\| + \|f - f_n^1\|}{2} = E_n f.$$

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 $\lambda \ne 1$. Если $\lambda = 1$, то $f - f_n^0 = f - f_n^1$, т.е. $f_n^0 = f_n^1$, Получили противоречие.

Пусть теперь $\lambda \ne 1$. 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 $L_\rho^p(a, b)$ ($\rho(x) > 0$ 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 линейно-независимые элементы этого пространства $f(x) = 1$ и $g(x) = x$. Имеем

$$\|f + g\| = \max_{x\in[0,1]}(1 + x) = 2,$$

$$\|f\| = 1, \quad \|g\| = \max_{x\in[0,1]} x = 1,$$

следовательно,

$$\|f + g\| = \|f\| + \|g\|.$$

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 $[a, b]$.

14.2 Приближения в гильбертовом пространстве

Пусть F — гильбертово пространство, $l_1, l_2, \ldots, l_n$ — система линейно-независимых элементов из F. 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 само 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 14.3 *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 $k=1,2,...,n$. В силу произвольности $f_n\in F_n$ немедленно получаем

$$(E_nf)^2=\|f\|^2-\sum_{k=1}^{n}|c_k^0|^2=\|f-\sum_{k=1}^{n}c_k^0l_k\|^2.$$

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 14.4 *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 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 $k = 1, 2, ..., n$. 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:

$$\Delta_n = \det((l_k, l_j)) \neq 0.$$

И решение системы имеет вид

$$\alpha_k^0 = \frac{\Delta_n^{(k)}}{\Delta_n},$$

следовательно,

$$f_n^0 = \sum_{k=1}^{n} \frac{\Delta_n^{(k)}}{\Delta_n} l_k.$$

Этим и завершается доказательство теоремы.

14.3 Примеры применения общих теорем

Приведем несколько примеров применения доказанных теорем.

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 $P_k(x)$ неизвестна, 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 отрезке $[a, b]$ 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. 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

$$n(l_k, l_j) = \sum_{l=0}^{n-1} u^l = \frac{u^n - 1}{u - 1} = \frac{e^{2\pi i(k-j)} - 1}{u - 1} = 0.$$

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

$$\alpha_k^0 = (f, e^{ikx}) = \frac{1}{n} \sum_{l=0}^{n-1} f(x_l) \cdot e^{-ikx_l}.$$

14.4 Наилучшие равномерные приближения полиномами

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 14.5 *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 непосредственно 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 14.6 *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 следующий 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

$$M = \max_{a\leq x\leq b} f(x) = f(x_1), \quad m = \min_{a\leq x\leq b} f(x) = f(x_2).$$

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 14.7 *(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 на менее $n + 2$ точек*

$$x_1 < x_2 < x_3 < \ldots < x_{n+2}$$

таких, что

$$P_n(x_j) - f(x_j) = \alpha(-1)^j \|P_n - f\|_{C[a,b]},\ j = 1, 2, \ldots, n + 2, \quad (20)$$

где $\alpha = const$, причем либо $\alpha = 1$, либо $\alpha = -1$.

Доказательство. Необходимость. Пусть $P_n = P_n^0$ — 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

$$r_n(x_j) = \alpha(-1)^j E_n f, \quad j = 1, 2, \ldots, m \ (\alpha = \text{const}, \ |\alpha| = 1).$$

Подчеркнем, что число m выбрано максимальным из всех возможных.

