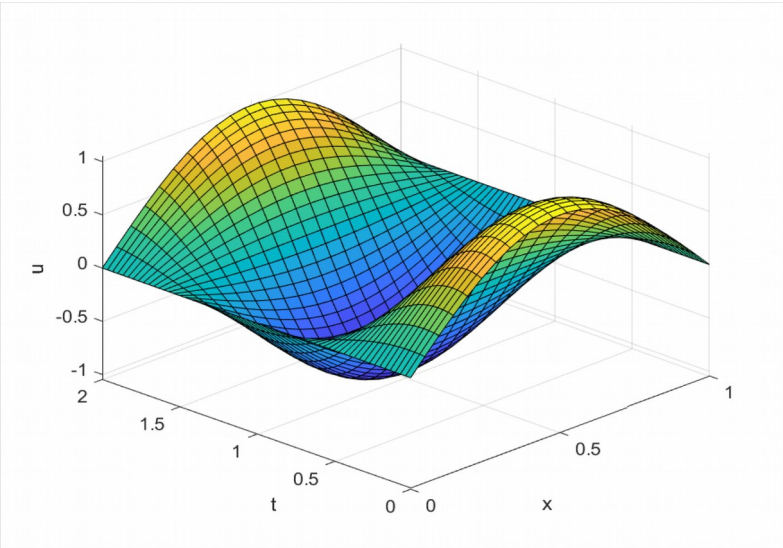
УНИВЕРСИТЕТ ИТМО

В.В. Залипаев, Д.Р. Гулевич

ЧИСЛЕННЫЕ МЕТОДЫ В ФИЗИКЕ И ТЕХНИКЕ



Санкт-Петербург
2020

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

УНИВЕРСИТЕТ ИТМО

В.В. Залипаев, Д.Р. Гулевич

ЧИСЛЕННЫЕ МЕТОДЫ В ФИЗИКЕ И ТЕХНИКЕ

УЧЕБНОЕ ПОСОБИЕ

РЕКОМЕНДОВАНО К ИСПОЛЬЗОВАНИЮ В УНИВЕРСИТЕТЕ ИТМО
по направлению подготовки 12.04.03
в качестве учебного пособия для реализации основных профессиональных образовательных программ высшего образования магистратуры.

Санкт-Петербург
2020

Гулевич Д.Р., Залипаев В.В., Численные методы в физике и технике– СПб: Университет ИТМО, 2020. – 211 с.

Рецензент: Мирошниченко Георгий Петрович, доктор физико-математических наук, профессор, профессор факультета лазерной фотоники и оптоэлектроники, Университета ИТМО.

В пособии излагаются основные принципы построения и исследования численных методов решения различных классов математических задач компьютерного моделирования. Наряду с традиционными разделами теории, такими как интерполирование и аппроксимация, численное интегрирование, методы решения задач Коши для обыкновенных дифференциальных уравнений, значительное место занимают разностные методы для уравнений в частных производных, а также методы Ритца и Галеркина, метод конечных элементов и метод граничных интегральных уравнений.

Содержание

1 Введение

В пособии излагаются основы численных методов решения задач алгебры, анализа, дифференциальных уравнений обыкновенных и в частных производных (задач математической физики). Изложенный материал предназначен для студентов университетов, будущих математиков, физиков и инженеров, специализирующихся в области прикладной математики, вычислительной физики и компьютерного моделирования. Изложение материала в пособие основано на курсе лекций, который читался авторами в течение ряда лет студентам факультетов физико-технического, фотоники и оптоинформатики, начиная с 2017 года, а также студентам инженерно-механического факультета Саутгемптонсого университета в Великобритании.

При изложении описания численных методов изучаются вопросы построения, применения и теоретического обоснования алгоритмов приближенного решения различных классов математических задач как основы компьютерного моделирования. В настоящее время большинство вычислительных алгоритмов уже ориентировано на использование быстродействующих компьютеров и компьютерных кластеров. Изучение численных методов можно охарактеризовать некоторыми особенностями. Для них характерна множественность, так как можно решать одну и ту же задачу разными методами. Это очень важно, так как в большинстве случаев в математическом и компьютерном моделировании точные решения задач отсутствуют и проверить полученное решение задачи одним численным методом можно только с помощью альтернативного алгоритма. Возникающие новые научные задачи и быстрое развитие вычислительной техники стимулируют переоценку мощности существующих алгоритмов и приводят к созданию новых. Авторы пособия поставили перед собой задачу собрать минимальный материал, достаточный для дальнейшей научной работы студентов и выпускников вузов в области применения и развития вычислительных методов.

Наибольшее внимание уделяется фундаментальным разделам численных методов – численному решению систем линейных алгебраических уравнений и разностным методам решений задач мате-

матической физики. Многие интересные и важные методы изложены недостаточно полно или совсем не вошли в пособие. За рамками остались такие этапы компьютерного моделирования, как построение математической модели, программирование и организация вычислений. Не вошли в материал изложения современные методы обработки сигналов. Для более глубокого и детального изученния теории численных методов авторы рекомендуют дополнительную литературу [1], [2], [3], [4], [5]. При изучении этих материалов следует обращаться к рекомендуемым ниже первоисточникам по курсу высшей математики, дифференциальному и интегральному исчислению и теории уранений в частных производных, например, [8], [9], [6], [7]. Для изученния теории численных методов на английском языке рекомендуются две хорошо известные и легко читаемые книги [10], [11].

Изложение материала численных методов в пособии начинается с описания традиционных вычислительных методов, таких как прямые и итерационные методы решения систем линейных алгебраических уравнений, интерполирование и аппроксимация функций, численное интегрирование, решение нелинейных уравнений, методы решения задачи Коши для обыкновенных дифференциальных уравнений. В качестве примера применения метода Ньютона рассмотрено численное решение обратной задачи определения диэлектрической проницаемости слоя по данным рассеяния. Далее обсуждается решение краевой задачи для обыкновенных дифференциальных уравнений второго порядка, полученное на основе метода конечных разностей. С помощью этого алгоритма описывается решение для построения дисперсионных зависимостей для задачи распространения волн в одномерном фотонном кристалле, хорошо известной в оптике, акустике и радиофизике. На основе этого алгоритма описана численная процедура нахождения спектра и волновых функций электронов в атоме водорода. Далее рассматриваются разностные методы решения смешанных задач математической физики для эллиптических, параболических и гиперболических уравнений в частных производных. Здесь изложение строится на простых и важнейших примерах уравнений Пуассона, теплопроводности и волнового уравнения. Большое внимание уделяется принципам построения разностных схем для широкого спектра задач, методам

решения сеточных уравнений, исследованию их устойчивости и сходимости, анализу погрешностей. Положительным моментом материала, изложенного в пособии, является описание методов Ритца и Галеркина как для краевых задач обыкновенных дифференциальных уравнений второго порядка, так и для для краевых задач дифференциальных уравнений второго порядка в частных производных эллиптического типа. После этого приводится введение в метод конечных элементов для уравнения Лапласа в прямоугольной области. В заключительной части присутствует кратко изложенный материал для хорошо известного и популярного в теории распространения волн в оптике, акустике и радиофизике, метода граничных интегральных уравнений Фредгольма. Рассматриваются с помощью этого метода задачи рассеяния плоской электромагнитной волны на бесконечном идеально-проводящем цилиндре и тонком проводе конечной длины. Главным достоинством пособия является приложение, в котором для всех двенадцати тем содержания по курсу численных методов приводятся описания заданий рекомендуемых лабораторных работ, которые были успешно реализованы и протестированы в системе MATLAB. Следует заметить, что предлагаемое пособие в некоторой степени аналогично следующим известным источникам [12], [13]. Но присутствие дополнения по лабораторным работам и последних трех тем в содержании дает явные преимущества по сравнению с материалом, представленным в [12], [13].

Пособие адресовано студентам Университета ИТМО, изучающим под руководством авторов курсы «Технологии программирования» и «Математические методы компьютерных технологий в научных исследованиях» в рамках направления «Фотоника и оптоинформатика» по подготовки бакалавров и магистров по программам «Оптические и квантовые технологии в коммуникациях» и «Квантовые коммуникации и фемтотехнологии» на факультете фотоники и оптоинформатики. Пособие может использоваться как для самостоятельного изучения численных методов студентами, аспирантами и научными работниками, так и служить руководством к решению задач, возникающих в научно-исследовательских бакалаврских, магистерских и аспирантских проектах, связанных с численными расчетами. В помощь к самостоятельному изучению мате-

риала в конце каждой главы в пособии приведены вопросы для самоконтроля. Последняя глава пособия полностью посвящена рекомендуемым к выполнению лабораторным работам.

2 Системы линейных алгебраических уравнений (прямые методы)

2.1 Пример из механики

Рассмотрим пример, в котором одномерный массив из трех типов частиц, взаимодействующих при помощи жестких пружин, описывается системой дифференциальных уравнений

$$\begin{cases} m_1\ddot{u}_n = -k(v_n + w_{n-1} - 2u_n) - f_n^{(1)}(t), \\ m_2\ddot{v}_n = -k(w_n + u_n - 2v_n) - f_n^{(2)}(t), \\ m_3\ddot{w}_n = -k(u_{n+1} + v_n - 2w_n) - f_n^{(3)}(t), \end{cases} \tag{1}$$

представляющей собой результат применения второго закона Ньютона, где u_n, v_n, w_n суть вертикальные смещения частиц, m_1, m_2, m_3 их массы, k является общей жесткостью пружин, $f_n^{(1)}$, $f_n^{(2)}$, $f_n^{(3)}$ суть внешние силы и $n = 0, \pm1, \pm2, ..., \pm\infty$. Происхождение этой системы уравнений можно проиллюстрировать с помощью аппроксимации конечными разностями хорошо известного уравнения колебаний струны (см. [6], [7])

$$\rho\ddot{u} = Tu_{xx} + F(t,x),$$

$$m\ddot{u}_n = \frac{T}{h}(u_{n+1} + u_{n-1} - 2u_n) + f_n(t), \quad m = \rho h, \quad f_n(t) = hF(t,x_n),$$

$$u_{xx}(x_n) \sim \frac{\frac{u_{n+1}-u_n}{h} - \frac{u_n-u_{n-1}}{h}}{h}, \qquad x_n = hn.$$

Мы предполагаем, что временная зависимость внешних сил имеет вид гармонической функции $\exp(-i\omega t)$ с частотой ω. Пользуясь трансляционной инвариантностью предположим, что

$$f_n^{(j)}(t) = f_j \exp(-i\omega t + iKLn),$$

где $j = 1, 2, 3$, K является «квазиимпульсом» - свободным параметром ($-\pi/L < K < \pi/L$), L - период массива. Мы ищем решение

системы в форме

$$\begin{cases} u_n = u\exp(-i\omega t + iKLn), \\ v_n = v\exp(-i\omega t + iKLn), \\ w_n = w\exp(-i\omega t + iKLn), \end{cases} \tag{2}$$

где $u,\ v,\ w$ суть неизвестные постоянные амплитуды. Подставляя эти формулы в систему дифференциальных уравнений (1), мы получаем систему 3×3 линейных алгебраических неоднородных уравнений для неизвестных $u,\ \ v,\ \ w$:

$$\begin{pmatrix} m_1\omega^2 + 2k & k & ke^{-iKL} \\ k & m_2\omega^2 + 2k & k \\ ke^{iKL} & k & m_3\omega^2 + 2k \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix}. \tag{3}$$

Решение этой неоднородной системы можно получить с помощью известных методов линейной алгебры, если определитель матрицы системы не равен нулю. Приравнивая определитель матрицы системы нулю для однородной задачи, мы получаем дисперсионные зависимости $\omega = \omega_s(K)$, $s = 1, 2, 3$, волн, распространяющихся вдоль рассматриваемого одномерного массива из трех частиц.

2.2 Метод Крамера

Рассмотрим сначала систему из трех линейных алгебраических уравнений с тремя неизвестными

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = f_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = f_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = f_3. \end{cases} \tag{4}$$

Эту систему можно решить методом Крамера:

$$x_1 = \frac{\det A_1}{\det A}, \qquad x_2 = \frac{\det A_2}{\det A}, \qquad x_3 = \frac{\det A_3}{\det A}, \tag{5}$$

если $\det A \neq 0$, где

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}, \quad A_1 = \begin{pmatrix} f_1 & a_{12} & a_{13} \\ f_2 & a_{22} & a_{23} \\ f_3 & a_{32} & a_{33} \end{pmatrix}, \quad A_2 = \begin{pmatrix} a_{11} & f_1 & a_{13} \\ a_{21} & f_2 & a_{23} \\ a_{32} & f_3 & a_{33} \end{pmatrix},$$

$$A_3 = \begin{pmatrix} a_{11} & a_{12} & f_1 \\ a_{21} & a_{22} & f_2 \\ a_{31} & a_{32} & f_3 \end{pmatrix}.$$

2.3 Пример применения метода Гаусса

Для решения системы (4) мы также можем использовать метод Гаусса - метод последовательных исключений неизвестных. Рассмотрим пример:

$$\begin{cases} x_1 + x_2 + 3x_4 = 4, \\ 2x_1 + x_2 - x_3 + x_4 = 1, \\ 3x_1 - x_2 - x_3 + 2x_4 = -3, \\ -x_1 + 2x_2 + 3x_3 - x_4 = 4, \end{cases}$$

Прямые исключения дают

$$\left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 2 & 1 & -1 & 1 & 1 \\ 3 & -1 & -1 & 2 & -3 \\ -1 & 2 & 3 & -1 & 4 \end{array}\right), \qquad \left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & -4 & -1 & -7 & -15 \\ 0 & 3 & 3 & 2 & 8 \end{array}\right),$$

$$\left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & 0 & 3 & 13 & 13 \\ 0 & 0 & 0 & -13 & -13 \end{array}\right).$$

Теперь рассмотрим обратные замены:

$$-13x_4 = -13, \quad \Rightarrow x_4 = 1,$$

$$3x_3 + 13x_4 = 13 \quad \Rightarrow x_3 = 0,$$

$$-x_2 - x_3 - 5x_4 = -7, \quad \Rightarrow x_2 = 2,$$

$$x_1 + x_2 + 3x_4 = 4 \quad \Rightarrow x_1 = -1.$$

2.4 Метод исключения Гаусса в общем случае

Рассмотрим систему линейных алгебраических уравнения

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = f_1, \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = f_2, \\ ....., \\ a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n = f_n, \end{cases} \tag{6}$$

который в матричной форме выглядит так: $Ax = f$, или

$$\begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ ... \\ f_n \end{pmatrix}. \tag{7}$$

Здесь $A = \| \ a_{ij} \ \|$ - матрица с индексами $i, j = 1, .., n$, $x = (x_1, x_2, ..., x_n)^T$ и $f = (f_1, f_2, ..., f_n)^T$ суть векторы. Вектор x неизвестен. Предположим, что $\det A \neq 0$.

Если n имеет большое значение, то методом Крамера потребуется много времени для выполнения вычислений, приблизительное количество операций имеет порядок $n!$. Напротив, метод Гаусса гораздо быстрее, так как ему нужно приблизительно $n^3/3$ операций. Для примера, если $n = 10$, тогда $10! = 3628800$ в сравнении с $10^3/3 \approx 333$. Таким образом метод исключений Гаусса более эффективен при численном анализе.

После n шагов по прямому устранению неизвестных мы получаем систему с верхней треугольной матрицей U

$$\begin{pmatrix} 1 & u_{12} & u_{13} & ... & u_{1n} \\ 0 & 1 & u_{23} & ... & u_{2n} \\ 0 & 0 & 1 & ... & u_{2n} \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ ... \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ ... \\ y_n \end{pmatrix}, \tag{8}$$

или в краткой форме $Ux = y$.

Что происходит после того, как мы сделали $k - 1$ шагов? У нас

есть блок уравнений

$$\begin{cases} a_{kk}^{(k-1)}x_k + a_{k,k+1}^{(k-1)}x_{k+1} + ... + a_{kn}^{(k-1)}x_n = f_k^{(k-1)}, \\ a_{k+1,k}^{(k-1)}x_k + a_{k+1,k+1}^{(k-1)}x_{k+1} + ... + a_{k+1,n}^{(k-1)}x_n = f_{k+1}^{(k-1)}, \\ ..., \\ a_{n,k}^{(k-1)}x_k + a_{n,k+1}^{(k-1)}x_{k+1} + ... + a_{nn}^{(k-1)}x_n = f_n^{(k-1)}. \end{cases}$$

Сделав следующий шаг, получим

$$\begin{cases} x_k + u_{k,k+1}x_{k+1} + ... + u_{kn}x_n = y_k, \\ 0 \ + a_{k+1,k+1}^{(k)}x_{k+1} + ... + a_{k+1,n}^{(k)}x_n = f_{k+1}^{(k)}, \\ ..., \\ 0 \ + a_{n,k+1}^{(k)}x_{k+1} + ... + a_{nn}^{(k)}x_n = f_n^{(k)}. \end{cases} \tag{9}$$

где

$$u_{kj} = \frac{a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}}, \qquad y_k = \frac{f_k^{(k-1)}}{a_{kk}^{(k-1)}}, \tag{10}$$

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - a_{ik}^{(k-1)}u_{kj}, \qquad i,j = k+1, ..., n, \tag{11}$$

$$f_i^{(k)} = f_i^{(k-1)} - a_{ik}^{(k-1)}y_k, \tag{12}$$

и $a_{1j}^{(0)} = a_{1j}, \quad j = 1, 2, ..., n$, $f_1^{(0)} = f_1$.
Эти формулы представляют собой первую часть алгоритма. Они описывают последовательное исключение неизвестных x_i. Неизвестные могут быть легко найдены из полученной системы с помощью обратных подстановок

$$x_i = y_i - \sum_{j=i+1}^{n} u_{ij}x_j \tag{13}$$

для $1 \le i \le n-1$, если $i = n$, мы имеем $x_n = y_n$. Эти формулы обеспечивают вторую часть алгоритма. Используя этот алгоритм, мы можем построить компьютерный код.

2.5 Трехдиагональные системы

Рассмотрим систему 4×4

$$\begin{pmatrix} b_1 & c_1 & 0 & 0 \\ a_1 & b_2 & c_2 & 0 \\ 0 & a_2 & b_3 & c_3 \\ 0 & 0 & a_3 & b_4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix}. \tag{14}$$

Это и есть пример трехдиагональной системы. Решение таких систем приводит к конкретным простым результатам по гауссовскому исключению неизвестных. Применим метод Гаусса для решения такой системы. Прямое исключение на каждом шаге приводит к системе следующего общего вида:

$$\begin{pmatrix} 1 & c_1/d_1 & 0 & 0 \\ 0 & 1 & c_2/d_2 & 0 \\ 0 & 0 & 1 & c_3/d_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}, \tag{15}$$

где

На первом этапе: $d_1 = b_1, \quad y_1 = f_1/d_1$.

На втором этапе: $d_2 = b_2 - a_1c_1/d_1, \quad y_2 = (f_1 - y_1a_1)/d_2$.

На третьем этапе: $d_3 = b_3 - a_2c_2/d_2, \quad y_3 = (f_3 - y_2a_2)/d_3$.

На четвертом этапе: $d_4 = b_4 - a_3c_3/d_3, \quad y_4 = (f_4 - y_3a_3)/d_4$.

Наконец, процедура обратной замены дает ответ

$$\begin{cases} x_4 = y_4, \\ x_3 = y_3 - x_4c_3/d_3, \\ x_2 = y_2 - x_3c_2/d_2, , \\ x_1 = y_1 - x_2c_1/d_1. \end{cases}$$

Ясно, что произойдет в общем случае для $n \times n$ трехдиагональной системы.

На первом этапе: $d_1 = b_1, \quad y_1 = f_1/d_1$.

На k-м этапе будем иметь: $d_k = b_k - a_{k-1}c_{k-1}/d_{k-1}, \quad y_k = (f_k - y_{k-1}a_{k-1})/d_k$.

Тогда обратные подстановки выглядят так:

$x_n = y_n, \qquad x_{k-1} = y_{k-1} - x_kc_{k-1}/d_{k-1}$.

2.6 LU разложение матриц

Рассмотрим систему $Ax = f$. В результате применения метода Гаусса мы получили систему $Ux = y$, где U есть верхняя треугольная матрица с единицами на главной диагонали. Вектор y зависит только от вектора f и матрицы A. Рассмотрим соотношение между y и f. В случае $n = 3$ мы имеем

$$\begin{cases} y_1 = f_1/a_{11}, \\ y_2 = (f_2 - y_1 a_{21})/a_{22}^{(1)}, \\ y_3 = (f_3 - y_1 a_{31} - y_2 a_{32}^{(1)})/a_{33}^{(2)}. \end{cases}$$

Отсюда следует

$$\begin{cases} f_1 = a_{11} y_1, \\ f_2 = y_1 a_{21} + y_2 a_{22}^{(1)}, \\ f_3 = y_1 a_{31} + y_2 a_{32}^{(1)} + y_3 a_{33}^{(2)}. \end{cases}$$

Понятно, что для произвольных n мы будем иметь

$$\begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ ... \\ f_n \end{pmatrix} = \begin{pmatrix} a_{11} & 0 & 0 & ... & 0 \\ a_{21} & a_{22}^{(1)} & 0 & ... & 0 \\ a_{31} & a_{32}^{(1)} & a_{33}^{(2)} & ... & 0 \\ ... & ... & ... & ... & \\ a_{n1} & a_{n2}^{(1)} & a_{n3}^{(2)} & ... & a_{nn}^{(n-1)} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ ... \\ y_n \end{pmatrix}. \qquad (16)$$

Таким образом, $f = Ly$, где L – нижняя треугольная матрица с элементами

$$l_{ij} = \begin{cases} a_{ij}^{(j-1)}, & i \geq j, \\ 0, & i < j, \end{cases}$$

которая может быть рассчитана с использованием формул базового алгоритма (10)-(12). Так как

$$Ax = f, \quad Ux = y, \quad Ly = f,$$

то получаем

$$Ax = Ly = LUx.$$

Следовательно, мы имеем $A = LU$, где L есть нижняя треугольная матрица и U - верхняя треугольная матрица. Это представление LU разложения для матрицы A часто называется LU факторизацией матрицы A. Приведем пример для 3×3 матрицы

$$\begin{pmatrix} 60 & 30 & 20 \\ 30 & 20 & 15 \\ 20 & 15 & 12 \end{pmatrix} = \begin{pmatrix} 60 & 0 & 0 \\ 30 & 5 & 0 \\ 20 & 5 & 1/3 \end{pmatrix} \begin{pmatrix} 1 & 1/2 & 1/3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Предположим, что для системы $Ax = f$ мы знаем матрицы L и U. Тогда мы сможем легко решить систему в два простых шага: первый - $Ly = f$ дает решение для y, второй - $Ux = y$ обеспечивает решение для x. Проблема состоит в получении матриц L и U. Матрица A имеет n^2 коэффициентов, тогда число коэффициентов в двух матрицах L и U определяется как $n(n+1)$, поэтому есть некоторая недоопределенность. Это дает нам свободу выбора некоторых коэффициентов в L или U.

Укажем некоторые конкретные варианты:

• Если метод исключения Гаусса косвенно выбирает $l_{ii} = 1$, то этот вариант называется *Метод Дулитла.*

• Представление с условием $u_{ii} = 1$ известно как *Метод Кроута.*

• Если $A = A^T$ (т. е. А является симметричной матрицей) и A является также положительно-определенной матрицей (т. е. имеет положительные собственные значения), то мы можем выбрать $u_{ij} = l_{ji}$. Этот вариант называется *Метод Холецкого.* В этом случае будем иметь

$$A = LU, \qquad A^T = U^T L^T = LU, \quad \rightarrow \quad L = U^T.$$

Заметим, что, строго говоря, LU разложение матрицы не всегда возможно. Теорема утверждает, что если все главные миноры матрицы A не равны нулю, то A представляется как $A = LU$ единственным образом. При этом диагональные элементы L отличны от нуля.

С другой стороны, все гауссовские преобразования могут быть записаны в матричной форме. Например, если $n = 3$, имеем

$$U = L_3 L_2 L_1 A,$$

где

$$L_1 = \begin{pmatrix} 1/a_{11} & 0 & 0 \\ -a_{21}/a_{11} & 1 & 0 \\ -a_{31}/a_{11} & 0 & 1 \end{pmatrix}, \quad L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/a_{22}^{(1)} & 0 \\ 0 & -a_{32}^{(1)}/a_{22}^{(1)} & 1 \end{pmatrix},$$

$$L_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/a_{33}^{(2)} \end{pmatrix}.$$

Здесь L_1 выполняет операцию исключения неизвестных в первом столбце, L_2 - во втором столбце, и так далее. Для матрицы L мы имеем

$$L = L_1^{-1} L_2^{-1} L_3^{-1},$$

где

$$L_1^{-1} = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & 1 & 0 \\ a_{31} & 0 & 1 \end{pmatrix}, \quad L_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & a_{22}^{(1)} & 0 \\ 0 & a_{32}^{(1)} & 1 \end{pmatrix}, \quad L_3^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & a_{33}^{(2)} \end{pmatrix}.$$

В общем случае для системы $n \times n$ у нас есть

$$U = L_n L_{n-1} ... L_2 L_1 A, \qquad L = L_1^{-1} L_2^{-1} ... L_{n-1}^{-1} L_n^{-1},$$

где матрицы L_k и L_k^{-1} $k = 1, ..., n$, похожи по структуре на матрицы, приведенные для случая $n = 3$.

2.7 Метод гауссовых исключений с перестановкой строк

Что делать, если на одном шаге ведущий элемент $a_{k+1,k+1}^{(k)} = 0$? Теорема утверждает, что если $\det A \neq 0$, существует матрица перестановок P, такая что матрица PA имеет все основные миноры, не равные нулю, следовательно $PA = LU$.

Основная идея исключения по Гауссу с перестановкой строкой заключается в том, что на каждом шаге процедуры исключения неизвестных мы осуществляем перестановку строк с i, $i = k+1, ..., n$, таким образом, что мы имеем старший элемент $a^{(k)}_{k+1,k+1}$ максимальный по модулю. Мы ищем строку i с максимальным $|a^{(k)}_{i,k+1}|$, $i = k+1, ..., n$ (см. (9)), а затем осуществляем перестановку между строками $k+1$ и i. Наконец, получим

$$L = L_n L_{n-1} P_{n-1,j_{n-1}} ... L_1 P_{1,j_1} A,$$

где P_{k,j_k} с $k = 1, ..., n-1$, является матрицей перестановки строк. Например, если $n = 3$,

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

обеспечивают перестановку между строками 1 и 2, 2 и 3, 1 и 3.

2.8 Численный расчет матрицы определителя и обратной матрицы

Используя LU разложение, легко вычислить определитель матрицы

$$\det A = \pm \det(PA) = \pm \det(LU) = \pm \det L \det U \\ = \pm \det L = \pm l_{11} l_{22} ... l_{nn}.$$

Если A^{-1} является матрицей, обратной по отношению к A, тогда $AA^{-1} = I$, где I - единичная матрица. Обозначим $X = A^{-1}$. На самом деле мы имеем n^2 систем линейных алгебраических уравнений

$$\sum_{k=1}^{n} a_{ik} x_{kj} = \delta_{ij},$$

или $Ax^{(j)} = \delta^{(j)}$, где $x^{(j)} = (x_{1j}, x_{2j}, ..., , x_{nj})^T$ и $\delta^{(j)} = (\delta_{1j}, ..., \delta_{nj})^T$, $i, j = 1, ..., n$. С помощью разложения $A = LU$ мы должны сначала

решить систему $Ly^{(j)} = \delta^{(j)}$ для $y^{(j)} = (y_{1j}, y_{2j}, ..., , y_{nj})^T$, и затем $Ux^{(j)} = y^{(j)}$ для $x^{(j)}$.

2.9 Разложение Холецкого для симметричных матриц

Метод Холецкого определяет $A = LL^T$, где A является симметричной матрицей и

$$L = \begin{pmatrix} l_{11} & 0 & 0 & ... & 0 \\ l_{21} & l_{22} & 0 & ... & 0 \\ l_{31} & l_{32} & l_{33} & ... & 0 \\ ... & ... & ... & ... & \\ l_{n1} & l_{n2} & l_{n3} & ... & l_{nn} \end{pmatrix}, \qquad L^T = \begin{pmatrix} l_{11} & l_{21} & l_{31} & ... & l_{n1} \\ 0 & l_{22} & l_{32} & ... & l_{n2} \\ 0 & 0 & l_{33} & ... & l_{n3} \\ ... & ... & ... & ... & \\ 0 & 0 & 0 & ... & l_{nn} \end{pmatrix}.$$

Если мы распишем его в компонентах для случая матриц 3×3, то получим:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ .. & a_{22} & a_{23} \\ .. & .. & a_{33} \end{pmatrix} = \begin{pmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ .. & l_{22}^2 + l_{21}^2 & l_{31}l_{21} + l_{22}l_{32} \\ .. & .. & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{pmatrix}.$$

В результате получим набор уравнений для l_{ij}:

$$\begin{aligned} l_{11}^2 &= a_{11} \quad \Rightarrow \; l_{11} = a_{11}^{1/2} \\ l_{11}l_{21} &= a_{12} \quad \Rightarrow \; l_{21} = \left(l_{11}^{-1}\right) a_{12} \\ l_{11}l_{31} &= a_{13} \quad \Rightarrow \; l_{31} = \left(l_{11}^{-1}\right) a_{13} \\ l_{22}^2 + l_{21}^2 &= a_{22} \quad \Rightarrow \; l_{22} = \left(a_{22} - l_{21}^2\right)^{1/2} \\ l_{31}l_{21} + l_{32}l_{22} &= a_{32} \quad \Rightarrow \; l_{32} = \left(l_{22}^{-1}\right)\left(a_{32} - l_{21}l_{31}\right) \\ l_{31}^2 + l_{32}^2 + l_{33}^2 &= a_{33} \quad \Rightarrow \; l_{33} = \left(a_{33} - l_{31}^2 - l_{32}^2\right)^{1/2} \end{aligned}$$

Этот алгоритм использует значения l_{ij}, которые уже получены для нахождения l_{ij} с более высокими i и j. Общая формула для $n \times n$ симметричной матрицы имеет вид

$$l_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2\right)^{1/2}; \tag{17}$$

$$l_{ji} = \left(l_{ii}^{-1}\right)\left(a_{ij} - \sum_{k=1}^{i-1} l_{ik}l_{jk}\right) \quad j = i+1, i+2, \ldots n. \tag{18}$$

Мы отмечаем, что (17) предполагает использование квадратного корня. Следовательно, факторизация будет существовать только для ограниченного класса матриц: это симметричные положительно определенные матрицы. Симметричная матрица A положительно определена, если выполняется условие $(x, Ax) > 0$ для всех ненулевых векторов x. Необходимое и достаточное условие для положительной определенности состоит в том, что все собственные значения положительны. Можно проверить матрицу на положительную определенность, пытаясь сделать вычисление факторизацией Холецкого. На самом деле трудным шагом здесь является знание того, что матрица положительно определенная! Это предполагает поиск собственных значений и может занять больше времени, чем время на решение уравнения! Существует теорема о том, что матрица положительно определена, если:

1. Диагональные члены a_{ii} все положительные, и
2. Матрица является *строго диагонально доминирующей*, т. е. если

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \quad i = 1, \ldots, n. \tag{19}$$

Вопросы для самоконтроля к главе 2

1. Перечислите примеры физических задач, где может возникнуть необходимость решения систем линейных алгебраических уравнений.

2. Перечислите группы методов решения систем линейных алгебраических уравнений.

3. Почему правило Крамера не является эффективным для построения численных решений систем линейных алгебраических уравнений больших размеров?

4. Как метод Гаусса связан с LU разложением матрицы?

5. Можно ли использовать метод Гаусса для нахождения обратной матрицы?

3 Системы линейных алгебраических уравнений (итерационные методы)

3.1 Итерационные методы Якоби, Зейделя и SOR метод

В отличие от прямых методов для решения системы линейных уравнений

$$Ax = \mathbf{f}, \tag{20}$$

которые предоставляют конечный алгоритм и приводят к формуле для точных ответов, итерационные методы состоят иногда из длинной последовательности вычислительных шагов, приводящей к приближенному решению проблемы. Выберем начальную итерацию

$$x = x^{(0)}. \tag{21}$$

Это, вероятнее всего, есть грубое приближение для создания новой итерации $x^{(1)}$. Повторим эту процедуру n раз, генерируя $x^{(n+1)}$ с помощью $x^{(n)}$. Остановимся, когда по какому-либо критерию итерация $x^{(n+1)}$ достаточно хороша - близка к истинному решению системы (20). Процедура итерационного процесса должна обеспечить сходимость итерационного процесса, а именно, приводить к тому, что $x^{(n+1)}$ расположено ближе к истинному решению, чем $x^{(n)}$, а также быть эффективной в смысле, что объем работы (компьютерное время и память) был сбалансирован между усилиями по воспроизведению следующего приближения и улучшением точности новой итерации в противоположность предыдущей. Итеративная процедура (в отличие от прямой процедуры) обычно не достигает точного результата в конечной последовательности вычислений. Чтобы определить, насколько близко n-я итерация приблизилась к точному решению, требуется определение нормы вектора и матрицы.

Определение: норма вектора x, есть неотрицательное число и обозначается как $||x||$, где $x = (x_1, x_2, ..., x_n)$ является вещественным или комплексным вектором из n компонент ($x \in \mathcal{R}_n, \mathcal{C}_n$). Норма вектора удовлетворяет условиям: $||x|| \geq 0$, и $x = 0 \Leftrightarrow ||x|| = 0$.

Для вещественных или комплексных λ имеем

$$||\lambda x|| = |\lambda| \cdot ||x||.$$

Для нормы двух векторов выполняется хорошо известное неравенство треугольника:

$$||x + y|| \leq ||x|| + ||y||.$$

Определение нормы не единственно, приведем некоторые примеры:

$$||x||_p = \left\{ \sum_{i=1}^{n} |x_i|^p \right\}^{1/p}, \quad p > 0, \tag{22}$$

для $p = 1$ имеем:

$$||x||_1 = \sum_{i=1}^{n} |x_i|, \tag{23}$$

случай $p = 2$ дает евклидову норму или длину вектора:

$$||x||_2 = \left\{ \sum_{i=1}^{n} |x_i|^2 \right\}^{1/2}, \tag{24}$$

и для $p = \infty$ получаем норму максимума:

$$||x||_\infty = ||x||_C = \max_{i=1,\ldots,n} |x_i|. \tag{25}$$

Важным моментом является то, что мы можем выбрать ту норму вектора, которая наиболее удобна для нас, чтобы проверить, какова будет разность между приближением и точным решением, и что независимо от того, какую норму мы выберем, мы получим более или менее тот же ответ на этот вопрос.

Мы также можем ввести понятие нормы для матриц. Они должны удовлетворять следующим условиям и неравенствам: $||A|| \geq 0$, и $||A|| = 0 \Leftrightarrow A = 0$;

$$||\lambda A|| = |\lambda| ||A||, \quad ||A + B|| \leq ||A|| + ||B||, \quad ||AB|| \leq ||A|| \cdot ||B||,$$

$$||Ax|| \leq ||A|| \cdot ||x||, \qquad \forall A, x. \tag{26}$$

Как и в случае с векторными нормами, существуют различные варианты. Норма матрицы порядка $m \times n$ может быть построена путем рассмотрения ее как вектора в пространстве $\mathcal{R}_{mn}$. Например, евклидова норма определяется формулой

$$||A||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}. \tag{27}$$

Наиболее общее определение матричной нормы формулируется с помощью векторной нормы:

$$||A|| = \max_{\mathbf{x} \neq 0} \frac{||Ax||}{||x||}. \tag{28}$$

Этот выбор матричной нормы дает для единичной матрицы I единичную норму: $||I|| = 1$. Рассмотрим некоторые примеры:

$$\begin{aligned} ||A||_1 &= \max_{j=1,\ldots,m} \left\{ \sum_{i=1}^{n} |a_{ij}| \right\}, \\ ||A||_\infty &= \max_{i=1,\ldots,n} \left\{ \sum_{j=1}^{n} |a_{ij}| \right\}, \\ ||A||_2 &= \left\{ \rho(A^* A) \right\}^{1/2}, \end{aligned} \tag{29}$$

где A^* – сопряженная матрица по отношению к матрице A размером $n \times n$, и $\rho(A)$ – спектральный радиус A:

$$\rho(A) = \max_{i=1,\ldots,n} |\lambda_i(A)|.$$

Это максимум из модулей собственных значений.

3.2 Общие итерационные методы

Напомним, что мы пытаемся решить общее уравнение $Ax = f$. Мы догадываемся, что $x^{(n)}$ есть некоторое приближение для точного

решения $\hat{x}$. Насколько они близки? Один из способов оценить это приближение состоит в том, чтобы определить невязку уравнения

$$r^{(n)} = f - Ax^{(n)}. \tag{30}$$

Если невязка равна нулю, то мы нашли решение. В противном случае мы продолжаем идти далее, уменьшая $||r^{(n)}||$. Эти идеи используются для наиболее распространенных итерационных методов. Подведем итог тому, что происходит в итеративной части процесса решения в этих случаях. Определим шаг

$$\Delta^{(n)} = x^{(n+1)} - x^{(n)},$$

где

$$M\Delta^{(n)} = r^{(n)}.$$

Единственная проблема теперь состоит в том, чтобы определить матрицу M.

Существует несколько наиболее простых методов решения этой проблемы:

1. В методе *Якоби* выбирается диагональ A таким образом:

$$\begin{cases} m_{ij} = 0 & i \neq j, \\ m_{ii} = a_{ii} & i = j. \end{cases} \tag{31}$$

2. В методе *Зейделя* выбирается верхняя треугольная часть A:

$$\begin{cases} m_{ij} = 0 & i > j, \\ m_{ij} = a_{ij} & i \leq j. \end{cases} \tag{32}$$

3. Метод SOR аналогичен методу Зейделя, но включает параметр ω, который улучшает сходимость (который должен быть определен из A):

$$\begin{cases} m_{ij} = 0 & i > j, \\ m_{ij} = a_{ij} & i < j, \\ m_{ii} = a_{ii}/\omega & i = j. \end{cases} \tag{33}$$

SOR означает **S**uccessive **O**ver **R**elaxation (метод верхней релаксации).

Рассмотрим более подробно вывод этих методов. Представим матрицу $A = ||a_{ij}||$ $(i, j = 1, ..., m,)$ в следующем виде

$$A = A_1 + D + A_2, \tag{34}$$

где A_1 есть нижняя треугольная матрица с нулями на диагонали и выше, D - диагональная матрица с той же диагональю, что и матрица A, A_2 есть верхняя треугольная матрица с нулями на диагонали и ниже. Очевидно, что

$$x = -D^{-1}A_1x - D^{-1}A_2x + D^{-1}f. \tag{35}$$

Тогда, учитывая это представление, метод Якоби можно записать в виде следующей итерационной формы:

$$x^{(n+1)} = -D^{-1}A_1x^{(n)} - D^{-1}A_2x^{(n)} + D^{-1}f, \tag{36}$$

$$Dx^{(n+1)} + (A_1 + A_2)x^{(n)} = f, \tag{37}$$

$$D(x^{(n+1)} - x^{(n)}) + Ax^{(n)} = f. \tag{38}$$

Метод Гаусса-Зейделя можно представить как

$$x^{(n+1)} = -D^{-1}A_1x^{(n+1)} - D^{-1}A_2x^{(n)} + D^{-1}f, \tag{39}$$

$$(D + A_1)x^{(n+1)} + A_2x^{(n)} = f, \tag{40}$$

$$(D + A_1)(x^{(n+1)} - x^{(n)}) + Ax^{(n)} = f. \tag{41}$$

Формулу (39) следует выписать подробно:

$$x_i^{(n+1)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(n+1)} - \sum_{j=i+1}^{m} \frac{a_{ij}}{a_{ii}} x_j^{(n)} + \frac{f_i}{a_{ii}}, \qquad i = 1, 2, ..., m. \tag{42}$$

Представим этот результат для $i = 1, 2$:

$$x_1^{(n+1)} = -\sum_{j=2}^{m} \frac{a_{ij}}{a_{11}} x_j^{(n)} + \frac{f_1}{a_{11}}, \tag{43}$$

$$x_2^{(n+1)} = -\frac{a_{21}}{a_{22}} x_1^{(n+1)} - \sum_{j=3}^{m} \frac{a_{2j}}{a_{22}} x_j^{(n)} + \frac{f_2}{a_{22}}. \quad (44)$$

Из этих соотношений следует, что итерационный процесс метода Зейделя явный, так как из первой формулы мы найдем значение $x_1^{(n+1)}$, а затем из второй формулы найдем значение $x_2^{(n+1)}$, подставив $x_1^{(n+1)}$ в правую часть, и так далее.

В более общем случае оба метода можно записать как

$$D(\frac{x^{(n+1)} - x^{(n)}}{\tau_n}) + Ax^{(n)} = f, \quad (45)$$

$$(D + A_1)(\frac{x^{(n+1)} - x^{(n)}}{\tau_n}) + Ax^{(n)} = f, \quad (46)$$

где $\tau_n > 0$ - параметр, который выбирается чтобы ускорить сходимость итерационного процесса. На основании вида явных итерационных процессов (45, 46) выписывается каноническая форма в виде

$$B_n(\frac{x^{(n+1)} - x^{(n)}}{\tau_n}) + Ax^{(n)} = f, \quad n = 0, 1, ..., N, \quad (47)$$

где матрица B_n размером $m \times m$ определяется некоторым образом. Если $B_n = B, \quad \tau_n = \tau$, то итерационный процесс называется стационарным.

Метод SOR (метод верхней релаксации) получается из итерационной схемы метода Гаусса-Зейделя введением дополнительного параметра ω:

$$(D + \omega A_1)\frac{x^{(n+1)} - x^{(n)}}{\omega} + Ax^{(n)} = f, \quad (48)$$

$$(I + \omega D^{-1} A_1)x^{(n+1)} = ((1 - \omega)I - \omega D^{-1} A_2)x^{(n)} + \omega D^{-1} f, \quad (49)$$

где I есть единичная матрица $m \times m$. Покомпонентная запись метода (49) имеет вид

$$x_i^{(n+1)} + \omega \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(n+1)} = (1 - \omega)x_i^{(n)} - \omega \sum_{j=i+1}^{m} \frac{a_{ij}}{a_{ii}} x_j^{(n)} + \omega \frac{f_i}{a_{ii}}, \quad (50)$$

$i = 1, 2, ..., m$. Отсюда следует, что при использовании (50) в обращении матрицы $(I + \omega D^{-1} A_1)$ нет необходимости.

Приведем пример явного метода простой итерации

$$\frac{x^{(n+1)} - x^{(n)}}{\tau} + Ax^{(n)} = f, \quad n = 0, 1, ..., N. \tag{51}$$

3.3 Теоремы о сходимости итерационных процессов

Давайте исследовать вопрос о сходимости стационарного итерационного метода

$$B(\frac{x^{(n+1)} - x^{(n)}}{\tau}) + Ax^{(n)} = f, \quad n = 0, 1, ..., N. \tag{52}$$

Пусть $y^{(n)} = x^{(n)} - \hat{x}, \quad y^{(0)} = x^{(0)} - \hat{x}$. Тогда имеем

$$B(\frac{y^{(n+1)} - y^{(n)}}{\tau}) + Ay^{(n)} = 0, \quad n = 0, 1, ..., N. \tag{53}$$

Сходимость стационарного итерационного процесса означает, что

$$\lim_{n \to \infty} y^{(n)} = 0, \tag{54}$$

или

$$\lim_{n \to \infty} ||y^{(n)}|| = 0. \tag{55}$$

Итерационный процесс можно представить в удобном виде:

$$y^{(n+1)} = Sy^{(n)} = S^{n+1} y^{(0)}, \qquad S = I - \tau B^{-1} A. \tag{56}$$

Теорема. Итерационный процесс (52) сходится, если все собственные значения матрицы S по модулю меньше единицы:

$$|s_i| < 1, \quad i = 1, 2, ..., m, \quad (\rho(S) < 1), \qquad ||S|| < 1.$$

Доказательство следует из неравенства

$$||y^{(n)}|| \leq ||S||^n \cdot ||y^{(0)}||.$$

Приведем еще ряд строгих результатов без доказательства.

Теорема. Итерационный процесс (52) сходится, если A – симметричная, положительно определенная матрица и $B - 1/2\tau A > 0$.

Положительная пределенность матрицы означает $(Cx, x) \geq \delta||x||^2$.

Следствие: Итерационный процесс метода Якоби сходится, если A – симметричная, положительно-определенная матрица с диагональным преобладанием:

$$a_{ii} > \sum_{j \neq i} |a_{ij}|, \quad i = 1, 2, ..., m.$$

Следствие: Итерационный процесс метода SOR сходится, если A – симметричная, положительно-определенная матрица и $0 < \omega < 2$.

Следствие: Итерационный процесс метода простой итерации сходится, если A – симметричная, положительно-определенная матрица и $I - 1/2\tau A > 0$.

Таким образом, итерационный метод будет сходиться тогда и только тогда, когда $||S|| < 1$. В этом случае матричный оператор S называется оператором сжатия. Необходимым и достаточным условием для этого является то, что все собственные значения S имеют модуль меньше единицы. В методе SOR мы выбираем параметр ω таким образом, чтобы минимизировать наибольшее собственное значение матрицы S. Это обеспечит быструю сходимость.