Замкнутое множество $E = \{x \in [a,b] : |r_n(x)| = E_n f\}$ представим в виде

$$E = \bigcup_{j=1}^{m} E_j,$$

где замкнутые множества E_j определены следующим образом:

$$E_1 = \{x \in [a, x_2) : \ r_n(x) = r_n(x_1)\},$$

$$E_j = \{x \in (x_{j-1}, x_{j+1}) : \ r_n(x) = r_n(x_j)\}, \qquad 2 \leq j \leq m-1,$$

$$E_m = \{x \in (x_{m-1}, b] : \ r_n(x) = r_n(x_m)\}.$$

Легко проверить (с учетом максимальности m), что определения множеств E_j корректны и эти множества не пусты, так как $x_j \in E_j$ и, кроме того,

$$a_{k+1} := \min\{x : x \in E_{k+1}\} > \max\{x : x \in E_k\} =: b_k$$

для всех $k = 1, 2, ..., m-1$. Следовательно, существуют точки

$$\xi_1 < \xi_2 < ... < \xi_{m-1},$$

удовлетворяющие условиям

$$b_k < \xi_k < a_{k+1} \ (k = 1, 2, ..., m-1).$$

Рассмотрим полином $s(x) = \lambda(x-\xi_1)(x-\xi_2)...(x-\xi_{m-1})$, выбрав знак постоянной λ из условия совпадения знаков $r_n(x_1)$ и $s(x_1)$.

Тогда $r_n(x)s(x) > 0$ для любого $x \in E$, и для достаточно малого $|\lambda| > 0$

$$\|r_n - s\|_{C[a,b]} = \|P_n^0 - s - f\|_{C[a,b]} < E_n f,$$

а это противоречит тому, что P_n^0 — полином наилучшего равномерного приближения.

Докажем теперь от противного достаточность условия (20). Предположим, что P_n удовлетворяет (20), 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 14.8 *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: $-E_n f \le f(x) - P_n^0(x) \le E_n f$ и $-E_n f \le f(x) - P_n^1(x) \le E_n f$.

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 14.7 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 $j = 1, 2, ..., n+2$, ($\alpha = 1$, 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

$$P_n^0(x_j) = P_n^1(x_j) \quad \text{для} \quad j = 1, 2, ..., n+2.$$

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 14.8.1 *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 $t = -x \in [-a,a]$

$$|P_n^0(-x) - f(x)| = |P_n^0(-x) - f(-x)| = |P_n^0(t) - f(t)| \le E_n f.$$

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 $\leq n-1$, 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 14.8.2 *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 $L^2[0;\pi]$.

15 Квадратурные формулы

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 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 точки $x_1, \ldots, x_n \in < a, b >$. 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

$$\int_a^b \rho(x) f(x)\,dx \approx \sum_{k=1}^{n} A_k f(x_k), \tag{21}$$

где A_k — некоторые вещественные числа. Предполагается, что коэффициенты A_k не зависят от f. Точки x_k в формуле (21) принято называть узлами.

Определение 15.1 *Пусть M — некоторое семейство функций, непрерывных на промежутке $< a, b >$. Говорят, что*

квадратурная формула (21) 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 (21) 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 $j = 0, 1, ..., m$.

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 (Кеплер и Торичелли (1664), формула Ньютона, изложенная в его письме Лейбницу (1676) и опубликованная Котесом (1722), Симпсон (1743)).

Прием, лежащий в основе всех классических квадратурных формул, состоит в замене подинтегральной функции некоторым ее приближением (например, интерполяционным полиномом или сплайном).

15.1 Интерполяционные квадратурные формулы

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 подинтегральную 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 15.1 *Квадратурная формула (21) 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 $A_k = p_k$ для всех $k = 1, 2, ..., n$.*

Доказательство. Предположим, что (21) 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 $\le n-1$, является тривиальным. Действительно, если F — полином степени $\le n-1$, 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 15.2 *Пусть* $\omega_n(x) = (x - x_1) \ldots (x - x_n)$, $n \ge 1$. *Если* $f \in C^n[a, b]$, *то существует точка* $\eta \in [a, b]$ *такая, что для погрешности интерполяционной квадратурной формулы справедлива оценка*

$$|R_n(f)| \le \frac{|f^{(n)}(\eta)|}{n!} \int_a^b \rho(x) |\omega_n(x)|\, dx.$$

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 при любом $x \in [a,b]$ 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 $f^{(n)}(x)$. 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.