3.4 Число обусловленности матрицы и оценки сходимости итераций

Пусть $\hat{x}$ является *точным* решение уравнения $Ax = f$. Определим невязку

$$r^{(n)} = f - Ax^{(n)} = A(\hat{x} - x^{(n)}). \tag{57}$$

Теперь, используя неравенство (26), сделаем следующие оценки:

$$||r^{(n)}|| \leq \|A\| \cdot \|\hat{x} - x^{(n)}\|, \tag{58}$$

$$||\hat{x} - x^{(n)}|| \leq \|A^{-1}\| \cdot \|r^{(n)}\|. \tag{59}$$

Поэтому будем иметь:

$$\frac{||r^{(n)}||}{||A||} \leq ||\hat{x} - x^{(n)}|| \leq \|A^{-1}\| \cdot \|r^{(n)}\|. \tag{60}$$

Аналогичным образом, используя систему уравнений $A\hat{x} = f$, получим:

$$\frac{||f||}{||A||} \leq ||\hat{x}|| \leq \|A^{-1}\| \cdot \|f\|. \tag{61}$$

Объединяя результаты (60-61), получаем неравенство, которое ограничивает относительную погрешность решения $x^{(n)}$:

$$\frac{1}{\|A\| \cdot \|A^{-1}\|} \frac{\|r^{(n)}\|}{\|f\|} \leq \frac{\|\hat{x} - x^{(n)}\|}{\|\hat{x}\|} \leq \|A\| \cdot \|A^{-1}\| \frac{\|r^{(n)}\|}{\|f\|}. \tag{62}$$

Таким образом, если выполняется неравенство

$$\frac{||r^{(n)}||}{||f||} \leq \epsilon, \tag{63}$$

то у нас есть оценка на то, на насколько близко расположено приближенное решение по отношению к точному. Вероятно, ограничение является завышенной оценкой. Как близко расположено приближенное решение по отношению к точному решению, существенно зависит от значения $||A||\ ||A^{-1}||$. Этот параметр называется числом обусловленности матрицы. Если оно очень велико, то, даже если $||r^{(n)}||$ мало, приближенное решение не будет близко расположено по отношению к точному. Проблема тогда становится некорректной. Величина числа обусловленности матрицы зависит от выбранной нормы. Если матрица A симметрична и мы воспользуемся нормой $||A||_2$, тогда

$$||A||\ ||A^{-1}|| = \frac{\max|\lambda_i(A)|}{\min|\lambda_i(A)|}. \tag{64}$$

Заметим, что если определитель очень мал и обратная матрица почти сингулярна, то число обусловленности матрицы – очень большое число. В таком случае, проблема вновь становится некорректной.

Вопросы для самоконтроля к главе 3

1. Что означает решить систему линейных алгебраических уравнений итерационным методом?

2. Какие виды норм матриц вам известны и как их вычислять?

3. Чем отличаются друг от друга итерационные методы Якоби, Зейделя и SOR решений систем линейных алгебраических уравнений?

4. Какое условие является критерием достижения заданной точности решений систем линейных алгебраических уравнений итерационными методами?

5. Сформулируйте примеры достаточных условий сходимости итерационных методов для решений систем линейных алгебраических уравнений итерационными методами?

6. Что означает число обусловленности матрицы и какой его смысл?

4 Интерполирование и аппроксимация функций

Задача интерполирования состоит в том, чтобы по значениям функции $f(x)$ в нескольких точках отрезка восстановить ее значения в остальных точках этого отрезка. Очевидно, что такая задача допускает сколь угодно много решений. Более детально можно сказать, что задача интерполирования возникает, например, в том случае, когда известны результаты измерения на эксперименте $y_k = f(x_k)$ $(x_k \in [a,b])$ некоторой физической величины $f(x)$ в точках $x_k,\ k = 0,1,...,n$, и требуется определить ее значения в любых других точках отрезка $x \in [a,b]$. Интерполирование используется также при сгущении сеток сложных специальных функций, когда вычисление значений $f(x)$ трудоемко. Иногда возникает необходимость приближенной замены, или аппроксимации данной функции другими функциями, которые легче вычислить. В частности, рассматривается задача о наилучшем приближении в нормированном пространстве H, когда заданную функцию $f(x) \in H$ требуется заменить линейной комбинацией $\varphi(x) \in H$ так, чтобы отклонение $||f - \varphi||$ было минимальным. Результаты и методы теории интерполирования и приближения функций нашли широкое применение в численном анализе, например, при выводе формул численного дифференцирования и интегрирования, при построении сеточных аналогов задач математической физики.

4.1 Интерполирование многочленами Лагранжа и Ньютона

Математически целесообразность задачи интерполирования и аппроксимации функций многочленами гарантируется известной в математическом анализе теоремой Вейерштрасса, которая утверждает, что для каждой непрерывной функции $f(x)$ на отрезке $[a,b]$ и любого, сколь угодно малого $\epsilon > 0$ найдется многочлен $P_n(x)$, такой что для всех $x \in [a,b]$ выполняется

$$|f(x) - P_n(x)| < \epsilon. \tag{65}$$

Пусть на отрезке $x \in [a, b]$ заданы точки $x_k,\ k = 0, 1, ..., n$ (узлы интерполирования), в которых нам известны значения функции $f(x)$. Задача интерполирования алгебраическим многочленом состоит в том, чтобы построить интерполяционный многочлен

$$L_n(x) = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n$$

такой, что $L_n(x_k) = f(x_k),\ k = 0, 1, ..., n$. Для любой непрерывной функции $f(x)$ эта задача имеет единственное решение, так как мы получем СЛАУ $(n + 1) \times (n + 1)$ для неизвестных $a_0, a_1, a_2, .., a_n$:

$$L_n(x_i) = a_0 + a_1 x_i + a_2 x_i^2 + ... + a_n x_i^n = f(x_i), \quad i = 0, 1, 2, ..., n.$$

Определитель этой системы (определитель Вандермонда) отличен от нуля.

Интерполяционная формула Лагранжа представляется в виде линейной комбинации

$$L_n(x) = \sum_{k=0}^{n} c_k(x) f(x_k). \tag{66}$$

Из условий интерполирования получаем

$$\sum_{k=0}^{n} c_k(x_i) f(x_k) = f(x_i), \quad i = 0, 1, 2, ..., n. \tag{67}$$

Следовательно, должны выполняться условия

$$c_k(x_i) = \begin{cases} 0, & i \neq k, \\ 1, & i = k. \end{cases} \tag{68}$$

Таким образом, коэффициенты $c_k(x)$ как многочлены степени n находятся по формулам

$$c_k(x) = \frac{\Pi_{j \neq k}(x - x_j)}{\Pi_{j \neq k}(x_k - x_j)}. \tag{69}$$

Часто интерполяционный многочлен Лагранжа записывается в виде

$$L_n(x) = \sum_{k=0}^{n} \frac{\omega(x)}{(x - x_k)\omega'(x_k)} f(x_k), \tag{70}$$

где

$$\omega(x) = (x - x_0)(x - x_1)...(x - x_{k-1})(x - x_k)(x - x_{k+1})...(x - x_n).$$

Теперь рассмотрим интерполяционный многочлен Ньютона. Введем в рассмотрение разделенные разности первого порядка:

$$f(x_i, x_j) = \frac{f(x_j) - f(x_i)}{x_j - x_i}, \qquad i \neq j, \qquad i, j = 0, 1, 2, ..., n. \tag{71}$$

Рассмотрим разделенные разности второго порядка:

$$f(x_0, x_1, x_2) = \frac{f(x_1, x_2) - f(x_0, x_1)}{x_2 - x_0},$$

$$f(x_1, x_2, x_3) = \frac{f(x_2, x_3) - f(x_1, x_2)}{x_3 - x_1}, ..., \tag{72}$$

$$f(x_{n-2}, x_{n-1}, x_n) = \frac{f(x_{n-1}, x_n) - f(x_{n-2}, x_{n-1})}{x_n - x_{n-2}}.$$

В общем случае для разделенной разности k-го порядка будем иметь

$$f(x_j, x_{j+1}, ..., x_{j+k}) = \sum_{i=j}^{j+k} \frac{f(x_i)}{\Pi_{m \neq i,\ m=j}^{j+k}(x_i - x_m)}. \tag{73}$$

В общем случае будем иметь

$$f(x_0, x_1, ..., x_n) = \\ = \sum_{i=0}^{n} \frac{f(x_i)}{(x_i - x_0)(x_i - x_1)...(x_i - x_{k-1})(x_i - x_{k+1})...(x_i - x_n)}. \tag{74}$$

Интерполяционным многочленом Ньютона называется многочлен

$$P_n(x) = f(x_0) + f(x_0, x_1)(x - x_0) + f(x_0, x_1, x_2)(x - x_0)(x - x_1) + ... \\ ... + f(x_0, x_1, ..., x_n)(x - x_0)(x - x_1)...(x - x_{n-1}). \tag{75}$$

Можно показать, что многочлен $P_n(x)$ тождественно равен интерполяционному многочлену Лагранжа $L_n(x)$.

Рассмотрим вопрос о погрешности интерполирования. Остаточный член интерполяционной формулы

$$r_n(x) = f(x) - L_n(x) \tag{76}$$

описывает погрешность интерполирования. В узлах интерполирования погрешность равна нулю. Оценим погрешность $r_n(x')$ в произвольной точке $x' \in [a, b]$. Рассмотрим вспомогательную функцию

$$g(x) = f(x) - L_n(x) - K\omega(x), \qquad K = const, \tag{77}$$

$$\omega(x) = (x - x_0)(x - x_1)...(x - x_n).$$

Выберем K так, чтобы $g(x') = 0$. Следовательно,

$$K = \frac{f(x') - L_n(x')}{\omega(x')}.$$

Пусть $f(x)$ – непрерывно дифференцируемая $n+1$ раз функция для $x \in [a, b]$. Функция $g(x)$ имеет не менее $n+2$ нулей для $[a, b]$. Значит, $g^{n+1}(x)$ имеет минимум один нуль для $[a, b]$, то есть, $g^{(n+1)}(\xi) = 0$, $\xi \in [a, b]$. Так как

$$g^{(n+1)}(x) = f^{(n+1)}(x) - K(n + 1)!,$$

то получаем, что

$$f^{(n+1)}(\xi) = K(n + 1)! = \frac{f(x') - L_n(x')}{\omega(x')}(n + 1)!.$$

Следовательно, погрешность интерполирования представляется как

$$f(x') - L_n(x') = \frac{f^{(n+1)}(\xi)}{(n + 1)!}\omega(x'). \tag{78}$$

А соответствующая оценка имеет вид

$$|f(x') - L_n(x')| \leq \frac{M_{n+1}}{(n + 1)!}|\omega(x')|, \tag{79}$$

где

$$M_{n+1} = \sup_{x \in [a,b]} |f^{(n+1)}(x)|.$$

Если $f(x)$ есть многочлен степени n, то интерполирование абсолютно точно.

Говорят, что интерполяционный процесс для функции $f(x)$ сходится в точке $x' \in [a,b]$, если существует предел

$$\lim_{n\to\infty} L_n(x') = f(x'). \tag{80}$$

Кроме поточечной сходимости, можно рассмотреть равномерную сходимость для $x \in [a,b]$

$$\lim_{n\to\infty} \sup_{x\in[a,b]} |f(x) - L_n(x)| = 0. \tag{81}$$

Свойство сходимости или расходимости интерполяционного процесса для функции $f(x)$ зависит как от выбора последовательности узловых сеток, так и от гладкости функции $f(x)$. Ограничимся формулировкой двух известных теорем. Первая утверждает, что какова бы не была последовательность узловых сеток, найдется непрерывная на $[a,b]$ функция $f(x)$ такая, что последовательность интерполяционных членов $L_n(x)$ не сходится равномерно к функции $f(x)$ на отрезке $[a,b]$. Вторая теорема говорит, что если функция $f(x)$ непрерывна на отрезке $[a,b]$, то найдется такая последовательность узловых сеток, для которой соответствующей интерполяционный процесс сходится равномерно при $x \in [a,b]$.

Величину $|\omega(x)|$, входящую в оценку погрешности, можно минимизировать за счет выбора узлов интерполирования. Эта задача решается с помощью многочленов Чебышева

$$T_{n+1}(x) = \frac{(b-a)^{n+1}}{2^{2n+1}} \cos((n+1)\arccos(\frac{2x-b-a}{b-a})). \tag{82}$$

Тогда узлы интерполирования являются нулями многочленов Чебышева:

$$x_k = \frac{a+b}{2} + \frac{a-b}{2}\cos\frac{(2k+1)\pi}{2(n+1)}, \qquad k = 0,1,2,...,n. \tag{83}$$

И тогда

$$\max_{x\in[a,b]} |\omega(x)| = \frac{(b-a)^{n+1}}{2^{2n+1}}, \tag{84}$$

а оценка погрешности примет вид

$$|f(x) - L_n(x)| \leq \frac{M_{n+1}}{(n+1)!}\frac{(b-a)^{n+1}}{2^{2n+1}}. \tag{85}$$

4.2 Интерполирование сплайнами

Интерполирование многочленом Лагранжа или Ньютона на всем отрезке $[a, b]$ с использованием большого числа узлов интерполяции часто приводит к плохому приближению, что объясняется сильным накоплением погрешностей в процессе вычислений. Кроме того, из-за расходимости процесса интерполяции увеличение числа узлов не всегда приводит к увеличению точности. Для того чтобы избежать больших погрешностей, весь отрезок $[a, b]$ разбивается на частичные отрезки и на каждом из частичных отрезков приближенно функция $f(x)$ заменяется многочленом невысокой степени (кусочно-полиномиальная интерполяция). Одним из способов интерполирования на всем отрезке является интерполирование с помощью сплайн-функции. Сплайн-функцией или сплайном называется кусочно-полиномиальная функция, определенная на отрезке $[a, b]$ и имеющая на нем некоторое число непрерывных производных. Преимуществом сплайнов перед обычной интерполяцией является их сходимость и устойчивость процесса вычислений. Рассмотрим распространенный в вычислительной практике случай, когда сплайн определяется с помощью многочлена третьей степени (кубический сплайн).

Пусть на отрезке $x \in [a, b]$ заданы точки $x_n,\ n = 0, 1, ..., N$ (узлы интерполирования), в которых нам известны значения функции $f(x)$. Сплайном, соответствующим данной функции $f(x)$ и узловой сетке $\{x_n\}_{n=0}^{N}$, называется функция $s(x)$, удовлетворяющая следующим условиям:

1. на каждом сегменте $[x_{n-1}, x_n]$, $n = 1, 2, ..., N$, функция $s(x)$ является многочленом третьей степени,

2. функция $s(x)$, а также ее первая и вторая производные непрерывны на $[a, b]$,

3. $s(x_n) = f(x_n)$ для $x_n,\ n = 0, 1, ..., N$.

Обозначим $f_n = f(x_n)$. На каждом отрезке $[x_{n-1}, x_n]$, $i =$

$1, 2, ..., N$, будем искать функцию $s(x) = s_n(x)$ в виде

$$s_n(x) = a_n + b_n(x - x_n) + \frac{c_n}{2}(x - x_n)^2 + \frac{d_n}{6}(x - x_n)^3, \tag{86}$$

$$x_{n-1} < x < x_n, \quad n = 1, 2, ..., N, \tag{87}$$

а коэффициенты

$$a_n = s_n(x_n), \quad b_n = s'_n(x_n), \quad c_n = s''_n(x_n), \quad d_n = s'''_n(x_n),$$

находятся аналитически. Так как $s(x_n) = f(x_n)$, то $a_n = f(x_n)$, $n = 1, 2, ..., N$. Доопределим $a_0 = f(x_0)$. Требование непрерывности $s(x)$ приводит к условиям

$$s_n(x_n) = s_{n+1}(x_n) = a_n, \qquad n = 1, 2, ..., N - 1.$$

Следовательно, мы получаем

$$a_n = a_{n+1} + b_{n+1}(x_n - x_{n+1}) + \frac{c_{n+1}}{2}(x_n - x_{n+1})^2 + \frac{d_{n+1}}{6}(x_n - x_{n+1})^3, \tag{88}$$

при $n = 0, 1, 2, ..., N - 1$. Обозначим $h_n = x_n - x_{n-1}$, тогда

$$h_n b_n - \frac{h_n^2}{2} c_n + \frac{h_n^3}{6} d_n = f_n - f_{n-1}, \tag{89}$$

для $n = 1, 2, ..., N$. Условия непрерывности первой производной $s'_n(x_n) = s'_{n+1}(x_n)$ приводит к уравнениям

$$c_n h_n - \frac{d_n}{2} h_n^2 = b_n - b_{n-1}, \qquad n = 2, ..., N. \tag{90}$$

А учитывая условия непрерывности второй производной $s''_n(x_n) = s''_{n+1}(x_n)$, получаем

$$d_n h_n = c_n - c_{n-1}, \qquad n = 2, ..., N. \tag{91}$$

Объединяя (89-91), получим систему $3N - 2$ уравнений относительно $3N$ неизвестных

$$\{b_n,\ c_n,\ d_n\}_{n=1}^{N}.$$

Недостающие два уравнения получим из условий $s''(a) = s''(b) = 0$, или $s_1''(x_0) = 0, \quad s_N''(x_n) = 0$. Таким образом, $c_1 - d_1 h_1 = 0$ и $c_N = 0$. Условие $c_1 - d_1 h_1 = 0$ совпадает с (91) при $n = 1$, и если считать $c_0 = 0$. Таким образом, мы получаем СЛАУ для всех коэффициентов кубического сплайна:

$$d_n h_n = c_n - c_{n-1}, \quad c_0 = 0, \quad c_N = 0, \quad n = 1, 2, ..., N, \tag{92}$$

$$c_n h_n - \frac{d_n}{2} h_n^2 = b_n - b_{n-1}, \quad n = 2, ..., N. \tag{93}$$

$$h_n b_n - \frac{h_n^2}{2} c_n + \frac{h_n^3}{6} d_n = f_n - f_{n-1}, \quad n = 1, 2, ..., N. \tag{94}$$

Исключая из системы b_n и d_n, получим уравнение для c_n

$$h_n c_{n-1} + 2(h_n + h_{n+1}) c_n + h_{n+1} c_{n+1} = 6\left(\frac{f_n - f_{n-1}}{h_n} - \frac{f_{n-1} - f_{n-2}}{h_{n-1}}\right), \quad n = 1, 2, ..., N-1, \quad c_0 = 0, \quad c_N = 0. \tag{95}$$

По найденным значениям c_n, получим b_n и d_n:

$$d_n = \frac{c_n - c_{n-1}}{h_n}, \qquad b_n = \frac{h_n}{2} c_n - \frac{h_n^2}{6} d_n + \frac{f_n - f_{n-1}}{h_n}, \qquad n = 1, 2, ..., N. \tag{96}$$

В отношении вопросов сходимости процесса интерполяции кубическими сплайнами приведем без доказательств следующую теорему.

Теорема. Пусть $s_h(x)$ - кубический сплайн, построенный для функции $f(x)$ класса $f \in C^{(4)}[a, b]$. Пусть $M_4 = ||f^{(4)}(x)||_{C[a,b]}$. Тогда справедливы оценки (см. [2])

$$||f(x) - s_h(x)||_{C[a,b]} \leq M_4 h^4, \tag{97}$$

$$||f'(x) - s_h'(x)||_{C[a,b]} \leq M_4 h^3, \tag{98}$$

$$||f''(x) - s_h''(x)||_{C[a,b]} \leq M_4 h^2. \tag{99}$$

Из этих оценок следует, что при $h \to 0$ ($N \to \infty$) последовательности $s_h(x)$, $s_h'(x)$ $s_h''(x)$ сходятся соответственно к функциям $f(x)$, $f'(x)$, $f''(x)$ на $[a, b]$.

4.3 Другие постановки задач интерполирования и приближения функций

Рассмотрим приближение рациональными функциями. Пусть значения функции $y = f(x)$ заданы в точках $x_i \in [a, b],\ i = 0, 1, ..., n$ $(y_i = f(x_i) = f_i)$. Требуется построить новую функцию

$$\varphi_{km}(x) = \frac{a_k x^k + a_{k-1} x^{k-1} + ... + a_0}{x^m + b_{m-1} x^{m-1} + ... + b_0}, \tag{100}$$

для которой

$$\varphi_{km}(x_i) = f(x_i), \qquad i = 0, 1, ..., n. \tag{101}$$

Получаем $n + 1$ уравнениий для $k + m + 1$ неизвестных. Пусть $n = k + m$. Тогда придем к СЛАУ относительно неизвестных $a_i,\ i = 0, 1, ..., k,\ b_i,\ i = 0, 1, ..., m - 1$:

$$\sum_{i=0}^{k} a_i x_j^i - f_j \sum_{i=0}^{m-1} b_i x_j^i = f_j x_j^m, \qquad j = 0, 1, 2, ..., k + m. \tag{102}$$

Рассмотрим пример дробно-линейной интерполяции. Пусть значения функции $f(x)$ заданы в трех узлах $x_{i-1} < x_i < x_{i+1}$. Требуется построить новую функцию

$$\varphi(x) = \frac{a_1 x + a_0}{x + b_0}, \qquad \varphi(x_j) = f(x_j), \quad j = i - 1, i, i + 1. \tag{103}$$

В этом случае получим СЛАУ трех уравнений относительно трех неизвестных $a_0,\ a_,\ b_0$

$$\begin{gathered} a_0 + a_1 x_{i-1} - b_0 f_{i-1} = x_{i-1} f_{i-1}, \\ a_0 + a_1 x_i - b_0 f_i = x_i f_i, \\ a_0 + a_1 x_{i+1} - b_0 f_{i+1} = x_{i+1} f_{i+1}. \end{gathered} \tag{104}$$

Используя приближение с помощью рациональных функций, необходимо следить, чтобы на отрезке интерполирования знаменатель выражения (100) не обращался в нуль. Другой проблемой является случай неудачного выбора узлов интерполирования, при котором выражение (100) вырождается в константу.

Рассмотрим пример двумерной интерполяции. На плоскости xOy заданы три точки $A_i(x_i, y_i)$, $i = 1, 2, 3$, не лежащие на одной прямой. Требуется, используя значения $u_i = u(x_i, y_i)$ функции $u(x, y)$ в этих точках, построить аппроксимацию производных $u_x,\ u_y$. Для решения этой задачи воспользуемся линейной интерполяцией

$$u(x, y) = a(x - x_1) + b(y - y_1) + c. \tag{105}$$

Очевидно, что $c = u(x_1, y_1)$ и $u_x = a,\ u_y = b$. Далее, получаем систему

$$a(x_2 - x_1) + b(y_2 - y_1) = u_2 - u_1,$$
$$a(x_3 - x_1) + b(y_3 - y_1) = u_3 - u_1,$$

и

$$a = \frac{\Delta_1}{\Delta}, \quad b = \frac{\Delta_2}{\Delta}, \tag{106}$$

$$\Delta_1 = \det \begin{pmatrix} u_2 - u_1 & y_2 - y_1 \\ u_3 - u_1 & y_3 - y_1 \end{pmatrix}, \qquad \Delta_2 = \det \begin{pmatrix} x_2 - x_1 & u_2 - u_1 \\ x_3 - x_1 & u_3 - u_1 \end{pmatrix},$$

$$\Delta = \det \begin{pmatrix} x_2 - x_1 & y_2 - y_1 \\ x_3 - x_1 & y_3 - y_1 \end{pmatrix}.$$

Общая задача интерполяции для функции $y = f(x)$, заданной на интервале $[a, b]$, формулируется следующим образом. Пусть на отрезке $[a, b]$ задана система линейно-независимых функций

$$\varphi_0(x),\ \varphi_1(x), ..., \varphi_n(x), \tag{107}$$

и введена сетка узлов

$$a = x_0 < x_1 < \ ... \ < x_n = b. \tag{108}$$

Образуем линейную комбинацию

$$\varphi(x) = c_0\varphi_0(x) + c_1\varphi_1(x) + ... + c_n\varphi_n(x) \tag{109}$$

с коэффициентами $c_0,\ c_1,\ ..., c_n$. Задача интерполирования функции $f(x)$ системой функций (107) на сетке (108) состоит в нахождении

коэффициентов $c_0,\ c_1,\ ...,c_n$, так чтобы выполнялись условия

$$\begin{cases} c_0\varphi_0(x_0) + c_1\varphi_1(x_0) + ... + c_n\varphi_n(x_0) = f(x_0), \\ c_0\varphi_0(x_1) + c_1\varphi_1(x_1) + ... + c_n\varphi_n(x_1) = f(x_1), \\ ..... \\ c_0\varphi_0(x_n) + c_1\varphi_1(x_n) + ... + c_n\varphi_n(x_n) = f(x_n). \end{cases} \tag{110}$$

Для того чтобы эта система имела единственное решение, необходимо и достаточно,чтобы определитель матрицы

$$A = \det \begin{pmatrix} \varphi_0(x_0) & \varphi_1(x_0) & ... & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & ... & \varphi_n(x_1) \\ ..... & ..... & ... & ..... \\ \varphi_0(x_n) & \varphi_1(x_n) & ... & \varphi_n(x_n) \end{pmatrix}. \tag{111}$$

был отличен от нуля. Необходимо, чтобы условие $\det A \neq 0$ выполнялось при любом расположении узлов сетки (112). Система функций $\varphi_k(x),\ k = 0, 1, ..., n$, на отрезке $[a, b]$ называется системой Чебышева, если условие $\det A \neq 0$ выполняется при любом расположении не совпадающих между собой узлов сетки (108). Функция $\varphi(x)$ называется обобщенным интерполяционным многочленом по системе Чебышева.

Рассмотрим задачу о наилучшем приближении функции, заданной таблично. Введем сетку узлов

$$a = x_0 < x_1 < \ ... \ < x_m = b, \tag{112}$$

и в общем случае $n \neq m$. Образуем вектор невязки (вектор погрешностей) $r_k = \varphi(x_k) - f(x_k)$ для $k = 0, 1, ..., m$. Рассмотрим нормы среднеквадратичного и равномерного приближений:

$$||r||^2 = \sum_{k=0}^{m} r_k^2 = \sum_{k=0}^{m} (\varphi(x_k) - f(x_k))^2, \tag{113}$$

$$||r||_C = \max_{x_k} |r_k| = \max_{x_k} |\varphi(x_k) - f(x_k)|. \tag{114}$$

Задача о наилучшем приближении функции $f(x)$, заданной таблично, состоит в нахождении коэффициентов $c_0,\ c_1,\ ...,c_n$, минимизирующих норму вектора r. Для нормы (113) получаем задачу

о наилучшем среднеквадратичном приближении функции, а для нормы (114) – задачу о равномерном приближении функции. Если $n = m$, то для обеих норм задача о наилучшем приближении функции совпадает с решением задачи интерполирования.

Построим наилучшее среднеквадратичное приближение функции для случая $n = 1,\ m = 2$, когда заданы $f_i = f(x_i)$ и $i = 0, 1, 2$. Тогда получим

$$\varphi(x) = c_0 + c_1(x - x_1),\ \ r(x) = \varphi(x) - f(x),\ \ r = (r(x_0), r(x_1), r(x_2)), \tag{115}$$

$$||r||^2 = F(c_0, c_1) = (c_0 - c_1 h_0 - f_0)^2 + (c_0 - f_1)^2 + (c_0 + c_1 h_1 - f_2)^2, \tag{116}$$

$$h_0 = x_1 - x_0, \quad h_1 = x_2 - x_1, \tag{117}$$

где $F(c_0, c_1)$ есть квадратичный функционал по отношению (c_0, c_1). Минимум нормы $||r||$ находим из условия экстремума функционала

$$\frac{\partial F}{\partial c_0} = \frac{\partial F}{\partial c_1} = 0,$$

которые приводят к системе уравнений

$$3c_0 + (h_1 - h_0)c_1 = f_0 + f_1 + f_2, \tag{118}$$

$$c_0(h_1 - h_0) + c_1(h_0^2 + h_1^2) = h_1 f_2 - h_0 f_0. \tag{119}$$

Отсюда получаем

$$c_0 = \alpha_1 f_0 + (1 - \alpha_1 - \alpha_2) f_1 + \alpha_2 f_2, \tag{120}$$

$$c_1 = \beta \frac{f_2 - f_1}{h_1} + (1 - \beta)\frac{f_1 - f_0}{h_0}, \tag{121}$$

где

$$\alpha_1 = \frac{h_1(h_0 + h_1)}{2(h_0^2 + h_1^2 + h_0 h_1)}, \qquad \alpha_2 = \frac{h_0(h_0 + h_1)}{2(h_0^2 + h_1^2 + h_0 h_1)},$$

$$\beta = \frac{h_1(h_0 + 2h_1)}{2(h_0^2 + h_1^2 + h_0 h_1)}.$$

Для равномерной сетки $h_0 = h_1 = h$ можно получить оценку (см. [2])

$$||f - \varphi|| = \frac{h^2 |f''(\zeta)|}{\sqrt{6}}, \quad \zeta \in [x_0, x_2]. \tag{122}$$

4.4 Наилучшее приближение в гильбертовом пространстве

Рассмотрим задачу о наилучшем приближении в вещественном гильбертовом пространстве H со скалярным произведением $(f,g)_H$ и нормой $||f||_H$. Типичным примером гильбертова пространства является пространство $L_2(a,b)$ вещественных функций $f(x)$, интегрируемых с квадратом на отрезке $[a,b]$ со скалярным произведениеим и нормой, определяемым по формулам

$$(f,g)_H = \int_a^b f(x)g(x)dx, \qquad ||f||_H^2 = \int_a^b |f(x)|^2 dx. \tag{123}$$

Пусть задана конечная система линейно независимых элементов $\varphi_k(x) \in H$, $k = 0,1,2,...,n$. Задача о наилучшем приближении в вещественном гильбертовом пространстве H состоит в том, чтобы для заданного элемента $f(x) \in H$ найти обобщенный многочлен

$$\varphi(x) = c_0\varphi_0(x) + c_1\varphi_1(x) + ... + c_n\varphi_n(x), \tag{124}$$

для которого отклонение

$$||f-\varphi||_H = (f-\varphi, f-\varphi)_H^{1/2} \tag{125}$$

является минимальным среди всех многочленов этого вида.

Перепишем равенство (125) в виде

$$||f-\varphi||_H^2 = \sum_{k,m=0}^{n} c_k c_m(\varphi_k,\varphi_m)_H - 2\sum_{k=0}^{n} c_k(f,\varphi_k)_H + ||f||_H^2. \tag{126}$$

Обозначим новую матрицу A с элементами

$$A_{km} = (\varphi_k,\varphi_m)_H, \qquad k,m = 0,1,2,...,n.$$

Теперь равенство (126) можно записать в более компактном виде

$$||f-\varphi||_H^2 = (A\bar{c},\bar{c}) - 2(\bar{f},\bar{c}) + ||f||_H^2, \tag{127}$$

где

$$\begin{aligned} &\bar{c} = (c_0, c_1, ..., c_n), \\ &\bar{f} = ((f,\varphi_0)_H, (f,\varphi_1)_H, ..., (f,\varphi_n)_H), \\ &k = 0,1,2,...,n. \end{aligned} \tag{128}$$

Отсюда видно, что задача о наилучшем приближении в вещественном гильбертовом пространстве H сводится к минимизации квадратичного функционала

$$F(\bar{c}) = (A\bar{c}, \bar{c}) - 2(\bar{f}, \bar{c}) \tag{129}$$

в координатном пространстве $\mathbf{R}^{n+1}$. Следует заметить, что симметричная матрица A является положительно определенной, так как при $f = 0$ из (127) следует $||\varphi||_H^2 = (A\bar{c}, \bar{c}) \geq 0$, и $(A\bar{c}, \bar{c}) > 0$, если $\bar{c} \neq 0$.

Проблема минимизации квадратичного функционала $F(\bar{c})$ сводится к решению некоторой СЛАУ. Об этом утверждает следующая теорема.

Теорема. Пусть A - симметричная положительно определенная матрица и $\bar{f}$ - заданный вектор. Тогда квадратичный функционал (129) имеет единственную точку минимума $\bar{c}_0$, удовлетворяющую системе линейных уравнений

$$A\bar{c}_0 = \bar{f}. \tag{130}$$

Доказательство основано на возможности сведения функционала $F(\bar{c})$ к форме

$$F(\bar{c}) = (A(\bar{c} - \bar{c}_0), \bar{c} - \bar{c}_0) - (A\bar{c}_0, \bar{c}_0) = \|\bar{c} - \bar{c}_0\|_A^2 - \|\bar{c}_0\|_A^2,$$

где мы ввели новую норму $\|\bar{c}\|_A^2 = (A\bar{c}, \bar{c})$. Если $A\bar{c}_0 = \bar{f}$, то минимум функционала имеет место при $\bar{c} = \bar{c}_0$.

Проанализируем функционал с произвольным и фиксированным $\bar{c}$, записанным в виде

$$F(\bar{c}_0 + t\bar{c}) = (A\bar{c}_0, \bar{c}_0) + 2t(A\bar{c}_0, \bar{c}) + t^2(A\bar{c}, \bar{c}) -$$

$$-2(\bar{f}, \bar{c}_0) - 2t(\bar{f}, \bar{c}),$$

где $\bar{c}_0$ является минимумом $F(\bar{c})$ при $t = 0$. Приравнивая производную от $F(\bar{c}_0 + t\bar{c})$ по параметру t к нулю, получаем

$$(A\bar{c}_0 - \bar{f}, \bar{c}) = 0, \quad \forall \bar{c}. \tag{131}$$

Таким образом, мы имеем $A\bar{c}_0 = \bar{f}$. Для второй производной по t получаем

$$\frac{d^2F}{dt^2} = 2(A\bar{c}, \bar{c}) = 2\|\bar{c}\|_A^2 > 0$$

для $\bar{c} \neq 0$. Это связано с тем фактом, что функционал $F(\bar{c})$ имеет минимум при $\bar{c} = \bar{c}_0$.

Таким образом, решение задачи о наилучшем приближении в вещественном гильбертовом пространстве H для функции $f(x)$ сводится к решению СЛАУ (130).

Оценим теперь величину отклонения $\|f - \varphi\|_H$ в случае, когда φ есть наилучшее приближение в вещественном гильбертовом пространстве H для функции $f(x)$. В этом случае справедливо

$$(f - \varphi, \varphi)_H = 0. \tag{132}$$

То есть, погрешность ортогональна элементу наилучшего приближения. Этот результат вытекает из следующих рассуждений:

$$(A\bar{c}_0, \bar{c}_0) = \|\varphi\|_H^2 = (\bar{f}, \bar{c}_0) = \sum_{k=0}^{n} c_k(f, \varphi_k)_H = (f, \varphi)_H, \tag{133}$$

и, следовательно, получаем

$$(f, \varphi)_H = \|\varphi\|_H^2. \tag{134}$$

Далее, из соотношения

$$\|f - \varphi\|_H^2 = \|f\|_H^2 - 2(f, \varphi)_H + \|\varphi\|_H^2 \tag{135}$$

и (132), для элемента наилучшего приближения φ получается важная оценка

$$\|f - \varphi\|_H^2 = \|f\|_H^2 - \|\varphi\|_H^2. \tag{136}$$

Заметим, что в задаче наилучшего приближения в вещественном гильбертовом пространстве H для функции $f(x)$ часто используется ортонормированная (ортонормальная) система функций

$$(\varphi_k, \varphi_m)_H = \begin{cases} 0, & k \neq m \\ 1, & k = m. \end{cases} \tag{137}$$

Тогда система (130) решается явно:

$$c_k = (f, \varphi_k)_H, \qquad k = 0, 1, ..., n. \tag{138}$$

Числа c_k называются коэффициентами Фурье разложения элемента $f(x) \in H$ по ортонормированной системе $\varphi_k(x)$, а обобщенный многочлен

$$\varphi = \sum_{k=0}^{n} c_k \varphi_k(x) \tag{139}$$

называется многочленом Фурье. Погрешность приближения определяется формулой

$$\|f - \varphi\|_H^2 = \|f\|_H^2 - \sum_{k=0}^{n} |c_k|^2. \tag{140}$$

Примером ортонормальной системы в $L_2(-\pi, \pi)$ является тригонометрическая система

$$\varphi_k(x) = \frac{e^{ikx}}{\sqrt{2\pi}}, \qquad k = 0, \pm 1, \pm 2, ... \tag{141}$$

Конечная или счетная система функций $\{\varphi\}_k$ называется линейно независимой, если для любого набора чисел $\{c\}_k$, $\sum_k |c_k| \neq 0$, невозможно тождество $\sum_k c_k \varphi(x) \equiv 0$. Важно, что ортонормальная система функций $\{\varphi\}_k$ состоит из линейно независимых функций. Всякая линейно независимая система функций $\{\psi_k\}_k$ из гильбертова пространства $L_2(D)$, $D \in R^n$, преобразуется в ортонормальную систему функций $\{\varphi\}_k$ с помощью следующего процесса ортогонализации Шмидта:

$$\varphi_1 = \frac{\psi_1}{\|\psi_1\|}, \quad \varphi_2 = \frac{\psi_2 - (\psi_2, \varphi_1)\varphi_1}{\|\psi_2 - (\psi_2, \varphi_1)\varphi_1\|}, ...,$$

$$\varphi_k = \frac{\psi_k - (\psi_k, \varphi_{k-1})\varphi_{k-1} - ... - (\psi_k, \varphi_1)\varphi_1}{\|\psi_k - (\psi_k, \varphi_{k-1})\varphi_{k-1} - ... - (\psi_k, \varphi_1)\varphi_1\|}. \tag{142}$$

Здесь и ниже индекс H у скалярных произведений опущен. Если в пространстве $L_2(-1, 1)$ ортогонализовать по Шмидту систему степеней $1, x, x^2, ...$, то получится система ортонормированных полиномов Лежандра.

Пусть система функций $\{\varphi\}_k$, $k = 1, 2, ...$ ортонормальна в $L_2(D)$ и $f(x) \in L_2(D)$. Формальный ряд с коэффициентами Фурье (f, φ_k)

$$\sum_{k=1}^{\infty}(f, \varphi_k)\varphi_k(x) \tag{143}$$

называется рядом Фурье функции $f(x)$ по ортонормальной системе функций $\varphi_k(x)$.

Если система функций φ_k, $k = 1, 2, ...$, ортонормальна в $L_2(D)$, то для любой функции $f(x) \in L_2(D)$ и любых чисел $a_1, a_2, ..., a_N$, $N = 1, 2, ...$, справедливо равенство (см. [6])

$$\|f - \sum_{k=1}^{N} a_k\varphi_k\|^2 = \|f - \sum_{k=1}^{N}(f, \varphi_k)\varphi_k\|^2 + \sum_{k=1}^{N}|(f, \varphi_k) - a_k|^2. \tag{144}$$

Из него вытекает неравенство

$$\|f - \sum_{k=1}^{N}(f, \varphi_k)\varphi_k\|^2 \leq \|f - \sum_{k=1}^{N} a_k\varphi_k\|^2, \tag{145}$$

которое еще раз демонстрирует смысл наилучшего приближения, полученного с помощью многочленов Фурье. Далее, полагая $a_k = 0$, $k = 1, 2, ..., N$ в (144), получим

$$\|f - \sum_{k=1}^{N}(f, \varphi_k)\varphi_k\|^2 = \|f\|^2 - \sum_{k=1}^{N}|(f, \varphi_k)|^2. \tag{146}$$

Из этого равенства следует неравенство Бесселя

$$\sum_{k=1}^{\infty}|(f, \varphi_k)|^2 \leq \|f\|^2. \tag{147}$$

Из этого неравенства и из теоремы Рисса-Фишера о полноте в $L_2(D)$ вытекает, что в силу

$$\|\sum_{k=n+1}^{n+m}(f, \varphi_k)\varphi_k(x)\|^2 = \sum_{k=n+1}^{n+m}|(f, \varphi_k)|^2$$

ряд Фурье (143) всегда сходится в $L_2(D)$ к некоторой функции из $L_2(D)$, но не обязательно к $f(x)$. Для того, чтобы ряд Фурье всегда сходится в $L_2(D)$ к $f(x)$, необходимо и достаточно, чтобы было выполнено равенство Парсеваля-Стеклова (уравнение замкнутости)

$$\sum_{k=1}^{\infty} |(f, \varphi_k)|^2 = \|f\|^2. \tag{148}$$

Если для любой $f(x) \in L_2(D)$ ее ряд Фурье по системе функций $\varphi_k(x) \in L_2(D)$, $k = 1, 2, ...$, сходится к $f(x)$, то эта система называется полной (замкнутой) $L_2(D)$. Для того, чтобы ортонормальная система $\varphi_k(x) \in L_2(D)$, $k = 1, 2, ...$, была полна в $L_2(D)$, необходимо и достаточно, чтобы для любой функции $f(x) \in L_2(D)$ было выполнено равенство Парсеваля-Стеклова.

4.5 Тригонометрическая интерполяция и дискретное преобразование Фурье

Если $f(x)$, $x \in [0, L]$, - периодическая функция с периодом L, то в тригонометрической интерполяции аппроксимация строится с помощью тригонометрического многочлена

$$T_n(x) = a_0 + \sum_{k=1}^{n} (a_k \cos \frac{2\pi k x}{L} + b_k \sin \frac{2\pi k x}{L}), \tag{149}$$

коэффициенты которого находятся из системы уравнений

$$T_n(x_j) = f(x_j), \quad j = 1, 2, ..., (2n + 1).$$

Для равномерной сетки узлов будем иметь $x_j = (j - 1)L/(2n)$.

Применяя дискретное преобразование Фурье, считаем, что функция $f(x)$, $x \in [0, L]$, задана в дискретной системе точек $x_k = kL/N$, $k = 0, 1, ..., N - 1$. Система векторов

$$f_m = e^{i\frac{2\pi m}{L} x_k}, \qquad m, k = 0, 1, ..., N - 1, \tag{150}$$

образует ортонормированную систему относительно скалярного произведения, с учетом комплексного сопряжения,

$$(f,g)=\frac{1}{N}\sum_{k=0}^{N-1}f(x_k)\bar{g}(x_k),\quad (f_{m_1},f_{m_2})=\delta_{m_1m_2}. \tag{151}$$

Тригонометрический многочлен

$$T_N(x)=\sum_{m=0}^{N-1}c_m f_m(x)=\sum_{m=0}^{N-1}c_m e^{i\frac{2\pi m}{L}x}\approx f(x),\quad x\in[0,L], \tag{152}$$

$$c_m=\frac{1}{N}\sum_{k=0}^{N-1}f(x_k)e^{-i\frac{2\pi mk}{N}},$$

принимает в точках x_k значения $f(x_k)$, то есть

$$T_N(x_k)=\sum_{m=0}^{N-1}c_m f_m(x_k)=f(x_k).$$

Причем этот тригонометрический многочлен дает наилучшее приближение к $f(x)$ в смысле указанной выше метрики (скалярного произведения) по сравнению с тригонометрическими многочленами с другими коэффициентами. Для вычисления коэффициентов Фурье c_m следует использовать хорошо известный численный метод быстрого дискретного преобразования Фурье. При этом целесообразно использовать значения для $N=2^n$.

4.6 Аппроксимация функций полиномами Чебышева

Поставим задачу: среди всех многочленов степени n со старшим коэффициентом 1 найти такой многочлен $T_n(x)$, для которого величина

$$\max_{x\in[-1,1]}|T_n(x)| \tag{153}$$

является минимальной. Многочлен, обладающий этим свойством, наименее уклоняющимся от нуля на отрезке $x\in[-1,1]$, называется

многочленом Чебышева и описывается формулой

$$T_n(x) = 2^{1-n} \cos(n \arccos x). \tag{154}$$

Экстремумы $T_n(x)$ наблюдаются в точках $x_k = \cos(k\pi/n)$, $k = 0, 1, ..., n$, причем

$$T_n(x_k) = (-1)^k 2^{1-n}. \tag{155}$$

Для многочлена Чебышева $T_n(x)$ существует ровно n действительных корней $T_n(x_k)$:

$$x_k = \cos(\frac{(2k+1)\pi}{2n}), \qquad k = 0, 1, ..., n-1. \tag{156}$$

Для отрезка $x \in [a, b]$, будем иметь

$$T_n(x) = \frac{(b-a)^n}{2^{2n-1}} \cos(n \arccos \frac{2x-b-a}{b-a}), \quad \max_{x\in[a,b]} |T_n(x)| = \frac{(b-a)^n}{2^{2n-1}}. \tag{157}$$

В дальнейшем, в этой секции, мы будем пользоваться упрощенным выражением для многочленов Чебышева:

$$T_n(x) = \cos(n \arccos x), \quad |x| \le 1. \tag{158}$$

Рассмотрим разложение в ряд Фурье по ортогональным многочленам Чебышева для функции $f(x)$, $x \in [-1, 1]$,

$$f(x) = \sum_{n=0}^{\infty} a_n T_n(x), \quad a_n = \frac{2}{\pi} \int_{-1}^{1} f(x) T_n(x) \frac{dx}{\sqrt{1-x^2}}, \quad n \neq 0, \tag{159}$$

$$a_0 = \frac{1}{\pi} \int_{-1}^{1} f(x) T_0(x) \frac{dx}{\sqrt{1-x^2}}.$$

Это разложение следует рассматривать в гильбертовом пространстве со скалярным произведением с весом

$$(f, g) = \int_{-1}^{1} f(x) g(x) \frac{dx}{\sqrt{1-x^2}}. \tag{160}$$

Для многочленов Чебышева выполняется условие ортогональности:

$$\int_{-1}^{1} T_n(x)T_m(x)\frac{dx}{\sqrt{1-x^2}} = \begin{cases} 0, & n \neq m \\ \frac{\pi}{2}, & n = m \neq 0, \\ \pi, & n = m = 0. \end{cases} \tag{161}$$

Более того, для многочленов Чебышева справедлива рекуррентная формула

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n = 1, 2, ..., \quad T_0(x) = 1, \quad T_1(x) = x. \tag{162}$$

Для отрезка ряда Фурье обозначим

$$f(x) \sim \sum_{n=0}^{N} a_n T_n(x) \equiv S_N(x), \qquad |x| \leq 1. \tag{163}$$

Существует алгоритм удобного и быстрого вычисления суммы $S_N(x)$ и ее производной $S'_N(x) = \sum_{n=0}^{N} a_n T'_n(x)$:

$$B_{N+1} = B_{N+2} = 0, \qquad C_N = C_{N+1} = 0,$$

$$B_n = 2xB_{n+1} - B_{n+2} + a_n, \quad n = N, N-1, ..., 0, \tag{164}$$

$$C_n = 2xC_{n+1} - C_{n+2} - 2B_{n+1}, \quad n = N-1, N-2, ..., 0,$$

$$S_N(x) = B_0 - xB_1, \qquad S'_N(x) = -B_1 - C_0 + xC_1.$$

При использовании многочленов Чебышева только четных степеней или только нечетных степеней

$$S_N(x) = \sum_{n=0}^{N} b_n T_{2n}(x), \qquad |x| \leq 1, \tag{165}$$

$$S_N(x) = \sum_{n=0}^{N} c_n T_{2n+1}(x), \qquad |x| \leq 1, \tag{166}$$

алгоритм изменится незначительно. В первом случае:

$$B_{N+1} = B_{N+2} = 0, \qquad C_N = C_{N+1} = 0,$$

$$B_n = 2(2x^2-1)B_{n+1} - B_{n+2} + b_n, \quad n = N, N-1, ..., 0, \qquad (167)$$
$$C_n = 2(2x^2-1)C_{n+1} - C_{n+2} - 8xB_{n+1}, \quad n = N-1, N-2, ..., 0,$$
$$S_N(x) = B_0 - (2x^2-1)B_1, \qquad S'_N(x) = -4xB_1 - C_0 + (2x^2-1)C_1.$$

Во втором случае:

$$B_{N+1} = B_{N+2} = 0, \qquad C_N = C_{N+1} = 0,$$

$$B_n = 2(2x^2-1)B_{n+1} - B_{n+2} + c_n, \quad n = N, N-1, ..., 0, \qquad (168)$$
$$C_n = 2(2x^2-1)C_{n+1} - C_{n+2} - 8xB_{n+1}, \quad n = N-1, N-2, ..., 0,$$
$$S_N(x) = x(B_0 - B_1), \qquad S'_N(x) = B_0 - B_1 - x(C_0 - C_1).$$

Сравним аппроксимации бесконечно дифференцируемой функции $f(x)$ на отрезке $[-1, 1]$, полученные с помощью тригонометрического многочлена ряда Фурье

$$f(x) \sim S_N^{(1)}(x) = \frac{a_0^{(1)}}{2} + \sum_{n=1}^{N}(a_n^{(1)} \cos(\pi n x) + b_n^{(1)} \sin(\pi n x)), \quad |x| \le 1, \qquad (169)$$

$$a_n^{(1)} = \int_{-1}^{1} f(x) \cos(\pi n x) dx, \quad n = 0, 1, ..., N, \qquad (170)$$

$$b_n^{(1)} = \int_{-1}^{1} f(x) \sin(\pi n x) dx, \quad n = 1, ..., N, \qquad (171)$$

и аппроксимации чебышевскими многочленами

$$f(x) \sim S_N^{(2)}(x) = \sum_{n=0}^{N} a_n^{(2)} T_n(x), \quad |x| \le 1,$$

$$a_n^{(2)} = \int_{-1}^{1} f(x) T_n(x) \frac{dx}{\sqrt{1-x^2}}. \qquad (172)$$

Оценим поведение коэффициентов $a_n^{(1)}$ и $b_n^{(1)}$ при $n \to +\infty$. С помощью интегрирования по частям, можно показать, что коэффициенты (170), (171) убывают как

$$a_n^{(1)} = \frac{(-1)^n}{(\pi n)^2}(f'(1) - f'(-1)) + O(n^{-3}), \qquad (173)$$

$$b_n^{(1)} = -\frac{(-1)^n}{\pi n}(f(1) - f(-1)) + O(n^{-2}), \tag{174}$$

т.е. недостаточно быстро чтобы получить хорошую аппроксимацию с небольшим числом членов в тригонометрическом многочлене. Если $f(1) - f(-1) \neq 0$, то из-за медленного убывания $b_n^{(1)}$ тригонометрический ряд Фурье не сходится равномерно.

Сделаем замену переменной $x = \cos t$, $t \in [0, \pi]$, для аппроксимации чебышевскими многочленами:

$$f(\cos t) \sim S_N^{(2)}(\cos t) = \sum_{n=0}^{N} a_n^{(2)} \cos nt, \tag{175}$$

$$a_0^{(2)} = \frac{1}{\pi} \int_0^{\pi} f(\cos t) dt, \tag{176}$$

$$a_n^{(2)} = \frac{2}{\pi} \int_0^{\pi} f(\cos t) \cos nt dt. \tag{177}$$

С помощью интегрирования по частям в последнем интеграле легко убедиться, что коэффициенты Чебышева при $n \to +\infty$ убывают сверх степенным образом. Это очень удобно, чтобы получить достаточно точную аппроксимацию с небольшим числом членов в чебышевском многочлене по сравнению с обычным тригонометрическим многочленом. Поэтому, аппроксимации чебышевскими многочленами очень часто используются на практике вычислений сложных спецфункций. Мы потратим машинное время один только раз для вычисления чебышевских коэффициентов. Зато в дальнейшем вычисления значений функций по алгоритмам (164), (167), (168), будут осуществляться мгновенно. Заметим, что для вычисления чебышевских коэффициентов по формулам (176), (177), следует использовать алгоритм быстрого дискретного преобразования Фурье

$$a_0^{(2)} \sim \frac{1}{N} \sum_{m=0}^{N-1} f(\cos t_m), \qquad t_m = \frac{\pi m}{N}, \tag{178}$$

$$a_n^{(2)} \sim \frac{2}{N} \sum_{m=0}^{N-1} f(\cos t_m) \cos(nt_m). \tag{179}$$

(a) (b)

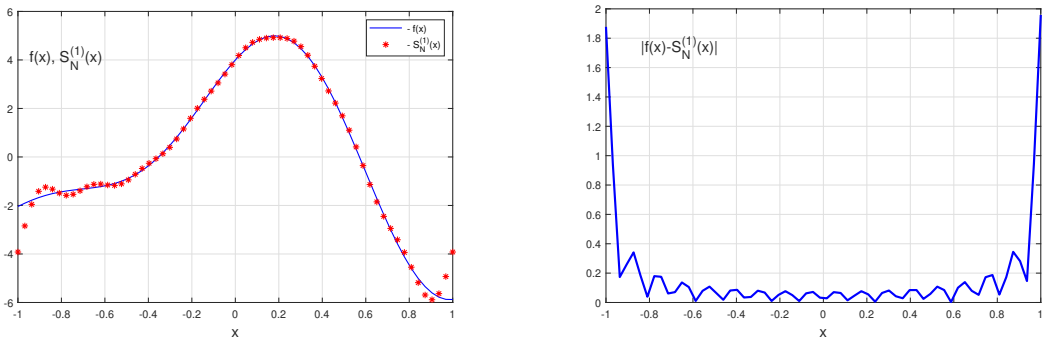


Рис. 1: Аппроксимация рядом Фурье: (a) - графики $f(x) = 4\cos(3x) + 2\sin(5x)$ и $S_N^{(1)}(x)$, (b) - $|f(x) - S_N^{(1)}(x)|$.

Рассмотрим пример аппроксимаций (169), (172) для функции $f(x) = 4\cos(3x) + 2\sin(5x)$, $x \in [-1, 1]$. На Рис. 1 для аппроксимации рядом Фурье для сравнения представлены графики $f(x)$ и $S_N^{(1)}(x)$ - слева, и $|f(x) - S_N^{(1)}(x)|$ - справа. Аналогично, на Рис. 2 для аппроксимации рядом полиномов Чебышева представлены графики $f(x)$ и $S_N^{(2)}(x)$ - слева, и $|f(x) - S_N^{(2)}(x)|$ - справа. При этом использовалось $N = 16$, а число узлов сетки полагалось равным 64. На Рис. 1 при $x = \pm 1$ отчетливо видны ошибки аппроксимации, так как функция $f(x)$ имеет разрыв на концах отрезка $x \in [-1, 1]$. Видно, что в целях достижения минимальной ошибки равномерного приближения функции $f(x)$, Рис. 2 демонстрирует преимущество в использовании аппроксимации рядом полиномов Чебышева.

Но очевидно, что разрывы функции $f(x)$ или ее производных внутри отрезка $|x| \leq 1$, а также быстрые осцилляции в поведении, приведут во всех случаях к более медленному убыванию любых коэффициентов Фурье. Следует привести пример аппроксимации хорошо известных в оптике и радиофизике интегралов Френеля, заимствованный из монографии [14]:

$$\int_0^x t^{-1/2} \cos t dt = x^{1/2} \sum_{n=0}^{N} a_n T_{2n}(\frac{x}{8}), \tag{180}$$

(a) (b)

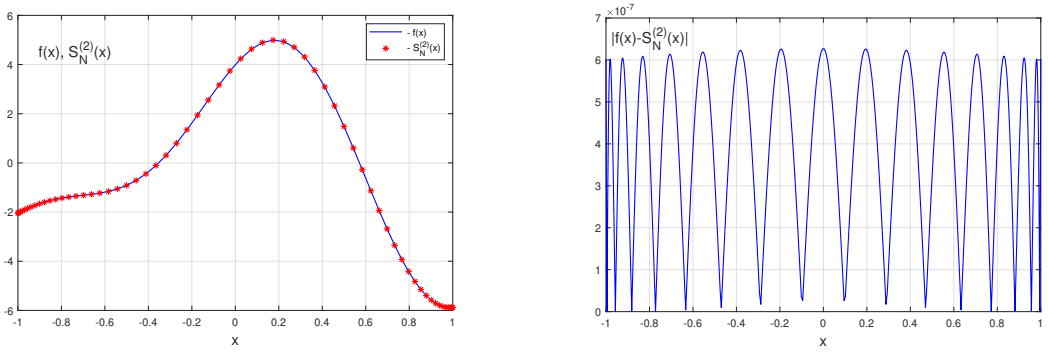


Рис. 2: Аппроксимация полиномами Чебышева: (a) - графики $f(x) = 4\cos(3x) + 2\sin(5x)$ и $S_N^{(2)}(x)$, (b) - $|f(x) - S_N^{(2)}(x)|$.

$$\int_0^x t^{-1/2} \sin t dt = x^{1/2} \sum_{n=0}^{N} b_n T_{2n+1}(\frac{x}{8}), \tag{181}$$

внутри отрезка $|x| \leq 8$. Коэффициенты Чебышева $a_n,\ b_n$, приведены в таблице

n	a_n	b_n
0	0.7643513866	0.6304140431
1	-0.4313554754	-0.4234451140
2	0.4328819997	0.3761717264
3	-0.2697331033	-0.1624948915
4	0.0841604532	0.0382225577
5	-0.0154652448	-0.0056456347
6	0.0018785542	0.0005745495
7	-0.0001626497	-0.0000428707
8	0.0000105739	0.0000024512
9	-0.0000005360	-0.0000001109
10	0.0000000218	0.0000000040
11	-0.0000000007	-0.0000000001

Вопросы для самоконтроля к главе 4

1. В чем заключаются достоинства и недостатки интерполяционных полиномов Лагранжа и Ньютона?

2. Почему интерполяция сплайнами оказалась более эффективной по сравнению с интерполяцией полиномами Лагранжа и Ньютона?

3. В чем заключается основной смысл наилучшего приближения функции в гильбертовом пространстве?

4. Как получить тригонометрическую интерполяцию функции?

5. Что называется гармоникой тригонометрического ряда Фурье?

6. В чем заключается основной смысл дискретного преобразования Фурье?

7. В чем преимущество интерполяции гладких функций полиномами Чебышева перед интерполяцией с помощью полиномов тригонометрического ряда Фурье?

5 Численные методы решения нелинейных уравнений и систем нелинейных уравнений

Пусть задана вещественно-значная функция $f(x)$. Требуется найти корни уравнения $f(x) = 0$, то есть нули функции $f(x)$. Уже на примере алгебраических многочленов известно,что нули функции $f(x)$ могут быть как вещественными, так и комплексными. Поэтому более точная постановка задачи состоит в нахождении корней уравнения $f(x) = 0$, расположенных в заданной области комплексной плоскости. Можно рассматривать также задачу нахождения вещественных корней, расположенных на заданном отрезке. Задача нахождения корней уравнения $f(x) = 0$ обычно решается в два этапа. На первом этапе изучается расположение корней, в общем случае на комплексной плоскости. Проводится их разделение, то есть выделяются области, содержащие только один корень. Кроме того изучается вопрос о кратности корней. Тем самым находятся некоторые начальные приближения для корней уравнения $f(x) = 0$. На втором этапе, используя заданные начальные приближения, строится итерационный процесс, позволяющий уточнить значения отыскиваемого корня. Не существует каких-то общих регулярных приемов решения задачи о расположении корней произвольной функции $f(x)$. Наиболее полно изучен вопрос о расположении корней алгебраических многочленов. Численные методы решений нелинейных уравнений являются, как правило, итерационными методами, которые предполагают задание достаточно близких к искомому решению начальных приближений.

Отметим два простых приема отделения действительных корней уравнения $f(x) = 0$. Предположим, что функция $f(x)$ определена и непрерывна на отрезке $[a, b]$. Первый прием состоит в том, что вычисляется таблица значений функции $f(x)$ в заданных точках $x_k \in [a, b]$, $k = 1, 2, ..., n$. Если обнаружится, что при некотором k числа $f(x_k)$ и $f(x_{k+1})$ имеют разные знаки, то это будет означать, что на интервале (x_k, x_{k+1}) функция $f(x)$ имеет по крайней мере один вещественный корень. Более точно можно сказать, что функция $f(x)$ имеет нечетное число корней на (x_k, x_{k+1}). Затем можно

разбить этот интервал на более мелкие интервалы и с помощью аналогичной процедуры уточнить расположение корня.

Более регулярным способом отделения вещественных корней является метод бисекции (метод деления пополам). Предположим, что на отрезке $[a, b]$ расположен лишь один корень x^* уравнения $f(x) = 0$. Тогда числа $f(a)$ и $f(b)$ имеют разные знаки. Пусть для определенности $f(a) > 0$, $f(b) < 0$. Положим $x_0 = (a+b)/2$ и вычислим $f(x_0)$. Если $f(x_0) < 0$, то искомый корень находится на интервале (a, x_0), если же $f(x_0) > 0$, то $x^* \in (x_0, b)$. Далее, из двух интервалов (a, x_0) и (x_0, b) выбираем тот, на границах которого функция $f(x)$ имеет различные знаки, находим точку x_1 - середину выбранного интервала, вычисляем $f(x_1)$ и повторяем указанный процесс. В результате получаем последовательность интервалов, содержащих искомый корень x^*, причем длина каждого последующего интервала вдвое меньше, чем предыдущего. Процесс заканчивается, когда длина вновь полученного интервала станет меньше заданного числа $\epsilon > 0$, и в качестве корня x^* приближенно принимается середина этого интервала.

Заметим, что если на отрезке $[a, b]$ имеется несколько корней, то указанный процесс сойдется к одному из корней, но заранее неизвестно, к какому именно. Можно использовать прием выделения корней: если корень $x = x^*$ кратности m найден, то рассматривается новая функция

$$g(x) = \frac{f(x)}{(x - x^*)^m},$$

и для нее повторяется процесс нахождения корня.

5.1 Метод простой итерации

Метод простой итерации состоит в том, что уравнение $f(x) = 0$ заменяется эквивалентным уравнением

$$x = s(x), \tag{182}$$

и итерации образуются следующим образом:

$$x_{n+1} = s(x_n), \quad n = 0, 1, 2, \ldots . \tag{183}$$

Начальное приближение x_0 задается. Для сходимости итерационного процесса большое значение имеет выбор функции $s(x)$. Эту функцию можно задавать различными способами. Часто она берется в виде

$$s(x) = x + \tau(x)f(x). \tag{184}$$

Функция $\tau(x)$ выбирается так, чтобы она не меняла знака на отрезке, где отыскивается корень. Ниже будет показано, что метод простой итерации сходится при правильном выборе начального приближения x_0 и если $|s'(x^*)| < 1$, где x^* есть корень уравнения. Заметим, что в форме метода простой итерации можно записать любой одношаговый итерационный метод. В частности, если $\tau(x) = \tau = const$, то получим метод релаксации

$$\frac{x_{n+1} - x_n}{\tau} = f(x_n), \qquad n = 0, 1, 2, ..., \tag{185}$$

для которого $s'(x) = 1 + \tau f'(x)$, и метод сходится при условии

$$-2 < \tau f'(x^*) < 0. \tag{186}$$

Если в некоторой окрестности корня выполняется условие

$$f'(x) < 0, \qquad 0 < m_1 < |f'(x)| < M_1, \tag{187}$$

то метод релаксации сходится при $\tau \in (0, 2/M_1)$.

5.2 Метод Ньютона

Пусть начальное приближение x_0 задано. Заменим функцию $f(x)$ отрезком ряда Тейлора

$$f(x) \sim f_1(x) = f(x_0) + f'(x_0)(x - x_0). \tag{188}$$

В качестве следующего приближения x_1 возьмем корень уравнения $f_1(x) = 0$. Получим

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{189}$$

Следующие приближения построим аналогичным способом. В результате мы получим метод Ньютона

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \qquad n = 1, 2, \dots . \tag{190}$$

Этот метод называют также методом касательных, так как новое приближение x_{n+1} является абциссой точки пересечения касательной, проведенной в точке $x_n, f(x_n)$ к графику функции $f(x)$, с осью OX. Метод Ньютона имеет квадратичную сходимость, то есть, в отличие от линейных задач, погрешность на следующей итерации пропорциональна квадрату погрешности на предыдущей итерации $|x_{n+1} - x^*| = O((x_n - x^*)^2)$. Такая быстрая сходимость метода Ньютона гарантируется лишь при очень хороших, близких к точному решению, начальных приближениях. Если начальное приближение выбрано неудачно, то метод может сходиться медленно или не сойтись вообще.

Итерационный метод секущих (метод хорд) получается из метода Ньютона следующим образом:

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n), \qquad n = 1, 2, \dots . \tag{191}$$

Этот метод является двухшаговым. В методе необходимо задавать $x_0,\ x_1$. Геометрическая интерпретация метода проста. Через точки $(x_{n-1}, f(x_{n-1}))$ и $(x_n, f(x_n))$ проводится прямая, абцисса точки пересечения этой прямой с осью OX и является новым приближением x_{n+1}.

Рассмотрим метод Ньютона в случае двух переменных. Требуется найти решение системы двух нелинейных уравнений

$$\begin{cases} f_1(x_1, x_2) = 0, \\ f_2(x_1, x_2) = 0. \end{cases} \tag{192}$$

Введем обозначения для этой системы: $f(x) = 0$, где $f(x) = (f_1(x), f_2(x))$, $x = (x_1, x_2)$. Пусть начальное приближение $x^{(0)} = (x_1^{(0)}, x_2^{(0)})$ задано. Заменим функцию $f(x)$ отрезком ряда Тейлора

$$f_1(x) \sim \bar{f}_1(x) = f_1(x^{(0)}) + f_{1x_1}(x^{(0)})(x_1 - x_1^{(0)}) + f_{1x_2}(x^{(0)})(x_2 - x_2^{(0)}),$$

$$f_2(x) \sim \bar{f}_2(x) = f_2(x^{(0)}) + f_{2x_1}(x^{(0)})(x_1 - x_1^{(0)}) + f_{2x_2}(x^{(0)})(x_2 - x_2^{(0)}).$$

В качестве следующего приближения $x^{(1)} = (x_1^{(1)}, x_2^{(1)})$ возьмем решение системы уравнений $\bar{f}_1(x) = 0$, $\bar{f}_2(x) = 0$. Получим линейную систему уравнений относительно неизвестных $x^{(1)} = (x_1^{(1)}, x_2^{(1)})$

$$\begin{cases} f_{1x_1}(x^{(0)})(x_1^{(1)} - x_1^{(0)}) + f_{1x_2}(x^{(0)})(x_2^{(1)} - x_2^{(0)}) = -f_1(x^{(0)}), \\ f_{2x_1}(x^{(0)})(x_1^{(1)} - x_1^{(0)}) + f_{2x_2}(x^{(0)})(x_2^{(1)} - x_2^{(0)}) = -f_2(x^{(0)}). \end{cases} \tag{193}$$

Решение этой системы дает вектор первой итерации, который представляется в виде

$$\begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \end{pmatrix} = \begin{pmatrix} x_1^{(0)} \\ x_2^{(0)} \end{pmatrix} - \begin{pmatrix} f_{1x_1}(x^{(0)}) & f_{1x_2}(x^{(0)}) \\ f_{2x_1}(x^{(0)}) & f_{2x_2}(x^{(0)}) \end{pmatrix}^{-1} \begin{pmatrix} f_1(x^{(0)}) \\ f_2(x^{(0)}) \end{pmatrix}. \tag{194}$$

Следующие приближения построим аналогичным способом. В результате мы получим метод Ньютона в двумерном случае для системы из двух нелинейных уравнений (192):

$$\begin{pmatrix} x_1^{(n+1)} \\ x_2^{(n+1)} \end{pmatrix} = \begin{pmatrix} x_1^{(n)} \\ x_2^{(n)} \end{pmatrix} - \begin{pmatrix} f_{1x_1}(x^{(n)}) & f_{1x_2}(x^{(n)}) \\ f_{2x_1}(x^{(n)}) & f_{2x_2}(x^{(n)}) \end{pmatrix}^{-1} \begin{pmatrix} f_1(x^{(n)}) \\ f_2(x^{(n)}) \end{pmatrix}, \tag{195}$$

$n = 1, 2, \ldots$. Более компактно метод Ньютона можно записать в виде

$$x^{(n+1)} = x^{(n)} - f'(x^{(n)})^{-1} f(x^{(n)}), \qquad n = 0, 1, \ldots . \tag{196}$$

Очевидно, что для сходимости метода необходимо, чтобы матрица $f'(x^{(n)})^{-1}$ существовала и ее норма $\|f'(x^{(n)})^{-1}\|$ была ограничена.

В общем случае m переменных требуется найти решение системы n нелинейных уравнений

$$\begin{cases} f_1(x_1, x_2, ..., x_m) = 0, \\ ..... \\ f_m(x_1, x_2 ..., x_m) = 0. \end{cases} \tag{197}$$

Если ввести обозначения для этой системы: $f(x) = 0$, где $f(x) = (f_1(x), f_2(x), ..., f_m(x))$, $x = (x_1, x_2, ..., x_m)$, то итерационный метод

Ньютона вновь записывается в виде (196), или более подробно

$$\begin{pmatrix} x_1^{(n+1)} \\ ..... \\ x_m^{(n+1)} \end{pmatrix} = \begin{pmatrix} x_1^{(n)} \\ ..... \\ x_m^{(n)} \end{pmatrix} - \begin{pmatrix} f_{1x_1}(x^{(n)}) & ..... & f_{1x_m}(x^{(n)}) \\ ..... & ..... & ..... \\ f_{mx_1}(x^{(n)}) & ..... & f_{mx_m}(x^{(n)}) \end{pmatrix}^{-1} \begin{pmatrix} f_1(x^{(n)}) \\ ..... \\ f_m(x^{(n)}) \end{pmatrix}, \tag{198}$$

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

5.3 Принцип сжимающих отображений

Пусть требуется решить систему нелинейных уравнений:

$$F(x) = 0, \quad F(x) = (f_1(x), f_2(x), ..., f_m(x)), \\ x = (x_1, x_2, ..., x_m) \in \mathbf{R}^m. \tag{199}$$

Здесь подразумевается нелинейное отображение $F : \mathbf{R}^m \to \mathbf{R}^m$. В более общем случае будем иметь $F : \mathbf{H} \to \mathbf{H}$, где $\mathbf{H}$ может быть $\mathbf{R}^m$ или некоторое функциональное нормированное пространство с функционалом нормы $\|x\|$, например, $H = C(D),\ D \subset \mathbf{R}^m$ (пространство непрерывных функций), или $H = L_2(D),\ D \subset \mathbf{R}^m$ (пространство квадратично-интегрируемых функций). В случае если $H = \mathbf{R}^m$, многие одношаговые итерационные методы для решения этой задачи можно записать в виде

$$B_n \frac{x^{(n+1)} - x^{(n)}}{\tau_n} + F(x^{(n)}), \qquad n = 0, 1, 2, ... , \tag{200}$$

где заданы: B_n - матрица $m \times m$, τ_n - числовой параметр ускорения сходимости. Для стационарных итерационных методов имеем $B_n = B$, $\tau_n = \tau$. Допустим, что систему $F(x) = 0$ можно переписать в виде

$$x = S(x), \tag{201}$$

а итерационный процесс представим как

$$x^{(n+1)} = S(x^{(n)}), \qquad n = 0, 1, ... \tag{202}$$

В случае если $H = \mathbf{R}^m$, получим, что $S(x) = x - \tau B^{-1} F(x)$.

Точка $x^* \in H$ называется неподвижной точкой оператора S, если $x^* = S(x^*)$, что эквивалентно решению $F(x) = 0$. Оператор S (линейный или нелинейный) является сжимающим оператором на некотором подпространстве $D \subset H$, если существует число $q \in (0,1)$ такое, что для любых $x_{1,2} \in D$ выполняется неравенство

$$\|S(x_1) - S(x_2)\| \leq q\|x_1 - x_2\|. \tag{203}$$

Сформулируем теорему, которая называется принципом сжимающих отображений и содержит условия сходимости метода простой итерации $x^{(n+1)} = S(x^{(n)})$ в линейном нормированном пространстве.

Теорема. Пусть оператор S определен на некотором множестве

$$\bar{U}_r(a) = \{x \in H : \|x - a\| \leq r\}, \tag{204}$$

и является сжимающим оператором на этом множестве с коэффициентом сжатия q, причем

$$\|S(a) - a\| \leq (1-q)r, \qquad 0 < q < 1. \tag{205}$$

Тогда в шаре $\bar{U}_r(a)$ оператор S имеет единственную неподвижную точку x^*, и итерационный метод $x^{(n+1)} = S(x^{(n)})$ сходится к x^* при любом $x^{(0)} \in \bar{U}_r(a)$. Для погрешности справедливы оценки:

$$\|x^{(n)} - x^*\| \leq q^n\|x^{(0)} - x^*\|, \tag{206}$$

$$\|x^{(n)} - x^*\| \leq \frac{q^n}{1-q}\|x^{(0)} - x^{(1)}\|. \tag{207}$$

Наметим основные шаги доказательства теоремы. Для простоты изложения предположим, что неподвижная точка x^* существует. Пусть $x^{(n)} \in \bar{U}_r(a)$, покажем что $x^{(n+1)} \in \bar{U}_r(a)$:

$$\|x^{(n+1)} - a\| \leq \|x^{(n+1)} - S(a)\| + \|S(a) - a\| \leq \|S(x^{(n)}) - S(a)\| +$$

$$+(1-q)r \leq q\|x^{(n)} - a\| + (1-q)r \leq r. \tag{208}$$

Рассмотрим следующие оценки:

$$\|x^{(n)} - x^*\| = \|S(x^{(n-1)}) - S(x^*)\| \leq q\|x^{(n-1)} - x^*\| =$$
$$= q\|S(x^{(n-2)}) - S(x^*)\| \leq\leq q^2\|x^{(n-2)} - x^*\| \quad ... \quad \leq q^n\|x^{(0)} - x^*\|. \tag{209}$$

Это доказывает сходимость итерационной последовательности. Далее,

$$\|x^{(0)} - x^{(n)}\| \leq \|x^{(0)} - x^{(1)}\| + \|x^{(1)} - x^{(2)}\| + ... + \|x^{(n-1)} - x^{(n)}\| \leq$$

$$\|x^{(0)} - x^{(1)}\|(1 + q + q^2 + ... + q^{n-1}). \tag{210}$$

Переходя к пределу при $n \to +\infty$, будем иметь

$$\|x^{(0)} - x^*\| \leq \|x^{(0)} - x^{(1)}\| \frac{1}{1-q}. \tag{211}$$

Таким образом, мы получаем, что

$$\|x^{(n)} - x^*\| \leq \|x^{(0)} - x^{(1)}\| \frac{q^n}{1-q}. \tag{212}$$

Последовательность $x^{(n)}$, $n = 0, 1, 2, ...$, является фундаментальной:

$$\|x^{(n)} - x^{(n+m)}\| \leq \|x^{(n)} - x^*\| + \|x^{(n+m)} - x^*\| \leq q^n \|x^{(0)} - x^*\|(1 + q^m) < \epsilon. \tag{213}$$

Но можно получить и другую оценку:

$$\|x^{(n)} - x^{(n+m)}\| \leq q^n \|x^{(0)} - x^{(m)}\| \leq \frac{q^n}{1-q} \|x^{(0)} - x^{(1)}\| < \epsilon. \tag{214}$$

Тогда из полноты линейного нормированного пространства H следует сходимость последовательности итерационного метода $x^{(n)}$, $n = 0, 1, 2, ...$ к неподвижной точке x^*.

Единственность следует из следующих оценок метода от противного, если предположить существование двух неповижных точек:

$$\|S(x_1^*) - S(x_2^*)\| = \|x_1^* - x_2^*\| \leq q\|x_1^* - x_2^*\|, \qquad 0 < q < 1. \tag{215}$$

Рассмотрим применение оценок этой теоремы к методу простой итерации в скалярном случае $x = s(x)$, если $s(x)$ является непрерывно-дифференцируемой функцией. Если $|s'(x)| \leq q < 1$ для любого $x \in \bar{U}_r(a)$, выполнено условие

$$|s(a) - a| \leq (1 - q)r, \tag{216}$$

и $x_0 \in \bar{U}_r(a)$, то итерационный метод $x_{n+1} = s(x_n)$ сходится и справедлива оценка

$$|x_n - x^*| \leq \frac{q^n}{1-q}|x_1 - x_0|. \tag{217}$$

Это утверждение следует из оценки сжимающего оператора

$$|x_{n+1} - x_{m+1}| = |s(x_n) - s(x_m)| = \\ = |s'(\bar{x})||x_n - x_m| \leq q|x_n - x_m|, \quad \bar{x} \in [x_n, x_m]. \tag{218}$$

В случае итерационного метода Ньютона

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{219}$$

для скалярного уравнения $f(x) = 0$ оценки сжимающего оператора получаются следующим образом:

$$|x_{n+1} - x_{m+1}| = |x_n - x_m - (\frac{f(x_n)}{f'(x_n)} - \frac{f(x_m)}{f'(x_m)})| =$$

$$= |x_n - x_m - (g(x_n) - g(x_m))| = |x_n - x_m||1 - g'(\bar{x})|, \qquad \bar{x} \in [x_n, x_m], \tag{220}$$

где

$$g(x) = \frac{f(x)}{f'(x)}, \qquad g'(x) = 1 - \frac{f(x)f''(x)}{f'(x)^2}.$$

Сжимающий оператор получим, если

$$|1 - g'(\bar{x})| \leq q < 1, \quad x \in \bar{U}_r(a), \tag{221}$$

или

$$|\frac{f(x)f''(x)}{f'(x)^2}| \leq q < 1, \quad x \in \bar{U}_r(a), \tag{222}$$

5.4 Сходимость метода Ньютона

В отличие от метода простой итерации, метод Ньютона обладает замечательной особенностью - он имеет квадратичную сходимость,

то есть он сходится, и погрешность $n+1$-й итерации пропорциональна квадрату погрешности на n-й итерации. Сформулируем соответствующую теорему. Пусть x^* - простой вещественный корень уравнения $f(x)=0$, и пусть $f'(x)\neq 0$ в окрестности

$$U_r(x^*)=\{x:|x-x^*|<r\}. \tag{223}$$

Предположим, что $f''(x)$ непрерывна в $U_r(x^*)$ и

$$0<m_1=\inf_{x\in U_r(x^*)}|f'(x)|, \qquad M_2=\sup_{x\in U_r(x^*)}|f''(x)|, \tag{224}$$

причем

$$\frac{M_2|x_0-x^*|}{2m_1}<1. \tag{225}$$

Тогда, если $x_0\in U_r(x^*)$, то метод Ньютона

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}, \quad n=0,1,2,\dots, \tag{226}$$

сходится, и для погрешности справедлива оценка

$$|x_n-x^*|<q^{2^n-1}|x_0-x^*|, \qquad q=\frac{M_2|x_0-x^*|}{2m_1}<1. \tag{227}$$

Рассмотрим доказательство теоремы. Так как

$$x_{n+1}-x^*=x_n-x^*-\frac{f(x_n)}{f'(x_n)}, \tag{228}$$

то

$$x_{n+1}-x^*=\frac{F(x)}{f'(x_n)}, \tag{229}$$

где

$$F(x)=(x-x^*)f'(x)-f(x), \qquad F(x^*)=0, \quad F'(x)=(x-x^*)f''(x). \tag{230}$$

Далее, воспользуемся тождеством

$$F(x_n)=F(x^*)+\int_{x^*}^{x_n}F'(t)dt=\int_{x^*}^{x_n}(t-x^*)f''(t)dt. \tag{231}$$

Воспользуемся формулой среднего значения

$$F(x_n) = f''(\xi)\int_{x^*}^{x_n}(t-x^*)dt = f''(\xi_n)\frac{(x_n-x^*)^2}{2}, \qquad \xi_n \in [x^*, x_n]. \tag{232}$$

Учитывая (229), получим

$$x_{n+1} - x^* = \frac{f''(\xi_n)(x_n-x^*)^2}{2f'(x_n)}, \tag{233}$$

то есть погрешность $n+1$-й итерации оказывается пропорциональной квадрату погрешности на n-й итерации. Далее очевидно, что

$$|x_1 - x^*| < \frac{M_2(x_0-x^*)^2}{2m_1} = q|x_0-x^*|, \tag{234}$$

что совпадает с оценкой (227) при $n=1$. Затем

$$|x_2 - x^*| = \frac{M_2(x_1-x^*)^2}{2m_1} < q^3|x_0-x^*|, \tag{235}$$

что совпадает с оценкой (227) при $n=2$. В конечном итоге, оценка (227) при любом n доказывается по индукции.

Приведем два важных дополнения. Пусть x^* является корнем кратности p функции $f(x)$:

$$f(x^*) = f'(x^*) = ... = f^{(p-1)}(x^*) = 0, \qquad f^{(p)}(x^*) \neq 0.$$

Пусть функции $f^{(p+1)}(x)$ непрерывна в окрестности $U_r(x^*)$. Тогда квадратичную сходимость имеет метод Ньютона с параметром p (см. [2])

$$x_{n+1} = x_n - p\frac{f(x_n)}{f'(x_n)}, \qquad n = 0, 1, 2, ... \ . \tag{236}$$

Пусть $z^* = x^* + iy^*$ является комплексным корнем кратности p аналитической функции $f(z)$ комплексного переменного z в области D ($z^* \in D$):

$$f(z^*) = f'(z^*) = ... = f^{(p-1)}(z^*) = 0, \qquad f^{(p)}(z^*) \neq 0.$$

Вновь квадратичную сходимость имеет метод Ньютона с параметром p (см. [2])

$$z_{n+1} = z_n - p\frac{f(z_n)}{f'(z_n)}, \qquad n = 0, 1, 2, ..., \qquad z_n \in D. \tag{237}$$

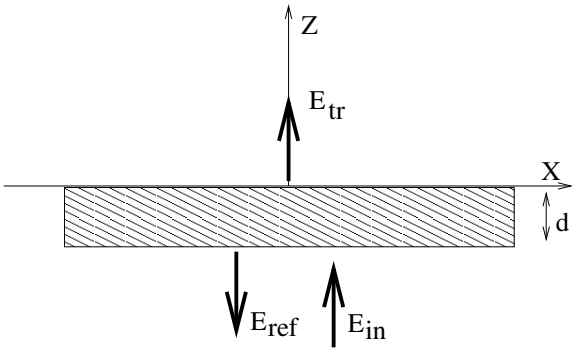


Рис. 3: Определение диэлектрической проницаемости слоя $\epsilon(\omega) = \epsilon_1(\omega) + i\epsilon_2(\omega)$ по данным рассеяния плоской электромагнитной волны.

5.5 Определение диэлектрической проницаемости слоя по данным рассеяния плоской электромагнитной волны.

Пусть через горизонтальный однородный слой диэектрика из диспергирующей среды с электродинамическими параметрами $\epsilon = \epsilon(\omega)$, $\mu = 1$ ($\omega = 2\pi f$ - частота) распространяется плоская электромагнитная волна (см. Рис. 3) вдоль оси Z снизу вверх. Пусть электрическая компонента поля направлена вдоль оси Y. Предполагается, что выше и ниже слоя находится ваккуум. Тогда, во временной области, ниже слоя $z < 0$, для электрической компоненты полного поля будем иметь сумму падающего и отраженного полей:

$$E(t, z) = E_{in}(t, z) + E_{ref}(t, z), \tag{238}$$

а выше слоя толщиной d для $z > 0$ присутствует только прошедшая волна

$$E(t, z) = E_{tr}(t, z). \tag{239}$$

Во временной области пусть падающая волна есть гауссов радиоимпульс с амплитудой E_0, распространяющийся вверх, а именно:

$$E_{in}(t, z) = E_0 f(t - \frac{z}{c}), \qquad f(t) = e^{-\frac{t^2}{4\tau_p^2} - i\omega_0 t}, \tag{240}$$

где c есть скорость света, τ_p - параметр, определяющий длительность импульса, $\omega_0 = 2\pi f_0$ - несущая частота. Спектр падающего

(a) (b)

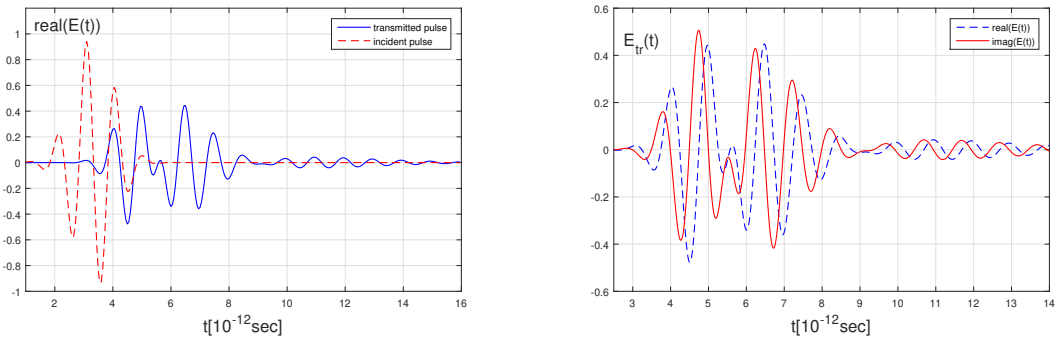


Рис. 4: Для выбранных физических параметров задачи во временной области представлены: (a) - вещественные части падающего и прошедшего радиоимпульсов, (b) - сравнение вещественной и мнимой частей прошедшего радиоимпульса.

радиоимпульса является гауссовой функцией

$$F(\omega) = \int_{-\infty}^{+\infty} e^{i\omega t} f(t) dt = 2\sqrt{\pi}\tau_p e^{-(\omega-\omega_0)^2\tau_p^2}. \tag{241}$$

Выше слоя для фиксированной точки наблюдения с координатой $z > 0$ для прошедшей волны будем иметь временной сигнал

$$E_{tr}(t, z) = E_{tr}(t - \frac{z}{c}) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-i\omega(t-\frac{z}{c})} F(\omega) T(\omega) d\omega, \tag{242}$$

где коэффициент прохождения $T(\omega)$ вычисляется по формуле (см. [15])

$$T(\omega) = \frac{4n}{(n+1)^2 e^{-ik_0 nd} - (n-1)^2 e^{ik_0 nd}}, \quad \sqrt{\epsilon} = n(\omega), \quad k_0 = \frac{\omega}{c}, \tag{243}$$

и $n = n_1(\omega) + in_2(\omega)$ есть индекс рефракции, а величина $n_2(\omega)$ называется коэффициентом экстинции или поглощения. Предположим, что диэлектрическая проницаемость слоя есть неизвестная

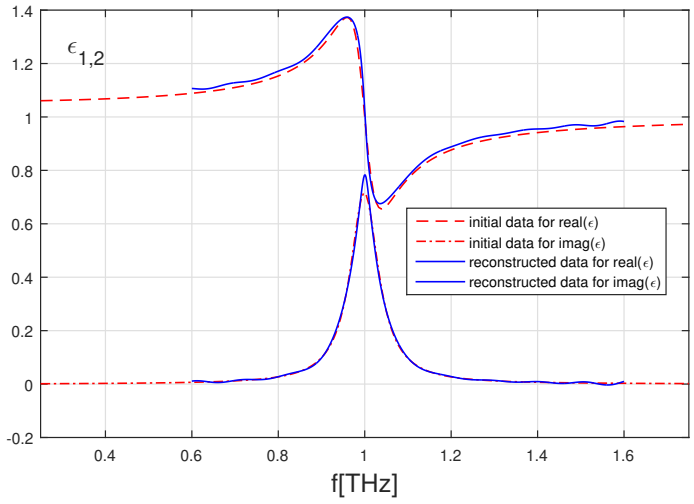


Рис. 5: Заданные и восстановленные вещественные и мнимые части диэлектрической проницаемости слоя $\epsilon(\omega) = \epsilon_1(\omega) + i\epsilon_2(\omega)$.

комплексная функция

$$\epsilon(\omega) = \epsilon_1(\omega) + i\epsilon_2(\omega),$$

подлежащая определению. Предположим, что нам известен временной сигнал $E_{tr}(t, z)$ прошедшей волны в точке наблюдения с координатой $z > 0$. Таким образом, мы приходим к решению обратной задачи определения параметров среды по данным рассеяния волн. Тогда мы в состоянии вычислить коэффициент прохождения $T(\omega)$ по формуле

$$T(\omega) = \frac{1}{F(\omega)} \int\limits_{-\infty}^{+\infty} e^{i\omega(t-\frac{z}{c})} E_{tr}(t, z)dt = \frac{1}{F(\omega)} \int\limits_{t_1}^{t_2} e^{i\omega(t-\frac{z}{c})} E_{tr}(t, z)dt \tag{244}$$

для некоторой полосы частот $\omega \in [\omega_{min}, \omega_{max}]$, учитывая, что прошедший импульс сосредоточен на отрезке $t \in [t_1, t_2]$. Следовательно, используя итерационную последовательность метода Ньютона

$$n^{(m+1)} = n^{(m)} - \frac{\Phi(n^{(m)})}{\Phi'(n^{(m)})}, \qquad m = 0, 1, 2, ..., \qquad n^{(0)} = 1, \tag{245}$$

для уравнения

$$\Phi(n) = \frac{4n}{(n+1)^2 e^{-ik_0 nd} - (n-1)^2 e^{ik_0 nd}} - T(\omega) = 0, \qquad (246)$$

для каждого значения $\omega \in [\omega_{min}, \omega_{max}]$ мы можем восстановить функцию $n(\omega)$, а значит, и $\epsilon(\omega) = n^2(\omega)$, то есть решить обратную задачу восстановления параметров среды. При этом мы не использовали информацию отраженного сигнала, которая могла бы быть использована для восстановления магнитной проницаемости среды $\mu(\omega) = \mu_1(\omega) + i\mu_2(\omega)$. В нашем случае для простоты мы предположили, что $\mu = 1$. Следует сказать, что для синтеза данных рассеяния (определение временного сигнала $E_{tr}(t, z)$ прошедшей волны в точке наблюдения с координатой $z > 0$) следует решить прямую задачу (242) с заданной функцией $\epsilon(\omega) = \epsilon_1(\omega) + i\epsilon_2(\omega)$.

Заметим, что описанная выше методика является строгой. Однако, если диэлектрический слой является толстым, то есть если в полосе $\omega \in [\omega_{min}, \omega_{max}]$ выполняется неравенство $k_0 d n_2(\omega) >> 1$, то формула (243) может быть упрощена:

$$T(\omega) \sim \frac{4ne^{ik_0 nd}}{(n+1)^2} \sim \frac{4n_1}{(n_1+1)^2} e^{ik_0 dn_1} e^{-k_0 dn_2}. \qquad (247)$$

Это приближение соответствует прямому прохождения падающей плоской волны через диэлектрический слой без каких-либо переотражений. Если в полосе $\omega \in [\omega_{min}, \omega_{max}]$ имеем $T(\omega) = A(\omega)e^{i\Phi(\omega)}$, $A(\omega) > 0$, то

$$\Phi(\omega) = k_0 d n_1(\omega), \qquad A(\omega) = \frac{4n_1}{(n_1+1)^2} e^{-k_0 dn_2}. \qquad (248)$$

Следовательно, минуя итерационный процесс метода Ньютона (246), мы с легкостью получаем

$$n_1(\omega) = \frac{\Phi(\omega)}{k_0 d}, \qquad n_2(\omega) = -\frac{1}{k_0 d} \log\left(A(\omega) \frac{(n_1+1)^2}{4n_1} \right). \qquad (249)$$

Следует заметить, что если радиоимпульс падающей волны задать в виде

$$f(t) = e^{-\frac{t^2}{4\tau_p^2}} cos(\omega_0 t), \qquad (250)$$

то прошедший сигнал $E_{tr}(t)$, вычисляемый по формуле (242), будет иметь только вещественные значения. Этот результат получается в следствие важноного свойства преобразования Фурье, а именно,

$$F(\omega)T(\omega) = F^*(-\omega)T^*(-\omega). \tag{251}$$

В результате формула (242) может быть записана в виде

$$E_{tr}(t,z) = E_{tr}(t-\frac{z}{c}) = \frac{1}{\pi}Re\int\limits_0^{+\infty} e^{-i\omega(t-\frac{z}{c})}F(\omega)T(\omega)d\omega. \tag{252}$$

В этой задаче мы будем использовать хорошо известную модель дисперсии диэлектрического слоя, состоящего из идентичных невзаимодействующих атомов (см. [16], глава 8):

$$\epsilon(\omega) = 1 + \omega_{pl}^2\sum_{j=1}^{J} F_j\left(\frac{\omega_j^2-\omega^2}{(\omega_j^2-\omega^2)^2+\omega^2\Gamma_j^2} + i\frac{\Gamma_j\omega}{(\omega_j^2-\omega^2)^2+\omega^2\Gamma_j^2}\right), \tag{253}$$

$$\hbar\omega_j = E_j - E_0, \quad j = 1,2,...,$$

где E_j суть энергетические уровни атома, ω_{pl} есть плазменная частота, F_j - силы осцилляторов, Γ_j - времена жизни энергетических уровней атомов. Рассмотрим упрощенную модель с одним слагаемым

$$\epsilon(\omega) = 1 + \omega_{pl}^2 F_1\left(\frac{\omega_1^2-\omega^2}{(\omega_1^2-\omega^2)^2+\omega^2\Gamma_1^2} + i\frac{\Gamma_1\omega}{(\omega_1^2-\omega^2)^2+\omega^2\Gamma_1^2}\right), \tag{254}$$

где мы выбрали следующие значения параметров

$$\omega_{pl} = 1.5THz, \quad \Gamma = 0.5THz, \quad \omega_1 = 2\pi\cdot 1THz, \quad F_1 = 1.$$

Далее, для падающего радиоимпульса и параметров z, d, мы использовали следующие значения

$$f_0 = 1THz, \quad \tau_p = 0.5\cdot 10^{-12}sec, \quad z = 10^{-3}m, \quad d = 0.5\cdot 10^{-3}m.$$

В результате численного моделирования для выбранных физических параметров задачи были получены результаты, представленные на Рис. 4: (a) – вещественные части падающего и прошедшего радиоимпульсов и (b) – сравнение вещественной и мнимой частей прошедшего радиоимпульса. Для сравнения на Рис. 5 изображены оригинальные (заданные) и восстановленные вещественные и мнимые части диэлектрической проницаемости слоя $\epsilon(\omega) = \epsilon_1(\omega) + i\epsilon_2(\omega)$. Видно, что восстановленные функции $\epsilon_{1,2}(\omega)$ мало отличаются от оригинальных.