Интерполяционную квадратурную формулу на $[a,b]$ для равномерной сетки шага $h = (b-a)/n$ с узлами

$$a = x_0,\ x_1 = a+h,\ x_2 = a+2h,\ \ldots,\ x_n = a+nh = b$$

принято называть формулой Ньютона-Котеса.

Поскольку число узлов равно $(n+1)$, то в этом случае интерполяционная квадратурная формула имеет вид

$$\int_a^b \rho(x)\, f(x)\, dx \approx \int_a^b \rho(x)\, L_{n+1}(f;x)\, dx = \sum_{k=0}^{n} c_k\ f(a+kh),$$

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 показывают, что

$$c_k = \frac{(-1)^{n-k}h}{k!(n-k)!}\int_0^n \rho(a+ht)\frac{t(t-1)...(t-n)}{t-k}\,dt.$$

Поскольку квадратурная формула Ньютона-Котеса точна для функции $f(x) \equiv 1$, имеем

$$\sum_{k=0}^{n} c_k = \int_a^b \rho(x)\,dx.$$

Отсюда следует, что если все коэффициенты $c_k \ge 0$, 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 n.

Поэтому 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 *формулы прямоугольников и трапеций, а также формула Симпсона являются составными квадратурными формулами.* Поясним этот факт подробнее на примере формулы трапеций для приближенного вычисления интеграла

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

Применим сначала квадратурную формулу Ньютона-Котеса на $[a, b]$ для сетки с двумя узлами $a = x_0,\ x_1 = b$. Имеем

$$\int_a^b f(x)\,dx \approx c_0 f(a) + c_1 f(b),$$

где

$$c_0 = \int_a^b \frac{x-b}{a-b}\,dx = \frac{b-a}{2}, \quad c_1 = \int_a^b \frac{a-x}{b-a}\,dx = \frac{b-a}{2}.$$

Получаем приближенную формулу

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2}\ [f(a)+f(b)],$$

которую принято называть **малой формулой трапеций**.

Общая (большая) формула трапеций строится так: сегмент $[a,b]$ делим на $n \geq 2$ равных частей точками

$$a = x_0,\ x_1 = a+h,\ \ldots, \quad x_k = a+kh,\ \ldots,\ x_n = a+nh = b,$$

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 $f_k = f(x_k)$, $h = (b-a)/n$, мы можем записать **большую формулу трапеций** в традиционной форме

$$\int_a^b f(x)\,dx \approx \frac{b-a}{n}\left[\frac{f_0+f_n}{2} + f_1 + f_2 + \ldots + f_{n-1}\right].$$

15.2 Оценки погрешности трех квадратурных формул

Формула трапеций

Из общей теоремы оценки 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 погрешность

$$R_2(f) = \int_a^b f(x)\,dx - \frac{b-a}{2}[f(a)+f(b)]$$

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 15.3 *Если $f \in C^2[a,b]$, то существует $\eta \in [a,b]$ 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 $f(x) = x^2$, сетку $x_k = kh$, $k = 0, 1, ..., n$, с шагом $h = 1/n$. Пользуясь известной формулой

$$1^2 + 2^2 + 3^2 + ... + (n-1)^2 = \frac{(n-1)n(2n-1)}{6},$$

легко вычисляем погрешность формулы трапеций для интеграла $\int_0^1 x^2\,dx$:

$$\int_0^1 x^2\,dx - \frac{1}{n}\left[\frac{1}{2} + \frac{1}{n^2} + \frac{2^2}{n^2} + ... + \frac{(n-1)^2}{n^2}\right] = -\frac{1}{6n^2}.$$

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

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

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

$$|R_n(x)| \le \omega\left(f, \frac{b-a}{n}\right) \cdot \int_a^b dx \le (b-a)\,\omega\left(f, \frac{b-a}{n}\right);$$

в частности, если $f \in \operatorname{Lip}\alpha$ $(0 < \alpha \le 1)$ с постоянной M, то

$$|R_n(f)| \le \frac{M(b-a)^{1+\alpha}}{n^\alpha} = O\left(\frac{1}{n^\alpha}\right);$$

2) пусть $f' \in \operatorname{Lip}\alpha$ $(0 < \alpha \le 1)$ 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

$$R_n(f) \le \frac{b-a}{4n} \cdot \frac{M_1(b-a)^\alpha}{n^\alpha} \cdot (b-a) =$$
$$= \frac{(b-a)^{2+\alpha}M_1}{4n^{1+\alpha}} = O\left(\frac{1}{n^{1+\alpha}}\right).$$

Квадратурные формулы прямоугольников

Малая формула прямоугольников для $f \in C[a,b]$ имеет вид

$$\int_a^b f(x)\,dx \approx (b-a)f(\xi), \quad \xi \in [a,b].$$

Наиболее употребительными являются три частных 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

$$\int_a^b f(x)\,dx = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} f(x)\,dx$$

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 прямоугольников 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 подинтегральной функции, мы можем получить порядковые оценки погрешности $R_n(f)$ для формул прямоугольников.

Теорема 15.4 *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

$$\frac{1}{n}\sum_{k=1}^{n} f(x_{k-1}) = \frac{1}{n}\left(0 + \frac{1}{n} + \frac{2}{n} + \ldots + \frac{n-1}{n}\right) = \frac{1}{2} - \frac{1}{2n},$$

и по формуле правых прямоугольников

$$\frac{1}{n}\sum_{k=1}^{n} f(x_k) = \frac{1}{n}\left(\frac{1}{n} + \frac{2}{n} + \ldots + \frac{n}{n}\right) = \frac{n(n+1)}{2n^2} = \frac{1}{2} + \frac{1}{2n}.$$

Для формулы средних прямоугольников справедлив удивительный факт: оценка погрешности по порядку оказывается такой же, какой она является для формулы трапеций.

Теорема 15.5 *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.

Квадратурная формула Симпсона

Стандартный путь построения формулы Симпсона состоит в замене подинтегральных 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: $A = f(a)$ силу условия $L_3(f;a) = f(a)$, равенство $L_3(f;b) = f(b)$ дает уравнение $f(b) = f(a) + B(b-a)$ 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

$$\int_a^b f(x)\,dx \approx$$

$$\approx \frac{b-a}{3n} \left[\frac{f_0 + f_n}{2} + 2(f_{1/2} + \ldots + f_{n-1/2}) + f_1 + \ldots + f_{n-1} \right].$$

Это и есть классическая формула Симпсона.

Суммируя погрешности малых формул для частичных отрезков, получаем погрешность большой формулы Симпсона: существуют точки $\eta_k \in [x_{k-1}, x_k]$ и $\eta \in [a, b]$ такие, что

$$|R_n(f)| \leq \frac{(b-a)^4}{192n^3} \cdot \left(\frac{1}{n} \sum_{k=1}^{n} |f'''(\eta_k)| \right) =$$

$$= \frac{(b-a)^4}{192n^3} |f'''(\eta)| = O\left(\frac{1}{n^3}\right).$$

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: $H_4(f;a) = f(a)$, $H_4(f;b) = f(b)$,

$$H_4(f;c) = f(c),\ H_4'(f;c) = f'(c).$$

Поскольку сумма кратностей равна $m = 1 + 2 + 1 = 4$, то существует единственный интерполяционный полином $H_4(f;x)$ степени $\leq m - 1 = 3$. Будем искать его в виде следующего полинома

$$H_4(f;x) = A + B(x-a) + C(x-a)(x-b) + D(x-a)(x-b)(x-c) =$$

$$= L_3(f;x) + D(x-a)(x-b)(x-c).$$

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 вычисления 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:

$$\int_a^b f(x)\,dx \approx \Phi_t(f; n) := \frac{b-a}{n} \left[\frac{f_0 + f_n}{2} + f_1 + f_2 + \ldots + f_{n-1} \right],$$

$$\int_a^b f(x)\,dx \approx \Phi_{mr}(f; n) := \frac{b-a}{n} [f_{1/2} + f_{3/2} \ldots + f_{n-1/2}].$$