Вопросы для самоконтроля к главе 5

1. Каковы постановка задачи и основные этапы решения систем нелинейных уравнений?

2. Что означает утверждение о квадратичной сходимости метода Ньютона?

3. Каковы достоинства и недостатки метода Ньютона?

4. Можно ли гарантировать единственность решения системы нелинейных уравнений?

5. Как зависит найденное решение от начального приближения?

6. Когда целесообразно остановить процесс вычисления решения системы нелинейных уравнений по методу Ньютона?

7. Сколько итераций потребуется для вычисления решения системы линейных уравнений по методу Ньютона с невырожденной матрицей Якоби, если пренебречь погрешностями вычислений?

8. В чем основной смысл принципа сжимающих отображений?

6 Методы оптимизации

6.1 Введение

Под оптимизацией понимают процесс выбора наилучшего варианта из всех возможных. С точки зрения инженерных-физических расчетов методы оптимизации позволяют выбрать наилучший вариант конструкции, наилучшее распределение ресурсов, и т. п. В процессе решения задачи оптимизации обычно необходимо найти оптимальные значения некоторых параметров, определяющих данную задачу. При решении инженерно-физических задач их принято называть проектными параметрами, а в экономических задачах их обычно называют параметрами плана. В качестве проектных параметров могут быть, в частности, значения линейных характеристик объекта, массы, температуры и т. п. Число n проектных параметров $x_1, x_2, ..., x_n$ характеризует размерность и степень сложности задач оптимизации.

Выбор оптимального решения или сравнение двух альтернативных решений проводится с помощью некоторой зависимой величины - функции, определяемой проектными параметрами. Эта величина называется целевой функцией или критерием качества. В процессе решения задачи оптимизации должны быть найдены такие значения проектных параметров, при которых целевая функция имеет минимум или максимум. Таким образом, целевая функция - это глобальный критерий оптимальности в математическмх моделях, с помощью которых описываются инженерно-физические и экономические задачи.

Целевую функцию можно записать в виде

$$u = f(x_1, x_2, ..., x_n). \tag{255}$$

Примерами целевой функции, встречающимися в инженерно - физических и экономических расчетах, могут служить прочность или масса конструкции, мощность установки, объем выпуска продукции, стоимость перевозок грузов, прибыль и т. п. В случае одного проектного параметра $n = 1$ целевая функция является функцией одной переменной, и ее график - некоторая кривая на плоскости. При $n = 2$ целевая функция является функцией двух переменных, и

ее графиком является поверхность. Следует отметить, что целевая функция не всегда может быть описана формулой. Иногда она может принимать только некоторые дискретные значения, задаваться в виде таблицы и т. п. Во все случаях она должна быть однозначной функцией проектных параметров.

Целевых функций может быть несколько. Например, при проектировании изделия машиностроения одновременно требуется обеспечить максимальную надежность, минимальную материалоемкость, максимальный полезный объем или грузоподъемность. Некоторые целевые функции могут оказаться несовместимыми. В таких случаях необходимо вводить приоритет той или иной целевой функции.

Можно выделить два типа задач оптимизации - безусловные и условные. Безусловная задача оптимизации состоит в отыскании максимума или минимума действительной функции от n действительных переменных и определении соответствующих значений аргументов на некотором множестве σ n-мерного пространства. Обычно рассматриваются задачи минимизации. К ним легко сводятся и задачи на поиск максимума путем замены знака целевой функции на противоположный.

Условные задачи оптимизации, или задачи с ограничениями, - это такие, при формулировке которых задаются некоторые условия (ограничения) на множестве σ. Эти ограничения задаются некоторой совокупностью некоторых функций, удовлетворяющих уравнениям или неравенствам.

В большинстве реальных задач оптимизации, представляющих практический интерес, целевая функция зависит от многих проектных параметров. Минимум дифференцируемой функции многих переменных

$$u = f(x_1, x_2, ..., x_n) \tag{256}$$

можно найти, исследуя ее значения в критических точках, которые определяются из решений системы уравнений

$$\frac{\partial u}{\partial x_1} = 0, \qquad \frac{\partial u}{\partial x_2} = 0, \quad ..... \quad \frac{\partial u}{\partial x_n} = 0. \tag{257}$$

Рассмотрим пример. Пусть требуется спроектировать контейнер в форме прямоугольного параллелепипеда объемом $V = 1\text{м}^3$, при-

чем необходимо израсходовать на его изготовление минимум материала. При постоянной толщине стенок последнее условие означает, что площадь полной поверхности контейнера S должна быть минимальной. Если обозначить через $x_{1,2,3}$ длины ребер контейнера, то задача сведется к минимизации функции

$$S(x_1, x_2, x_3) = 2(x_1x_2 + x_2x_3 + x_1x_3). \tag{258}$$

Эта функция в данном случае является целевой, а условие $V = 1$ - ограничением-равенством, которое позволяет исключить один параметр:

$$V = x_1x_2x_3 = 1, \qquad x_3 = \frac{1}{x_1x_2}, \tag{259}$$

$$S(x_1, x_2) = 2(x_1x_2 + \frac{1}{x_1} + \frac{1}{x_2}). \tag{260}$$

Задача свелась к минимизации функции двух переменных. В результате решения задачи будут найдены значения проектных параметров $x_{1,2}$, а затем и x_3. В приведенном примере фактически получилась задача безусловной оптимизации для целевой функции $S(x_1, x_2)$, поскольку ограничение-равенство было использовано для исключения параметра x_3.

В соответствии с условием экстремума (257) получим систему двух уравнений

$$\frac{\partial S}{\partial x_1} = 2(x_2 - \frac{1}{x_1^2}) = 0, \qquad \frac{\partial S}{\partial x_2} = 2(x_1 - \frac{1}{x_2^2}) = 0. \tag{261}$$

Отсюда находим $x_1 = x_2 = 1$м, $x_3 = (x_1x_2)^{-1} = 1$м. Для найденных значений $x_1 = x_2 = 1$ находим матрицу вторых производных

$$\frac{\partial^2 S}{\partial x_1^2} = \frac{4}{x_1^3} = 4, \qquad \frac{\partial^2 S}{\partial x_2^2} = \frac{4}{x_2^3} = 4, \qquad \frac{\partial^2 S}{\partial x_1 \partial x_2} = 2. \tag{262}$$

Собственные значения матрицы вторых производных суть 2 и 6, что означает минимум целевой функции $S(x_1, x_2)$ в точке $x_1 = x_2 = x_3 = 1$. Таким образом оптимальной формой контейнера является куб, длина ребра которого равна 1м.

Рассмотренный метод можно использовать лишь для дифференцируемой целевой функции. Но и в этом случае могут возникнуть

новые серьезные трудности при решении системы нелинейных уравнений условия экстремума (257). Во многих случаях никакой формулы для целевой функции не существует, а имеется лишь возможность определения ее значений в произвольных точках рассматриваемой области с помощью некоторого вычислительного алгоритма или путем физических измерений. Задача состоит в приближенном определении наименьшего значения функции во всей области при известных ее значениях в отдельных точках.

Для решения подобной задачи в области проектирования G, в которой ищется минимум целевой функции

$$u = f(x_1, x_2, ..., x_n), \tag{263}$$

можно ввести дискретное множество точек (узлов) путем разбиения интервалов изменения параметров $x_1, x_2, ..., x_n$, на части с шагами $h_1, h_2, ..., h_n$. В полученных узлах можно вычислить значения целевой функции, и среди этих значений найти наименьшее. Следует отметить, что такой метод может быть использован для функции одной переменной. Но в многомерных задачах оптимизации, где число проектных параметров достигает пяти и более, этот метод потребовал бы слишком большого объема вычислений.

6.2 Метод покоординатного спуска

Пусть требуется найти наименьшее значение целевой функции

$$u(M) = f(x_1, x_2, ..., x_n). \tag{264}$$

В качестве начального приближения выберем в n-мерном пространстве точку M_0 с координатами $(x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)})$. Зафиксируем все координаты функции $u(M)$, кроме первой. Тогда

$$u = f(x_1, x_2^{(0)}, ..., x_n^{(0)}) \tag{265}$$

есть функция одной переменной x_1. Решая одномерную задачу оптимизации для этой функции, мы от точки M_0 переходим к точке M_1 с координатами $(x_1^{(1)}, x_2^{(0)}, ..., x_n^{(0)})$, в которой функция u принимает наименьшее значение по координате x_1 при фиксированных

остальных координатах. Это первый шаг процесса оптимизации, состоящий в спуске по координате x_1.

Зафиксируем теперь все координаты, кроме x_2 и рассмотрим функцию

$$u = f(x_1^{(1)}, x_2, x_3^{(0)}, ..., x_n^{(0)}). \tag{266}$$

Снова решая одномерную задачу оптимизации для этой функции, находим ее наименьшее значение при $x_2 = x_2^{(1)}$, то есть в точке

$$M_2 = (x_1^{(1)}, x_2^{(1)}, x_3^{(0)}, ..., x_n^{(0)}).$$

Аналогично проводится спуск по координатам $x_3, x_4, ..., x_n$, а затем процедура снова повторяется от x_1 и до x_n, и т.д. В результате этого процесса получается последовательность точек $M_0, M_1, ...$, в которых значения целевой функции составляет монотонно убывающую последовательность

$$f(M_0) > f(M_1) > f(M_2) > .... .$$

На любом k-м шаге этот процесс можно прервать, и значение $f(M_k)$ принимается в качестве наименьшего значения целевой функции в рассматриваемой области. Таким образом, метод покоординатного спуска сводит задачу о нахождении наименьшего значения функции многих переменных к многократному решению одномерных задач оптимизации по каждому параметру. Данный метод легко проиллюстрировать геометрически для случая функции двух переменных $z = f(x, y)$, описывающую некоторую поверхность в трехмерном пространстве. Процесс оптимизации происходит следующим образом. Точка $M_0(x_0, y_0)$ описывает начальное приближение. Проводя спуск по координате x, попадаем в точку $M_1(x_1, y_0)$. Далее, двигаясь параллельно оси ординат, придем в точку $M_2(x_1, y_1)$, и т.д.

Важным здесь является вопрос о сходимости рассматриваемого процесса оптимизации. Другими словами, будет ли последовательность значений целевой функции $f(M_0), f(M_1), ...$, сходиться к наименьшему ее значению в данной области? Это зависит от вида самой функции и выбора начального приближения. Для функции двух переменных очевидно, что этот метод неприменим в случае наличия изломов в линиях уровня целевой функции. Это соответствует так называемому "оврагу"на поверхности. Здесь возможен

случай, когда спуск по одной координате приводит на "дно оврага". Тогда любое движение вдоль другой координаты ведет к возрастанию целевой функции, соответствующему подъему на "берег оврага". Поскольку поверхности типа оврага встречаются в инженерно-физической практике, то при использовании метода покоординатного спуска следует убедиться, что решаемая задача не имеет этого недостатка.

Для гладких функций при удачно выбранном начальном приближении, в некоторой окрестности минимума, процесс сходится к минимуму. К достоинствам метода покоординатного спуска следует отнести возможность использования простых алгоритмов одномерной оптимизации.

6.3 Метод градиентного спуска

При решении инженерно-физических задач мы нередко сталкиваемся с явлениями, сходными с решением проблемы на нахождение минимума. Мы заключаем, что направление наискорейшего спуска соответствует направлению наибольшего убывания функции. Известно, что направление наибольшего возрастания функции двух переменных $u = f(x, y)$ характеризуется ее градиентом (см. [23])

$$\mathrm{grad}u(M) = \frac{\partial u}{\partial x}(M)\bar{e}_1 + \frac{\partial u}{\partial y}(M)\bar{e}_2, \tag{267}$$

где $\bar{e}_{1,2}$ - единичные векторы (орты) в направлении координатных осей. Следовательно, направление, противоположное градиентному, укажет путь, ведущий вниз к минимуму целевой функции вдоль наиболее крутой линии. Методы, основанные на выборе пути оптимизации с помощью градиента целевой функции, называются градиентными.

Идея метода градиентного спуска состоит в следующем. Выбираем некоторую начальную точку и вычисляем в ней градиент рассматриваемой целевой функции. Делаем шаг в направлении, обратном градиентному. В результате приходим в точку, значение целевой функции в которой обычно меньше первоначального. Если это условие не выполнено, то есть значение целевой функции не изменилось либо даже возросло, то нужно уменьшить шаг. В новой

точке процедуру повторяем: вычисляем градиент целевой функции и снова делаем шаг в обратном к нему направлении. Процесс продолжается до получения наименьшего значения целевой функции. Момент окончания поиска наступит тогда, когда движение из полученной точки с любым шагом приводит к возрастанию значения целевой функции. Строго говоря, если минимум целевой функции достигается внутри рассматриваемой области, то в этой точке градиент равен нулю, что также может служить сигналом для окончания процесса оптимизации.

В описанном методе градиентного спуска требуется вычислять на каждом шаге процесса оптимизации градиент целевой функции

$$\text{grad}u = (\frac{\partial u}{\partial x_1}(M), \frac{\partial u}{\partial x_2}(M), ..., \frac{\partial u}{\partial x_n}(M)). \qquad (268)$$

Формулы для частных производных в явном виде можно получить лишь в случае, когда целевая функция задана аналитически. В противном случае эти частные производные вычисляются с помощью численного дифференцирования в приближении конечных разностей:

$$\frac{\partial u}{\partial x_i}(M) \equiv \frac{1}{\Delta x_i}[u(x_1, ..., x_i + \Delta x_i, ..., x_n) - u(x_1, ..., x_i, ..., x_n)], \quad (269)$$

где $i = 1, 2, ..., n$.

При использовании метода градиентного спуска в задачах оптимизации основной объем вычислений приходится обычно на вычисление градиента целевой функции в каждой точки траектории спуска. Поэтому целесообразно уменьшить количество таких точек без ущерба для самого решения. Это достигается в некоторых методах, являющихся модификациями метода градиентного спуска. Одним из них является метод наискорейшего спуска. Согласно этому методу, после определения в начальной точке направления спуска, противоположного градиенту целевой функции, в этом направлении делают не один шаг, а двигаются до тех пор, пока целевая функция убывает, достигая таким образом минимума в некоторой точке. В этой новой точке снова определяют направление спуска с помощью градиента и ищут новую точку минимума целевой функции, и т. д.

В этом методе спуск происходит гораздо более крупными шагами, и градиент целевой функции вычисляется в меньшем числе точек.

Заметим, что метод наискорейшего спуска сводит многомерную задачу оптимизации к последовательности одномерных задач на каждом щаге оптимизации, как и в случае покоординатного спуска. Разница состоит в том, что здесь направление одномерной оптимизации определяется градиентом целевой функции, тогда как в случае покоординатного спуска направление одномерной оптимизации проводится на каждом шаге вдоль одного из координатных направлений.

Вопросы для самоконтроля к главе 6

1. В чем заключаются основные задачи оптимизации?

2. В чем состоит основная идея метода покоординатного спуска для нахождения минимума целевой функции?

3. Как применяется метод градиентного спуска для нахождения минимума целевой функции?

7 Численное интегрирование

7.1 Введение

В математическом моделировании есть много обстоятельств, когда нам нужно оценить численно интеграл. Большинство геометрических и физических величин в математике, физике и других прикладных науках представляются в виде определенных, повторных и многократных интегралов. Известно, что многие задачи математической и теоретической физике, допускающие точные решения для обыкновенных дифференциальных уравнений и дифференциальных уравнений в частных производных, представляются в интегральной форме, которая требует численных оценок. Поэтому в этой секции мы приведем описание простейших и основных методов вычисления определенных интегралов.

7.2 Формула прямоугольников

Основная идея оценки определенного интеграла состоит в том, чтобы заменить интеграл на конечную сумму

$$\int_a^b f(x)dx \approx \sum_{i=0}^{n} A_i f(x_i). \tag{270}$$

Это приближенное равенство называется квадратурной формулой. Точки x_i называются узлами, а A_i – коэффициенты квадратурной формулы. Сначала мы выберем простейшую квадратурную формулу, которая называется прямоугольной формулой. Введем равномерную сетку для отрезка $[a, b]$ с шагом h

$$\Omega_h = \{x_i = a + ih, \quad i = 0, 1, ..., n, \quad h = (b - a)/n\}. \tag{271}$$

Ясно, что

$$\int_a^b f(x)dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)dx. \tag{272}$$

Заменим интеграл

$$\int_{x_{i-1}}^{x_i} f(x)dx$$

на $f(x_{i-1/2})h$, где $x_{i-1/2} = x_i - h/2$. Геометрически это означает, что мы заменяем площадь криволинейной трапеции площадью соответствующего прямоугольника (см. [23]). Итак, мы получаем формулу

$$\int_{x_{i-1}}^{x_i} f(x)dx \approx f(x_{i-1/2})h,$$

которая известна как формула прямоугольников для частичного сегмента $[x_{i-1}, x_i]$. Используя ряд Тейлора (см. [23]), мы можем легко оценить соответствующую ошибку

$$R_i = \int_{x_{i-1}}^{x_i} (f(x) - f(x_{i-1/2}))dx = \int_{x_{i-1}}^{x_i} \frac{(x - x_{i-1/2})^2}{2} f''(\zeta_i)dx,$$

так как

$$f(x) = f(x_{i-1/2}) + f'(x_{i-1/2})(x - x_{i-1/2}) + f''(\zeta_i)\frac{(x - x_{i-1/2})^2}{2},$$

где $\zeta_i \in [x_{i-1}, x_i]$. Если $M_{2,i} = \max |f''(x)|$ для $x \in [x_{i-1}, x_i]$, мы имеем оценку для ошибки

$$|R_i| \leq M_{2,i} \int_{x_{i-1}}^{x_i} \frac{(x - x_{i-1/2})^2}{2} dx = \frac{h^3}{24} M_{2,i}. \tag{273}$$

В результате суммирования вкладов для всех частичных отрезков получается составная прямоугольная формула, которая выглядит так:

$$\int_a^b f(x)dx = h \sum_{i=0}^{n} f(x_{i-1/2}) + R, \tag{274}$$

где оценка ошибки R равна

$$|R| \leq \frac{h^2(b-a)}{24} M_2, \tag{275}$$

где $M_2 = \max_{x \in [a,b]} |f''(x)|$.

7.3 Формулы Ньютона-Котеса

Рассмотрим теперь полиномиальную интерполяцию Лагранжа для функции $f(x)$:

$$p(x) = \sum_{i=0}^{n} f(x_i) l_i(x), \qquad l_i(x) = \prod_{j=0,\ j\neq i}^{n} \frac{x - x_j}{x_i - x_j}, \qquad i, j = 0, 1, 2, ..., n, \tag{276}$$

с узлами x_i и соответствующими значениями функции $f(x_i)$, и подставим $p(x)$ в подынтегральное выражение. Итак, мы получаем

$$\int_a^b f(x)dx \approx \int_a^b p(x)dx = \sum_{i=0}^{n} f(x_i) \int_a^b l_i(x)dx = \sum_{i=0}^{n} A_i f(x_i), \tag{277}$$

где

$$A_i = \int_a^b l_i(x)dx.$$

Формулы такого типа называются **формулами Ньютона-Котеса**, если узлы сетки для сегмента $[a, b]$ равномерно распределены.

7.4 Правило трапеций

Простейший случай возникает, если положить $n = 1$ и в качестве узлов взять $x_0 = a$, $x_1 = b$. Следовательно

$$l_0(x) = \frac{b - x}{b - a}, \qquad l_1(x) = \frac{x - a}{b - a}, \qquad p_1(x) = f(a)l_0(x) + f(b)l_1(x).$$

как следствие,

$$A_0 = \int_a^b l_0(x)dx = \frac{b - a}{2} = \int_a^b l_1(x)dx = A_1.$$

Соответствующая квадратурная формула:

$$\int_a^b f(x)dx = \frac{b - a}{2} \cdot [f(a) + f(b)] + R. \tag{278}$$

Это правило известно как **правило трапеции** (см. [23]), где величина погрешности

$$R = -f''(\zeta)(b-a)^3/12,$$

может быть получена спомощью интегрирования оценки

$$f(x) - p_1(x) = f''(\zeta)(x-a)(x-b)/2, \qquad \zeta \in [a,b].$$

Если сегмент $[a,b]$ разбивается на частичные сегменты, мы получим составное правило трапеции с произвольным образом неравномерно расположенными узлами:

$$\int_a^b f(x)dx \approx \frac{1}{2}\sum_{i=1}^{n}(x_i - x_{i-1}) \cdot [f(x_{i-1}) + f(x_i)]. \tag{279}$$

С равномерным разбиением сегмента $[a,b]$ с шагом $h = (b-a)/n$ мы получим

$$\int_a^b f(x)dx = \frac{h}{2}\left(f(a) + 2\sum_{i=1}^{n-1} f(a+ih) + f(b)\right) + R, \tag{280}$$

и оценка погрешности выглядит так:

$$|R| \le \frac{(b-a)h^2 M_2}{12}, \tag{281}$$

где $M_2 = \max|f''(x)|$ at $x \in [a,b]$.

7.5 Правило Симпсона

Рассмотрим формулу

$$\int_0^1 f(x)dx = A_0 f(0) + A_1 f(\frac{1}{2}) + A_2 f(1) \tag{282}$$

с неопределенными коэффициентами $A_0,\ A_1,\ A_2$. Мы определяем эти коэффициенты, предполагая, чтобы эта формула была точной для пробных функций $f(x) = 1,\ x,\ x^2$. Таким образом, получаем

систему трех уравнений относительно неизвестных коэффициентами $A_0,\ A_1,\ A_2$,

$$1 = \int_0^1 dx = A_0 + A_1 + A_2,$$

$$\frac{1}{2} = \int_0^1 x dx = \frac{1}{2}A_1 + A_2,$$

$$\frac{1}{3} = \int_0^1 x^2 dx = \frac{1}{4}A_1 + A_2.$$

Решение системы есть $A_0 = 1/6,\ A_1 = 2/3,\ A_2 = 1/6$. Аналогичные вычисления для произвольного сегмента $[a, b]$ приводят к **правилу Симпсона** (см. [23]):

$$\int_a^b f(x)dx = \frac{b-a}{6}\left[f(a) + 4f(\frac{a+b}{2}) + f(b)\right] + R, \qquad (283)$$

где погрешность имеет вид

$$R = -\frac{f^{(4)}(\zeta)}{90}\left[\frac{b-a}{2}\right]^5, \qquad \zeta \in [a, b].$$

Заметим, что правило Симпсона является точной формулой для любого кубического многочлена, так как $f^{(4)}(x) = 0$. Формулу погрешности можно получить, анализируя представление

$$\int_a^{a+2h} f(x)dx \approx \frac{h}{3}\left[f(a) + 4f(a+h) + f(a+2h)\right]$$

где $h = (b-a)/2$. Применяя ряд Тейлора до членов порядка $O(h^5)$ для правой части равенства, получим

$$2hf(a) + +2h^2 f'(a) + \frac{4}{3}h^3 f''(a) + \frac{2}{3}h^4 f'''(a) + \frac{100}{3 \cdot 5!}h^5 f^{(4)}(\zeta).$$

Такая же процедура, применяемая к левой части, дает

$$2hf(a) + +2h^2 f'(a) + \frac{4}{3}h^3 f''(a) + \frac{2}{3}h^4 f'''(a) + \frac{32}{5!}h^5 f^{(4)}(\zeta).$$

Разность этих результатов дает желаемую формулу для погрешности:

$$R = -\frac{f^{(4)}(\zeta)}{90}h^5.$$

Составное правило Симпсона, использующее четное количество подинтервалов с равномерным распределением узлов

$$x_i = a + ih, \qquad h = (b-a)/n, \qquad i = 0, 1, ..., n,$$

выглядит так:

$$\int_a^b f(x)dx = \frac{h}{3}\left[f(a) + 2\sum_{i=2}^{n/2} f(x_{2i-2}) + 4\sum_{i=1}^{n/2} f(x_{2i-1}) + f(b)\right] + R, \tag{284}$$

где оценка погрешности дается выражением

$$|R| \leq \frac{M_4}{180}(b-a)h^4, \qquad M_4 = \max_{x\in[a,b]} |f^{(4)}(x)|. \tag{285}$$

Эта формула очень популярна в компьютерном моделировании.

7.6 Адаптивная квадратура

Адаптивные квадратурные методы предназначены для вычисления определенных интегралов, автоматически принимая во внимание специфический характер поведения подынтегрального выражения, например, быстрые осцилляции. В идеальном варианте пользователь поставляет во входные данные программы только подынтегральную функцию $f(x)$, интервал $[a, b]$ и точность ϵ для вычисления интеграла

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

Затем программа делит интервал на отрезки переменной длины, так что численное интегрирование по ним даст результаты приемлемой точности. Основная идея заключается в том, что если выполнение правила Симпсона на заданном подинтервале недостаточно точно, этот интервал будет разделен на две равные части, и правило

Симпсона будет использоваться на каждой половине. Эта процедура будет повторяться с целью получения приближения к интегралу с той же точностью по всем задействованным отрезкам. В итоге мы вычислим интеграл с n раз применением правила Симпсона

$$\int_a^b f(x)dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)dx = \sum_{i=1}^{n} (S_i + e_i) = \sum_{i=1}^{n} S_i + \sum_{i=1}^{n} e_i, \quad (286)$$

где S_i есть результат вычисления по формуле Симпсона для сегмента $[x_{i-1}, x_i]$, и e_i является локальной ошибкой вычислений. Если

$$|e_i| \leq \epsilon(x_i - x_{i-1})/(b-a),$$

то общая ошибка будет ограничена:

$$\left| \sum_{i=1}^{n} e_i \right| \leq \sum_{i=1}^{n} |e_i| \leq \frac{\epsilon}{b-a} \sum_{i=1}^{n} (x_i - x_{i-1}) = \epsilon. \quad (287)$$

Вот краткое описание метода Рунге о том, как оценивать локальную ошибку. Предположим, что с использованием правила Симпсона мы имеем два результата расчета с n и $2n$ для частичного сегмента $[x_{i-1}, x_i]$ $(h = (x_i - x_{i-1})/n)$

$$\int_{x_{i-1}}^{x_i} f(x)dx = S_h(x_{i-1}, x_i) + c_i h^4$$

и

$$\int_{x_{i-1}}^{x_i} f(x)dx \approx S_{h/2}(x_{i-1}, x_i) + c_i \frac{h^4}{2^4}.$$

Исключая неизвестный коэффициент c_i, получим

$$\int_{x_{i-1}}^{x_i} f(x)dx - S_{h/2}(x_{i-1}, x_i) \approx \frac{1}{2^4} \left(\int_{x_{i-1}}^{x_i} f(x)dx - S_h(x_{i-1}, x_i) \right),$$

и тогда

$$\int_{x_{i-1}}^{x_i} f(x)dx - S_{h/2}(x_{i-1}, x_i) \approx \frac{S_{h/2}(x_{i-1}, x_i) - S_h(x_{i-1}, x_i)}{2^4 - 1}.$$

Это приближенное равенство позволяет получить оценку локальной ошибки.

7.7 Квадратура Гаусса

Предположим, что квадратурная формула с неравномерным расстоянием между узлами является точной для любого многочлена степени $2n-1$, тогда

$$\int_a^b x^s dx = \sum_{i=1}^{n} c_i x_i^s, \qquad s = 0, 1, ..., 2n-1. \tag{288}$$

Таким образом, мы имеем $2n$ нелинейных уравнений для $2n$ неизвестных $x_1, ..., x_n$ и $c_1, ..., c_n$. Эта проблема имеет единственное решение. Результатом является формула, называемая гауссовой квадратурой. Например, рассмотрим вычисление интеграла

$$\int_{-1}^{1} f(x)dx.$$

Тогда $n = 2$, $s = 0, 1, 2, 3$, и соответствующая система имеет вид

$$c_1 + c_2 = 2, \quad c_1x_1 + c_2x_2 = 0, \quad c_1x_1^2 + c_2x_2^2 = 2/3, \quad c_1x_1^3 + c_2x_2^3 = 0.$$

Ее решение есть $c_1 = c_2 = 1$, $x_1 = -\frac{1}{\sqrt{3}}$ и $x_2 = \frac{1}{\sqrt{3}}$. Наконец, мы получаем гауссову квадратурную формулу для этого случая

$$\int_{-1}^{1} f(x)dx \approx f(-\frac{1}{\sqrt{3}}) + f(\frac{1}{\sqrt{3}}),$$

которая является точным результатом для произвольного многочлена со степенью меньшей 4.

7.8 Вычисление определенных интегралов методом Монте-Карло

Пусть требуется вычислить интеграл

$$I = \int_0^1 \varphi(t)dt. \tag{289}$$

Пусть t является равномерно распределенная случайная величина, $p(t)$ - плотность распределения вероятности этой случайной величины:

$$\varphi(t) = \begin{cases} 0, & t < 0, \\ 1, & 0 < t < 1, \\ 0, & t > 1. \end{cases} \tag{290}$$

Тогда математическое ожидание случайной функции $\varphi(t)$ определяется равенством

$$M[\varphi(t)] = \int_0^1 \varphi(t)p(t)dt = \int_0^1 \varphi(t)dt. \tag{291}$$

Найдем приближенное значение математического ожидания. Пусть в результате N испытаний получено N значений случайной величины равномерного распределения $t_1, t_2, ..., t_N$. Тогда приближенное значение $M[\varphi(t)]$, согласно теореме Чебышева (см. [23]), определится из равенства

$$M[\varphi(t)] \approx \frac{1}{N} \sum_{n=1}^{N} \varphi(t_n). \tag{292}$$

Следовательно, получаем приближенное значение для интеграла I_1

$$I = \int_0^1 \varphi(t)dt \approx \frac{1}{N} \sum_{n=1}^{N} \varphi(t_n). \tag{293}$$

Пусть требуется вычислить интеграл

$$I_1 = \int_a^b f(x)dx. \tag{294}$$

Перейдем к новой переменной t с помощью равенства $x = a+(b-a)t$. Тогда

$$I_1 = \int_a^b f(x)dx = (b-a)\int_0^1 f(a+(b-a)t)dt = (b-a)\int_0^1 \varphi(t)dt. \tag{295}$$

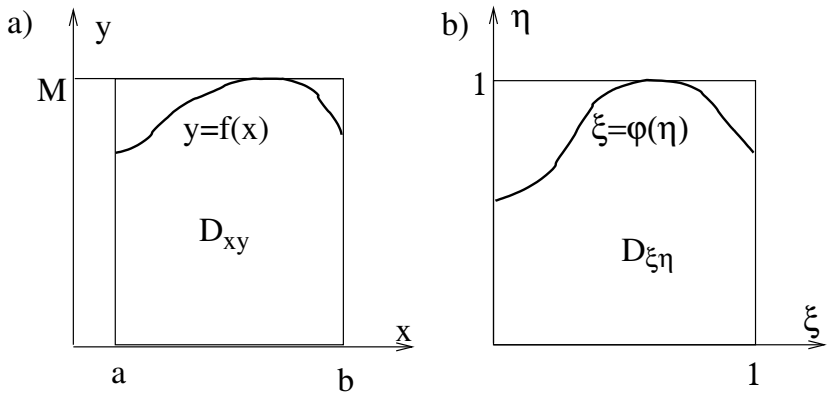


Рис. 6: Вычисление определенных интегралов методом Монте-Карло

Повторяя сказанное выше, получаем приближенное значение для интеграла I_1

$$I_1 = \int_a^b f(x)dx \approx \frac{b-a}{N}\sum_{n=1}^{N}\varphi(t_n) \approx \frac{b-a}{N}\sum_{n=1}^{N} f(x_n), \qquad (296)$$

где $x_n = a + (b-a)t_n$.

Существует другой способ вычисления определенных интегралов, основанный на использовании метода Монте-Карло. Из геометрического смысла определенного интеграла следует,что интеграл

$$I_1 = \int_a^b f(x)dx. \qquad (297)$$

выражает площадь криволинейной трапеции, ограниченной линиями $x = a, x = b,\ y = 0, y = f(x)$, если функция непрерывна и неотрицательна на отрезке $[a, b]$. Рассмотрим прямоугольник $x = a, x = b,\ y = 0, y = M,\ M = \sup_{x\in[a,b]} f(x)$. Если функция $f(x)$ удовлетворяет неравенству $f(x) \geq 0$ не во всех точках отрезка $[a, b]$, то следует воспользоваться тождеством

$$I_1 = \int_a^b (f(x) + h)dx - h(b-a), \qquad (298)$$

так что $f(x) + h \geq 0$ во всех точках отрезка $[a, b]$. Пусть ниже $0 \leq f(x) \leq M$.

Преобразуем прямоугольник $a \le x \le b$, $0 \le y \le M$, в квадрат $0 \le \xi \le 1$, $0 \le \eta \le 1$ (см. Рис. 6), с помощью формул

$$x = a + (b-a)\xi, \qquad y = M\eta. \tag{299}$$

Тогда получим

$$I_1 = (b-a)M \int_0^1 \varphi(\xi)d\xi, \qquad \varphi(\xi) = \frac{1}{M} f(a + (b-a)\xi). \tag{300}$$

Таким образом, область криволинейной трапеции $D_{x,y}$ переходит в $D_{\xi,\eta}$.

Рассмотрим множество случайных точек (ξ_1, η_1), (ξ_2, η_2),...,(ξ_N, η_N), равномерно распределенных на единичном квадрате. Пусть в область $D_{\xi,\eta}$ попадет n точек. Так как точки (ξ_k, η_k), $k = 1, 2, ..., N$, распределены равномерно, то имеет место предел по вероятности

$$\lim_{N\to\infty} \frac{n}{N} = \int_0^1 \varphi(\xi)d\xi, \tag{301}$$

и, таким образом, мы получаем вторую приближенную формулу метода Монте-Карло

$$I_1 = \int_a^b f(x)dx \approx (b-a)M\frac{n}{N}. \tag{302}$$

7.9 Вычисление двойных интегралов методом Монте-Карло

Пусть теперь требуется вычислить двойной интеграл

$$I_2 = \int\limits_{D_{xy}} f(x,y)dxdy, \tag{303}$$

где область D_{xy} определяется неравенствами $a \le x \le b$, $\varphi_1(x) \le y \le \varphi_2(x)$, и

$$c = \min_{x\in[a,b]} \varphi_1(x) > 0, \qquad d = \max_{x\in[a,b]} \varphi_2(x) > c.$$

Пусть непрерывные кривые $y = \varphi_{2,1}(x)$ расположены одна над другой и не пересекаются.

Перейдем к новым переменным с помощью равенств $x = a+(b-a)\xi$, $y = c+(d-c)\eta$. При таком преобразовании область D_{xy} переходит в область $D_{\xi\eta}$, содержащуюся в единичном квадрате $0 \le \xi \le 1$, $0 \le \eta \le 1$ (см. Рис. 7). Вновь рассмотрим множество случайных точек (ξ_1, η_1), (ξ_2, η_2), ..., (ξ_N, η_N), равномерно распределенных на единичном квадрате. Пусть в область $D_{\xi,\eta}$ попадет n точек (ξ_k, η_k), $k = 1, 2, ..., n$. Очевидно, что в область D_{xy} попадет n точек (x_k, y_k) с координатами

$$x_k = a + (b-a)\xi_k, \qquad y_k = c + (d-c)\eta_k, \qquad k = 1, 2, ..., n.$$

По теореме о среднем (см. [23]) имеем

$$\int\limits_{D_{xy}} f(x,y)dxdy = f(\bar{x}, \bar{y})S, \tag{304}$$

где $(\bar{x}, \bar{y}) \in D_{xy}$, а S - площадь области. За приближенное значение $f(\bar{x}, \bar{y})$ возьмем среднее арифметическое значений функции $f(x, y)$ в n случайных точках, попавших в область D_{xy}

$$f(\bar{x}, \bar{y}) \approx \frac{1}{n}\sum_{k=1}^{n} f(x_k, y_k). \tag{305}$$

Таким образом, мы получаем

$$\int\limits_{D_{xy}} f(x,y)dxdy \approx \frac{S}{n}\sum_{k=1}^{n} f(x_k, y_k). \tag{306}$$

Так как точки (ξ_k, η_k), $k = 1, 2, ..., N$, распределены равномерно, то имеет место предел по вероятности

$$\lim_{N\to\infty} \frac{n}{N} = \frac{S}{(b-a)(d-c)}. \tag{307}$$

Следовательно, будем иметь

$$S \approx (b-a)(d-c)\frac{n}{N}, \tag{308}$$

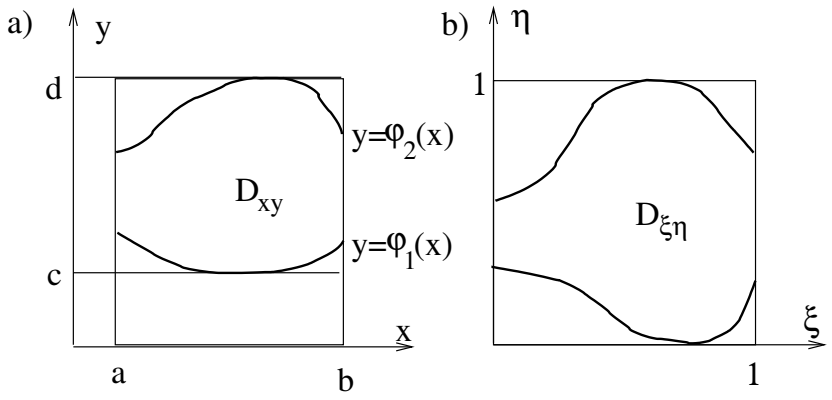


Рис. 7: Вычисление двойных интегралов методом Монте-Карло

и таким образом, в итоге мы получаем

$$\int\limits_{D_{xy}} f(x,y)dxdy \approx \frac{(b-a)(d-c)}{N}\sum_{k=1}^{n} f(x_k,y_k). \qquad (309)$$

Рассмотрим второй способ вычисления двойных интегралов, основанный на использовании метода Монте-Карло. Пусть

$$M = \max_{(x,y)\in D_{xy}} f(x,y),$$

и $f(x,y) \geq 0$ для всех $(x,y) \in D_{xy}$. Двойной интеграл, как известно, выражает объем цилиндрического тела Ω_{xyz}, определенного неравенствами

$$a \leq x \leq b, \qquad \varphi_1(x) \leq y \leq \varphi(x), \qquad 0 \leq z \leq f(x,y). \qquad (310)$$

Это цилиндрическое тело расположено внутри параллелепипеда

$$a \leq x \leq b, \qquad c \leq y \leq d, \qquad 0 \leq z \leq M. \qquad (311)$$

Преобразуем этот параллелепипед в куб $0 \leq \xi \leq 1$, $0 \leq \eta \leq 1$, $0 \leq \zeta \leq 1$, с помощью формул

$$x = a + (b-a)\xi, \qquad y = c + (d-c)\eta, \qquad z = M\zeta. \qquad (312)$$

Таким образом, область криволинейной трапеции $D_{x,y}$ переходит в $D_{\xi,\eta}$, а цилиндрическое тело Ω_{xyz} преобразуется в $\Omega_{\xi\eta\zeta}$, и мы получим

$$I_2 = (b-a)(d-c)M \int\limits_{D_{\xi,\eta}} \varphi(\xi,\eta)d\xi d\eta, \tag{313}$$

$$\varphi(\xi,\eta) = \frac{1}{M} f(a+(b-a)\xi, c+(d-c)\eta).$$

Рассмотрим множество случайных точек (ξ_1,η_1,ζ_1), (ξ_2,η_2,ζ_2),...,(ξ_N,η_N,ζ_N), равномерно распределенных на единичном кубе. Пусть в область $\Omega_{\xi\eta\zeta}$ попадет n точек. Так как эти точки (ξ_k,η_k,ζ_k), $k=1,2,...,N$ распределены равномерно, то имеет место предел по вероятности

$$\lim_{N\to\infty} \frac{n}{N} = \int\limits_{D_{\xi,\eta}} \varphi(\xi,\eta)d\xi d\eta, \tag{314}$$

и, таким образом, мы получаем вторую приближенную формулу метода Монте-Карло для двойных интегралов:

$$I_2 = \int\limits_{D_{xy}} f(x,y)dxdy \approx (b-a)(d-c)M\frac{n}{N}. \tag{315}$$

Заметим, что формула, аналогичная соотношениям (302) и (315), имеет место и для m-кратных интегралов:

$$I_m = \int\limits_{D_{x_1...x_m}} f(x_1,x_2,...,x_m)dx_1dx_2...dx_m \approx M\frac{n}{N}\prod_{k=1}^{m}(b_k-a_k), \tag{316}$$

где область $D_{x_1...x_m}$ принадлежит m-мерному параллелепипеду, координаты точек которого удовлетворяют m неравенствам

$$a_k \le x_k \le b_k, \qquad k=1,2,...,m, \tag{317}$$

а функция $f(x_1,x_2,...,x_m)$ непрерывна в области $D_{x_1...x_m}$ и удовлетворяет условию

$$0 \le f(x_1,x_2,...,x_m) \le M. \tag{318}$$

Вывод формул (302), (315), (316) основан на использовании понятия сходимости по вероятности. Поэтому соотношение n/N тем устойчивее, чем больше N. Это означает, что для любого сколь угодно малого числа $\epsilon > 0$ вероятность неравенства $|I - I_N| < \epsilon$, где I есть точное значение интеграла, а I_N является его приближенным значением, найденным методом Монте-Карло, стремится к единице с увеличением N. Тем не менее может случиться, что и при очень больших N окажется, что $|I - I_N| > \epsilon$. Последнее обстоятельство на практике встречается редко.

Вопросы для самоконтроля к главе 7

1. Каковы основные формулы численного интегрирования и какова соответствующая погрешность вычислений?

2. В чем состоит основная идея метода адаптивной квадратуры?

3. В чем заключаются основные этапы применения метода Монте-Карло к вычислению определенных и кратных интегралов?

8 Численные методы решения задачи Коши для ОДУ первого порядка

8.1 Введение

Дифференциальное уравнение первого порядка описывает связь между неизвестной функцией $y(t)$ и его производной $y'(t)$. В общем виде дифференциальное уравнение первого порядка имеет вид

$$y'(t) = f(t, y(t)). \tag{319}$$

Решение дифференциального уравнения является функция $y(t)$, которая удовлетворяет соотношению (319). Аналитические методы позволяют построить семейство решений. Начальное условие

$$y(0) = y_0 \tag{320}$$

используется, чтобы выделить единственного представителя этого семейства. Дифференциальное уравнение (319) и начальное условие (320) вместе формулируют начальную задачу Коши. Поиск аналитических решений мы не рассматриваем, но замечаем, что аналитическое решение может быть очень сложным и довольно трудным для понимания. Например, достаточно простое уравнение

$$y' = t^3 - y^2 \tag{321}$$

имеет решение, содержащее функцию Бесселя.

Численные решения необходимы, когда аналитическое решение не может быть найдено в выражениях элементарных функций, что часто происходит на практике. Таким образом, вместо формулы мы получаем таблицу приблизительных значений. Использование приближенных численных методов неотделимо от изучения ошибок в этих приближениях. В качестве предварительного изучения дифференциальных уравнений мы отмечаем, что решение дифференциального уравнения обязательно связано с интегрированием. Например, пусть нам дано уравнение:

$$\frac{dy}{dt} = f(t), \tag{322}$$

с $f(t)$, которая является известной функцией, и граничным условием

$$y(0) = y_0. \tag{323}$$

Это уравнение легко решить:

$$y(t) = y_0 + \int_0^t f(\tau)d\tau. \tag{324}$$

Точность численного решения данного дифференциального уравнения, очевидно, будет зависеть от точности численного интегрирования.

8.2 Существование и единственность решений задачи Коши

Рассмотрим ОДУ с начальным условием

$$\frac{dy(t)}{dt} = f(y,t), \qquad y(0) = y_0, \qquad 0 \le t \le T, \tag{325}$$

где T является верхним пределом интегрирования уравнения. Если решение этого уравнения не единственно, то численный метод может выбирать один из нескольких вариантов разных решений. Например, для

$$\frac{dy(t)}{dt} = y \quad y_0 = 1, \tag{326}$$

существует единственное решение $y = e^t$. Однако для

$$\frac{dy(t)}{dt} = y^{1/2} \quad y_0 = 0, \tag{327}$$

существуют два решения $y \equiv 0$ и $y = t^2/4$, которые удовлетворяют тем же начальным условиям $(0, 0)$. Известен критерий, гарантирующий единственность решения ОДУ, включающий в себя условие Липшица: существует постоянная $L > 0$, такая, что $\forall y_{1,2} \in \mathcal{R}$ выполняется условие

$$|f(t, y_1) - f(t, y_2)| \le L|y_1 - y_2|. \tag{328}$$

Оказывается, если частная производная $\frac{\partial f}{\partial y}$ существует и ограничена в рассматриваемой области, то

$$L = \sup_{(t,y)} |\frac{\partial f}{\partial y}(t,y)|.$$

Это условие явно не выполняется в нашем втором примере.

Рассмотрим еще раз начальную задачу Коши (325). Решение этой задачи эквивалентно нахождению решения для следующего интегрального уравнения:

$$y(t) = y + 0 + \int_0^t f(\tau, y(\tau))d\tau. \tag{329}$$

Построим решение интегрального уравнения с помощью итерационного процесса

$$y_{n+1}(t) = S(y_n(t)) = \int_0^t f(\tau, y_n(\tau))d\tau, \qquad n > 0. \tag{330}$$

Покажем, что оператор $S(y)$ есть оператор сжатия (см. параграф 5.3.):

$$\|y_{n+1}(t) - y_{m+1}(t)\| = \sup_t |y_{n+1}(t) - y_{m+1}(t)| \leq$$
$$\leq \int_0^t |f(\tau, y_n(\tau)) - f(\tau, y_m(\tau))|d\tau. \tag{331}$$

Пусть функция $f(t,y)$ удовлетворяет условию Липшица (328). Тогда получим

$$\|y_{n+1}(t) - y_{m+1}(t)\| \leq L \int_0^t |y_n(\tau) - y_m(\tau)|d\tau \leq LT\|y_n(t) - y_m(t)\|. \tag{332}$$

Если $q = LT < 1$, то $S(y)$ есть оператор сжатия. Следовательно, его неподвижная точка $y = S(y)$ дает единственное решение начальной задачи Коши (325), и мы имеем следующую оценку (см. параграф 5.3.):

$$\|y_n(t) - y(t)\| \leq \frac{q^n}{1-q}\|y_1(t) - y_0(t)\|. \tag{333}$$

Теперь мы готовы дискретизировать наше дифференциальное уравнение, но для этого мы должны обсудить, как перейти от непрерывных к дискретным приближениям для производных.

8.3 Аппроксимация производных конечными разностями

Важнейшим шагом в дискретизации дифференциального уравнения при написании численных алгоритмов является замена производной ее некоторым приближением. Предположим, что мы имеем функцию $y(x)$, заданную на регулярной сетке точек $\{x_i\}_{i=1}^n$, где $y(x_i) = y_i$, и длина шага есть $x_{i+1} - x_i = h$. Если параметр h есть достаточно малая величина, тогда

$$\frac{dy(x_i)}{dx} \approx \frac{y_{i+1} - y_i}{h} \approx \frac{y_i - y_{i-1}}{h} \approx \frac{y_{i+1} - y_{i-1}}{2h}.$$

Теперь мы можем определить оператор *правой конечной разности*

$$\Delta_+ y_i(x_i) = \frac{y_{i+1} - y_i}{h},$$

и аналогично, оператор *левой конечной разности*

$$\Delta_- y_i(x_i) = \frac{y_i - y_{i-1}}{h}.$$

Производная второго порядка аппроксимируется как

$$\frac{d^2y}{dx^2} \approx \Delta_0^2 y_i(x_i) = \frac{y_{i+1} + y_{i-1} - 2y_i}{h^2}.$$

8.4 Модифицированный метод Эйлера

Рассмотрим начальную задачу Коши для ОДУ первого порядка

$$y' = f(x, y), \qquad y(x_0) = y_0, \tag{334}$$

и предположим, что имеет место однозначность решения. Используя формулу Тейлора

$$y(x_0 + h) = y(x_0) + hy'(x_0) + \frac{y''(\xi)h^2}{2}, \qquad x_0 < \xi < x_0 + h,$$

мы получим приближение

$$y(x_0 + h) = y(x_0) + hf(x_0, y_0) + O(h^2).$$

Будем использовать этот метод итеративно, продолжая решение в точку $x = x_0+2h$ после того, как было вычислено значение $y(x_0+h)$. А затем рассматривается продолжение в точку $x = x_0 + 3h$, и так далее. Таким образом, мы можем написать алгоритм для метода Эйлера в следующем виде:

$$y_{n+1} = y_n + hf(x_n, y_n). \tag{335}$$

Проблема с использованием этого простого метода заключается в отсутствии достаточной точности, требующей чрезвычайно малого размера шага. В простом методе Эйлера мы используем наклон графика функции, определенный в начале интервала, y'_n, для определения приращения функции. Этот метод был бы правильным, только если бы функция была линейной. Вместо этого нам нужен правильный усредненный наклон в интервале. Этот результат можно получить усреднением наклонов на обоих концах интервала

$$y_{n+1} = y_n + h\frac{f(x_n, y_n) + f(x_{n+1}, y_{n+1})}{2}. \tag{336}$$

Это должно дать улучшенную оценку для y_n в x_n. Однако мы не можем использовать этот подход для определения y_{n+1}. Лучшее решение имеет вид:

$$y_{n+1} = y_n + \frac{h}{2}(f(x_n, y_n) + f(x_{n+1}, y_n + hf(x_n, y_n))), \tag{337}$$

которое называется модифицированный метод Эйлера. Таким образом, этот метод работает путем оценивания или прогнозирования значения y_{n+1} с помощью простого метода Эйлера, которое также называется методом прогноза и коррекции (predictor-corrector) Эйлера.

Мы можем найти ошибку для модифицированного метода Эйлера, сравнив ее с рядом Тейлора:

$$y_{n+1} = y_n + y'_n h + \frac{1}{2}y''_n h^2 + \frac{1}{6}y'''(\xi_n)h^3, \qquad x_n < \xi_n < x_{n+1}.$$

Заменяя вторую производную на конечно-разностное приближение, получим

$$y_{n+1} = y_n + y'_n h + \frac{1}{2}h^2 \frac{y'_{n+1} - y'_n}{h} + O(h^3),$$

и поэтому будем иметь

$$y_{n+1} = y_n + \frac{h}{2}(y'_{n+1} + y'_n) + O(h^3).$$

Это показывает, что ошибка для одного шага модифицированного метода Эйлера есть величина порядка $O(h^3)$. Это локальная ошибка. Происходит накопление локальных ошибок шаг за шагом, так что полная ошибка для всего интервала расчета, называемая глобальной ошибкой, есть величина порядка $O(h^2)$.

8.5 Метод Рунге-Кутта

Метод Эйлера недостаточно точен. Гораздо большую точность можно получить более эффективно с помощью группы методов, названных в честь двух немецких математиков Рунге и Кутта. Рассмотрим вывод простого метода второго порядка. Здесь приращение для y представляет собой среднее значение двух оценок, которые мы называем k_1 и k_2. Таким образом, для уравнения

$$\frac{dy}{dx} = f(x, y)$$

мы имеем

$$y_{n+1} = y_n + ak_1 + bk_2, \tag{338}$$

$$k_1 = hf(x_n, y_n), \qquad k_2 = hf(x_n + \alpha h, y_n + \beta k_1).$$

Наша задача состоит в том, чтобы построить схему нахождения четырех параметров a, b, α, β. Представляя y_{n+1} в выражении (338) с помощью ряда Тейлора, получаем

$$y_{n+1} = y_n + hf(x_n, y_n) + \frac{h^2}{2}f'(x_n, y_n) + ..., \tag{339}$$

или,

$$y_{n+1} = y_n + hf_n + h^2\left(\frac{1}{2}f_x + \frac{1}{2}f_y f\right) + ... \ . \tag{340}$$

С другой стороны, используя правую часть (338), будем иметь

$$y_{n+1} = y_n + ahf_n + bhf[x_n + \alpha h, y_n + \beta h f_n].$$

Разложив правую часть $f(x, y)$ в ряд Тейлора в окрестности (x_n, y_n), получим

$$y_{n+1} = y_n + ahf_n + bh(f_n + f_x(x_n, y_n)\alpha h + f_y(x_n, y_n)\beta h f_n),$$

или, меняя порядок членов, приводит к выражению

$$y_{n+1} = y_n + h(a + b)f_n + h^2(f_x(x_n, y_n)\alpha b + f_y(x_n, y_n)\beta b f_n).$$

Этот результат идентичен разложению в ряд Тейлора (340), если

$$a + b = 1, \qquad \alpha b = \frac{1}{2}, \qquad \beta b = \frac{1}{2}.$$

Заметим, что эти три уравнения должны быть удовлетворены с помощью четырех неизвестных. Мы можем выбрать одно значение произвольно. Например, если мы возьмем $a = b = \frac{1}{2}, \quad \alpha = \beta = 1$, то мы получаем модифицированный метод Эйлера.

Методы Рунге-Кутта четвертого порядка наиболее широко используются в численном анализе. Соответствующие построения формул метода Рунге-Кутта четвертого порядка производятся аналогичным образом. Наиболее часто используемый алгоритм определяется как

$$y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \tag{341}$$

$$k_1 = hf(x_n, y_n), \qquad k_2 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1),$$

$$k_3 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_2), \qquad k_4 = hf(x_n + h, y_n + k_3).$$

Локальная ошибка для методов Рунге-Кутта четвертого порядка есть величина порядка $O(h^5)$; глобальная ошибка - $O(h^4)$. Это гораздо более эффективный вычислительный метод по сравнению с модифицированным методом Эйлера.

8.6 Введение в многошаговые методы

Модифицированный метод Эйлера и методы Рунге-Кутта для решения начальной задачи Коши для ОДУ первого порядка являются одношаговыми методами, поскольку они не используют никаких предыдущих значений $y(x)$, когда решение численно продолжается из x и $x+h$. Если $x_0, x_1, ..., x_n$ являются шагами по оси x, тогда y_{n+1} зависит только от y_n, и информация о значениях $y_{n-1}, ..., y_0$ не используется.

Более эффективная процедура может быть разработана, если на каждом этапе используются некоторые предварительные значения решения. Приведенный здесь принцип заключается в следующем: для решения начальной задачи

$$\frac{dy}{dx} = f(x,y), \qquad y(x_0) = y_0,$$

мы полагаем, что

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} f(x, y(x))dx.$$

Интеграл справа может быть аппроксимирован численной квадратурной формулой, и результатом будет формула для построения приближенного решения шаг за шагом.

8.7 Формула Адамса-Бэшфорта

Предположим, что полученная формула имеет следующий тип:

$$y_{n+1} = y_n + af_n + bf_{n-1} + cf_{n-2} + ..., \tag{342}$$

где $f_i = f(x_i, y_i)$. Уравнение этого типа называется формулой Адамса-Бэшфорта. Вот формула Адамса-Бэшфорта пятого порядка , основанная на равномерной сетке с узлами в точках $x_i = x_0 + ih, \;\; i = 0, 1, ..., n$:

$$y_{n+1} = y_n + \frac{h}{720}[1901f_n - 2774f_{n-1} + 2616f_{n-2} - 1274f_{n-3} + 251f_{n-4}]. \tag{343}$$

Рассмотрим вывод этой формулы. Начнем с приближения интеграла

$$\int_{x_n}^{x_{n+1}} f(x, y(x))dx \approx h[Af_n + Bf_{n-1} + Cf_{n-2} + Df_{n-3} + Ef_{n-4}].$$

Коэффициенты A, B, C, D, E определяются тем, что это уравнение является точным всякий раз, когда подынтегральное выражение является многочленом степени меньше 5. Для простоты будем считать, что $x_n = 0$, $x_{n-1} = -1$, $x_{n-2} = -2$, $x_{n-3} = -3$, $x_{n-4} = -4$, и $h = 1$. В качестве базиза возьмем следующие пять многочленов:

$$\begin{array}{lll} p_0(x) = 1, & p_1(x) = x, & p_2(x) = x(x+1), \\ p_3(x) = x(x+1)(x+2), & p_4(x) = x(x+1)(x+2)(x+3). & \end{array}$$

Когда они подставляются в уравнение

$$\int_0^1 p_m(x)dx \approx Ap_m(0) + Bp_m(-1) + Cp_m(-2) + Dp_m(-3) + Ep_m(-4)$$

для $m = 0, 1, 2, 3, 4$, мы получим систему для определения A, B, C, D, E:

$$\begin{cases} A + B + C + D + E = 1, \\ -B - 2C - 3D - 4E = 1/2, \\ 2C + 6D + +12E = 5/6, \\ -6D - 24E = 9/4, \\ 24E = 251/30. \end{cases}$$

Коэффициенты формулы Адамса-Бэшфорта получаются с помощью обратных подстановок.

8.8 Формула Адамса-Молтона

В численной практике формулы Адамса-Бэшфорта редко используются сами по себе. Они используются вместе с другими формулами для повышения точности. Предположим, что мы используем численную квадратурную формулу, которая включает f_{n+1}. Тогда уравнение (342) будет иметь форму

$$y_{n+1} = y_n + af_{n+1} + bf_n + cf_{n-1} + \dots . \tag{344}$$

Аналогичным образом получается формула, известная как формула Адамса-Молтона пятого порядка :

$$y_{n+1} = y_n + \frac{h}{720}[251f_{n+1} + 646f_n - 264f_{n-1} + 106f_{n-2} - 19f_{n-3}]. \quad (345)$$

Она не может использоваться непосредственно, так как y_{n+1} присутствует по обе стороны уравнения. Однако очень эффективный алгоритм, называемый методом прогноза и коррекции (predictor-corrector), использует формулу Адамса-Бэшфорта для прогнозирования условного значения для y^*_{n+1}, а затем формулу Адамса-Молтона для вычисления скорректированного значения y_{n+1}. Итак, в (345) мы оцениваем f_{n+1} как $f_{n+1}(x_{n+1}, y^*_{n+1})$, используя прогнозируемое значение y^*_{n+1}, полученное из формулы Адамса-Бэшфорта. При использовании метода прогноза и коррекции, для запуска метода на самом первом шаге необходимо задать значения $y_0,\ y_1,\ y_2,\ y_3,\ y_4$. Но поскольку изначально известно только y_0, то метод Рунге-Кутта идеален для получения $y_1,\ y_2,\ y_3,\ y_4$. Обычно формулы одного и того же порядка используются вместе.

8.9 Порядок аппроксимации линейных многошаговых методов

Рассмотрим линейный многошаговый метод в общем виде

$$a_k y_n + a_{k-1} y_{n-1} + ... + a_0 y_{n-k} = h[b_k f_n + b_{k-1} f_{n-1} + ... + b_0 f_{n-k}] \quad (346)$$

для решения начальной задачи

$$y' = f(x, y), \qquad y(x_0) = x_0. \quad (347)$$

Это называется k-ступенчатым методом. Коэффициенты a_i, b_i заданы. Как прежде, y_i обозначает аппроксимацию решения при $x_i = x_0 + ih$, и $f_i = f(x_i, y_i)$. Эта формула используется для вычисления y_n, при условии, что $y_{n-k}, y_{n-k+1}, ..., y_{n-1}$, известны. Предположим, что $a_k \neq 0$. Если $b_k = 0$, этот метод называется явным, и y_n может быть вычислен напрямую. В противном случае метод называется неявным.

Точность численного решения во многом определяется порядком аппроксимации используемого алгоритма. Порядок определяется погрешностью аппроксимации с помощью формулы Тейлора. Следующий линейный функционал и определяет порядок аппроксимации алгоритма:

$$L(y) = \sum_{i=0}^{k} [a_i y(x_0 + ih) - hb_i y'(x_0 + ih)]. \tag{348}$$

Используя разложение в ряд Тейлора для $y(ih)$ и $y'(ih)$ для $i = 0, 1, ..., k$,

$$y(x_0 + ih) = \sum_{j=0}^{+\infty} \frac{(ih)^j}{j!} y^{(j)}(x_0), \qquad y'(x_0 + ih) = \sum_{j=0}^{+\infty} \frac{(ih)^j}{j!} y^{(j+1)}(x_0),$$

мы получаем

$$L(y) = d_0 y(x_0) + d_1 h y'(x_0) + d_2 h^2 y''(x_0) + ... \ , \tag{349}$$

где

$$d_0 = \sum_{i=0}^{k} a_i, \qquad d_1 = \sum_{i=0}^{k} (ia_i - b_i), \qquad d_2 = \sum_{i=0}^{k} (i^2 a_i/2 - ib_i), ...$$

$$d_j = \sum_{i=0}^{k} \left(\frac{i^j}{j!} a_i - \frac{i^{j-1}}{(j-1)!} b_i \right).$$

Если $d_0 = d_1 = ... = d_m = 0$, тогда

$$L(y) = d_{m+1} h^{m+1} y^{(m+1)}(x_0) + O(h^{m+2})$$

представляет локальную ошибку ограничения, а метод имеет порядок m. Например, в случае

$$y_n - y_{n-2} = \frac{1}{3} h (f_n + 4 f_{n-1} + f_{n-2}),$$

мы имеем,что a_1, a_2, a_3 суть $-1, 0, 1,$ и b_1, b_2, b_3 это $1/3, 4/3, 1/3$. В результате получаем

$$d_0 = a_0 + a_1 + a_2 = 0, \qquad d_1 = -b_0 + (a_1 - b_1) + (2a_2 - b_2) = 0,$$

$$d_2 = d_3 = d_4 = 0, \qquad d_5 = -1/90.$$

Следовательно, порядок метода равен 4, а локальная ошибка аппроксимации есть величина порядка $O(h^5)$.

8.10 Устойчивость линейных многошаговых методов

Рассмотрим снова многошаговый метод

$$a_k y_n + a_{k-1} y_{n-1} + ... + a_0 y_{n-k} = h[b_k f_n + b_{k-1} f_{n-1} + ... + b_0 f_{n-k}] \quad (350)$$

для решения начальной задачи Коши

$$y' = f(x, y), \qquad y(x_0) = x_0. \quad (351)$$

Связанными с многошаговым методом являются два многочлена:

$$p(z) = a_k z^k + a_{k-1} z^{k-1} + ... + a_0, \qquad q(z) = b_k z^k + b_{k-1} z^{k-1} + ... + b_0.$$

Анализ покажет, что некоторые важные свойства многошагового метода зависит от расположения корней многочленов $p(z)$ и $q(z)$.

Сходимость метода означает, что приближенное решение $y(h, x)$, полученное с использованием шага h, удовлетворяет следующему пределу:

$$\lim_{h \to 0} y_n = y(x_n)$$

для фиксированного $x_n = x_0 + nh$ $(n \to +\infty)$, при условии, что начальное значение соответствует начальному условию $y_0 = y(x_0)$. Этот метод стабилен, если все корни $p(z)$ удовлетворяют следующим условиям: все корни $p(z)$ лежат в круге $|z| \leq 1$, и каждый корень, равный по модулю 1, простой. Кроме того, дополнительно пусть выполняются условия $p(1) = 0$ и $p'(1) = q(1)$.

Теорема. Для сходимости многошагового метода необходимо и достаточно, чтобы метод был стабилен. Мы проиллюстрируем утверждение этой теоремы на простых примерах. Предположим, что существует корень $|\lambda| > 1$. Рассмотрим начальную задачу

$$y' = 0, \qquad y(0) = 0.$$

Тогда решение определяется уравнением

$$a_k y_n + a_{k-1} y_{n-1} + ... + a_0 y_{n-k} = 0.$$

Это линейное разностное уравнение. Очевидно, что должно быть $y_n \to 0$ при $n \to \infty$. Мы ищем решение в виде $y_n = hz^n$. Тогда получим для z характеристическое уравнение $p(z) = 0$. Если корень $z = \lambda$ такой, что $|\lambda| > 1$, соответствующее решение ведет себя так:

$$|y_n| = |h\lambda^n| = h|\lambda|^n \to +\infty$$

для фиксированного $x = x_0 + nh$ при $h \to 0$ и $n \to +\infty$. Это расходящееся частичное решение приведет к расходимости соответствующего численного решения. Если $|\lambda| = 1$ и $p'(\lambda) = 0$, то имеет место следующее решение $nh\lambda^n$, и

$$|y_n| = |hn\lambda^n| = hn|\lambda|^n \to nh,$$

для фиксированных $x = x_0 + nh$ как $h \to 0$ и $n \to +\infty$. Это противоречит тому, что $y_n \to 0$ при $n \to \infty$.

Рассмотрим теперь начальную задачу

$$y' = 0, \qquad y(0) = 1.$$

Решение $y = 1$, и $y_n = 1$ для любого n. Таким образом, должно быть $p(1) = 0$. Рассмотрим дополнительно другую начальную задачу

$$y' = 1, \qquad y(0) = 0.$$

Точное решение $y = x$ и $y_n = hn$. Уравнение многошагового метода преобретает вид

$$a_k y_n + a_{k-1} y_{n-1} + ... + a_0 y_{n-k} = h[b_k + b_{k-1} + ... + b_0].$$

Следовательно, необходимо, чтобы для $n = k$ выполнялось условие

$$a_k k + a_{k-1}(k-1) + ... + a_1 = b_k + b_{k-1} + ... + b_0,$$

то есть $p'(1) = q(1)$.

Рассмотрим **пример Милна**. Метод Милна определяется формулой

$$y_n - y_{n-2} = \frac{h}{3}(f_n + 4f_{n-1} + f_{n-2}).$$

Это неявный метод, и для него мы получаем:

$$p(z) = z^2 - 1, \qquad q(z) = (z^2 + 4z + 1)/3.$$

Простые корни $p(z)$ суть ± 1. Более того, $p'(z) = 2z$, $p'(1) = q(1) = 2$. Таким образом, выполняются условия стабильности для сформулированной теоремы. В результате метод Милна сходится.

8.11 Глобальная ошибка линейных многошаговых методов

Рассмотрим начальную задачу Коши

$$x' = f(t, x), \qquad x(0) = x_0, \qquad 0 \le t \le T > 0.$$

Как получить оценку глобальной ошибки при численном решении исходной задачи для дифференциального уравнения? Разность $|x(t_n) - x_n|$ есть глобальная ошибка. Это не просто сумма всех локальных ошибок. Ключевым моментом здесь является понимание того, как два решения различаются в некоторый момент, если они запущены с разными начальными условиями, так как каждый шаг в построении численного решения должен использовать в качестве начального значения приближенную величину, вычисленную на предыдущем шаге.

Рассмотрим начальную задачу Коши

$$x' = f(t, x), \qquad x(0) = s, \qquad 0 \le t \le T > 0.$$

Предположим, что $f_x = \frac{\partial f}{\partial x}(t, x)$ непрерывна и удовлетворяет условию $f_x \le \lambda$ для $0 \le t \le T > 0$. Мы хотели бы знать, как решение $x = x(t, s)$ зависит от s. Определим $u(t) = \frac{\partial x}{\partial s}(t, s)$. Мы можем получить дифференциальное уравнение - уравнение в вариациях - для u, дифференцируя по s исходную задачу

$$u'(t) = f_x(t, x)u, \qquad u(0) = 1, \qquad 0 \le t \le T > 0.$$

Заметим, что если $f_x \leq \lambda$ для $0 \leq t \leq T > 0$, то решение уравнения в вариациях удовлетворяет неравенству

$$|u(t)| \leq e^{\lambda t}, \qquad 0 \leq t \leq T > 0.$$

Доказательство: у нас есть

$$u'/u = f_x = \lambda - \alpha(t), \qquad \alpha(t) \geq 0.$$

Интегрируя это неравенство, получим

$$\ln |u| = \lambda t - \int_0^t \alpha(\tau) d\tau,$$

$$|u(t)| = \exp\left\{\lambda t - \int_0^t \alpha(\tau) d\tau\right\} \leq e^{\lambda t},$$

потому что

$$\int_0^t \alpha(\tau) d\tau \geq 0.$$

Используя это неравенство, легко показать, что если начальная задача задана с двумя начальными значениями s и $s + \delta$, то их решения различаются по t как

$$|x(t,s) - x(t,s+\delta)| = \left|\frac{\partial}{\partial s} x(t, s+\theta\delta)\right| |\delta| = |u(t)||\delta| \leq e^{\lambda t}|\delta|, \quad 0 < \theta < 1.$$

Теорема о глобальной ошибке. Если все локальные ошибки вычисления при $t_1, t_2, ..., t_n$ не превышает δ по абсолютной величине, то глобальная ошибка не превышает величины $\delta(1 - e^{n\lambda h})(1 - e^{\lambda h})^{-1}$.

Доказательство: Пусть при вычислении x_1 получилась ошибка δ_1. При вычислении x_2 ошибка

$$|\delta_1| e^{\lambda h} + |\delta_2|,$$

где первый член в правой части - это ошибка в начальном условии, а второе слагаемое - новая ошибка вычисления на этом шаге. Далее, при вычислении x_3 ошибка будет

$$(|\delta_1| e^{\lambda h} + |\delta_2|) e^{\lambda h} + |\delta_3|,$$

и так далее. Наконец, если $|\delta_i| \leq \delta$, $i = 1, 2, ..., n$, то мы получаем для глобальной ошибки

$$|\delta_1|e^{n\lambda h} + |\delta_2|e^{(n-1)\lambda h} + ... + |\delta_n| \leq \delta\frac{1 - e^{n\lambda h}}{1 - e^{\lambda h}}.$$

Если все локальные ошибки имеют порядок $|\delta_i| = O(h^{m+1})$, то глобальная ошибка есть величина порядка $O(h^m)$.

8.12 Линейные дифференциальные уравнения и интегральное уравнение Вольтерра

Существует фундаментальная связь между линейными ОДУ и интегральным уравнением Вольтерра. Действительно, решение любого линейного ОДУ вида

$$\frac{d^n u}{dx^n} + a_1(x)\frac{d^{n-1}u}{dx^{n-1}} + ... + a_n(x)u = F(x) \tag{352}$$

с непрерывными коэффициентами при начальных условиях

$$u(0) = c_0, \quad u'(0) = c_1, \quad u^{(n-1)} = c_{n-1}, \tag{353}$$

может быть сведено к решению некоторого интегрального уравнения Вольтерра второго рода

$$\varphi(x) + \int_0^x K(x, y)\varphi(y)dy = f(x). \tag{354}$$

Продемонстрируем эту связь в деталях. Полагаем

$$D^n u = \frac{d^n u}{dx^n} = \varphi(x), \tag{355}$$

и далее последовательно получаем

$$D^{-1}\varphi(x) = \int_0^x \varphi(y)dy, \tag{356}$$

$$D^{-2}\varphi(x) = D^{-1}(D^{-1}\varphi(x)) = \int_0^x (x - y)\varphi(y)dy, \tag{357}$$

.....,

$$D^{-n}\varphi(x) = D^{-1}(D^{-n+1}\varphi(x)) = \frac{1}{(n-1)!}\int_0^x (x-y)^{n-1}\varphi(y)dy. \quad (358)$$

Принимая во внимание начальные условия, получаем

$$\frac{d^{n-1}u}{dx^{n-1}} = c_{n-1} + D^{-1}\varphi(x), \quad (359)$$

$$\frac{d^{n-2}u}{dx^{n-2}} = c_{n-1}x + c_{n-2} + D^{-2}\varphi(x), \quad (360)$$

.....,

$$u = c_{n-1}\frac{x^{n-1}}{(n-1)!} + c_{n-2}\frac{x^{n-2}}{(n-2)!} + ... + c_0 + D^{-n}\varphi(x). \quad (361)$$

Теперь просуммируем эти уравнения, добавив $D^n u = \frac{d^n u}{dx^n} = \varphi(x)$. В результате мы получим интегральное уравнение (354), в котором следует положить

$$K(x,y) = \sum_{m=1}^{n} a_m(x)\frac{(x-y)^{n-1}}{(n-1)!}, \quad (362)$$

$$f(x) = F(x) - c_{n-1}a_1(x) - (c_{n-1}x + c_{n-2})a_2(x) - \left(c_{n-1}\frac{x^{n-1}}{(n-1)!} + c_{n-2}\frac{x^{n-2}}{(n-2)!} + ... + c_0\right)a_n(x). \quad (363)$$

Таким образом, решая интегральное уравнение (354) с указанными функциями ядра $K(x,y)$ и неоднородного члена $f(x)$ и подставляя выражение, полученное для $\varphi(x)$, в уравнение (361) для u, мы получим единственное решение уравнения (352), удовлетворяющее указанным начальным данным.

Предположим, что в интегральном уравнениия Вольтерра второго рода

$$\varphi(x) = \int_0^x K(x,y)\varphi(y)dy + f(x) \quad (364)$$

$f(x) \in C([0,a])$ и ядро $K(x,y)$ непрерывно в замкнутом треугольнике $0 \le y \le x \le a > 0$. В таком случае $|K(x,y)| \le M$, и интегральный оператор

$$(Kf)(x) = \int_0^x K(x,y) f(y) dy \tag{365}$$

переводит $C([0,a])$ в $C([0,a])$. Определим последовательные приближения по формуле

$$\varphi^{(0)}(x) = f(x), \qquad \varphi^{(p)}(x) = K\varphi^{(p-1)}(x) + f = \sum_{k=0}^{p} K^k f, \quad p = 1, 2, \dots . \tag{366}$$

Итерация $K^P f \in C([0,a])$ и удовлетворяет оценке

$$|(K^p f)(x)| \le \|f\|_C \frac{(Mx)^p}{p!}, \qquad x \in [a,b], \qquad p = 0, 1, 2, \dots , \tag{367}$$

которая доказывается по методу индукции. Оценка верна при $p = 0$. Предполагая, что оценка верна для $p-1$, докажем ее для p:

$$\begin{aligned} |(K^p f)(x)| &= \left| \int_0^x K(x,y)(K^{p-1} f)(y) dy \right| \\ &\le M\|f\|_C M^{p-1} \int_0^x \frac{y^{p-1}}{(p-1)!} = \|f\|_C \frac{(Mx)^p}{p!}. \end{aligned} \tag{368}$$

Из оценки (367) вытекает, что ряд последовательных итераций (366) (ряд Неймана) мажорируется на $[0,a]$ сходящимся числовым рядом

$$\|f\|_C \sum_{k=0}^{\infty} \frac{(Ma)^k}{k!} = \|f\|_C e^{Ma}, \tag{369}$$

и потому сходится равномерно по x на $[0,a]$, определяя непрерывную функцию $\varphi(x)$. Таким образом, последовательные приближения $\varphi^{(p)}(x)$ при $p \to +\infty$ равномерно сходятся к функции $\varphi(x)$:

$$\varphi(x) = \lim_{p \to \infty} \varphi^{(p)}(x).$$

При этом справедлива оценка

$$\|\varphi(x)\|_C \leq \|f\|_C e^{Ma}.$$

Переходя к пределу при $p \to \infty$ в формулах ряда последовательных итераций (366) и пользуясь равномерной сходимостью $\varphi^{(p)}(x)$ к $\varphi(x)$ по x на $[0, a]$, заключаем, что построенная функция $\varphi(x)$ удовлетворяет интегральному уравнению единственным образом.

Для частного случая линейного ОДУ второго порядка

$$u'' + a_1(x)u' + a_2(x) = F(x), \qquad u(0) = c_0, \quad u'(0) = c_1, \qquad (370)$$

будем иметь

$$K(x, y) = a_1(x) + a_2(x)(x - y), \quad f(x) = F(x) - c_1 a_1(x) - (c_1 x + c_0)a_2(x), \qquad (371)$$

$$u(x) = c_0 + c_1 x + \int_0^x (x - y)\varphi(y)dy. \qquad (372)$$

Рассмотрим пример для начальной задачи Коши $u'' + u = 1$, $u(0) = 0$, $u'(0) = 1$, на отрезке $x \in [0, 10]$. На Рис. 8 для сравнения приводятся графики точного и прилиженного решений. Приближенное решение полученно с помощью интегрального уравнения Вольтерра для 50 итераций. Отчетливо видно хорошее совпадение обоих результатов.

8.13 ОДУ высокого порядка и системы ОДУ

Задача Коши для ОДУ n-го порядка с начальными условиями при $x = x_0$ выглядит так:

$$y^{(n)} = f(x, y, y', y'', ..., y^{(n-1)}), \qquad y = y_0, \quad y' = y'_0, ..., y^{(n-1)} = y_0^{(n-1)}.$$

Для системы ОДУ будем иметь:

$$\begin{cases} y'_1 = f_1(x, y_1, y_2, ..., y_n), \\ y'_2 = f_2(x, y_1, y_2, ..., y_n), \\ ....., \\ y'_n = f_n(x, y_1, y_2, ..., y_n), \end{cases}$$

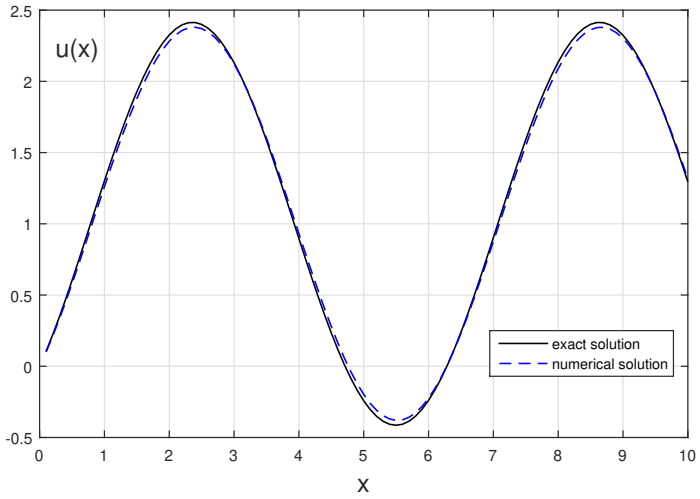


Рис. 8: Сравнение точного и прилиженного решений для начальной задачи $u'' + u = 1$, $u(0) = 0$, $u'(0) = 1$, на отрезке $x \in [0, 10]$.

с учетом начальных условий $y_1 = y_1^{(0)}, ..., y_n = y_n^{(0)}$ при $x = x_0$. ОДУ n-го порядка может быть преобразовано в систему из n уравнений первого порядка:

$$\begin{cases} y_1' = y_2, \\ y_2' = y_3, \\ ....., \\ y_n' = f_n(x, y_1, y_2, ..., y_n), \end{cases}$$

и при x_0 начальные условия аналогичны $y_1 = y_0, \quad y_2 = y_0', ..., \ y_n = y_0^{(n-1)}$.

Рассмотрим особый вид системы ОДУ

$$\begin{cases} \dot{x}_1(t) = f_1(x_1, x_2, ..., x_n), \\ \dot{x}_2(t) = f_2(x_1, x_2, ..., x_n), \\ ....., \\ \dot{x}_n(t) = f_n(x_1, x_2, ..., x_n), \end{cases}$$

и при $t = t_0$ мы имеем начальные значения $x_1 = x_1^{(0)}, ..., x_n = x_n^{(0)}$. Система в этой форме называется автономной. Этот класс систем ОДУ хорошо известен в механике, где $x_1, ..., x_n$ суть координаты

механической системы и t означает время. Система может быть записана в векторной форме

$$\dot{X} = F(X),$$

где неизвестный вектор X опрелеляется как $(x_1(t), ..., x_n(t))^T$. Одношаговая процедура Рунге-Кутта для системы уравнений первого порядка наиболее легко записывается в том случае, когда наша система является автономной:

$$X(t+h) = X(t) + \frac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4),$$

где

$$F_1 = hF(X), \qquad F_2 = hF(X + \frac{1}{2}F_1),$$

$$F_3 = hF(X + \frac{1}{2}F_2), \qquad F_4 = hF(X + F_3).$$

Многошаговые методы также применяются к системам уравнений. В качестве примера мы приводим векторную форму метода predictor-corrector Адамса Бэшфорта-Молтона для автономной системы

$$X^*(t+h) = X(t) + \frac{h}{720}[1901F(X(t)) -$$

$$2774F(X(t-h)) + 2616F(X(t-2h)) - 1274F(X(t-3h)) + 251(X(t-4h))],$$

$$X(t+h) = X(t) + \frac{h}{720}[251F^*(X(t+1)) +$$

$$646F(X(t)) - 264F(X(t-h)) + 106F(X(t-2h)) - 19(X(t-3h))].$$

Как и в случае ОДУ первого порядка, для запуска многошагового численного метода следует воспользоваться методом Рунге-Кутта четвертого порядка:

$$X(t_0 + h), \quad X(t_0 + 2h), \quad X(t_0 + 3h), \quad X(t_0 + 4h).$$

Вопросы для самоконтроля к главе 8

1. Как формулируется постановка задачи Коши для ОДУ первого порядка и теорема единственности?

2. Каковы основные детали метода Рунге-Кутта численного решения задачи Коши для ОДУ первого порядка?

3. В чем состоит основная идея линейных многошаговых методов численного решения задачи Коши для ОДУ первого порядка и каков соответствующий порядок аппроксимации?

4. Какова оценка глобальной ошибки численного решения задачи Коши для ОДУ первого порядка, полученного с помощью линейных многошаговых методов?

5. Каковы основные детали метода Рунге-Кутта численного решения задачи Коши для системы ОДУ первого порядка?

9 Краевые задачи ОДУ второго порядка

9.1 Постановка задачи

Многие часто встречаемые в приложениях проблемы математической физики сводятся к нахождению решения ОДУ второго порядка

$$Lu = -\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)u = f(x), \tag{373}$$

которые должны быть интегрированы на промежутке $a \leq x \leq b$ при выполнении граничных условий

$$h_1u'(a) - h_2u(a) = 0, \qquad H_1u'(b) + H_2u(b) = 0, \tag{374}$$

где h_1, h_2, H_1, H_2, являются константами. Это означает, что мы имеем дело с краевой (граничной) задачей для ОДУ второго порядка. Простейшей проблемой такого типа является задача статического прогиба проволоки, которая в состоянии равновесия занимает сегмент $a \leq x \leq b$ оси x. Элемент $(x, x + dx)$ находится под влиянием восстанавливающей силы $q(x)udx$ и внешней силы $f(x)dx$, а напряжения $p(x)$ и $p(x + dx)$ прикладываются к концам x и $x + dx$ элемента соответственно. В общем случае граничные условия (374) соответствуют упруго фиксированным концам проволоки. В частном случае, когда, например, $h_1 \neq 0$, $h_2 = 0$, конец $x = a$ проволоки свободен; если, однако, $h_1 = 0$, $h_2 \neq 0$, то его конец жестко закреплен.

Это уравнение также описывает малые по амплитуде стационарные продольные колебания упругого стержня. В этом случае $p = ES$ и $q = -\omega^2\rho S$, где $E(x)$ - модуль Юнга, $S(x)$ - квадрат поперечного сечения стержня, ω частота, ρ плотность, $h_1 = H_1 = E$ и $h_2 = H_2$ - коэффициент жесткости в соответствии с законом Гука.

Общий вид краевой задачи выглядит так:

$$y'' = f(x, y, y'), \qquad y(a) = A, \qquad y(b) = B, \qquad a \leq x \leq b. \tag{375}$$

В общем случае мы не можем гарантировать существование решения задачи (375), просто предполагая, что $f(x_1, x_2, x_3)$ имеет хорошие свойства. Однако для задачи

$$-y'' = f(x, y), \qquad y(0) = A, \qquad y(1) = B, \qquad 0 \leq x \leq 1, \tag{376}$$

мы можем сформулировать **теорему**: задача (376) имеет единственное решение в $C[0,1]$, если $f(x,y)$ удовлетворяет неравенству Липшица

$$|f(x,y_1)-f(x,y_2)| \le L|y_1-y_2|, \qquad 0<L<1,$$

в бесконечной полосе $0 \le x \le 1$, $-\infty < y < +\infty$.

Доказательство основано на интегральном представлении решения задачи (376) с помощью функции Грина (см. [6])

$$y(x) = c_1 + c_2 x + \int_0^1 g(x,t) f(t,y(t))dt = y_0(t) + \int_0^1 g(x,t) f(t,y(t))dt,$$

где c_1 и c_2 выбраны таким образом, чтобы $y_0(x) = c_1 + c_2 x$ удовлетворяет граничным условиям в (376), а $g(x,t)$ является функцией Грина

$$g(x,t) = \begin{cases} x(1-t), & 0<x<t<1, \\ t(1-x), & 0<t<x<1. \end{cases}$$

Построим решение интегрального уравнения с помощью итерационного процесса

$$y_{n+1}(x) = S(y_n(x)) = y_0(x) + \int_0^1 g(x,t) f(t,y_n(t))dt, \qquad n>0. \quad (377)$$

Покажем, что оператор $S(y)$ есть оператор сжатия (см. параграф 5.3.):

$$\begin{aligned} \|y_{n+1}(x) - y_{m+1}(x)\| &= \sup_{x\in[0,1]} |y_{n+1}(x) - y_{m+1}(x)| \\ &\le \int_0^1 |g(x,t)||f(t,y_n(t)) - f(t,y_m(t))|dt. \quad (378) \end{aligned}$$

Пусть функция $f(x,y)$ удовлетворяет условию Липшица (328) с константой L. Тогда получим

$$\begin{aligned} \|y_{n+1}(x) - y_{m+1}(x)\| &\le GL \int_0^1 |y_n(t) - y_m(t)|dt \\ &\le GL\|y_n(x) - y_m(x)\|, \qquad G = \sup_{(x,t)} |g(x,t)|. \quad (379) \end{aligned}$$

Если $q = GL < 1$, то $S(y)$ есть оператор сжатия. Следовательно его неподвижная точка $y = S(y)$ дает единственное решение рассматриваемой краевой задачи, и мы имеем следующую оценку (см. параграф 5.3.):

$$\|y_n(x) - y(x)\| \leq \frac{q^n}{1-q}\|y_1(x) - y_0(x)\|. \tag{380}$$

9.2 Метод стрельбы

Рассмотрим краевую задачу

$$y'' = f(x, y, y'), \qquad y(a) = A, \qquad y(b) = B, \qquad a \leq x \leq b. \tag{381}$$

Один естественный способ рассмотреть эту проблему - решить соответственную начальную задачу с предположением относительно соответствующего начального значения $y'(a)$. Тогда мы можем интегрировать уравнение для получения приближенного решения, надеясь, что $y(b) = B$. Если нет, то угаданное значение $y'(a)$ может быть изменино, и мы можем попробовать еще раз. Процесс называется **метод стрельбы**, и есть способы сделать это систематически в виде численного алгоритма.

Обозначим угадываемое значение $y'(a)$ как параметр z, так что соответствующая начальная задача имеет вид

$$y'' = f(x, y, y'), \qquad y(a) = A, \qquad y'(a) = z, \qquad a \leq x \leq b. \tag{382}$$

Решение этой проблемы есть функция $y = y(x, z)$. Цель состоит в том, чтобы выбрать z таким образом чтобы $y(b, z) = B$. Положим

$$\phi(z) = y(b, z) - B,$$

так что наша цель - найти такое z, при котором $\phi(z) = 0$, или просто решить уравнение $\phi(z) = 0$. Метод Ньютона здесь применим, и для итераций по z мы имеем

$$z_{n+1} = z_n - \frac{\phi(z_n)}{\phi'(z_n)}.$$

Для того чтобы определить $\phi'(z)$, мы вводим новую функцию

$$u(x,z) = \frac{\partial y}{\partial z}.$$

Дифференцируя по z все уравнения в (382), получим начальную задачу

$$u'' = f_y(x,y,y')u + f_{y'}(x,y,y')u', \qquad u(a) = 0, \qquad u'(a) = 1. \quad (383)$$

Последнее дифференциальное уравнение называется уравнением в вариациях. Его можно решить, например, многошаговым методом. Тогда мы получим

$$\phi'(z) = u(b,z).$$

И таким образом, это позволяет нам использовать метод Ньютона, чтобы найти корень для $\phi(z)$.

Соответствующий алгоритм работает следующим образом. Во-первых, получим решение $y = y(x,z_0)$ из

$$y'' = f(x,y,y'), \qquad y(a) = A, \qquad y'(a) = z_0, \qquad a \le x \le b.$$

Тогда, подставляя $y = y(x,z_0)$ и $y' = y'(x,z_0)$ в дифференциальное уравнение (383), мы решаем соответствующую начальную задачу с $u(a) = 0,\ u'(a) = 1$. Таким образом, мы имеем $\phi(z_0) = y(b,z_0) - B$ и $\phi'(z_0) = u(b,z_0)$. Следовательно, мы знаем $z_1 = z_0 - \phi(z_0)/\phi'(z_0)$. Теперь получим решение $y = y(x,z_1)$ из

$$y'' = f(x,y,y'), \qquad y(a) = A, \qquad y'(a) = z_1, \qquad a \le x \le b.$$

Подставляя $y = y(x,z_1)$ и $y' = y'(x,z_1)$ в дифференциальное уравнение (383), и мы вновь решаем соответствующую начальную задачу с $u(a) = 0,\ u'(a) = 1$. Таким образом, мы имеем $\phi(z_1) = y(b,z_1) - B$ и $\phi'(z_1) = u(b,z_1)$. Следовательно, мы знаем $z_2 = z_1 - \phi(z_1)/\phi'(z_1)$. Этот итерационный процесс следует продолжить далее, до тех пор, пока не будет выполнено условие $|\phi(z_n)| = |y(b,z_n) - B| < \epsilon$, где ϵ задается точностью численного расчета.