Очевидно, формула Симпсона

$$\int_a^b f(x)\,dx \approx$$

$$\approx \frac{b-a}{3n}\left[\frac{f_0+f_n}{2} + 2(f_{1/2}+\ldots+f_{n-1/2}) + f_1+\ldots+f_{n-1}\right]$$

получается следующим образом

$$\int_a^b f(x)\,dx \approx \frac{\Phi_t(f;n) + 2\Phi_{mr}(f;n)}{3}.$$

Это дает нам третью, легко запоминающуюся, интерпретацию формулы Симпсона как одну третью часть суммы формулы трапеций и удвоенной формулы средних прямоугольников.

16 Квадратурные формулы Гаусса

До сих пор мы рассматривали квадратурные формулы с произвольными узлами. При любом выборе узлов интерполяционная квадратурная формула

$$\int_a^b \rho(x) f(x)\,dx \approx \sum_{k=1}^{n} A_k\, f(x_k) \tag{22}$$

с коэффициентами

$$A_k = p_k := \int_a^b \rho(x) \frac{\omega_n(x)\,dx}{(x-x_k)\omega_n'(x_k)} \tag{23}$$

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

$$\int_a^b \rho(x)x^m dx = \sum_{k=1}^{n} p_k \; x_k^m \tag{24}$$

для любого $m = 0, 1, \ldots, 2n-1$.

Интерполяционные квадратурные формулы вида (22), точные 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 справедливо следующее утверждение:

ни при каком выборе узлов $x_1, \ldots, x_n \in < a, b >$ *и коэффициентов* A_k *квадратурная формула вида (22) не может быть точной для всех полиномов степени* $2n$.

Доказательство от противного: если существует квадратурная формула вида (22), точная на полиномах степени $2n$, то для функции

$$f(x) = \omega_n^2(x), \qquad \omega_n(x) = \prod_{k=1}^{n}(x - x_k),$$

являющейся полиномом степени $2n$, мы получаем противоречивое соотношение

$$0 < \int_a^b \rho(x)\omega_n^2(x)\, dx = \sum_{k=1}^{n} A_k \omega_n^2(x_k) = 0.$$

16.1 Структура квадратурных 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 16.1 *Квадратурная формула (22) точна для любого полинома степени* $\leq 2n-1$ *тогда и только тогда, когда выполняются следующие два условия:*

1) полином $\omega_n(x) = \prod_{k=1}^{n}(x - x_k)$ *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 (23).*

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). Рассмотрим произвольный 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 равенство

$$\int_a^b \rho(x)q(x)\omega_n(x)dx = \sum_{k=1}^{n} A_k\, q(x_k)\ \omega_n(x_k) = 0.$$

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 $x_1, x_2, ..., x_n$, для которой $\omega_n(x)$ удовлетворяет условию 1) этой теоремы. Окончательное утверждение вытекает из двух последующих теорем.

Теорема 16.2 *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

$$P_n(x) = b_0 + b_1 x + ... + b_{n-1} x^{n-1} + x^n$$

требуемое условие ортогональности можно записать в виде равенств

$$\int_a^b \rho(x)\,(b_0 + b_1 x + ... + b_{n-1} x^{n-1})\, x^j\, dx = -\int_a^b \rho(x)\, x^{n+j}\, dx,$$

которые должны выполняться для всех $j = 0, 1, ..., n-1$. Очевидно, эти интегральные равенства представляют собой систему линейных алгебраических уравнений относительно неизвестных коэффициентов $b_0,\ b_1,\ ...\ , b_{n-1}$. Достаточно показать, что соответствующая однородная система уравнений

$$\int_a^b \rho(x)\,(a_0 + ... + a_{n-1} x^{n-1})\, x^j\, dx = 0\,(j = 0, 1, ..., n-1) \qquad (25)$$