Существуют разные варианты метода стрельбы, называемые многократной прострелкой. Основная стратегия здесь состоит в

том, чтобы разделить данный интервал $[a, b]$ на отрезки и попытаться решить глобальную проблему по кусочкам. Опишем, что было бы сделано, если бы интервалы были разделены на две части, $[a, c]$ и $[c, b]$. На двух сегментах мы решаем начальные задачи для построения двух функций y_1 и y_2:

$$y'' = f(x, y, y'), \qquad y(a) = A, \qquad y'(a) = z_1, \qquad a \le x \le c,$$

$$y'' = f(x, y, y'), \qquad y(b) = B, \qquad y'(b) = z_2, \qquad c \le x \le b.$$

Обратите внимание, что z_1 и z_2 являются параметрами, находящимися в нашем распоряжении. Функцию y_1 требуется найти только на интервале $[a, c]$, и y_2 - только на интервале $[c, b]$. Параметры z_1 и z_2 теперь нужно итерировать до тех пор, пока кусочно-непррывная функция

$$y(x) = \begin{cases} y_1(x), & a < x < c, \\ y_2(x), & c < x < b, \end{cases}$$

не обеспечит решение проблемы. Таким образом, мы требуем непрерывности для $y(x)$ и $y'(x)$ в точке c: $y_1(c) - y_2(c) = 0$, $y_1'(c) - y_2'(c) = 0$. Эти два условия должны выполняться по завершению итерационного процесса нахождения z_1 и z_2. Это будет сделано методом Ньютона в размерности 2. Многократная прострелка с k сегментами будет включать в себя k функций.

9.3 Метод конечных разностей

Рассмотрим теперь краевую (граничную) задачу для линейного ОДУ второго порядка:

$$y'' + p(x)y' + q(x)y = f(x), \qquad y(a) = A, \qquad y(b) = B, \qquad a \le x \le b. \tag{384}$$

Другой подход к решению проблемы состоит в дискретизации интервала $x \in [a, b]$, за которым следует использование приближенных формул конечных разностей для производных

$$y'(x_i) = \frac{y_{i+1} - y_{i-1}}{2h} - \frac{h^2}{6} y'''(\xi_1^{(i)}),$$

$$y''(x_i) = \frac{y_{i+1} + y_{i-1} - 2y_i}{h^2} - \frac{h^2}{12} y^{(4)}(\xi_2^{(i)}), \qquad (385)$$

$$x_{i-1} < \xi_{1,2}^{(i)} < x_{i+1},$$

где $x_i = a+ih$, $i = 0, 1, 2, ..., n+1$, $h = (b-a)/(n+1)$. Таким образом, дискретный аналог (384) есть система линейных алгебраических уравнений

$$y_{i-1}(hp_i-2)+y_i(4-2q_ih^2)+y_{i+1}(-hp_i-2) = -2f_ih^2, \quad y_0 = A, \quad y_{n+1} = B, \qquad (386)$$

где $p_i = p(x_i)$, $q_i = q(x_i)$, $f_i = f(x_i)$. Например, если $n = 4$, соответствующая матричная форма дается формулой

$$\begin{pmatrix} 4-2q_1h^2 & -hp_1-2 & 0 & 0 \\ hp_2-2 & 4-2q_2h^2 & -hp_2-2 & 0 \\ 0 & hp_3-2 & 4-2q_3h^2 & -hp_3-2 \\ 0 & 0 & hp_4-2 & 4-2q_4h^2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \end{pmatrix}$$

$$F_1 = -2f_1h^2 - A(hp_1 - 2), \qquad F_2 = -2f_2h^2,$$
$$F_3 = -2f_3h^2, \qquad F_4 = -2f_4h^2 + B(hp_4 + 2).$$

Эта система тридиагональна и может быть решена специальным гауссовским алгоритмом. В общем случае мы имеем $n \times n$ систему $TY = F$, где T является трехдиагональной $n \times n$ матрицей, $Y = (y_1, ..., y_n)^T$ и $F = (F_1, ..., F_n)^T$.

В анализе ошибок с использованием приближений (385), подставляя $y_i = \bar{y}(x_i) + e_i$, $i = 1, ..., n$, в (386), где $(e_1, ..., e_n)$ есть вектор ошибок, а $\bar{y}(x)$ предполагается быть точным решением, мы получаем систему линейных алгебраических уравнений

$$e_{i-1}(hp_i - 2) + e_i(4 - 2q_ih^2) + e_{i+1}(-hp_i - 2) = 2h^2(\frac{h^2}{12}\bar{y}^{(4)}(\xi_2^{(i)}) + p_i\frac{h^2}{6}\bar{y}^{(3)}(\xi_1^{(i)})), \qquad (387)$$

$$e_0 = 0, \qquad e_{n+1} = 0,$$

которая может быть представлена в компактной форме

$$Te = h^4G$$

с вектором $G = (g_1, ..., g_n)^T$ в правой части, где

$$g_i = 2(\frac{1}{12}\bar{y}^{(4)}(\xi_2^{(i)}) + p_i\frac{1}{6}\bar{y}^{(3)}(\xi_1^{(i)})), \qquad i = 1, 2, ..., n.$$

Если $\det T \neq 0$, мы имеем

$$e = h^4 T^{-1} G,$$

и поэтому,

$$||e|| \leq h^4 ||T^{-1}|| \cdot ||G||.$$

Таким образом, если $h \to 0$, тогда $||e|| \to 0$.

9.4 Одномерный фотонный кристалл

Рассмотрим теперь задачу для линейного ОДУ второго порядка с периодическим коэффициентом:

$$y'' + (\frac{\omega}{c})^2 q(x) y = 0, \qquad q(x + d) = q(x), \tag{388}$$

где ω есть частота, c - скорость света, а периодическая функция $q(x)$ с периодом d есть квадрат индекса рефракции (коэффициента преломления). Эта задача представляет собой процесс распространения волн для одномерного фотонного кристалла, описываемого уравнением Гельмгольца. В качестве функции y может быть одна из компонент электрического или магнитного полей. Мы ищем решения Блоха-Флоке, удовлетворяющие условию квази-периодичности

$$y(x + d) = e^{ikd} y(x), \qquad -\frac{\pi}{d} < k < \frac{\pi}{d}, \tag{389}$$

где k является квази-импульсом. Пусть для простоты изложения $c = 1$.

Вновь воспользуемся приближенными формулами конечных разностей для производных (385), где $x_i = ih$, $i = 0, 1, 2, ..., n + 1$, $h = d/(n + 1)$. Таким образом, учитывая (389), представим задачу

(388) в виде однородной системы линейных алгебраических уравнений

$$\begin{gathered} -2y_0 + y_1(4 - 2\omega^2 q_1 h^2) - 2y_2 = 0, \\ -2y_{i-1} + y_i(4 - 2\omega^2 q_i h^2) - 2y_{i+1} = 0, \qquad i = 2, ..., n, \\ -2y_n + y_{n+1}(4 - 2\omega^2 q_{n+1} h^2) - 2y_{n+2} = 0, \end{gathered} \tag{390}$$

где $q_i = q(x_i)$, $y_0 = \exp(-ikd)y_{n+1}$, $y_{n+2} = \exp(ikd)y_1$. Например, если $n = 4$, соответствующая матричная форма дается формулой $T(\omega, k)Y = 0$, где

$$T(\omega, k) = \begin{pmatrix} Q_1 & -2 & 0 & 0 & -2e^{-ikd} \\ -2 & Q_2 & -2 & 0 & 0 \\ 0 & -2 & Q_3 & -2 & 0 \\ 0 & 0 & -2 & Q_4 & -2 \\ -2e^{ikd} & 0 & 0 & -2 & Q_5 \end{pmatrix}, \tag{391}$$

и мы ввели обозначение $Q_i = 4 - 2\omega^2 q_i h^2$. Определитель эрмитовой матрицы $T(\omega, k)$ есть вещественная величина. Система $T(\omega, k)Y = 0$ имеет нетривиальное решение, если $\det(T(\omega, k)) = 0$ - так называемое дисперсионное уравнение. Разрешая это уравнение относительно ω, получаем дисперсионные кривые $\omega = \omega_s(k)$, $s = 1, 2, 3, ...$, описывающие зонный спектр задачи. Следовательно, в одномерном фотонном кристалле могут распространяться волны только с теми частотами ω, которые подчиняются закону дисперсии $\omega = \omega_s(k)$, $s = 1, 2, 3, ...$. Так как дисперсионные кривые в нашем случае суть четные функции параметра k, то достаточно рассмотреть интервал $0 < k < \pi/d$. Рассмотрим пример:

$$q(x) = \sin(\frac{2\pi}{d}x) + q_0, \quad q_0 > 0.$$

Пусть $d = 2$, $q_0 = 2$. На Рис. 9 для этого случая демонстрируется график зависимости $\det(T(\omega, k))$ для фиксированного значения $k = \pi/4$. Отчетливо видны четыре нуля для функции $\det(T(\omega, k))$. На Рис. 10 показаны четыре первые нижние дисперсионные зависимости $\omega = \omega_s(k)$, $s = 1, 2, 3, 4$. Между второй и третьей кривыми отчетливо виден горизонтальный "band gap". Первая нижняя дисперсионная зависимость называется акустической, следующая - оптическая ветвь.

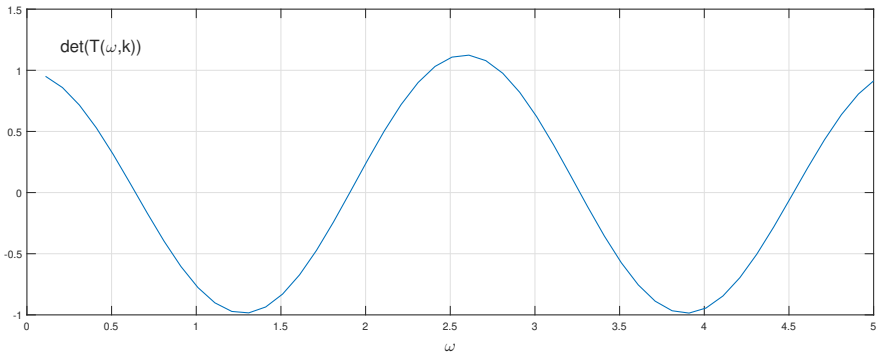


Рис. 9: График зависимости $\det(T(\omega, k))$ для фиксированного значения $k = \pi/4$ и $q(x) = \sin(\pi x) + 2$.

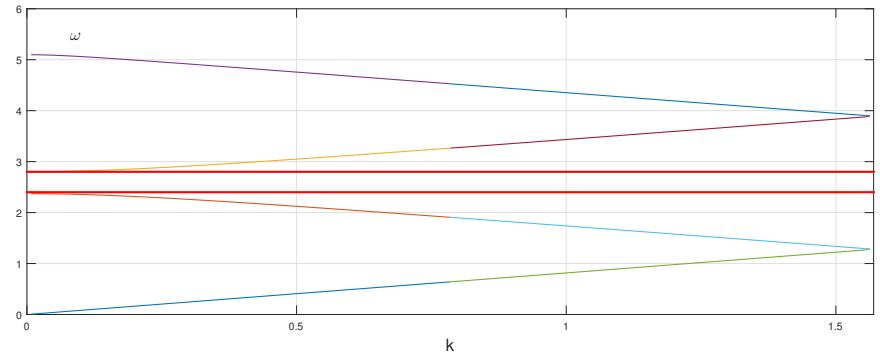


Рис. 10: Дисперсионные зависимости $\omega = \omega_s(k)$, $s = 1, 2, 3, 4$, одномерного фотонного кристалла для $q(x) = \sin(\pi x) + 2$.

Рассмотрим другой пример из квантовой механики - движение электрона в кулоновском центральном поле с потенциальной энергией

$$U(r) = -\frac{Ze^2}{r}, \tag{392}$$

где e - заряд электрона. Эта задача имеет большое значение в теории атома водорода ($Z = 1$) и других многократно ионизированных атомов (He^+, Li^{++}), и так далее. Стационарные состояния движение электрона в результате разделения переменных для трехмерного уравнения Шредингера в сферических координатах можно представить в виде (см. [18], [19])

$$\psi(\bar{r}) = f(r)Y_{lm}(\theta, \varphi), \tag{393}$$

где Y_{lm} обозначают сферические функции с орбитальным квантовым числом $l = 0, 1, 2, ...$ ($m = -l, -l+1, ..., 0, ..., l-1, l$), а радиальная волновая функция $R(r) = rf(r)$ удовлетворяет радиальному уравнению Шредингера

$$\frac{d^2R}{dr^2} + \left(\frac{2mE}{\hbar^2} + \frac{2mZe^2}{\hbar^2 r} - \frac{l(l+1)}{r^2}\right)R = 0, \tag{394}$$

где m - масса электрона. Переходя к безразмерным величинам с помощью боровского радиуса a

$$r' = \frac{r}{a}, \qquad E' = \frac{E}{E_a}, \qquad a = \frac{\hbar^2}{me^2} \sim 5.29210^{-9}\text{cm} \tag{395}$$

$$E_a = \frac{e^2}{a} = \frac{me^4}{\hbar^2} = 27.21eV, \tag{396}$$

получим

$$\frac{d^2R}{dr'^2} + \left(2E' + \frac{2Z}{r'} - \frac{l(l+1)}{r'^2}\right)R = 0. \tag{397}$$

Ниже мы будем пользоваться (397) с безразмерными величинами r и E, опустив штрихи. Для этого уравнения мы рассмотрим приближенную спектральную задачу на дискретный спектр $U_{min} < \{E_{kl}\} < 0$, $k = 0, 1, 2, ...$, на отрезке $r \in [0, r_{max}]$, вместо

$r \in [0, \infty]$, с учетом краевых условий $R(0) = R(r_{max}) = 0$, где значение r_{max} следует выбрать достаточно большим. Здесь для $l > 0$ оценка U_{min} есть минимальное значение эффективного потенциала

$$U_{eff}(r) = -\frac{2Z}{r} + \frac{l(l+1)}{r^2}. \tag{398}$$

Собственные фунции R_{kl} этой спектральной задачи для дискретного множества $\{E_{kl}\}$ экспоненциально убывают на бесконечности $r \to \infty$, и таким образом $R_{kl} \in L_2(0, \infty)$.

Снова воспользуемся приближенными формулами конечных разностей для производных в точках $r_i = ih$, $i = 0, 1, 2, ..., n+1$, $h = r_{max}/(n+1)$. Таким образом, дискретный аналог (397) есть система линейных алгебраических уравнений, представляющей собой матричную задачу на собственные значения и векторы, а именно:

$$-2R_{i-1} + R_i(4 - 2q_i h^2 - 4Eh^2) - 2R_{i+1} = 0, \qquad R_0 = 0, \qquad R_{n+1} = 0, \tag{399}$$

где

$$q_i = \frac{2Z}{r_i} - \frac{l(l+1)}{r_i^2}.$$

Например, соответствующая матричная форма

$$\hat{Q}R = 4Eh^2 R \tag{400}$$

для трехдиагональной матрицы $\hat{Q}$ с отличными от неля элементами $\hat{Q}_{ii} = 4 - 2q_i h^2$, $\hat{Q}_{i,i+1} = \hat{Q}_{i,i-1} = -2$, и $R = (R_1, R_2, ..., R_n)^T$, если $n = 4$, дается формулой

$$\begin{pmatrix} 4 - 2q_1h^2 & -2 & 0 & 0 \\ -2 & 4 - 2q_2h^2 & -2 & 0 \\ 0 & -2 & 4 - 2q_3h^2 & -2 \\ 0 & 0 & -2 & 4 - 2q_4h^2 \end{pmatrix} \begin{pmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \end{pmatrix} = 4Eh^2 \begin{pmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \end{pmatrix}. \tag{401}$$

Следует сказать, что спектральная задача (397) для кулоновского потенциала для $r \in (0, \infty)$ имеет точное решение

$$E_{kl} = -\frac{Z^2}{2(k+l+1)^2}, \tag{402}$$

а радиальные функции выражаются через полиномы Лагерра.

На Рис.11 показаны графики зависимости функции плотности вероятности $|R_{kl}(r)|^2$ электрона для атома водорода ($Z = 1$) от нормированного расстояния для значений $k = 0, 1, 2$: (a) - состояния с $l = 0$, (b) - состояния с $l = 1$. Результаты расчетов получены на основе решения спектральной матричной задачи (399), полученной с помощью применения метода конечных разностей для задачи (397), полагая $r_{max} = 50,\ n = 500$. При этом самые нижние собственные значения матрицы $\hat{Q}$ дают нижние значения спектра энергии E_{kl}. В приведенном численном примере мы получаем для состояний с $l = 0$ приближенные значения спектра энергии электрона для атома водорода:

$$E_{00} = -13.5711eV, \quad E_{10} = -3.3991eV, \quad E_{20} = -1.5112eV.$$

При этом точные значения таковы:

$$E_{00} = -13.6050eV, \quad E_{10} = -3.4012eV, \quad E_{20} = -1.5117eV.$$

Аналогичным образом, для состояний с $l = 1$ получены приближенные значения спектра энергии

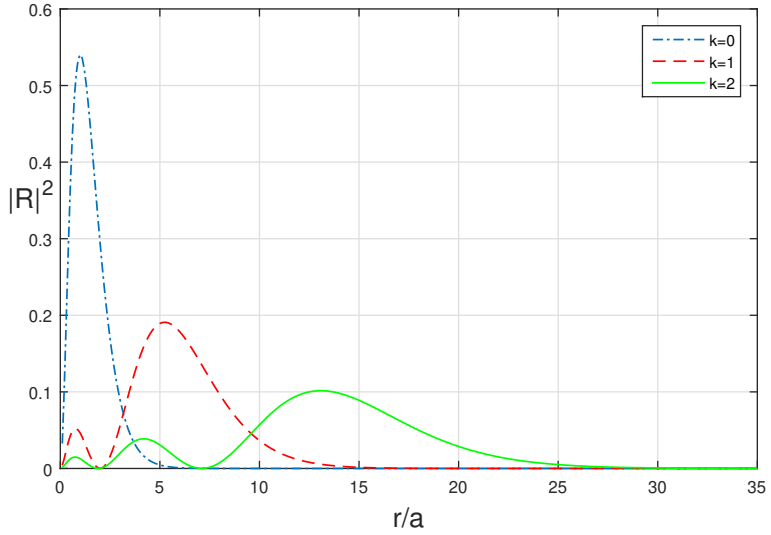
$$E_{01} = -3.4019eV, \quad E_{11} = -1.5119eV, \quad E_{21} = -0.8496eV,$$

которые очень близки к соответствующим точным значениям

$$E_{01} = -3.4012eV, \quad E_{11} = -1.5117eV, \quad E_{21} = -0.8503eV.$$

В заключение следует отметить, что рассмотренный кратко метод конечных разностей для задачи (397) выглядит достаточно простым, быстрым, стабильным и точным. Точное решение для убывающих на бесконечности потенциалов центрального поля существует только для кулоновского потенциала, а метод конечных разностей в этом случае всегда предоставит эффектиное приближенное решение.

(a)



(b)

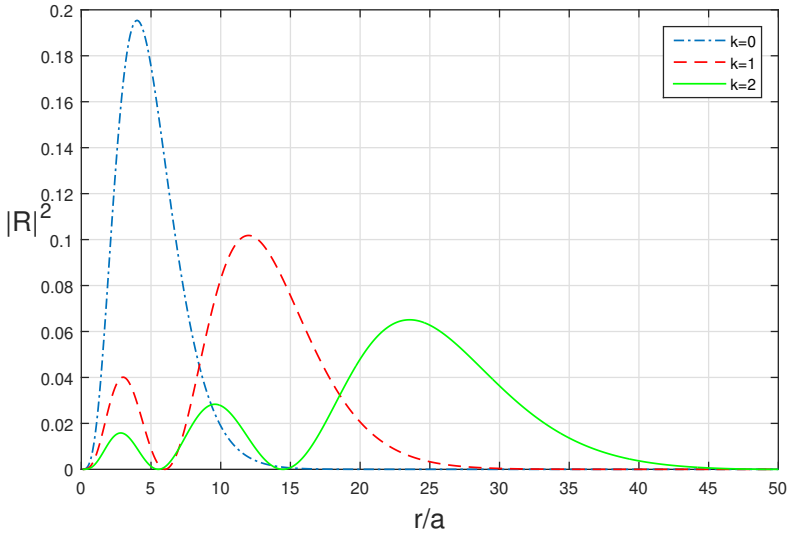


Рис. 11: Функция плотности вероятности $|R_{kl}(r)|^2$ электрона для атома водорода ($Z = 1$) для значений $k = 0, 1, 2$: (a) - состояния с $l = 0$ (b) - состояния с $l = 1$.

9.5 Метод Ритца (случай 1D)

Методы Ритца, Галеркина, наименьших квадратов широко используются при решении граничных задач для дифференциальных

уравнений, как обыкновенных, так и в частных производных. Предположим, что мы решаем следующую задачу:

$$Au(x) = f(x), \tag{403}$$

в которой A - линейный оператор

$$Au = -\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)u = f(x),$$

$f(x) \in L_2[a,b]$ - заданная функция, и $u(x) \in C^2[a,b]$ - неизвестная функция, определяемая из уравнения (403) и граничных условий

$$u(a) = 0, \qquad u(b) = 0.$$

Предположим, что выполняются неравенства $p(x) \geq p_0 > 0$, $q(x) \geq 0$ для $x \in [a,b]$. Это означает, что оператор A симметричен и положительно определен в вещественном гильбертовом пространстве $L_2[a,b]$, поскольку, используя интегрирование по частям для скалярного произведения в $L_2[a,b]$, мы получаем

$$< Au, u >= \int_a^b Au \cdot u dx = -\int_a^b (pu')' u dx + \int_a^b qu^2 dx$$
$$= \int_a^b p(u')^2 dx + \int_a^b qu^2 dx \geq 0 \tag{404}$$

для произвольных u. Скалярное произведение $< Au, u >= \|u\|_A^2$ называется энергетической нормой элемента u относительно оператора A (энергия элемента).

Рассмотрим линейный квадратичный функционал

$$J(u) =< Au, u > -2 < f, u >= \int_a^b p(u')^2 dx + \int_a^b qu^2 dx - 2\int_a^b fu dx. \tag{405}$$

Теорема утверждает, что если уравнение $Au(x) = f(x)$ имеет решение $u_0(x)$, это решение минимизирует функционал $J(u)$. Обратно, если существует элемент u_0, который минимизирует функционал

$J(u)$, то этот элемент удовлетворяет уравнению $Au(x) = f(x)$. Доказательство хорошо известно (см. [9], том 4(1)), и основано на возможности сведения функционала $J(u)$ к форме

$$J(u) =< A(u-u_0), u-u_0 > - < Au_0, u_0 >= \|u-u_0\|_A^2 - \|u_0\|_A^2,$$

и анализа минимума функции $g(t)$ по параметру t

$$g(t) = J(u_0(x) + tv(x)) =< Au_0, u_0 > +2t < Au_0, v > +t^2 < Av, v >$$
$$-2 < f, u_0 > -2t < f, v > .$$

Основная идея метода Ритца заключается в замене граничной задачи для $Au(x) = f(x)$ проблемой минимизации интегрального функционала $J(u)$, то есть нахождения функции u, дающей минимум функционала. Предположим, что мы выбираем базисные функции

$$u_1, u_2, ..., u_n, ... \in L_2[a, b].$$

Следовательно, мы ищем решение в виде

$$u_n(x) = \sum_{m=1}^{n} c_m u_m(x), \tag{406}$$

где $c_1, c_2, ..., c_n$, суть неизвестные константы, подлежащие определению. Подставляя эту сумму в функционал $J(u)$, мы получаем новый функционал

$$J(u_n) = J(c_1, c_2, ..., c_n) = \sum_{m,k=1}^{n} c_m c_k < Au_m, u_k > -2\sum_{m=1}^{n} c_m < f, u_m >, \tag{407}$$

который является квадратичной формой по отношению к неизвестным константам $c_1, c_2, ..., c_n$. Для поиска минимального элемента $J(u_n) = J(c_1, c_2, ..., c_n)$ мы требуем, чтобы выполнялись следующие условия:

$$\frac{\partial}{\partial c_m} J(c_1, c_2, ..., c_n) = 0, \qquad m = 1, 2, ..., n.$$

Таким образом, мы приходим к линейной $n \times n$ системе алгебраических уравнений для неизвестных констант $c_1, c_2, ..., c_n,$, то есть

$$\sum_{m=1}^{n} c_m < Au_m, u_k >=< f, u_k >, \qquad k = 1, 2, ..., n, \tag{408}$$

которая называется системой Ритца. После того как мы найдем неизвестные константы $c_1, c_2, ..., c_n$, выражение (406) дает нам приближенное решение.

Рассмотрим следующий простой пример:

$$-u'' = \cos x, \qquad u(0) = 0, \qquad u(\pi) = 0.$$

Эта граничная задача имеет точное решение $u = \cos x + 2x/\pi - 1$. Решим эту проблему, используя метод Ритца. Сначала выберем базисные функции:

$$\sin 2x, \quad \sin 4x, ..., \quad \sin 2nx.$$

Здесь мы учитываем, что решение нечетно по отношению к точке $x = \pi/2$. Следовательно, мы ищем решение в виде

$$u_n(x) = \sum_{m=1}^{n} c_m \sin 2mx.$$

Для матричных элементов $< Au_m, u_k >$ в системе (408) мы получаем

$$< A\varphi_m, \varphi_k >= \int_0^\pi 4mk \cos 2mx \cos 2kx dx = \begin{cases} 0, & m \neq k, \\ 2\pi m^2, & m = k. \end{cases}$$

Затем, для вектора правой части

$$< f, \varphi_m >= \int_0^\pi \cos x \sin 2m dx = \frac{1}{2} \int_0^\pi \sin(2m-1)x dx +$$

$$\frac{1}{2} \int_0^\pi \sin(2m+1)x dx = \frac{4m}{4m^2-1}.$$

Следовательно, система Ритца с диагональной матрицей определяется формулой

$$\begin{pmatrix} 2\pi 1^2 & 0 & ... & 0 \\ 0 & 2\pi 2^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & 0 & 2\pi n^2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ ... \\ c_n \end{pmatrix} = \begin{pmatrix} \frac{4}{4\cdot 1^2-1} \\ \frac{4\cdot 2}{4\cdot 2^2-1} \\ ... \\ \frac{4n}{4\cdot n^2-1} \end{pmatrix}.$$

Таким образом, мы сразу получаем ответ

$$c_m = \frac{2}{\pi m(4m^2-1)}, \qquad m = 1, 2, ...,$$

и приближение к решению дается формулой

$$u_n = \frac{2}{\pi} \sum_{m=1}^{n} \frac{\sin 2mx}{m(4m^2-1)}.$$

В случае $n \to +\infty$ мы получим бесконечный ряд Фурье, который сходится к точному решению $u = \cos x + 2x/\pi - 1$ в вещественном гильбертовом пространстве $L_2[a, b]$.

Вопросы для самоконтроля к главе 9

1. Как формулируется постановка краевой задачи для ОДУ второго порядка и теорема единственности?

2. Каковы основные детали метода стрельбы для построения численного решения краевой задачи для ОДУ второго порядка?

3. В чем состоит основная идея применения метода конечных разностей численного решения краевой задачи для ОДУ второго порядка и каков соответствующий порядок аппроксимации?

4. Как получается дисперсионная диаграмма одномерного фотонного кристалла?

5. В чем состоит основная трудность нахождения численного решения задачи определения спектра атома водорода?

6. Какова основная идея использования метода метода Ритца для построения численного решения краевой задачи для ОДУ второго порядка?

10 Краевые и начально-краевые задачи уравнений в частных производных второго порядка

10.1 Введение в краевые и начально-краевые задачи уравнений в частных производных второго порядка

Многие важные проблемы математической физики сводятся к квазилинейным уравнениям в частных производных второго порядка вида

$$A\frac{\partial^2 u}{\partial x^2} + 2B\frac{\partial^2 u}{\partial x \partial y} + C\frac{\partial^2 u}{\partial y^2} + f(x, u, \nabla u) = 0. \tag{409}$$

В зависимости от постоянных коэффициентов A, B, C эти уравнения делятся на три категории. Если $B^2 - AC < 0$, то это уравнение эллиптического типа. Для неравенства $B^2 - AC > 0$ рассматриваемое уравнение принадлежит гиперболическому типу. И если $B^2 - AC = 0$, то это уравнение параболического типа.

В следующем разделе мы подробно рассмотрим уравнение теплопроводности (параболический тип)

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + f(x, t), \tag{410}$$

где u есть температура, которая зависит от координаты x и времени t, D является коэффициентом теплопроводностьи.

Другим примером является волновое уравнение (гиперболический тип)

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2} + f(x, t), \tag{411}$$

которое описывает малой амплитуды колебания струны. Здесь c - скорость распространения волн. В многомерном пространстве координат в эти два уравнения входит оператор Лапласа

$$\frac{\partial u}{\partial t} = D\Delta u + f(x, t), \qquad \frac{\partial^2 u}{\partial t^2} = c^2\Delta u + f(x, t),$$

который в случае двумерного координатного пространства задается формулой

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = u_{xx} + u_{yy}.$$

Без временной зависимости оба уравнения переходят в уравнение Пуассона, которое хорошо известно в теории электростатики и магнитостатики,

$$\Delta u = -f(x, y), \tag{412}$$

где u является функцией (x, y). Соответствующее однородное уравнение называется уравнением Лапласа

$$\Delta u = 0. \tag{413}$$

В следующем разделе мы обратим особое внимание на граничную задачу для (412), рассматриваемую внутри области Ω на плоскости X, Y, где на решение u накладываются одно из граничных условий

$$u|_{\partial\Omega} = \phi, \quad \left.\frac{\partial u}{\partial n}\right|_{\partial\Omega} = \psi, \tag{414}$$

граничных условий Дирихле или Неймана, соответственно, где обозначение $\partial\Omega$ означает границу области Ω.

10.2 Задача Дирихле для уравнений типа Пуассона

Внутренняя задача Дирихле для уравнений Пуассона формулируется следующим образом:

$$\begin{cases} u_{xx} + u_{yy} = f(x, y), & \text{in} \quad \Omega, \\ u|_{\partial\Omega} = \varphi(x, y), & \text{on} \quad \partial\Omega. \end{cases} \tag{415}$$

Эта проблема имеет единственное решение, если область Ω имеет достаточно гладкую границу, а f и φ являются непрерывными функциями (см. [6], [9](том 4(2))). Мы будем решать эту задачу

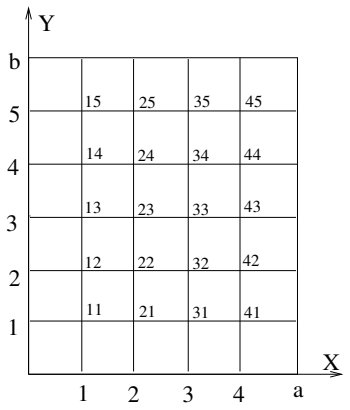


Рис. 12: Задача Дирихле на прямоугольнике $0 < x < a, \quad 0 < y < b$

численно с помощью метода конечных разностей. Чтобы проиллюстрировать этот метод, мы рассмотрим задачу Дирихле на прямоугольнике $0 < x < a, \quad 0 < y < b$ (см. Рис. 12).

Такой подход к численному решению задачи (415) использует конечно-разностные аппроксимации производных

$$y''(x) = \frac{1}{h^2}[y(x+h)+y(x-h)-2y(x)] - \frac{h^2}{12}y^{(4)}(\xi), \quad \xi \in (x-h, x+h). \tag{416}$$

Прежде всего, в прямоугольнике установим сеть узловых точек для $\bar{\Omega}$:

$$(x_i, y_j) = (ih_1, jh_2), \qquad 0 \le i \le n+1, \qquad 0 \le j \le m+1,$$

$$h_1 = \frac{a}{n+1}, \qquad h_2 = \frac{b}{m+1}.$$

Далее, дифференциальное уравнение в (415) в точках сетки (x_i, y_j) заменяется его конечно-разностным аналогом:

$$\frac{1}{h_1^2}[u_{i-1,j}+u_{i+1,j}-2u_{i,j}]+\frac{1}{h_2^2}[u_{i,j-1}+u_{i,j+1}-2u_{i,j}] = f_{ij}, \quad f_{ij} = f(x_i, y_j),$$

или

$$u_{i-1,j} + u_{i+1,j} - 2u_{i,j} + \alpha[u_{i,j-1} + u_{i,j+1} - 2u_{i,j}] = h^2 f_{ij}, \tag{417}$$

где $\alpha = (h_1/h_2)^2$ и $h = h_1$. Здесь значения $u_{i,j}$, как решение дискретной задачи (конечно-разностного уравнения), представляют собой приближение для точного решения непрерывной задачи $u(x_i, y_j)$.

Заметим, что значения $u_{i,j}$ известны, когда $i = 0$ или $n + 1$, и $j = 0$ или $m+1$, так как это заданные граничные значения в задаче (415) в точках границы области Ω:

$$u_{0,j} = \phi(x_0, y_j) = \phi(0, y_j), \qquad u_{i,0} = \phi(x_i, y_0) = \phi(x_i, 0), \tag{418}$$

$$u_{n+1,j} = \phi(x_{n+1}, y_j) = \phi(a, y_j), \qquad u_{i,m+1} = \phi(x_i, y_{m+1}) = \phi(x_i, b). \tag{419}$$

Таким образом, это означает, что мы будем решать неоднородную систему линейных алгебраических уравнений с неизвестными $u_{i,j}$ и $1 \le i \le n$, $1 \le j \le m$. Для простоты рассмотрим случай для $n = 2$ и $m = 3$. Тогда мы получим 6×6 систему

$$\begin{cases} u_{01} - 2u_{11} + u_{21} + \alpha[u_{10} - 2u_{11} + u_{12}] = h^2 f_{11}, \\ u_{02} - 2u_{12} + u_{22} + \alpha[u_{11} - 2u_{12} + u_{13}] = h^2 f_{12}, \\ u_{03} - 2u_{13} + u_{23} + \alpha[u_{12} - 2u_{13} + u_{14}] = h^2 f_{13}, \\ u_{11} - 2u_{21} + u_{31} + \alpha[u_{20} - 2u_{21} + u_{22}] = h^2 f_{21}, \\ u_{12} - 2u_{22} + u_{32} + \alpha[u_{21} - 2u_{22} + u_{23}] = h^2 f_{22}, \\ u_{13} - 2u_{23} + u_{33} + \alpha[u_{22} - 2u_{23} + u_{24}] = h^2 f_{23}. \end{cases} \tag{420}$$

Неизвестные величины в этой задаче можно упорядочить поразному. Мы выбираем способ, который известен как естественный порядок:

$$U = [u_{11}, u_{12}, u_{13}, u_{21}, u_{22}, u_{23}]^T.$$

Система имеет вид

$$AU = F, \tag{421}$$

где

$$A = \begin{pmatrix} -2(1+\alpha) & \alpha & 0 & 1 & 0 & 0 \\ \alpha & -2(1+\alpha) & \alpha & 0 & 1 & 0 \\ 0 & \alpha & -2(1+\alpha) & 0 & 0 & 1 \\ 1 & 0 & 0 & -2(1+\alpha) & \alpha & 0 \\ 0 & 1 & 0 & \alpha & -2(1+\alpha) & \alpha \\ 0 & 0 & 1 & 0 & \alpha & -2(1+\alpha) \end{pmatrix},$$

$$F = \begin{pmatrix} h^2 f_{11} - u_{01} - \alpha u_{10} \\ h^2 f_{12} - u_{02} \\ h^2 f_{13} - u_{03} - \alpha u_{14} \\ h^2 f_{21} - u_{31} - \alpha u_{20} \\ h^2 f_{22} - u_{32} \\ h^2 f_{23} - u_{33} - \alpha u_{24} \end{pmatrix},$$

где в выражениях компонент вектора F присутствуют только граничные значения $u_{i,j}$. В общем случае $nm \times nm$ система имеет большое число нулей, потому что каждое уравнение будет содержать не более пяти неизвестных. Структура матрицы A такова. Это квадратная матрица, которая состоит из $n \times n$ квадратных блоков. Каждый блок представляет собой матрицу размера $m \times m$. Диагональные блоки размером $m \times m$ описываются так:

$$\begin{pmatrix} -2(1+\alpha) & \alpha & 0 & 0 \\ \alpha & -2(1+\alpha) & \alpha & 0 \\ 0 & \alpha & -2(1+\alpha) & \alpha \\ 0 & 0 & \alpha & ..... \end{pmatrix}.$$

Ниже и выше главной блочной диагонали мы имеем две диагонали из блоков единичной матрицы размером $m \times m$. Остальные блоки матрицы A совпадают с нулевыми матрицами $m \times m$.

Существует другой, более простой способ решения конечно-разностной схемы (417-419). Это итеративная процедура подобна итерационному методу Зайделя. Следующая версия итерационного метода очень эффективна в практике численного анализа (следует из формулы (417)):

$$(\frac{2}{h_1^2} + \frac{2}{h_2^2})u_{ij} = \frac{u_{i-1,j} + u_{i+1,j}}{h_1^2} + \frac{u_{i,j-1} + u_{i,j+1}}{h_2^2} - f_{ij}. \qquad (422)$$

Это уравнение используется для обновления u_{ij}. Когда это уравнение используется, значение, полученное из правой части, заменяет старое значение u_{ij} на новое. Важно, чтобы в этом обновлении только внутренние точки сетки области Ω были задействованы. Для начальной итерации внутренних точек сетки можно взять нули. Заметим, что матрица A не сохраняется в памяти компьютера, так как итеративный метод не требует этого. Таким образом, в алгоритме множество значений $u_{i,j}$ использует весь $(n+1) \times (m+1)$ массив элементов.

Рассмотрим пример решения краевой задачи для уравнения Лапласа

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad (x,y) \in \Omega, \qquad u|_{\partial\Omega} = \phi, \quad (x,y) \in \partial\Omega,$$

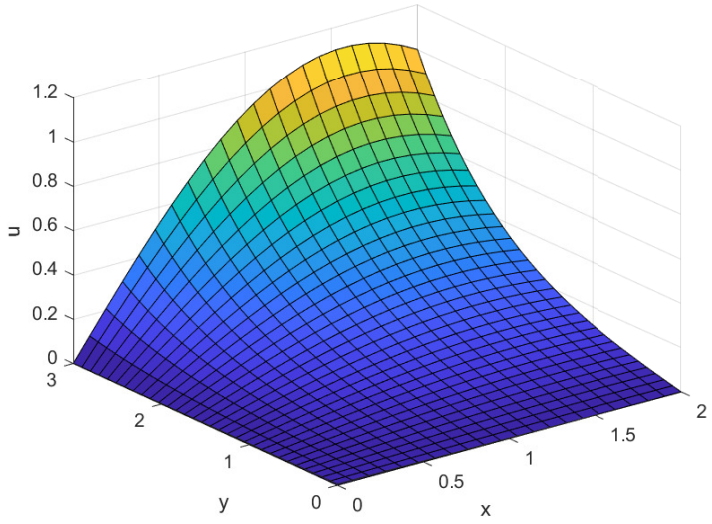


Рис. 13: Точное решение краевой задачи для уравнения Лапласа в прямоугольнике $a = 2, b = 3$.

используя следующие граничные условия:

$$\phi(x, 0) = 0, \quad \phi(x, b) = \sin(x)/\sin(a), \quad x \in [0, a],$$

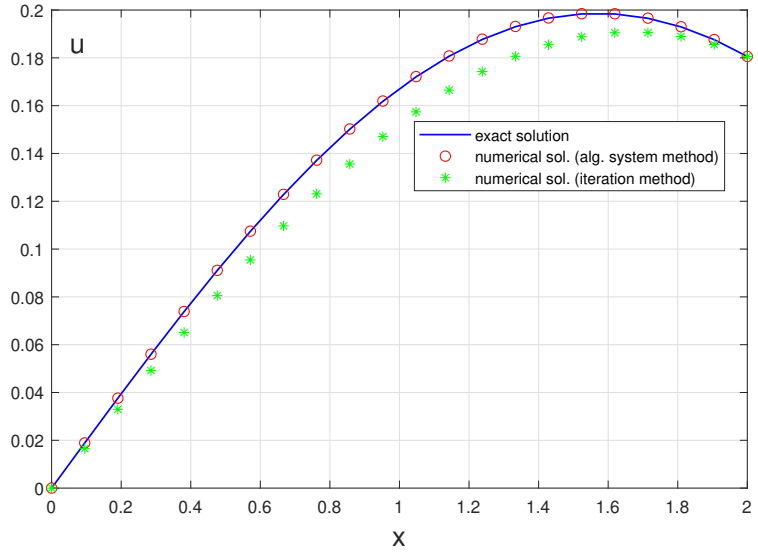
$$\phi(0, y) = 0, \quad \phi(a, y) = \sinh(y)/\sinh(b), \quad y \in [0, b],$$

где область Ω есть прямоугольник со сторонами $x \in [0, a]$ и $y \in [0, b]$. Эта задача имеет точное решение, которое можно получить методом разделения переменных,

$$u_{ex}(x, y) = \frac{\sin(x)\sinh(y)}{\sin(a)\sinh(b)}.$$

На Рис. 13 для прямоугольника $a = 2, b = 3$ построено точное решение рассматриваемой краевой задачи для уравнения Лапласа. На Рис. 14 (a) и (b) сравниваются графики точного решение $u(x, b/2), x \in [0, a]$ и $u(a/2, y), y \in [0, b]$) с численными решениями, полученными методами решения алгебраической системы и итерационным, для $n = 20, m = 30$. При этом использовалось число итераций, равное $N = 200$. В этом случае графики абсолютной ошибки численных решений показаны на Рис. 15 (a) и (b).

(a)



(b)

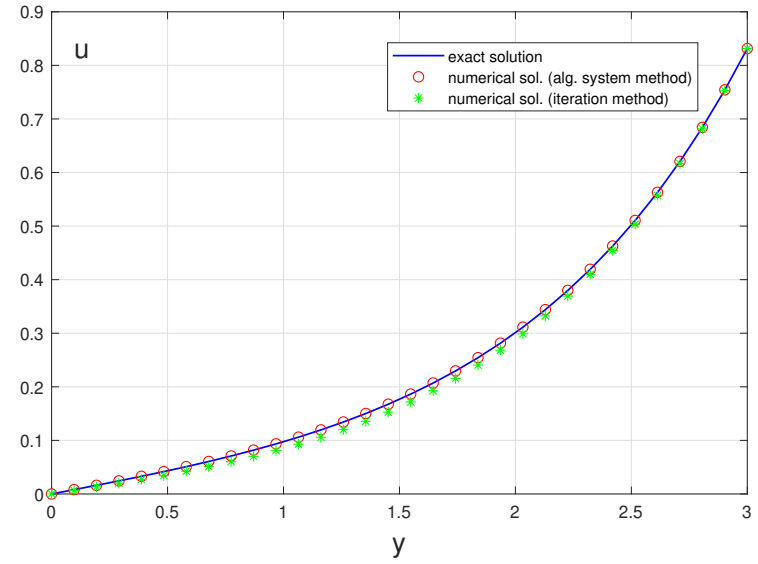
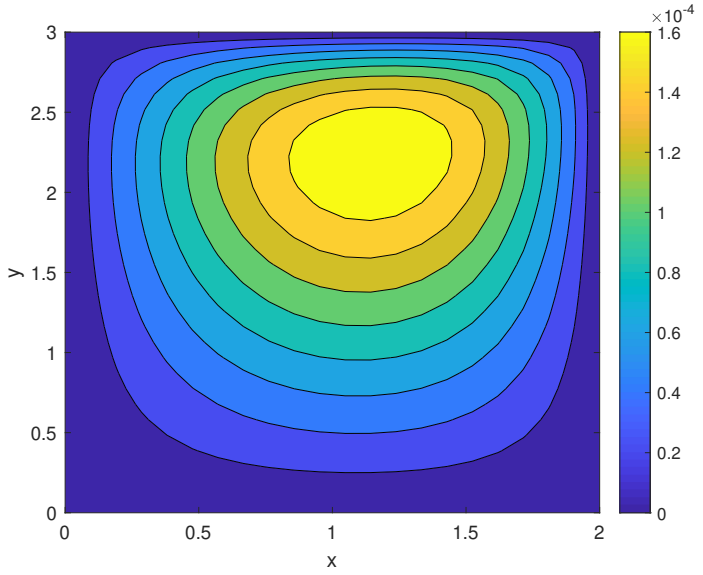


Рис. 14: Сравнение графиков точного решения $u(x, b/2), x \in [0, a]$, и $u(a/2, y), y \in [0, b]$), краевой задачи для уравнения Лапласа с численными решениями, полученными методами решения алгебраической системы - a и итерационным - (b), для $n = 20, m = 30$.

(a)



(b)

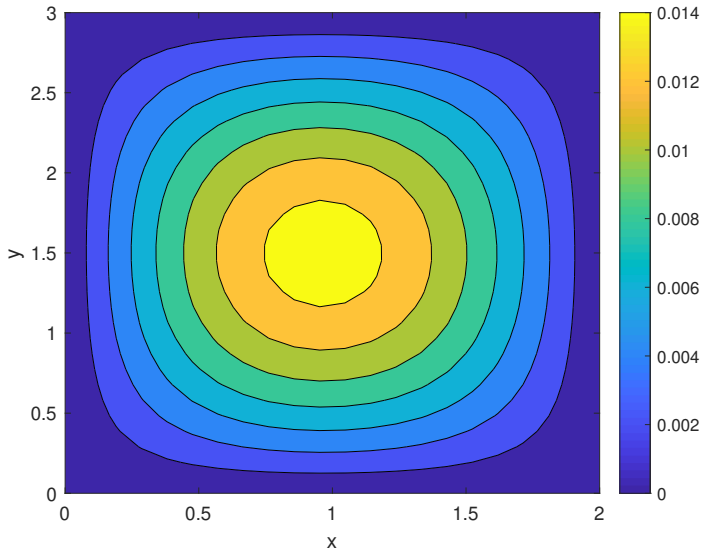


Рис. 15: Распределение абсолютной ошибки численных решений краевой задачи для уравнения Лапласа, полученных методами решения алгебраической системы - (a) и итерационным - (b), для $n = 20, m = 30$.

В анализе погрешностей мы вводим ошибку

$$e_{ij} = u_{ij} - \hat{u}_{ij},$$

где $\hat{u}_{ij} = u(x_i, y_j)$ являются значениями точного решения. Подставляя

$$u_{ij} = e_{ij} + \hat{u}_{ij}$$

в разностное уравнение (417), получим

$$\frac{1}{h_1^2}[e_{i-1,j} + e_{i+1,j} - 2e_{i,j}] + \frac{1}{h_2^2}[e_{i,j-1} + e_{i,j+1} - 2e_{i,j}] =$$

$$= f_{ij} - \frac{1}{h_1^2}[\hat{u}_{i-1,j} + \hat{u}_{i+1,j} - 2\hat{u}_{i,j}] - \frac{1}{h_2^2}[\hat{u}_{i,j-1} + \hat{u}_{i,j+1} - 2\hat{u}_{i,j}].$$

Далее, используя формулу с разностной аппроксимацией (416), получим

$$\frac{1}{h_1^2}[e_{i-1,j} + e_{i+1,j} - 2e_{i,j}] + \frac{1}{h_2^2}[e_{i,j-1} + e_{i,j+1} - 2e_{i,j}] =$$

$$= f_{ij} - [u_{xx}(x_i, y_j) + \frac{h_1^2}{12}u_{xxxx}(\xi_i, y_i)] - [u_{yy}(x_i, y_j) + \frac{h_2^2}{12}u_{yyyy}(\mathbf{x}_i, \zeta_j)] =$$

$$= -\frac{h_1^2}{12}u_{xxxx}(\xi_i, y_i) - \frac{h_2^2}{12}u_{yyyy}(\mathbf{x}_i, \zeta_j)$$

так как

$$u_{xx}(x_i, y_j) + u_{yy}(x_i, y_j) = f_{ij}.$$

Следовательно, ошибка удовлетворяет разностному уравнению (417)

$$e_{i-1,j} + e_{i+1,j} - 2e_{i,j} + \alpha[e_{i,j-1} + e_{i,j+1} - 2e_{i,j}]$$
$$= -\frac{h_1^4}{12}u_{xxxx}(\xi_i, y_i) - \frac{h_1^2 h_2^2}{12}u_{yyyy}(\mathbf{x}_i, \zeta_j), \quad (423)$$

и нулевым граничным условиям. Используя выражения для системы (421), для вектора ошибки E, построенного аналогичным образом, получим представление в виде решения системы $AE = F_h$,

где F_h есть вектор, составленный из значений правой части (423). Следовательно, будем иметь

$$||E|| \leq \|A^{(-1)}\|\|F_h\|.$$

Это означает, что в пределе $n \to +\infty$ и $m \to +\infty$ ($h_{1,2} \to 0$) норма ошибки стремится к нулю.

10.3 Уравнение теплопроводности (УТ)

Модельная задача для уравнения теплопроводности в одномерном случае вместе со вспомогательными условиями выглядит следующим образом:

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} + f(x,t), \tag{424}$$

с начальным условием

$$u(x,0) = g(x), \qquad a \leq x \leq b,$$

и с граничными условиями

$$u(a,t) = \phi_1(t), \qquad u(b,t) = \phi_2(t), \qquad t \geq 0,$$

где u есть температура, которая зависит от координаты x и времени t. При этом D есть коэффициент теплопроводности. Для простоты предположим, что

$$f(x,t) = 0, \qquad \phi_1(t) = \phi_2(t) = 0, \qquad D = 1, \qquad a = 0, \qquad b = 1.$$

Таким образом, мы собираемся подробно рассмотреть следующую начально-краевую задачу

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad u(x,0) = g(x), \qquad 0 \leq x \leq 1, \tag{425}$$

$$u(0,t) = u(1,t) = 0, \qquad t \geq 0. \tag{426}$$

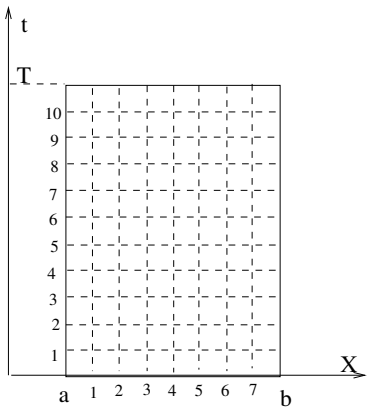


Рис. 16: Начально-краевая задача для уравнения теплопроводности

Использование конечно-разностного метода предполагает дискретизацию области - прямоугольника:

$$t_j = jk, \quad j \geq 0, \qquad x_i = ih, \quad 0 \leq i \leq n+1, \tag{427}$$

где переменные t и x имеют разный шаг дискретизации k и h, и для шага по переменной x будем иметь

$$h = \frac{1}{n+1}.$$

Наша цель - вычислить приближенные значения функции решения u в так называемых узловых точках прямоугольной сетки (t_j, x_i). Если $0 \leq t \leq T$, то $k = T/(m+1)$, $j = 0, 1, ..., m+1$.

10.4 Решение УТ явным методом

Заменим непрерывные производные конечными разностями

$$g'(t) = \frac{1}{k}[g(t+k) - g(t)] - \frac{k}{2}g''(s),$$

$$f''(x) = \frac{f(x+h) + f(x-h) - 2f(x)}{h^2} - \frac{h^2}{12}f^4(\xi).$$

В результате дискретизации и этой замены мы получим конечно-разностное уравнение

$$\frac{u_{i-1,j} - 2u_{ij} + u_{i+1,j}}{h^2} = \frac{u_{i,j+1} - u_{ij}}{k}. \tag{428}$$

Далее, начальное распределение температуры $g(x)$ дает нам значения u_{ij} на нижней границе прямоугольной области (см. Рис. 13):

$$g(x_i) = u(x_i, 0) = u_{i0}.$$

Граничные условия обеспечивают значения u_{ij} на вертикальных частях границы области (см. Рис. 13):

$$u_{0j} = u_{n+1,j} = 0.$$

Конечно-разностное уравнение можно записать в виде

$$u_{i,j+1} = \frac{k}{h^2}(u_{i-1,j} - 2u_{ij} + u_{i+1,j}) + u_{i,j},$$

или, вводя обозначение $s = k/h^2$,

$$u_{i,j+1} = su_{i-1,j} + (1 - 2s)u_{ij} + su_{i+1,j}. \tag{429}$$

Так как это уравнение дает новые значения $u_{i,j+1}$ явно в терминах предыдущих значений $u_{i+1,j}, \quad u_{ij}, \quad u_{i-1,j}$, то метод, основанный на этом уравнении, называется явным методом (явная разностная схема).

Рассмотрим анализ устойчивости явного разностного алгоритма (429). Это конечно-разностное уравнение, определяющее численный процесс, можно эффективно интерпретировать с использованием матричных и векторных обозначений. Его можно записать так:

$$U_{j+1} = AU_j, \tag{430}$$

где

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

$$U_j = \begin{pmatrix} u_{1j} \\ u_{2j} \\ u_{3j} \\ ... \\ u_{nj} \end{pmatrix}, \quad U_0 = \begin{pmatrix} g(x_1) \\ g(x_2) \\ g(x_3) \\ ... \\ g(x_n) \end{pmatrix}.$$

Заметим, что граничные условия $u_{0j} = u_{n+1,j} = 0$ учитываются. Следовательно, мы имеем

$$U_j = AU_{j-1} = A^2U_{j-2} = ... = A^jU_0. \tag{431}$$

С точки зрения физических процессов распространения тепла от источников в нашем случае поле температуры должно затухать экспоненциально со временем. В результате должно выполняться условие

$$\lim_{t\to+\infty} u(x,t) = 0,$$

таким образом, мы потребуем

$$\lim_{j\to+\infty} U_j = \lim_{j\to+\infty} A^jU_0 = 0$$

для всех векторов U_0. Если явный алгоритм удовлетворяет этому условию, он называется устойчивым. С другой стороны, мы имеем

$$||U_j|| = ||A^jU_0|| \le ||A||^j||U_0||,$$

и указанное выше предельное условие выполняется, если $||A|| < 1$. Матрица A симметрична, неравенство $||A|| < 1$ эквивалентно тому, что все собственные значения A должны удовлетворять $|\lambda_m| < 1$. Чтобы завершить предыдущий анализ, мы используем следующее представление для A:

$$A = I - sB, \tag{432}$$

где I - единичная матрица, и

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

Собственные значения матрицы A можно представить как $\lambda_m = 1 - s\mu_m$, где μ_m являются собственными значениями B. Более того, числа μ_m можно записать в виде:

$$\mu_m = 2(1 - \cos\theta_m), \qquad \theta_m = \frac{m\pi}{n+1}, \qquad 1 \le m \le n.$$

Следовательно, неравенство $|\lambda_m| < 1$ имеет место, если

$$-1 < 1 - 2s(1 - \cos\theta_m) < 1.$$

Обратите внимание, что $s > 0$. Таким образом, получаем, что $|\lambda_m| < 1$ эквивалентно условию

$$s < \frac{1}{1 - \cos\theta_m},$$

или

$$s < \frac{1}{2}. \tag{433}$$

Этот результат означает, что если $s < 1/2$, или $k/h^2 < 1/2$, явный алгоритм устойчив. Это серьезное ограничение заставляет метод быть очень медленным, так как шаг k по переменной t должен выбираться очень маленьким, удовлетворяющим неравенству $k < h^2/2$.

Существует еще один способ получить условие стабильности. Это применение метода Фурье. Основная идея заключается в следующем. Будем искать решение разностного уравнения (429) в виде

$$u_{lj} = e^{i\alpha lh} q^j.$$

Подставляя это решение в разностное уравнение и сокращая фактор $e^{i\alpha lh} q^j$, мы получаем

$$q = se^{-i\alpha h} + 1 - 2s + se^{i\alpha h} = 1 - 4s\sin^2(\frac{\alpha h}{2}).$$

Явный метод устойчив, если решение удовлетворяет условию

$$\lim_{j\to+\infty} u_{lj} = 0.$$

Таким образом, $|q| < 1$, или

$$-1 < 1 - 4s\sin^2(\frac{\alpha h}{2}) < 1.$$

Это приводит к ограничению

$$s < \frac{1}{2\sin^2(\frac{\alpha h}{2})},$$

и мы должны иметь $s < 1/2$ для стабильности.

10.5 Решение УТ неявным методом

Мы продолжаем изучать модельную задачу уравнения теплопроводности. Используя обозначения, введенные в предыдущем разделе, и конечно-разностные формулы, мы можем аппроксимировать исходное уравнение следующим образом:

$$\frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2} = \frac{u_{i,j} - u_{i,j-1}}{k}. \tag{434}$$

Это уравнение и дает неявный алгоритм. Оно приводит к соотношению

$$u_{i,j-1} = -su_{i-1,j} + (1+2s)u_{ij} - su_{i+1,j},$$

и, следовательно, мы имеем

$$AU_j = U_{j-1}, \tag{435}$$

где

$$A = \begin{pmatrix} 1+2s & -s & 0 & \dots & 0 \\ -s & 1+2s & -s & \dots & 0 \\ 0 & -s & 1+2s & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1+2s \end{pmatrix}, \qquad U_j = \begin{pmatrix} u_{1j} \\ u_{2j} \\ u_{3j} \\ \dots \\ u_{nj} \end{pmatrix}.$$

Формально решение разностного уравнения дается формулой

$$U_j = A^{-1}U_{j-1} = A^{-j}U_0.$$

Вопрос устойчивости для неявного алгоритма решается легко, так как в этом случае матрица A представляется в виде

$$A = I + sB,$$

и ее собственные значения выражаются с помощью следующих формул:

$$\lambda_m = 1 + 2s(1 - \cos\theta_m), \qquad \theta_m = \frac{m\pi}{n+1}, \qquad 1 \le m \le n.$$

Поскольку они, очевидно, удовлетворяют $\lambda_m > 1$, собственные значения A^{-1}, лежащие в интервале $(0,1)$ для произвольных значений

$s > 0$, и как результат $||A^{-1}|| < 1$. Таким образом, мы заключаем, что предложенный метод устойчив при всех значениях h и k, так как

$$||U_j|| = ||A^{-j}U_0|| \leq ||A^{-1}||^j||U_0||, \qquad \lim_{j\to+\infty} ||U_j|| = 0.$$

10.6 Метод Кранка-Николсона

Можно объединить неявные и явные методы в более общую формулу, содержащую параметр $\theta \in [0, 1]$. Результат имеет слудующий вид:

$$\frac{\theta}{h^2}(u_{i-1,j} - 2u_{ij} + u_{i+1,j}) + \frac{1-\theta}{h^2}(u_{i-1,j-1} - 2u_{i,j-1} + u_{i+1,j-1}) =$$

$$\frac{1}{k}(u_{i,j} - u_{i,j-1}). \tag{436}$$

Мы видим, что когда $\theta = 0$, эта формула дает явную схему. Когда $\theta = 1$, формула сводится к неявной схеме, рассмотренной выше. Случай $\theta = 1/2$ приводит к числовой процедуре, известной как метод Кранка-Николсона. В этом случае разностное уравнение дается формулой

$$-su_{i-1,j} + (2+2s)u_{ij} - su_{i+1,j} = su_{i-1,j-1} + (2-2s)u_{i,j-1} + su_{i+1,j-1}. \tag{437}$$

Это уравнение имеет векторную форму

$$(2I + sB)U_j = (2I - sB)U_{j-1}.$$

Следовательно, решение определяется формулой

$$U_j = (2I + sB)^{-1}(2I - sB)U_{j-1}.$$

Можно показать, что все собственные значения λ_m матрицы

$$(2I + sB)^{-1}(2I - sB)$$

удовлетворяют неравенству $|\lambda_m| < 1$ для произвольных $s > 0$. Таким образом, метод Кранка-Николсона устойчив для всех значений параметра $s = k/h^2$.

Следует отметить, что предыдущие явные и неявные методы имеют порядок приближений $O(k)+O(h^2)$, тогда как метод Кранка-Николсона имеет порядок $O(k^2)+O(h^2)$. Из предыдущего анализа видно, что мы имеем

$$\frac{u_{i-1,j+1}-2u_{ij+1}+u_{i+1,j+1}+u_{i-1,j}-2u_{i,j}+u_{i+1,j}}{2h^2}=\frac{u_{i,j+1}-u_{i,j}}{k}.$$

Затем перейдем к следующему приближению

$$\frac{1}{2}(u_{xx}(j+1)+\frac{h^2}{12}u_{xxxx})+\frac{1}{2}(u_{xx}(j)+\frac{h^2}{12}u_{xxxx})=u_t+\frac{k}{2}u_{tt}.$$

Так как

$$\frac{1}{2}u_{xx}(j+1)=\frac{1}{2}u_{xx}(j)+\frac{1}{2}ku_{xxt}+O(k^2),$$

и

$$u_t=u_{xx},\qquad u_{tt}=u_{xxt},$$

то видно, что метод Кранка-Николсона имеет погрешность порядка $O(k^2)+O(h^2)$.

Рассмотрим пример решения начально-краевой задачи для уравнения теплопроводности

$$\begin{gathered}\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2},\\ u(x,0)=\sin(\pi x),\qquad 0\le x\le 1,\\ u(0,t)=u(1,t)=0,\qquad t\ge 0.\end{gathered}$$

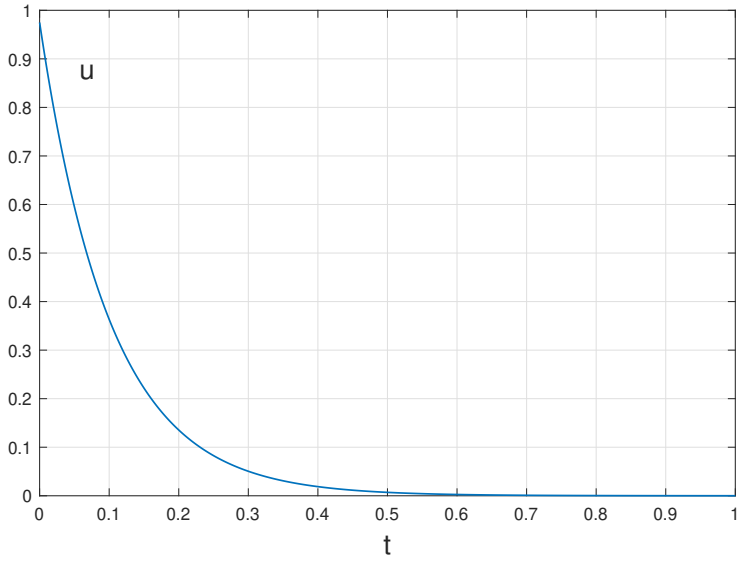
Пусть в данном случае метод конечных разностей реализован тремя способами: явная схема, неявная схема и метод Крэнка-Николсона с параметром $\theta=0$ (явная схема), $\theta=1$ (неявная схема), $\theta=1/2$, для $T=1$. Точное решение задачи получается методом разделения переменных и имеет вид функции

$$u_{ex}(x,t)=\sin(\pi x)\exp(-\pi^2 t),$$

экспоненциально убывающей функции по времени (см. Рис. 17 (а) для $T=1$). Графики среднеквадратичной ошибки

$$\delta u_j=\sqrt{\frac{1}{n}\sum_{i=1}^{n}|u_{ex}(i,j)-u(i,j)|^2}$$

(a)



(b)

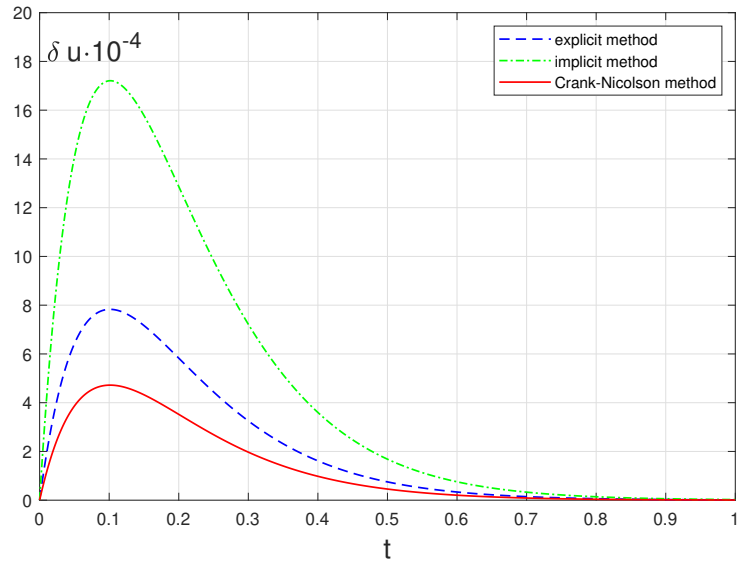


Рис. 17: (a) - точное решение начально-краевой задачи для уравнения теплопроводности, (b) - графики среднеквадратичной ошибки для всех трех численных методов решения начально-краевой задачи.

для всех трех численных методов решения этой задачи показаны на Рис. 17 (b). При этом использовалась сеть $n = 20, m = 1000$. Хорошо видно, что метод Крэнка-Николсона дает меньшую ошибку по сравнению с двумя другимим методами.

10.7 Анализ ошибок явного метода

Стабильность, конечно, не единственный критерий, используемый при выборе размера шагов k и h в этих методах. В общем можно сказать, что чем меньше мы берем значения k и h, тем точнее дискретизированная задача будет аппроксимировать исходное дифференциальное уравнение. Требуется теорема, гарантирующая, что решение дискретной задачи сходится к решению исходной непрерывной задачи, когда $k \to 0$ и $h \to 0$, то есть ошибка как разность между обоими решениями стремится к нулю. Ниже мы рассмотрим соответствующий анализ ошибок для явного метода.

Для анализа погрешностей вводится ошибка

$$e_{ij} = u_{ij} - \hat{u}_{ij},$$

где $\hat{u}_{ij} = u(x_i, y_j)$ являются значениями точного решения. Подставляя

$$u_{ij} = e_{ij} + \hat{u}_{ij}$$

в разностное уравнение явного алгоритма (429)

$$u_{i,j+1} = su_{i-1,j} + (1 - 2s)u_{ij} + su_{i+1,j},$$

мы получаем

$$e_{i,j+1} = se_{i-1,j}+(1-2s)e_{ij}+se_{i+1,j}-s[u_{i-1,j}+u_{i+1,j}-2u_{i,j}]+[u_{i,j+1}-u_{i,j}].$$

Применение формулы приближения конечных разностей

$$f''(x) = \frac{1}{h^2}[f(x+h) + f(x-h) - 2f(x)] - \frac{h^2}{12}f^{(4)}(\xi),$$

$$g'(t) = \frac{1}{k}[g(t+k) - g(t)] - \frac{k}{2}g''(\tau),$$

дает следующий результат:

$$e_{i,j+1} = se_{i-1,j} + (1-2s)e_{ij} + se_{i+1,j} - s[h^2 u_{xx}(x_i, t_j) + \frac{h^4}{12} u_{xxxx}(\xi_i, t_j)] +$$

$$[ku_t(x_i, t_j) + \frac{k^2}{2} u_{tt}(x_i, \tau_j)].$$

Воспользовавшись тем, что $sh^2 = k$ и $u_{xx} = u_t$, последнее уравнение можно записать в виде

$$e_{i,j+1} = se_{i-1,j} + (1-2s)e_{ij} + se_{i+1,j} - kh^2[\frac{1}{12} u_{xxxx}(\xi_i, t_j) - \frac{s}{2} u_{tt}(x_i, \tau_j)]. \quad (438)$$

Ограничим наш анализ, рассматривая компактное множество

$$S = \{(x, t) : 0 \le x \le 1, \quad 0 \le t \le T\}.$$

Предположим, что для $(x, t) \in S$ мы имеем следующие оценки:

$$M = \frac{1}{12} \sup |u_{xxxx}(\xi_i, t_j)| + \frac{s}{2} \sup |u_{tt}(x_i, \tau_j)|.$$

Введем вектор ошибки

$$E_j = \begin{pmatrix} e_{1j} \\ e_{2j} \\ e_{3j} \\ \dots \\ e_{nj} \end{pmatrix},$$

и положим

$$||E_j|| = \max |e_{ij}|, \qquad 1 \le i \le n.$$

Наконец, предположим, что $1 - 2s \ge 0$. Тогда из уравнения (438), получим

$$|e_{i,j+1}| \le s|e_{i-1,j}| + (1-2s)|e_{ij}| + s|e_{i+1,j}| + kh^2 M \le s||E_j|| +$$

$$(1-2s)||E_j|| + s||E_j|| + kh^2 M = ||E_j|| + kh^2 M,$$

и поэтому

$$||E_{j+1}|| \le ||E_j|| + kh^2 M \le ||E_{j-1}|| + 2kh^2 M \le ||E_0|| + (j+1)kh^2 M,$$

или

$$||E_j|| \leq ||E_0|| + jkh^2M.$$

Тогда, потому что $t \leq T$ и $||E_0|| = 0$, будем иметь

$$||E_j|| \leq Th^2M.$$

Таким образом, поскольку $h \to 0$, $||E_j|| \to 0$, то численное решение сходится к точному решению при условии, что $s < 1/2$ и функции u_{xxxx} и u_{tt} непрерывны.

10.8 Волновое уравнение

В качестве модельной задачи для волнового уравнения рассмотрим одномерную начально-краевую задачу вместе со вспомогательными условиями:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} + f(x,t), \tag{439}$$

с начальным условием

$$u(x,0) = u_0(x), \qquad u_t(x,0) = u_1(x), \qquad a \leq x \leq b,$$

и с граничными условиями

$$u(a,t) = \phi_1(t), \qquad u(b,t) = \phi_2(t), \qquad t \geq 0.$$

Решение u зависит от координаты x и времени t и описывает распространение волны со скоростью c. Для простоты мы снова предполагаем, что

$$f(x,t) = 0, \qquad \phi_1(t) = \phi_2(t) = 0, \qquad c = 1, \qquad a = 0, \qquad b = 1.$$

Тогда начально-краевая задача задается следующим образом:

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \qquad u(x,0) = u_0(x), \qquad u_t(x,0) = u_1(x), \qquad 0 \leq x \leq 1,$$
$$u(0,t) = u(1,t) = 0, \qquad t \geq 0. \tag{440}$$

Использование конечно-разностного метода предполагает дискретизацию области:

$$t_j = jk, \quad j \geq 0, \qquad x_i = ih, \quad 0 \leq i \leq n+1, \tag{441}$$

где переменные t и x имеют разный размер шага k и h, (см. Рис. 11)

$$h = \frac{1}{n+1}.$$

Наша цель - вычислить приближенные значения решения u в узловых точках прямоугольной сетки (t_j, x_i). Если $0 \leq t \leq T$, то $k = T/(m+1)$, $j = 0, 1, ..., m+1$.

Следующим шагом является замена непрерывных производных конечными разностями по следующим формулам:

$$f''(x) \approx \frac{f(x+h) + f(x-h) - 2f(x)}{h^2}.$$

В результате дискретизации и замены мы получим конечно-разностное уравнение

$$\frac{u_{i-1,j} - 2u_{ij} + u_{i+1,j}}{h^2} = \frac{u_{i,j-1} - 2u_{ij} + u_{i,j+1}}{k^2}. \tag{442}$$

Начальные условия дают нам

$$u_0(x_i) = u(x_i, 0) = u_{i0},$$

$$u_1(x_i) = u_t(x_i, 0) \approx \frac{u(x_i, k) - u(x_i, 0)}{k} = \frac{u_{i1} - u_{i0}}{k}.$$

Это означает, что мы немедленно получаем значения

$$u_{i1} = u_1(x_i)k + u_0(x_i).$$

Граничные условия дают

$$u_{0j} = u_{n+1,j} = 0.$$

Конечно-разностное уравнение можно записать в виде

$$u_{i,j+1} = su_{i-1,j} + 2(1-s)u_{ij} + su_{i+1,j} - u_{i,j-1}, \tag{443}$$

где $s = k^2/h^2$, и, следовательно, мы имеем

$$U_{j+1} = AU_j - U_{j-1}, \tag{444}$$

где

$$A = \begin{pmatrix} 2(1-s) & s & 0 & ... & 0 \\ s & 2(1-s) & s & ... & 0 \\ 0 & s & 2(1-s) & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 2(1-s) \end{pmatrix}, \qquad U_j = \begin{pmatrix} u_{1j} \\ u_{2j} \\ u_{3j} \\ ... \\ u_{nj} \end{pmatrix}.$$

Так как это уравнение дает новые значения $u_{i,j+1}$ явно в терминах предыдущих значений $u_{i+1,j}$, u_{ij}, $u_{i-1,j}$, $u_{i,j-1}$, то метод, основанный на этом уравнении, называется явным методом. Принимая во внимание граничные и начальные условия, этот алгоритм позволяет нам оценивать численно u_{ij} в прямоугольнике $0 \le x \le 1, \qquad 0 \le t \le T$.

Используя метод Фурье для конечно-разностного уравнения (443), получим условие устойчивости. Будем искать решение разностного уравнения в виде

$$u_{lj} = e^{i\alpha lh} q^j, \qquad l = 0, 1, 2, ..., n+1.$$

Подставляя это решение в разностное уравнение и сокращая фактор $e^{i\alpha lh} q^j$, мы получим

$$q^2 = q(se^{-i\alpha h} + 2(1-s) + se^{i\alpha h}) - 1 = 2q(1 - 2s\sin^2(\frac{\alpha h}{2})) - 1,$$

или

$$q^2 - 2q(1 - 2s\sin^2(\frac{\alpha h}{2})) + 1 = 0.$$

Явный метод устойчив, если решение удовлетворяет условию ограниченности решения

$$\lim_{j \to +\infty} u_{lj} \neq \infty,$$

которое следует из физических принципов теории распространения волн. Таким образом, мы потребуем, чтобы корни этого квадратного уравнения удовлетворяли условию $|q_{1,2}| \le 1$. Принимая во внимание что $q_1 q_2 = 1$, становится ясно, что неравенство $|q_{1,2}| \le 1$ имеет

место, если оба корня $q_{1,2}$ являются комплексно-сопряженными

$$q_{1,2} = e^{\pm i\gamma}.$$

Это выполняется, если

$$[1 - 2s\sin^2(\frac{\alpha h}{2})]^2 \leq 1,$$

что приводит к ограничению

$$s \leq \frac{1}{\sin^2(\frac{\alpha h}{2})}.$$

Таким образом, мы должны потребовать для стабильности описанного выше явного метода (443) выполнения условия $s \leq 1$, то есть

$$k \leq h. \tag{445}$$

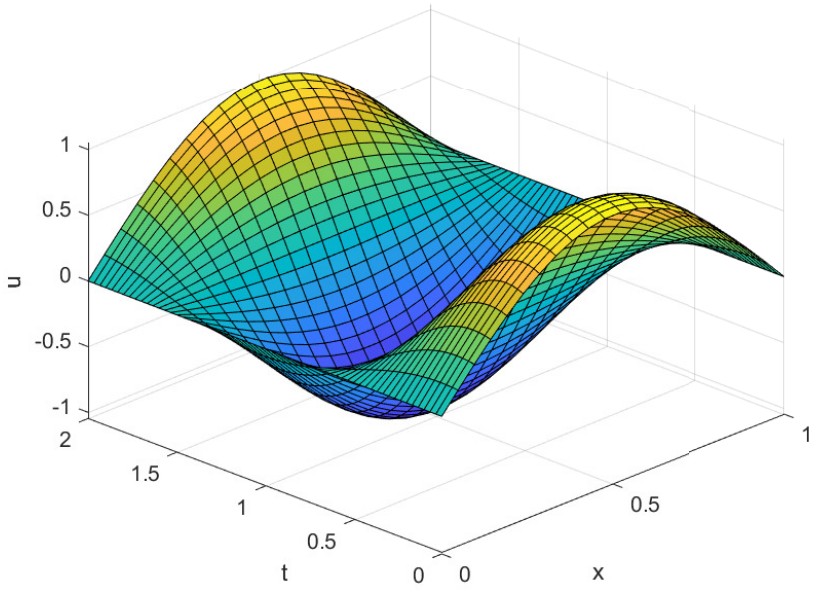
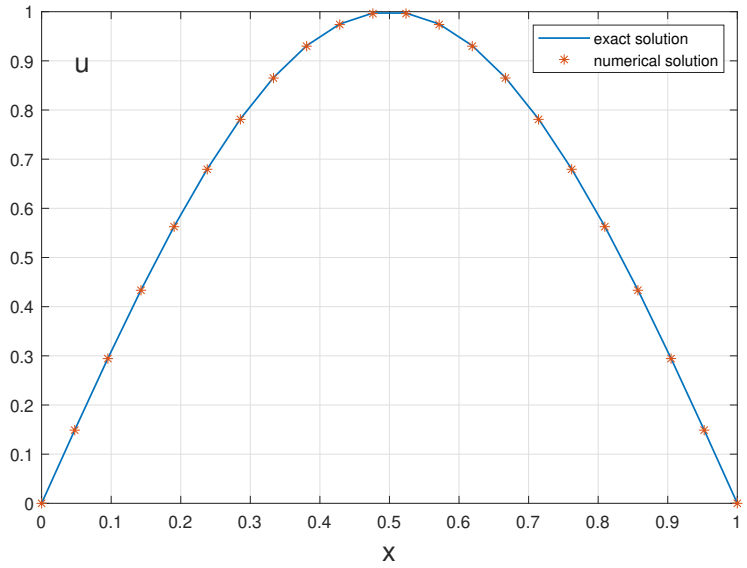


Рис. 18: Точное решение начально-краевой задачи для волнового уравнения.

(a)



(b)

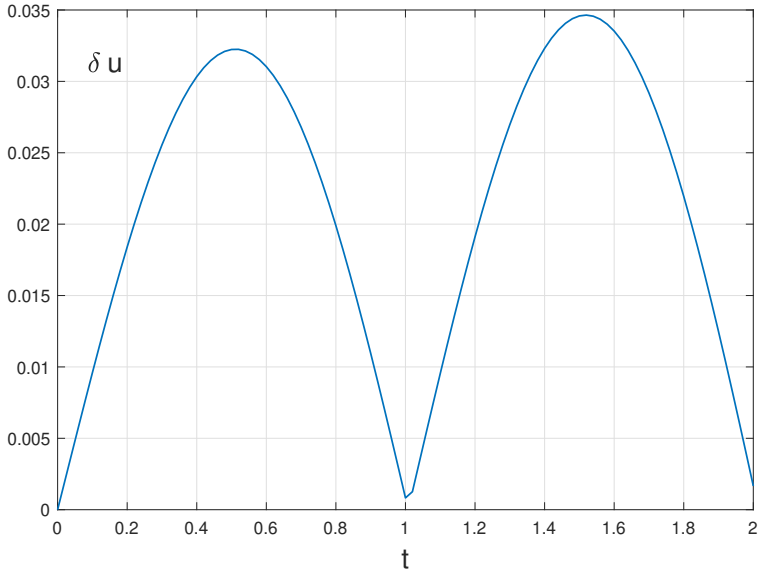


Рис. 19: (a) - сравнение графиков точного $u_{ex}(x,T)$ и численного $u(x,T)$ решений начально-краевой задачи для волнового уравнения для $T = 2$, (b) - графики среднеквадратичной ошибки для численного решения.

Рассмотрим пример решения начально-краевой задачи для волнового уравнения

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \qquad u(x,0) = \sin(\pi x), \qquad u_t(x,0) = \sin(\pi x),$$

$$0 \le x \le 1, \qquad u(0,t) = u(1,t) = 0, \qquad t \ge 0,$$

методом конечных разностей (явная схема). Точное решение получается методом разделения переменных и имеет вид функции

$$u_{ex}(x,t) = \sin(\pi x)(\cos(\pi t) + \frac{1}{\pi}\sin(\pi t)),$$

осциллирующей по времени (см. Рис. 18 для $T = 2$). Сравнение графиков точного $u_{ex}(x,T)$ и численного $u(x,T)$ решений для $T = 2$ изображено на Рис. 19 (a). График среднеквадратичной ощибки

$$\delta u_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |u_{ex}(i,j) - u(i,j)|^2}$$

для численного решения этой задачи показан на Рис. 19 (b). При этом использовалась сеть $n = 20, m = 100$. Таким образом условие стабильности было выполнено.

Вопросы для самоконтроля к главе 10

1. Как формулируется постановка краевой задачи Дирихле для уравнения Пуассона и теорема единственности?

2. Каковы основные детали применения метода конечных разностей численного решения краевой задачи Дирихле для уравнения Пуассона и каков соответствующий порядок аппроксимации?

3. Каковы основные детали применения метода конечных разностей численного решения начально-краевой задачи для уравнения теплопроводности и каков соответствующий порядок аппроксимации?

4. В чем заключается принципиальное отличие использования явного, неявного и Кранка-Николсона численных методов решения начально-краевой задачи для уравнения теплопроводности?

5. Что можно сказать о стабильности численного решения начально-краевой задачи для уравнения теплопроводности?

6. Каковы основные детали применения метода конечных разностей численного решения начально-краевой задачи для волнового уравнения и каков соответствующий порядок аппроксимации (явный метод)?

7. Что можно сказать о стабильности численного решения начально-краевой задачи для волнового уравнения?

11 Методы Ритца, Галеркина и введение в метод конечных элементов для дифференциальных уравнений в частных производных

11.1 Методы Ритца и Галеркина

Рассмотрим применение методов Ритца и Галеркина для решения краевых задач для линейных дифференциальных уравнений в частных производных второго порядка элиптического типа. Предположим, что мы имеем операторное уравнение вида

$$Au(x) = f(x) \tag{446}$$

где A - линейный оператор

$$Au = -div\Big(p(x)\nabla u(x)\Big) + q(x)u = f(x),$$

$$x = (x_1, ..., x_n) \in \Omega, \quad \Omega \subset \mathbf{R}^n, \quad n = 2, 3,$$

$f(x)$ является заданной функцией и $u(x)$ - решение, которое должно быть определено из уравнения и граничного условия Дирихле:

$$u|_{\partial\Omega} = 0.$$

Предположим, что $p(x) \geq p_0 > 0$, $q(x) \geq 0$, $p \in C^1(\Omega)$, $q \in C(\Omega)$ для $x \in \Omega$, и $f \in L_2(\Omega)$. Оператор определен для функций u, которые принадлежат множеству $D_A = C^2(\Omega) \cap C^1(\bar{\Omega})$ и $Au \in L_2(\Omega)$ (см. [9], том 4(1,2)). Это означает, что оператор A симметричен и положителен в вещественном гильбертовом пространстве $H = L_2(\Omega)$, так как, используя интегрирование по частям, мы имеем

$$< Au, u >= \int_\Omega Au \cdot u dx = \int_\Omega (-div(p\nabla u)u + qu^2)dx =$$

$$= \int_\Omega (p|\nabla u|^2 + qu^2)dx \geq 0$$

для произвольных $u \in D_A$. Скалярное произведение $< Au, u >$ называется энергией элемента u относительно оператора A и определяет энергетическую норму элемента (его энергию)

$$\|u\|_A^2 =< Au, u > .$$

На самом деле область определения оператора A шире и является пространством Соболева $W_2^1(\Omega)$ (см. [9], том 4(1,2)). Оно плотно по отношению к гильбертову пространству $H = L_2(\Omega)$. Таким образом, $D_A = W_2^1(\Omega) = H_A$, и это гильбертово пространство называется энергетическим пространством $H_A = W_2^1(\Omega)$. Можно доказать (см. [9], том 4(1,2)), что оператор A положительно определен в $H_A = W_2^1(\Omega)$, а именно,

$$< Au, u >\geq a^2\|u\|^2, \quad a > 0, \tag{447}$$

или

$$\|u\|_A \geq a\|u\|, \tag{448}$$

где $\|u\|$ является нормой в $L_2(\Omega)$.

Рассмотрим линейный квадратичный функционал

$$J(u) =< Au, u > -2 < f, u >= \int_\Omega (p|\nabla u|^2 + qu^2)dx - 2\int_\Omega fudx. \tag{449}$$

Утверждается (см. [9], том 4(1,2)), что функционал является непрерывным и ограничен снизу. Если уравнение $Au(x) = f(x)$ имеет решение u_0, то это решение минимизирует функционал $J(u)$ (утверждение необходимости). Обратное утверждение гласит, что если существует элемент u_0, который минимизирует функционал $J(u)$, то этот элемент удовлетворяет уравнению $Au(x) = f(x)$ (утверждение достаточности). Доказательство необходимости также основано на возможности сведения функционала $J(u)$ к форме

$$J(u) =< A(u - u_0), u - u_0 > - < Au_0, u_0 >= \|u - u_0\|_A^2 - \|u_0\|_A^2,$$

где $Au_0 = f$. Для доказательства достаточности проанализируем функционал с произвольным $v \in D_A$

$$J(u_0(x) + tv(x)) =< Au_0, u_0 > +2t < Au_0, v > +t^2 < Av, v > -$$

$$-2 < f, u_0 > -2t < f, v >,$$

где $u_0 \in H_A$ является минимумом $J(u)$. Приравнивая производную от $J(u_0(x) + tv(x))$ по параметру t к нулю в точке $t = 0$, получаем

$$< Au_0 - f, v >= 0, \quad \forall v \in D_A. \tag{450}$$

Это уравнение согласуется с методом Галеркина. Таким образом, мы имеем $Au_0 = f$. Вторая производная по t приводит к

$$\frac{d^2J}{dt^2} = 2 < Av, v >= 2\|v\|_A^2 > 0$$

для $v \neq 0$. Это связано с минимумом $J(u)$.

Основная идея метода Ритца заключается в замене граничной задачи для $Au(x) = f(x)$ проблемой минимизации функционала $J(u)$. Предположим, что мы выбрали линейные независимые базисные функции $\varphi_1, \varphi_2, ..., \varphi_n, ...$ в $H_A(\Omega)$. Следовательно, мы ищем решение в виде

$$u_n(x) = \sum_{m=1}^{n} c_m \varphi_m(x), \tag{451}$$

где $c_1, c_2, ..., c_n$ суть неизвестные константы. Подставляя эту сумму в функционал

$$J(u) =< Au, u > -2 < f, u >,$$

мы получаем выражение

$$\begin{aligned} J(u_n) &= J(c_1, c_2, ..., c_n) \\ &= \sum_{m,k=1}^{n} c_m c_k < A\varphi_m, \varphi_k > -2\sum_{m=1}^{n} c_m < f, \varphi_m >, \end{aligned} \tag{452}$$

которое является квадратичной по отношению к неизвестным константам $c_1, c_2, ..., c_n$. Для поиска минимального элемента $J(u_n) = J(c_1, c_2, ..., c_n)$ мы потребуем, чтобы

$$\frac{\partial}{\partial c_m} J(c_1, c_2, ..., c_n) = 0, \qquad m = 1, 2, ..., n.$$

Таким образом, мы приходим к линейной $n \times n$ системе алгебраических уравнений для неизвестных констант $c_1, c_2, ..., c_n$:

$$\sum_{m=1}^{n} c_m < A\varphi_m, \varphi_k >=< f, \varphi_k >, \qquad k = 1, 2, ..., n, \tag{453}$$

$$< A\varphi_m, \varphi_k >= \int_\Omega (p\nabla\varphi_m \cdot \nabla\varphi_k + q\varphi_m\varphi_k)dx, \qquad < f, \varphi_k >= \int_\Omega f\varphi_k dx,$$

которая называется системой Ритца. После того как мы найдем неизвестные константы $c_1, c_2, ..., c_n$, выражение (451) дает нам приближенное решение.

Стоит отметить, что для метода Галеркина-Петрова, который является более общим, чем метод Ритца, мы имеем

$$< Au_n - f, \psi_k >=< A(\sum_{m=1}^{n} c_m\varphi_m) - f, \psi_k >, \quad k = 1, 2, ..., n, \tag{454}$$

$$\sum_{m=1}^{n} c_m < A\varphi_m, \psi_k >=< f, \psi_k >, \qquad k = 1, 2, ..., n, \tag{455}$$

$$< A\varphi_m, \psi_k >= \int_\Omega (p\nabla\varphi_m \cdot \nabla\psi_k + q\varphi_m\psi_k)dx, \qquad < f, \psi_k >= \int_\Omega f\psi_k dx,$$

где линейно независимые функции $\psi_k \in H_A$ называются тестовыми функциями и вообще отличаются от набора линейных независимых базисных функций $\varphi_k \in H_A$. Если оба набора идентичны, то процедура Галеркина-Петрова превращается в приближение метода Ритца.

В соответствии с неравенством Шварца и использованием (448) мы могли бы получить некоторые полезные оценки, а именно:

$$| < Au_0, u > | = | < f, u > | \leq \|f\|\|u\| \leq \frac{1}{a}\|f\|\|u\|_A,$$

поскольку мы предполагаем, что $Au_0 = f$. Пусть $u = u_0$, то мы получим очень важную оценку

$$a\|u_0\| \leq \|u_0\|_A \leq \frac{1}{a}\|f\|. \tag{456}$$

Если мы попытаемся решить проблему нахождения приближения $Au_n = f_n$, где

$$u_n = \sum_{m=1}^{n} c_m \varphi_m, \tag{457}$$

то для $A(u_0 - u_n) = f - f_n$ мы получим

$$\|u_0 - u_n\|_A \le \frac{1}{a}\|f - f_n\| = \frac{1}{a}\|f - Au_n\|, \tag{458}$$

$$\|u_0 - u_n\| \le \frac{1}{a^2}\|f - Au_n\|. \tag{459}$$

Это означает, что если

$$\lim_{n\to\infty} Au_n = f,$$

тогда

$$\lim_{n\to\infty} u_n = u_0$$

в смысле энергетической нормы $\|u\|_A$ и в смысле нормы L_2.

11.2 Введение в метод конечных элементов

Рассмотрим следующую двухмерную задачу Дирихле в $\Omega \subset \mathbf{R}^2$:

$$-\Delta u = f(M), \quad M(x,y) \in \Omega, \tag{460}$$

$$u|_S = g(M), \quad M(x,y) \in S, \tag{461}$$

где граница $S = \partial\Omega$ достаточно гладкая. Мы разделим проблему на две задачи $u(M) = u^{(1)} + u^{(2)}$, где

$$-\Delta u^{(1)} = f(M), \qquad u^{(1)}|_S = 0, \tag{462}$$

и

$$-\Delta u^{(2)} = 0, \qquad u^{(2)}|_S = g(M). \tag{463}$$

Проблема (462) с однородным граничным условием может быть решена, например, с помощью описанного выше метода Ритца или Галеркина-Петрова. Здесь мы сосредоточимся на решении задачи

Дирихле (463) с использованием простейшего варианта метода конечных элементов. Ниже мы предполагаем, что $w = u^{(2)}$. Используя интегрирование по частям, мы получим важное интегральное соотношение

$$0 =< -\Delta w, \psi >= -\int_\Omega \Delta w \psi dxdy = \int_\Omega \nabla w \cdot \nabla \psi dxdy - \int_S \psi \frac{\partial w}{\partial n} ds. \tag{464}$$

Таким образом, метод Галеркина для $A = -\Delta$ в этом случае можно записать следующим образом. Приближение к решению w представим, как и в случае применения метода Ритца, в виде

$$w_n(x) = \sum_{m=1}^{n} c_m \varphi_m(x). \tag{465}$$

Неизвестные коэффициенты c_m следует искать из линейной алгебраической системы

$$\sum_{m=1}^{n} c_m \int_\Omega (\nabla \varphi_m \cdot \nabla \varphi_k) dxdy = \sum_{m=1}^{n} c_m \int_S \varphi_k \frac{\partial \varphi_m}{\partial n} ds, \qquad k = 1, 2, ..., n, \tag{466}$$

где $\psi = \varphi_k$.

Решим краевую задачу с заданным краевым условием Дирихле с функцией $g(x)$:

$$-\Delta w = 0, \qquad M(x, y) \in \Omega, \qquad w|_S = g(M), \qquad M(x, y) \in S, \tag{467}$$

используя аппроксимацию метода конечных элементов для простого случая квадрата $0 < x, y < a$ ($a = 2$), как показано в Рис. 14. В результате триангуляции мы имеем четыре элемента - треугольники $\{T_m\}_{m=1}^4$ и пять узлов $\{E_k\}_{k=1}^5$. Каждый треугольник имеет три вершины, в совокупности мы имеем множество вершин $\{N_j^{(m)}\}$ для $m = 1, ..., 4$ и $j = 1, 2, 3$ (см. Рис. 14). Для каждого треугольника мы определяем три полинома

$$P_j^{(m)}(x, y) = a_j^{(m)} + b_j^{(m)} x + c_j^{(m)} y$$

таким образом, что

$$P_j^{(m)}(N_i^{(m)}) = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

где $i = 1, 2, 3$.