имеет единственное решение $a_0 = a_1 = ... = a_{n-1} = 0$. С этой целью умножим j-тое уравнение (25) на a_j и просуммируем по $j = 0, 1, ..., n-1$. Будем иметь равенства

$$\sum_{j=0}^{n-1} a_j \int_a^b x^j\, \rho(x) \sum_{k=0}^{n-1} a_k x^k\, dx =$$

$$= \int_a^b \rho(x) \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} a_k\, a_j\, x^k\, x^j\, dx =$$

$$= \int_a^b \rho(x) \left| \sum_{k=0}^{n-1} a_k\, x^k \right|^2 dx = 0.$$

Отсюда с учетом неотрицательности подинтегральной 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 $a_0 + a_1 x + ... + a_{n-1} x^{n-1}$ 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 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

$$\|F\| = \sqrt{\int_a^b \rho(x) |F(x)|^2 dx}.$$

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 $a_0, a_1, ..., a_n$ такие, что хотя бы один из коэффициентов a_k отличен от нуля, но полином $a_0 + a_1 x + ... + a_{n-1} x^{n-1}$ 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, во-первых, линейной комбинацией элементов $\{1, x, x^2, \ldots, x^n\}$, в которую входит с ненулевым коэффициентом элемент x^n, и во-вторых, 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.

Отметим, что 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 16.3 *Все нули ортогонального полинома P_n вещественны, 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 сначала, 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 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 $(j = 0, 1, ..., n - 1)$ 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 приводит 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.

16.2 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 $\omega_n(x) = (x-x_1)...(x-x_n)$ *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 $(k = 1, \ldots, n)$. 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 наивысшего 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 16.4 *Для любой функции* $f \in C[a, b]$

$$|\psi_n(f)| \le 2(E_{2n-1} f) \int_a^b \rho(x) dx,$$

где $E_{2n-1} f$ — *наилучшее равномерное приближение* f *полиномами степени* $\le 2n - 1$.

Доказательство. Для произвольного полинома $Q(x)$ степени $\le 2n - 1$ 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:

$$|\psi_n(f)| = \left| \int_a^b \rho(x)[f(x) - Q(x)] dx - \sum_{k=1}^{n} p_k\, [f(x_k) - Q(x_k)] \right| \le$$

$$\leq \|f(x) - Q(x)\|_{C[a,b]} \left\{ \int_a^b \rho(x)dx + \sum_{k=1}^{n} p_k \right\} =$$

$$= 2\|f(x) - Q(x)\|_{C[a,b]} \int_a^b \rho(x)dx.$$

Отсюда в силу произвольности полинома $Q(x)$ степени $\leq 2n-1$ вытекает утверждение теоремы.

Теорема 16.5 *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 $H_n(f;x_k) = f(x_k)$, $H_n'(f;x_k) = f'(x_k)$ $(k = 1,2,...,n)$. 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 $\psi_n(f)$.

16.3 Явный вид формул для специальных весов

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. Мы дадим краткое описание наиболее употребительных ортогональных полиномов и соответствующих им квадратурных формул 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 значений $n = 512$. Соответствующая квадратурная формула

$$\int_{-1}^{1} f(x)dx \approx \sum_{k=1}^{n} p_k\, f(x_k), \qquad p_k = \int_{-1}^{1} \frac{L_n(x)dx}{(x-x_k)L_n'(x_k)}\, dx,$$

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

$$x_k = \cos\left(\frac{(2k-1)\pi}{2n}\right) \quad (k = 1, ..., n).$$

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 $x_k = \cos\frac{k\pi}{n+1}$ $(k = 1, ..., n)$, а квадратурная формула также имеет явный вид:

$$\int_{-1}^{1} \sqrt{1-x^2} f(x)dx \approx \frac{\pi}{n+1} \sum_{k=1}^{n} \sin^2 \frac{k\pi}{n+1} f\left(\cos\frac{k\pi}{n+1}\right).$$

4) Пусть $\rho(x) = (1-x)^\alpha(1+x)^\beta$ на отрезке $[-1,1]$. Фиксированные 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 ортогональна 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 $H_n(x) = e^{x^2}\left(e^{-x^2}\right)^{(n)}$.

17 Дополнительные вопросы

17.1 Об интегрировании периодических функций

Рассмотрим 2π-периодическую, непрерывную функцию $f(x)$. Понятно, что в этом случае для вычисления интеграла

$$\int_0^{2\pi} \rho(x)\, f(x)\, dx$$

можно использовать приведенные ранее квадратурные формулы.

Для периодических функций естественной является также приближенная формула, получаемая заменой функции его 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

$$T_n(f;x_0) = f(x_0),\ T_n(f;x_1) = f(x_1),\ \ldots,\ T_n(f;x_{2n}) = f(x_{2n})$$

на сетке с $2n+1$ узлами $x_0, x_1, \ldots, x_{2n} \in [0, 2\pi]$, $0 < |x_i - x_j| < 2\pi$, $i \neq j$. 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

$$t_k(x) = \frac{\prod_{j=0, j\neq k}^{2n} \sin\frac{x-x_j}{2}}{\prod_{j=0, j\neq k}^{2n} \sin\frac{x_k-x_j}{2}},\ k = 0, 1, ..., 2n.$$

Будем иметь

$$\int_0^{2\pi} \rho(x)\, f(x)\, dx \approx \sum_{k=0}^{2n} q_k\, f(x_k), \tag{26}$$

где

$$q_k = \int_0^{2\pi} \rho(x)\, t_k(x)\, dx\ \ (k = 0, 1, ..., 2n).$$

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 (26) будет точна для любого тригонометрического полинома F степени $\leq n$.

17.2 Интегралы от быстро осциллирующих функций

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 стандартных квадратурных формул может привести к ошибочным результатам. Например, пусть узлы $x_1, x_2, ..., x_n$ 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 неудовлетворительному 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 пользуются формулами 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 $x_1, x_2, ..., x_n \in [a,b]$. 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 $\int_a^b l_j(x)\varphi(x)\,dx$.

17.3 Несобственные интегралы

Если подинтегральная функция f не ограничена на отрезке $[a,b]$, т. е. интеграл является несобственным, то непосредственное применение квадратурных формул может привести к сколь угодно большим ошибкам. Для приближенного вычисления несобственных интегралов нужны предварительные преобразования интеграла. 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

$$J = \int_0^1 \frac{\ln x}{1+x^2} dx.$$

Заменой переменных $x = t^k$ с постоянной $k > 1$ получаем

$$J = k^2 \int_0^1 \frac{t^{k-1} \ln t}{1+t^{2k}} dt.$$

Новая подинтегральная функция

$$g(t) = \frac{t^{k-1} \ln t}{1+t^{2k}}$$

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 подинтегральную 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

$$\int_0^1 \rho(x)g(x)dx \approx \sum_{k=1}^{n} A_k\, g(x_k).$$

Понятно, что число подобных приемов можно увеличить.

Для вычисления несобственных интегралов вида

$$\int_0^\infty f(x)\,dx$$

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 интеграла по лучу $[A, +\infty)$.

18 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 интерполяционным полиномом Лагранжа, найдите приближенное значение $\log 70$ и дайте оценку погрешности в двух случаях: заданы a) $\log 1$, $\log 10$; 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 Чебышева $T_n(t)$ удовлетворяет следующему дифференциальному уравнению

$$(1-t^2)T_n''(t) - tT_n'(t) + n^2 T_n(t) = 0.$$

6. Для функции $f(x) = \sin \pi x$ и узлов $\{0, 1/4, 1/3, 1/2\}$ запишите интерполяционный полином в форме 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 интерполяционный полином Лагранжа для равноотстоящих узлов $x_0, x_1, \ldots, x_n$, причем $a = x_0, \quad b = x_n$ 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 при разбиении $x_k = kh$, $h = 1/n$, $k = 0, 1, ..., n$. Дайте оценку погрешности приближения.

15. При доказательстве теоремы Вейерштрасса по методу Лебега нам встретилась система линейных алгебраических уравнений

$$a_1 - \sum_{j=2}^{m} a_j = k_1,$$

$$\sum_{j=1}^{s} a_j - \sum_{j=s+1}^{m} a_j = k_s, \quad s = 2, ..., m-1,$$

$$\sum_{j=1}^{m} a_j = k_m$$

относительно коэффициентов $a_1, a_2, ..., a_m$ при заданных $k_1, k_2, ..., k_m$. Покажите, что решение этой системы можно записать в явном виде.

16. Для функции $f(x) = x^3$ постройте полином наилучшего равномерного приближения степени n на отрезке $[0, 1]$ для всех $n = 0, 1, 2, 3, ...$

17. Для функции $f(x) = x^3$ постройте полином наилучшего приближения первой и второй степени в пространстве $L_2[0, 1]$.

18. Пусть функция $f(x) = x^3$ задана в точках $1, 2, 3$. Найдите полином наилучшего среднеквадратичного приближения.

19. Для интеграла

$$\int_0^1 xf(x)\,dx$$

постройте квадратурную формулу с двумя узлами, точную для всех полиномов:

a) первой степени, b) второй степени.

20. Найдите алгебраический порядок точности квадратурной формулы

$$\int_0^1 f(x)\,dx \approx \frac{f(0) + 4f(1/2) + f(1)}{6}.$$

21. Для интеграла

$$\int_0^1 xf(x)\,dx$$

постройте квадратурную формулу Гаусса с двумя узлами.

22. 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.

23. Вычислите интеграл

$$\int_{-1}^1 \frac{dx}{\sqrt{1-x^4}}$$

с помощью квадратурной формулы Гаусса с двумя узлами.

24. С точностью $\varepsilon = 0,01$ вычислите интеграл

$$\int_0^1 e^x \sin 100x\,dx.$$

25. С точностью $\varepsilon = 0,01$ вычислите несобственный интеграл

$$\int_0^1 \frac{dx}{\sqrt{x(1-x)(x+1)}}.$$

26. 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.

19 Рекомендуемая литература

Список литературы

[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 университета. 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.* 3-е изд., перераб. и доп. 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 математики.* Учебное пособие. Под ред. Г. И. Марчука. Москва: Наука. 1980.

[8] Канторович Л. В., Крылов В. И. *Приближенные методы высшего анализа.* Москва-Ленинград: ГИТТЛ, 1949.

[9] Крылов В. И., Бобков В. В., Монастырный П. И. *Вычислительные методы.* Т.1, Москва: Наука, 1976. То же. Т. 2. Москва: Наука, 1977.

[10] Натансон И. П. *Конструктивная теория функций.* Москва: Гостехиздат. 1949.

[11] Рябенький В. С. *Введение в вычислительную математику.* Серия "Физтеховский учебник". Москва: Физматлит. 2008.

[12] Самарский А. А., Гулин А. В. *Численные методы.* Учебное пособие для вузов. Москва: Наука. Гл. ред. физ.-мат. лит. 1989.

[13] Срочко В. А. *Численные методы. Курс лекций.* Учебное пособие для вузов. Санкт-Петербург: Изд-во ЛАНЬ. 2010.

[14] Стечкин С. Б., Субботин Ю. Н. *Сплайны в вычислительной математике.* Москва: Наука. Гл. ред. физ.-мат. лит. 1976.

[15] Тихомиров В. М. *Теория приближений.* "Современные проблемы математики. Фундаментальные направления. Т. 14 (Итоги науки и техники, ВИНИТИ АН СССР)" Москва. 1987, с. 103-260.

[16] Тыртышников Е. Е. *Методы численного анализа.* Москва. 2006.