Таким образом, для треугольника T_1 мы получим

$$P_1^{(1)} = 1 - \frac{1}{2}x - \frac{1}{2}y, \quad P_2^{(1)} = \frac{1}{2}x - \frac{1}{2}y, \quad P_3^{(1)} = y, \tag{468}$$

так как

$$\begin{cases} a_1^{(1)}1 + b_1^{(1)}0 + c_1^{(1)}0 = 1, \\ a_1^{(1)}1 + b_1^{(1)}2 + c_1^{(1)}0 = 0, \\ a_1^{(1)}1 + b_1^{(1)}1 + c_1^{(1)}1 = 0, \end{cases} \quad \begin{cases} a_2^{(1)}1 + b_2^{(1)}0 + c_2^{(1)}0 = 0, \\ a_2^{(1)}1 + b_2^{(1)}2 + c_2^{(1)}0 = 1, \\ a_2^{(1)}1 + b_2^{(1)}1 + c_2^{(1)}1 = 0, \end{cases}$$

$$\begin{cases} a_3^{(1)}1 + b_3^{(1)}0 + c_3^{(1)}0 = 0, \\ a_3^{(1)}1 + b_3^{(1)}2 + c_3^{(1)}0 = 0, \\ a_3^{(1)}1 + b_3^{(1)}1 + c_3^{(1)}1 = 1. \end{cases}$$

Аналогично, для треугольника $T_{2,3,4,}$ мы будем иметь

$$P_1^{(2)} = \frac{1}{2}x - \frac{1}{2}y, \quad P_2^{(2)} = -1 + \frac{1}{2}x + \frac{1}{2}y, \quad P_3^{(2)} = 2 - x, \tag{469}$$

для треугольника T_2,

$$P_1^{(3)} = -1 + \frac{1}{2}x + \frac{1}{2}y, \quad P_2^{(3)} = -\frac{1}{2}x + \frac{1}{2}y, \quad P_3^{(3)} = 2 - y, \tag{470}$$

для треугольника T_3,

$$P_1^{(4)} = -\frac{1}{2}x + \frac{1}{2}y, \quad P_2^{(4)} = 1 - \frac{1}{2}x - \frac{1}{2}y, \quad P_3^{(4)} = x, \tag{471}$$

наконец, для треугольника T_4.

Введем следующие элементы базиса - функции $\varphi_k(M)$, $k = 1, 2, 3, 4, 5$, каждая связана с соответствующим узлом E_k. А именно, мы имеем

$$\varphi_1(M) = \begin{cases} P_1^{(1)}(M), & M \in T_1, \\ P_2^{(4)}(M), & M \in T_4, \\ 0, & M \in T_2 \cup T_3, \end{cases} \quad \varphi_2(M) = \begin{cases} P_2^{(1)}(M), & M \in T_1, \\ P_1^{(2)}(M), & M \in T_2, \\ 0, & M \in T_4 \cup T_3, \end{cases} \tag{472}$$

$$\varphi_3(M) = \begin{cases} P_2^{(2)}(M), & M \in T_2, \\ P_1^{(3)}(M), & M \in T_3, \\ 0, & M \in T_1 \cup T_4, \end{cases} \quad \varphi_4(M) = \begin{cases} P_2^{(3)}(M), & M \in T_3, \\ P_1^{(4)}(M), & M \in T_4, \\ 0, & M \in T_1 \cup T_2, \end{cases} \tag{473}$$

$$\varphi_5(M) = \begin{cases} P_3^{(1)}(M), & M \in T_1, \\ P_3^{(2)}(M), & M \in T_2, \\ P_3^{(3)}(M), & M \in T_3, \\ P_3^{(4)}(M), & M \in T_4. \end{cases} \tag{474}$$

Аппроксимация решению строится как

$$w_n(x) = \sum_{m=1}^{5} c_m \varphi_m(x), \tag{475}$$

и, используя граничное условие $w|_S = g(M),\ M(x,y) \in S$, в узлах $E_{1,2,3,4}$, мы сразу получаем

$$c_1 = g(E_1), \quad c_2 = g(E_2), \quad c_3 = g(E_3), \quad c_4 = g(E_4). \tag{476}$$

Последнее неизвестное c_5 можно найти из соотношения (466)

$$\sum_{m=1}^{5} c_m \int_\Omega (\nabla\varphi_m \cdot \nabla\varphi_5) dxdy = \sum_{m=1}^{5} c_m \int_S \varphi_5 \frac{\partial\varphi_m}{\partial n} ds. \tag{477}$$

Оценив коэффициенты этого соотношения

$$\int_\Omega (\nabla\varphi_k \cdot \nabla\varphi_5) dxdy = -1, \qquad k = 1,2,3,4, \tag{478}$$

$$\int_\Omega |\nabla\varphi_5|^2 dxdy = 4, \tag{479}$$

$$\int_S \varphi_5 \frac{\partial\varphi_m}{\partial n} ds = 0, \qquad k = 1,2,3,4,5, \tag{480}$$

мы получим из (477) значение c_5, то есть среднее значение по отношению к уже найденным угловым значениям $w(E_k) = g(E_k)$ с $k = 1,2,3,4$. А именно, мы имеем

$$c_5 = \frac{c_1 + c_2 + c_3 + c_4}{4} = \frac{1}{4} \sum_{k=1}^{4} g(E_k). \tag{481}$$

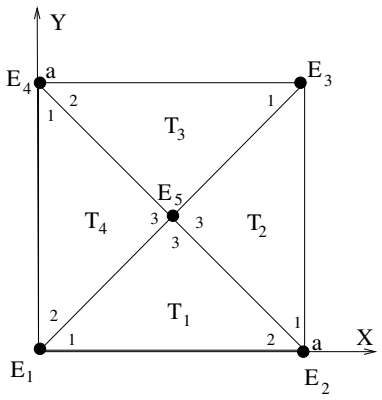


Рис. 20: Применение метода конечных элементов для решения задачи Дирихле для квадрата $0 < x < a, \quad 0 < y < a$.

Стоит отметить, что этот идеальный результат можно было бы получить независимо с помощью известной теоремы о гармонических функциях, утверждающей, что для любой гармонической функции w в области Ω следующий криволинейный интеграл вдоль границы области с нормальной производной для w всегда равен нулю, а именно

$$\int_S \frac{\partial w}{\partial n} ds = 0. \tag{482}$$

Если $S = U_{m=1}^4 S_m$, тогда

$$0 = \int_S \frac{\partial w}{\partial n} ds = \sum_{m=1}^5 c_m \int_{S_m} \frac{\partial \varphi_m}{\partial n} ds = 2c_1 + 2c_2 + 2c_3 + 2c_4 - 8c_5. \tag{483}$$

Таким образом, мы снова получили результат (481).

Заметим, что если необходимо решить краевую задачу с краевым условием Неймана

$$-\Delta w = 0, \qquad M(x,y) \in \Omega, \qquad \left.\frac{\partial w}{\partial n}\right|_S = g(M), \qquad M(x,y) \in S, \tag{484}$$

то следует использовать другой вариант приближения метода конечных элементов. Соответствующее приближение к решению w, как обычно, отыскиваем в виде

$$w_n(x) = \sum_{m=1}^n c_m \varphi_m(x). \tag{485}$$

Но неизвестные коэффициенты c_m должны быть найдены из линейной алгебраической системы

$$\sum_{m=1}^{n} c_m \int_{\Omega} (\nabla\varphi_m \cdot \nabla\varphi_k) dx = \int_S \varphi_k g(M) ds, \qquad k = 1, 2, ..., n. \quad (486)$$

Следует сказать, что предложенное описание элементарного применения метода конечных элементов несложно обобщить для любой прямоугольной области для задач Дирихле и Неймана для уравнения Пуассона.

Вопросы для самоконтроля к главе 11

1. Каковы основные детали применения методов Ритца и Галеркина для построения численного решения краевой задачи Дирихле для уравнения Пуассона?

2. Каковы основные детали применения метода конечных элементов для построения численного решения краевой задачи Дирихле для уравнения Пуассона?

3. В чем состоит принципиальное отличие использования метода конечных элементов от применения метода Ритца?

12 Метод граничных интегральных уравнений

12.1 Метод интегральных уравнений Фредгольма

Многие задачи математической физики (оптики, радиофизики, акустики) сводятся к линейным интегральным уравнениям вида

$$\int_D K(x,y)\varphi(y)dy = f(x), \tag{487}$$

$$\varphi(x) = \lambda \int_D K(x,y)\varphi(y)dy + f(x), \tag{488}$$

относительно искомой функции $\varphi(x)$ в области $D \in \mathbf{R}^n$, $n = 1, 2, 3$. Уранения (487-488) называются интегральными уравнениями Фредгольма первого и второго родов, соответственно. Заданные функции $K(x,y)$ и $f(x)$ называются ядром и неоднородным членом интегральных уравнений, а λ - спектральный параметр. Интегральное уравнение (488) при $f = 0$

$$\varphi(x) = \lambda \int_D K(x,y)\varphi(y)dy, \tag{489}$$

называется однородным интегральным уравнением Фредгольма второго рода. Интегральные уравнения Фредгольма второго рода

$$\psi(x) = \mu \int_D K^*(x,y)\psi(y)dy + g(x), \tag{490}$$

$$\psi(x) = \mu \int_D K^*(x,y)\psi(y)dy, \tag{491}$$

где $K^*(x,y) = \bar{K}(y,x)$, называются союзными к уравнениям (488-489), соответственно. Ядро $K^*(x,y)$ называется эрмитово сопряженным ядром к ядру $K(x,y)$. Запишем все эти интегральные уравнения в компактной операторной форме:

$$K\varphi = f, \qquad \varphi = \lambda K\varphi + f, \qquad \varphi = \lambda K\varphi, \tag{492}$$

$$K^*\psi = f, \qquad \psi = \mu K\psi + g, \qquad \psi = \mu K^*\psi. \tag{493}$$

То комплексное значение λ, при котором однородное интегральное уравнение (489) имеет ненулевые решения из $L_2(D)$, называется характеристическим числом ядра $K(x,y)$, а соотвествующие решения - собственными функциями этого ядра, соответствующими этому характеристическому числу. Таким образом, характеристические числа ядра $K(x,y)$ и собственные значения оператора K взаимно обратны, а их собственные функции совпадают.

Рассмотрим интегральные уравнения Фредгольма с непрерывным ядром. Предположим, что D - ограниченная область, функция $f(x)$ непрерывна в $\bar{D}$, и ядро $K(x,y)$ непрерывно в $\bar{D}\times\bar{D}$ (такие ядра будем называть непрерывными). Напомним определение норм в $L_2(D)$, $C(\bar{D})$, и скалярного произведения в $L_2(D)$:

$$\|f\|^2 = \int_D |f(x)|^2 dx = \sqrt{(f,f)}, \quad f\in L_2(D), \tag{494}$$

$$\|f\|_C = \max_{x\in\bar{D}} |f(x)|, \quad f\in C(\bar{D}), \tag{495}$$

$$(f,g) = \int_D f(x)g(x)dx, \quad f,g\in L_2(D). \tag{496}$$

Интегральный оператор K с непрерывным ядром $K(x,y)$ переводит $L_2(D)$ в $C(\bar{D})$, $L_2(D)$ в $L_2(D)$, $C(\bar{D})$ в $C(\bar{D})$, и ограничен, причем

$$\|Kf\|_C \le M\sqrt{V}\|f\|, \quad f\in L_2(D), \tag{497}$$

$$\|Kf\| \le MV\|f\|, \quad f\in L_2(D), \tag{498}$$

$$\|Kf\|_C \le MV\|f\|_C, \quad f\in C(\bar{D}), \tag{499}$$

где

$$M = \max_{x\in\bar{D},y\in\bar{D}} |K(x,y)|, \qquad V = \int_D dy.$$

Докажем неравенство (497). Применяя неравенство Коши-Буняковского (Шварца), получим

$$\|Kf\|_C = \max_{x\in\bar{D}} |(Kf)(x)| = \max_{x\in\bar{D}} \left|\int_D K(x,y)f(y)dy\right| \le$$

$$\leq \max_{x\in\bar{D}}(\int_D |K(x,y)|^2 dy)^{1/2} \int_D (|f(y)|^2 dy)^{1/2} \leq M\sqrt{V}\|f\|.$$

Аналогично доказываются (498), (499).

Имеет место взаимно однозначное соответствие между непрерывными ядрами и соответствующими им интегральными операторами. Будем искать решение уравнения (488) методом последовательных приближений:

$$\varphi^{(0)}(x) = f(x),$$

$$\varphi^{(n)}(x) = \lambda \int_D K(x,y)\varphi^{(n-1)}(y)dy + f(x)$$

$$= \lambda K\varphi^{(n-1)} + f = \sum_{m=0}^{n} \lambda^m K^m f(x), \quad n > 0. \quad (500)$$

Функции $K^m f$ называются итерациями функции f. Используя оценку (499), получим

$$\|K^n f\|_C \leq \|KK^{n-1}f\|_C \leq MV\|K^{n-1}f\|_C \leq ... \leq (MV)^n\|f\|_C. \quad (501)$$

Из этой оценки следует, что ряд

$$\sum_{m=0}^{+\infty} \lambda^m K^m f(x), \quad x \in \bar{D}, \quad (502)$$

называемый рядом Неймана, мажорируется положительным числовым рядом

$$\|f\|_C \sum_{m=0}^{+\infty} |\lambda|^m (MV)^m = \frac{\|f\|_C}{1-|\lambda|MV}, \quad (503)$$

сходящимся в круге $|\lambda| < 1/(MV)$. Поэтому при таких λ ряд (502) сходится равномерно по $x \in \bar{D}$, определяя тем самым непрерывную на $\bar{D}$ функцию $\varphi(x)$. Это означает

$$\lim_{n\to\infty} \varphi^{(n)}(x) = \varphi(x) = \sum_{m=0}^{+\infty} \lambda^m K^m f(x), \quad (504)$$

и справедлива оценка

$$\|\varphi\|_C \leq \frac{\|f\|_C}{1-|\lambda|MV}. \tag{505}$$

Более того, функция $\varphi(x)$ удовлетворяет интегральному уравнению (488)

$$\varphi(x) = \lim_{n\to\infty} \varphi^{(n)}(x) = \lambda \int_D K(x,y) \lim_{n\to\infty} \varphi^{(n-1)}(y)dy + f(x) = \tag{506}$$

$$= \lambda \int_D K(x,y)\varphi(y)dy + f(x).$$

Можно показать, что решение в этом случае единственно. Это следует из оценки для решения $\varphi_0(x)$ однородного уравнения (489)

$$\|\varphi_0(x)\| \leq |\lambda|MV\|\varphi_0(x)\|.$$

Другими словами, в круге $|\lambda| < 1/(MV)$ существует и ограничен обратный оператор $(I-\lambda K)^{-1}$, где I есть единичный оператор.

Метод последовательных приближений может эффективно использоваться в численных расчетах для нахождения приближенного решения интегрального уравнения Фредгольма второго рода (488).

Рассмотрим важный случай, когда интегральное уравнение

$$\varphi(x) = \lambda \int_G K(x,y)\varphi(y)dy + f(x) \tag{507}$$

решается точно. Это случай интегрального уравнения с вырожденным ядром:

$$K(x,y) = \sum_{m=1}^{n} p_m(x)q_m(y), \tag{508}$$

где системы функций $\{p_m(x)\}_{m=1}^n$ и $\{q_m(x)\}_{m=1}^n$ являются линейно независимыми. В этом случае мы будем иметь

$$\varphi(x) = \lambda \sum_{m=1}^{n} p_m(x) \int_G q_m(y)\varphi(y)dy + f(x) = \lambda \sum_{m=1}^{n} c_m p_m(x) + f(x), \tag{509}$$

где

$$c_m = \int_G q_m(y)\varphi(y)dy.$$

Умножим равенство (509) на $q_k(x)$, $k = 1, 2, ..., n$, и проинтегрируем его по области G. В результате мы получаем систему линейных алгебраических уравнений для неизвестных чисел c_m:

$$c_k = \lambda \sum_{m=1}^{n} c_m \int_G p_m(x)q_k(x)dx + \int_G f(x)q_k(x)dx = \lambda \sum_{m=1}^{n} A_{km}c_m + f_k. \tag{510}$$

Здесь мы ввели новые обозначения

$$A_{km} = \int_G p_m(x)q_k(x)dx, \qquad f_k = \int_G f(x)q_k(x)dx. \tag{511}$$

С использованием обозначений вектров $\bar{c} = (c_1, c_2, ..., c_n)^T$ и $\bar{f} = (f_1, f_2, ..., f_n)^T$ эту систему линейных алгебраических уравнений можно представить в матричном виде:

$$\bar{c} = \lambda A\bar{c} + \bar{f}. \tag{512}$$

Важно заметить, что интегральное уравнение (507) и (512) эквивалентны друг другу. Действительно, если вектор $\bar{c}$ есть решение (512), то функция

$$\varphi(x) = \lambda \sum_{m=1}^{n} c_m p_m(x) + f(x) \tag{513}$$

удовлетворяет интегральному уравнению, а именно:

$$\begin{aligned}\varphi(x) - \lambda \sum_{m=1}^{n} p_m(x) \int_G q_m(y)\varphi(y)dy - f(x) = \lambda \sum_{m=1}^{n} c_m p_m(x) + f(x) \\ - \lambda \sum_{m=1}^{n} p_m(x) \int_G q_m(y)(\lambda \sum_{k=1}^{n} c_k p_k(y) + f(y))dy - f(x) = \\ \lambda \sum_{m=1}^{n} p_m(x)(c_m - \lambda \sum_{k=1}^{n} A_{mk}c_k - f_m) = 0. \end{aligned} \tag{514}$$

Обозначим через $D(\lambda)$ определитель системы (512)

$$D(\lambda) = \det(I - \lambda A), \tag{515}$$

и через M_{km} алгебраические дополнения матрицы $I - \lambda A$. Пусть $D(\lambda) \neq 0$, тогда по теореме Крамера решение системы (512) единственно и определяется формулой

$$c_k = \frac{1}{D(\lambda)} \sum_{m=1}^{n} M_{km}(\lambda) f_m, \qquad k = 1, 2, ..., n. \tag{516}$$

Подставляя найденное решение в формулу (513) и вспоминая определение вектора $\bar{f}$, мы получим точное решение интегрального уравнения (507) в виде

$$\varphi(x) = \frac{1}{D(\lambda)} \sum_{k,m=1}^{n} M_{km}(\lambda) p_k(x) \int_G q_m(y) f(y) dy + f(x). \tag{517}$$

С другой стороны, это решение можно записать в форме некоторого интегрального представления, хорошо известного в теории интегральных уравнений,

$$\varphi(x) = \lambda \int_G R(x, y, \lambda) f(y) dy + f(x), \tag{518}$$

с оператором резольвенты, у которого ядро $R(x, y, \lambda)$ имеет вид

$$R(x, y, \lambda) = \frac{1}{D(\lambda)} \sum_{k,m=1}^{n} M_{km}(\lambda) p_k(x) q_m(x).$$

Если ядро $K(x, y)$ оператора интегрального уравнения (507) есть непрерывная функция, то оно может быть аппроксимировано достаточно точно,

$$K(x, y) \sim \sum_{n_1, n_2=1}^{N_1, N_2} T_{n_1 n_2} p_{n_1}(x) q_{n_2}(y), \tag{519}$$

с помощью систем линейно независимых непрерывных функций $\{p_m(x)\}_{m=1}^{N_1}$ и $\{q_m(x)\}_{m=1}^{N_2}$, где T_{n_1, n_2} суть постоянные. В этом случае

мы вновь приходим к интегральному уравнения с вырожденным ядром, для которого мы аналогичным образом получим систему (512) в виде

$$c_m = \lambda \sum_{n_1,n_2=1}^{N_1,N_2} A_{mn_1} T_{n_1 n_2} c_{n_2} + f_m, \qquad m = 1, 2, ..., n_2, \tag{520}$$

или

$$\bar{c} = \lambda B \bar{c} + \bar{f}, \quad B = AT, \tag{521}$$

где были использованы обозначения (511) и

$$B_{mn_2} = \sum_{n_1=1}^{N_1} A_{mn_1} T_{n_1 n_2}.$$

В заключение этой секции введения в теорию интегральных уравнений приведем без доказательства основные положения Альтернативы Фредгольма для разрешимости интегральных уравнений Фредгольма второго рода (488) с непрерывным ядром:

Если интегральное уравнение (488) разрешимо в $C(\bar{D})$ при любом неоднородном члене $f \in C(\bar{D})$, то союзное к нему уравнение (490) разрешимо в $C(\bar{D})$ при любом неоднородном члене $g \in C(\bar{D})$, причем эти решения единственны (первая теорема Фредгольма).

Если интегральное уравнение (488) разрешимо в $C(\bar{D})$ не при любом неоднородном члене $f \in C(\bar{D})$, то:

1) однородные уравнения (489) и (491) имеют одинаковое конечное число линейно независимых решений (вторая теорема Фредгольма);

2) для разрешимости уравнения (488) необходимо и достаточно, чтобы неоднородный член был ортогонален ко всем решениям союзного однородного уравнения (491) (третья теорема Фредгольма);

3) в каждом круге $|\lambda| < R$ может находиться лишь конечное число характеристических чисел ядра $K(x, y)$ (четвертая теорема Фредгольма).

Эти основные положения Альтернативы Фредгольма для разрешимости интегральных уравнений Фредгольма второго рода (488) с непрерывным ядром очень важны в математическом и численном

моделировании краевых задач эллиптических уравнений в математической физике. Они гарантируют существование и единственность решения изучаемой математической модели. Заметим, что интегральные уравнения Фредгольма первого рода (487) тоже эффективно используются в численном анализе при моделировании краевых задач эллиптических уравнений. Но ядро интегрального уравнения должно быть при этом сингулярно. В следующей секции мы рассмотрим примеры применения интегральных уравнений Фредгольма обоих типов для решения двух разных задач рассеяния электромагнитных волн.

12.2 Рассеяние плоской электромагнитной волны на бесконечном идеально-проводящем цилиндре

Рассмотрим одну из важнейших задач классической электродинамики - рассеяние плоской электромагнитной волны на бесконечном идеально-проводящем (металлическом) цилиндре. Пусть цилиндр направлен вдоль оси Z, а волновой вектор лежит в плоскости XY. Пусть сечение цилиндра представляет собой произвольный гладкий выпуклый контур в плоскости XY (см. Рис. 15). Тогда рассматриваемая задача рассеяния сведется к двухмерной внешней краевой задаче для уравнения Гельмгольца (см. [17]). В этом случае исходная проблема разбивается на две задачи двух разных поляризаций падающей плоской волны. В случае TE-поляризации будем иметь следующую краевую задачу для функции $u(x, y)$, представляющую E_z компоненту полного поля:

$$E_x = E_y = 0, \quad E_z = u(x, y), \tag{522}$$

$$H_x = -\frac{i}{\omega} u_y, \quad H_y = \frac{i}{\omega} u_x, \quad H_z = 0, \tag{523}$$

$$(\Delta + k^2)u(x, y) = 0, \quad u|_S = 0, \quad u(x, y) = u_{in} + u_{sc}, \tag{524}$$

$$u_{in} = e^{i(\alpha_0 x + \beta_0 y)} = e^{ikr\cos(\theta - \theta_0)}, \tag{525}$$

где вектора $\mathbf{E} = (E_x, E_y, E_z)$, $\mathbf{H} = (H_x, H_y, H_z)$ представляют электрическую и магнитную компоненты поля, Δ - оператор‘ Лапласа

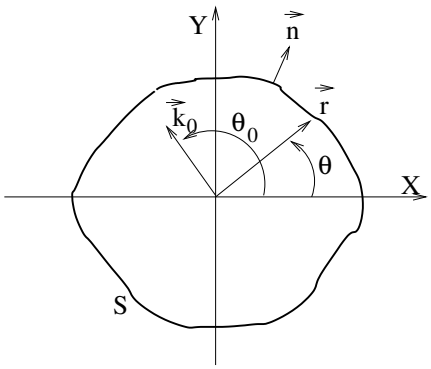


Рис. 21: Рассеяние плоской электромагнитной волны на бесконечном идеально-проводящем цилиндре. Показано сечение цилиндра в плоскости XY.

в плоскости XY, $k = \omega/c$ есть волновое число, S - боковая поверхность цилиндра, в полярных координатах r, θ параметр θ_0 задает направление распространения плоской волны. Здесь выполняется краевое условие Дирихле, так как электрическая компонента полного поля на поверхности цилиндра должна обращаться в ноль.

В случае TM-поляризации будем иметь другую краевую задачу для функции $u(x, y)$, представляющую H_z компоненту полного поля:

$$H_x = H_y = 0, \quad H_z = u(x, y), \tag{526}$$

$$E_x = \frac{i}{\omega} u_y, \quad E_y = -\frac{i}{\omega} u_x, \quad E_z = 0, \tag{527}$$

$$(\Delta + k^2)u(x, y) = 0, \quad \frac{\partial u}{\partial n}|_S = 0, \quad u(x, y) = u_{in} + u_{sc}. \tag{528}$$

Здесь выполняется краевое условие Неймана с производной по внешней нормали к поверхности цилиндра $\frac{\partial u}{\partial n}|_S$, так как

$$0 = (\mathbf{n} \times \mathbf{E})|_S = (n_x E_y - n_y E_x)|_S = -\frac{i}{\omega}(n_x u_x + n_y u_y)|_S.$$

В обоих задачах требуется найти рассеянное поле $u_{sc}(x, y)$.

В строгой математической постановке, гарантирующей существование и единственность решения краевой задачи для компоненты рассеянного поля $u_{sc}(x, y)$, потребуется удовлетворение условию

излучения Зоммерфельда (см. [20])

$$\lim_{r\to+\infty}\sqrt{r}(\frac{\partial u_{sc}}{\partial r}-iku_{sc})=0, \tag{529}$$

что эквивалентно тому, что на бесконечно большом расстоянии от цилиндра в плоскости XY рассеянное поле ведет себя подобно уходящей на бесконечность цилиндрической волне

$$u_{sc}(M)=\frac{e^{ikr}}{\sqrt{kr}}F(\theta)+O((kr)^{-1}) \tag{530}$$

с диаграммой рассеяния $F(\theta)$, где $M=(x,y)$.

В соответствии с теорией интегральных уравнений, применяемых для решения краевых задач рассеяния для уравнения Гельмгольца (см. [6]), в случае TE-поляризации будем искать решение в виде потенциала двойного слоя криволинейного интеграла:

$$u_{sc}(M)=\frac{\pi i}{2}\int_S\frac{\partial}{\partial n'}H_0^{(1)}(k|\mathbf{QM}|)\nu(Q)ds, \tag{531}$$

где $\mathbf{n}'$ есть единичный вектор нормали в точке интегрирования Q, $|\mathbf{QM}|$ - расстояние между точками M и Q (длина вектора $\mathbf{QM}$), ds - дифференциал длины дуги, $H_0^{(1)}(kR)$ - функция Ханкеля первого рода порядка 0, и $\nu(Q)$ - искомая плотность ($Q=(x',y')$). В случае TM-поляризации будем искать решение в виде потенциала простого слоя

$$u_{sc}(M)=\frac{\pi i}{2}\int_S H_0^{(1)}(k|\mathbf{QM}|)\mu(Q)ds, \tag{532}$$

где $\mu(Q)$ есть искомая плотность.

Если воспользоваться известной формулой разложения функции Ханкеля по плоским волнам (см. [21], [22])

$$H_0^{(1)}(k\sqrt{x^2+y^2})=\frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{e^{iwx+iv|y|}}{v}dw,\qquad v=\sqrt{k^2-w^2}, \tag{533}$$

то формулы для рассеянного поля (531) и (532) в области $y > y(s)$ можно представить в виде разложения по плоским волнам

$$u_{sc}(M) = \int\limits_{-\infty}^{+\infty} A(w) e^{iwx+iv|y|} dw, \tag{534}$$

где для амплитуды углового спектра $A(w)$ получим

$$A(w) = \frac{1}{2} \int_S \nu(s) e^{-iwx(s)-ivy(s)} (\frac{w}{v} n_x(s) + n_y(s)) ds$$

в случае TE-поляризации и

$$A(w) = \frac{i}{2v} \int_S \mu(s) e^{-iwx(s)-ivy(s)} ds$$

в случае TM-поляризации. Здесь предполагается использование параметрического задания контура поверхности цилиндра $x = x(s),\ y = y(s)$ с параметром длины дуги.

Рассмотрим дальнюю зону для точки наблюдения $M = (r\cos\theta, r\sin\theta)$ для рассеянного поля

$$|\mathbf{QM}| = r - (x(s)\cos\theta + y(s)\sin\theta) + O(r^{-1}), \tag{535}$$

$$r >> \max_{\theta\in[0,2\pi]} \sqrt{x(s)^2 + y(s)^2}. \tag{536}$$

Воспользуемся известной асимптотикой функции Ханкеля (см. [21], [22])

$$H_0^{(1)}(kR) = e^{ikR-i\pi/4} \sqrt{\frac{2}{\pi kR}} (1 + O((kR)^{-1})). \tag{537}$$

Тогда для формул рассеянного поля (531) и (532) в дальней зоне мы получим представление (530) в виде цилиндрической волны, уходящей на бесконечность, где для диаграммы рассеянного поля получим следующие представления

$$F(\theta) = -k\sqrt{\frac{\pi}{2}} e^{-i\pi/4} \int_S \nu(s) e^{-ik(x(s)\cos\theta+y(s)\sin\theta)} (n_x(s)\cos\theta + (n_y(s)\sin\theta) ds \tag{538}$$

в случае TE-поляризации и

$$F(\theta) = -k\sqrt{\frac{\pi}{2}} e^{-i\pi/4} \int_S \mu(s) e^{-ik(x(s)\cos\theta+y(s)\sin\theta)} ds \tag{539}$$

в случае TM-поляризации.

Для потенциалов двойного и простого слоев хорошо известны их свойства (см. [6]). При переходе через поверхность (контур) S потенциал простого слоя непрерывен, но его производная по нормали претерпевает скачок. При переходе через поверхность (контур) S потенциал двойного слоя претерпевает скачок. Введем обозначения

$$V^{(1)}(M) = \frac{\pi i}{2} \int_S \frac{\partial}{\partial n'} H_0^{(1)}(k|\mathbf{QM}|)\nu(Q)ds, \tag{540}$$

$$V^{(0)}(M) = \frac{\pi i}{2} \int_S H_0^{(1)}(k|\mathbf{QM}|)\mu(Q)ds. \tag{541}$$

Если мы обозначим предельные значения извне и изнутри по отношению к поверхности цилиндра для потенциала двойного слоя как $V_{\pm}^{(1)}(M)$, а для нормальной производной простого слоя как $\frac{\partial}{\partial n}V_{\pm}^{(0)}(M)$, где $\mathbf{n}$ есть единичный вектор нормали в точке наблюдения M, то справедливы следующие соотношения (см. [6]):

$$V_{\pm}^{(1)}(M) = \pm\pi\nu(M) + V_{PV}^{(1)}(M), \tag{542}$$

$$\frac{\partial}{\partial n}V_{\pm}^{(0)}(M) = \mp\pi\mu(M) + \frac{\partial}{\partial n}V_{PV}^{(0)}(M). \tag{543}$$

Учитывая эти предельные соотношения, в случае TE-поляризации для искомой плотности $\nu(Q)$ потенциала двойного слоя получим интегральное уравнение Фредгольма второго рода (см. [6]):

$$\nu(M) = -\frac{u_{in}(M)}{\pi} - \frac{i}{2} \int_S \nu(Q) \frac{\partial}{\partial n'} H_0^{(1)}(k|\mathbf{QM}|)ds, \tag{544}$$

или

$$\nu(M) = -\frac{u_{in}(M)}{\pi} + \frac{ik}{2} \int_S \nu(Q) H_1^{(1)}(k|\mathbf{QM}|)\cos\varphi ds, \tag{545}$$

$$\cos\varphi = \mathbf{n}(Q) \cdot \frac{\mathbf{MQ}}{|\mathbf{MQ}|}. \tag{546}$$

В случае TM-поляризации для искомой плотности $\mu(Q)$ потенциала простого слоя также получим интегральное уравнение Фредгольма второго рода:

$$\mu(M) = \frac{1}{\pi}\frac{\partial}{\partial n}u_{in}(M) - \frac{ik}{2}\int_S \mu(Q)H_1^{(1)}(k|\mathbf{QM}|)\cos\psi ds, \qquad (547)$$

$$\cos\psi = \mathbf{n}(M)\cdot\frac{\mathbf{QM}}{|\mathbf{QM}|}. \qquad (548)$$

Следует сказать, что в обоих случаях интегральные уравнения имеют ограниченные ядра. Это очень важное обстоятельство для численного анализа. Обсуждение теоремы существования и единственности для интегральных уравнений (546) и (548) можно найти в [6], [17].

Заметим, что если искать решение в случае TE-поляризации (задача Дирихле) в виде потенциала простого слоя

$$u_{sc}(M) = \frac{\pi i}{2}\int_S H_0^{(1)}(k|\mathbf{QM}|)\nu(Q)ds, \qquad (549)$$

то мы получим интегральное уравнение Фредгольма первого рода с сингулярным ядром

$$u_{in}(M) + \frac{\pi i}{2}\int_S H_0^{(1)}(k|\mathbf{QM}|)\nu(Q)ds = 0, \qquad (550)$$

которое хорошо известно в классической электродинамике и акустике и часто используется в численных расчетах.

Если контур границы цилиндра S задается параметрически $x = x(\theta)$, $y = y(\theta)$, где $0 \le \theta < 2\pi$, то интегральные уравнения (546) и (548) представляются в виде

$$\nu(\theta) = \nu_0(\theta) + \int_0^{2\pi} K^{(1)}(\theta,\theta')\nu(\theta')d\theta', \qquad (551)$$

$$\nu_0(\theta) = -\frac{e^{i(\alpha_0 x(\theta)+\beta_0 y(\theta))}}{\pi},$$

$$K^{(1)}(\theta,\theta') = \frac{ik}{2}H_1^{(1)}(k|\mathbf{QM}|)\cos\varphi(\theta,\theta')\sqrt{x'(\theta')^2 + y'(\theta')^2},$$

$$|\mathbf{QM}| = \sqrt{(x(\theta) - x(\theta'))^2 + (y(\theta) - y(\theta'))^2},$$

$$\cos\varphi(\theta,\theta') = \frac{(x(\theta') - x(\theta))n_x(\theta') + (y(\theta') - y(\theta))n_y(\theta')}{\sqrt{(x(\theta) - x(\theta'))^2 + (y(\theta) - y(\theta'))^2}},$$

$$\mu(\theta) = \mu_0(\theta) + \int_0^{2\pi} K^{(2)}(\theta,\theta')\mu(\theta')d\theta', \tag{552}$$

$$\mu_0(\theta) = \frac{e^{i(\alpha_0 x(\theta) + \beta_0 y(\theta))}}{\pi} i(n_x(\theta)\alpha_0 + n_y(\theta)\beta_0),$$

$$K^{(2)}(\theta,\theta') = -\frac{ik}{2} H_1^{(1)}(k|\mathbf{QM}|)\cos\psi(\theta,\theta')\sqrt{x'(\theta')^2 + y'(\theta')^2},$$

$$\cos\psi(\theta,\theta') = \frac{(x(\theta) - x(\theta'))n_x(\theta) + (y(\theta) - y(\theta'))n_y(\theta)}{\sqrt{(x(\theta) - x(\theta'))^2 + (y(\theta) - y(\theta'))^2}}.$$

Хорошо известно, что в компьютерном моделировании интегральные уравнения (551) и (552) решаются методом моментов. В результате мы получаем конечномерные системы линейных алгебраических уравнений. Один из простейших вариантов метода моментов реализуется с помощью разложения искомых функций ν, μ в стандартный ряд Фурье по ортонормированным функциям

$$\nu(\theta) = \sum_{n=-N}^{N} \nu_n \frac{e^{in\theta}}{\sqrt{2\pi}}, \qquad \mu(\theta) = \sum_{n=-N}^{N} \mu_n \frac{e^{in\theta}}{\sqrt{2\pi}}. \tag{553}$$

Подставляя эти разложения в интегральные уравнения (551) и (552) и проектируя результаты на элемент $\frac{e^{-im\theta}}{\sqrt{2\pi}}$, получим соответствующие системы линейных алгебраических уравнений:

$$\nu_m = \nu_m^{(0)} + \sum_{n=-N}^{N} K_{mn}^{(1)}\nu_n, \qquad \mu_m = \mu_m^{(0)} + \sum_{n=-N}^{N} K_{mn}^{(2)}\mu_n, \tag{554}$$

$m = -N, ..., 0, ...N$, где

$$\nu_m^{(0)} = \int_0^{2\pi} \nu_0(\theta)\frac{e^{-im\theta}}{\sqrt{2\pi}}d\theta, \qquad \mu_m^{(0)} = \int_0^{2\pi} \mu_0(\theta)\frac{e^{-im\theta}}{\sqrt{2\pi}}d\theta,$$

$$K_{mn}^{(1,2)} = \frac{1}{2\pi} \int\limits_0^{2\pi} d\theta \int\limits_0^{2\pi} d\theta' K^{(1,2)}(\theta, \theta') e^{i(n\theta' - m\theta)}.$$

Рассмотрим ниже более подробно случай TE-поляризации, так как численный анализ для TM-поляризации полностью аналогичен. Рассмотрим самый простой случай кругового цилиндра с радиусом a. Тогда для интегрального уравнения (551) получим

$$\mathbf{n}(M) = (\cos\theta, \sin\theta), \qquad |\mathbf{QM}| = a\sqrt{2 - 2\cos(\theta - \theta')},$$

$$\cos\varphi(\theta, \theta') = \frac{a(1 - \cos(\theta - \theta'))}{a\sqrt{2 - 2\cos(\theta - \theta')}} = \frac{1}{\sqrt{2}}\sqrt{1 - \cos(\theta - \theta')}.$$

Воспользуемся простейшим вариантом метода моментов для интегрального уравнения (551), вычисляя интеграл с помощью формулы трапеций (кусочно-линейная аппроксимация решения). Для равномерной сетки

$$\theta_n = h(n-1), \qquad h = \frac{2\pi}{N}, \qquad n = 1, 2, ..., N+1, \qquad \nu_n = \nu(\theta_n), \quad (555)$$

мы получим систему из $N+1$ линейных алгебраических уравнений

$$\nu_n = \nu_0(\theta_n) + \frac{h}{2}\left(K^{(1)}(\theta_n, \theta_1)\nu_1 + 2\sum_{m=2}^{N} K^{(1)}(\theta_n, \theta_m)\nu_m + K^{(1)}(\theta_n, \theta_{N+1})\nu_{N+1} \right), \quad (556)$$

где

$$K^{(1)}(\theta_n, \theta_m) = \frac{ika}{2\sqrt{2}} H_1^{(1)}(ka\sqrt{2 - 2\cos(\theta_n - \theta_m)})\sqrt{1 - \cos(\theta_n - \theta_m)}.$$

Получив решение системы (556) в виде вектора $\{\nu_n\}_{n=1}^{N+1}$, мы в состоянии вычислить рассеянное поле в любой точке наблюдения $M = (r\cos\theta, r\sin\theta)$ вне цилиндра по формуле (531)

$$u_{sc}(M) = -\frac{\pi ik}{2} \int_S H_1^{(1)}(k|\mathbf{QM}|)\cos\varphi\nu(Q)ds, \quad (557)$$

где

$$|\mathbf{QM}| = \sqrt{(r\cos\theta - a\cos\theta')^2 + (r\sin\theta - a\sin\theta')^2} \\ = \sqrt{r^2 + a^2 - 2ra\cos(\theta - \theta')}, \quad (558)$$

$$\cos\varphi = -\frac{(r\cos\theta - a\cos\theta')\cos\theta' + (r\sin\theta - a\sin\theta')\sin\theta'}{\sqrt{(r\cos\theta - a\cos\theta')^2 + (r\sin\theta - a\sin\theta')^2}} \\ = \frac{r\cos(\theta - \theta') - a}{|\mathbf{QM}|}, \quad (559)$$

Окончательно для рассеянного поля получаем

$$u_{sc}(M) = -\frac{\pi ika}{2}\int\limits_0^{2\pi} H_1^{(1)}(k|\mathbf{QM}|)\frac{r\cos(\theta - \theta') - a}{|\mathbf{QM}|}\nu(\theta')d\theta'. \quad (560)$$

Этот интеграл можно вновь вычислить по формуле трапеций. Пусть

$$u_{sc}(M) = -\frac{\pi ika}{2}\int\limits_0^{2\pi} K(r,\theta,\theta')\nu(\theta')d\theta'. \quad (561)$$

Тогда будем иметь

$$u_{sc}(M) = \\ -\frac{\pi ikah}{4}\left(K(r,\theta,\theta_1)\nu_1 + 2\sum_{m=2}^{N} K(r,\theta,\theta_m)\nu_m + K(r,\theta,\theta_{N+1})\nu_{N+1}\right). \quad (562)$$

Для рассеянного поля в дальней зоне для цилиндра кругового сечения также можно получить представление в виде цилиндрической волны (530), уходящей на бесконечность и описываемой следующей диаграммой рассеяния (см. (538))

$$F(\theta) = -ka\sqrt{\frac{\pi}{2}}e^{-i\pi/4}\int\limits_0^{2\pi}\nu(\theta')e^{-ika\cos(\theta-\theta')}\cos(\theta - \theta')d\theta'. \quad (563)$$

И этот интеграл следует вычислить по формуле трапеций. Введем обозначение

$$\Psi(\theta,\theta') = e^{-ika\cos(\theta-\theta')}\cos(\theta-\theta'). \tag{564}$$

Тогда диаграмма рассеяния может быть вычислена по формуле

$$F(\theta) = \\ -ka\sqrt{\frac{\pi}{2}}e^{-i\pi/4}\frac{h}{2}\left(\Psi(\theta,\theta_1)\nu_1 + 2\sum_{m=2}^{N}\Psi(\theta,\theta_m)\nu_m + \Psi(\theta,\theta_{N+1})\nu_{N+1}\right). \tag{565}$$

С другой стороны, в классической электродинамике и акустике хорошо известно строгое решение задачи рассеяния плоской волны на бесконечном цилиндре с граничным условием Дирихле (нормальное падение). Это решение легко получить для рассматриваемого случая TE-поляризации. Вновь воспользуемся полярными координатами, тогда будем иметь

$$u_{in} = e^{i\mathbf{k}\cdot\mathbf{r}} = e^{ikr\cos(\theta-\theta_0)} = \sum_{m=-\infty}^{+\infty} i^m J_m(kr)e^{im(\theta-\theta_0)}, \tag{566}$$

где $J_m(kr)$ есть функция Бесселя. Будем искать решение для рассеянного поля в виде ряда Фурье

$$u_{sc} = \sum_{m=-\infty}^{+\infty} A_m H_m^{(1)}(kr)e^{im\theta}, \tag{567}$$

с неизвестными коэффициентами A_m, где $H_m^{(1)}(kr)$ есть функция Ханкеля первого рода порядка m. Коэффициентами A_m находятся из краевого условия

$$A_m = -\frac{i^m J_m(ka)e^{-im\theta_0}}{H_m^{(1)}(ka)}, \tag{568}$$

и, следовательно, мы получаем точное решение для рассеянного поля в виде бесконечного ряда и его конечную аппроксимацию

$$u_{sc} = -\sum_{m=-\infty}^{+\infty}\frac{i^m J_m(ka)e^{-im\theta_0}}{H_m^{(1)}(ka)}H_m^{(1)}(kr)e^{im\theta} \sim$$

$$- \sum_{m=-M}^{M} \frac{i^m J_m(ka) e^{-im\theta_0}}{H_m^{(1)}(ka)} H_m^{(1)}(kr) e^{im\theta}. \qquad (569)$$

Следует заметить, что при условии $ka \sim 1$ или $ka << 1$ коэффициенты A_m очень быстро убывают с ростом $|m|$, так как функции Бесселя $J_m(z)$ и Неймана $N_m(z)$ ($H_m^{(1)}(z) = J_m(z) + iN_m(z)$) имеют асимптотическое поведение при $|m| \to \infty$ (см. [21], [22]):

$$J_m(z) \sim \frac{1}{\sqrt{2\pi m}} (\frac{ez}{2m})^m, \qquad N_m(z) \sim -\sqrt{\frac{2}{\pi m}} (\frac{2m}{ez})^m. \qquad (570)$$

Однако при $ka > 10$ сходимость ряда (569) просто исчезает. Рассмотрим поведение решения (569) в дальней зоне. Для этого воспользуемся асимптотическим разложением для функции Ханкеля (см. [21], [22]):

$$H_m^{(1)}(z) = \sqrt{\frac{2}{\pi z}} e^{i(z - \frac{\pi}{4}(2m+1))} (1 + O(z^{-1})).$$

Таким образом, для точного решения рассеянного поля в дальней зоне для цилиндра кругового сечения мы вновь получаем представление в виде цилиндрической волны (530), уходящей на бесконечность, и описываемой диаграммой рассеяния в виде бесконечного ряда и его конечной аппроксимации:

$$F_{ex}(\theta) = -\sqrt{\frac{2}{\pi}} \sum_{m=-\infty}^{+\infty} \frac{i^m J_m(ka)}{H_m^{(1)}(ka)} e^{im(\theta - \theta_0) - i\frac{\pi}{4}(2m+1)} \sim$$

$$\sim -\sqrt{\frac{2}{\pi}} \sum_{m=-M}^{M} \frac{i^m J_m(ka)}{H_m^{(1)}(ka)} e^{im(\theta - \theta_0) - i\frac{\pi}{4}(2m+1)}. \qquad (571)$$

На Рис. 16-18 приводятся численные результаты расчетов для случая TE-поляризации рассеяния плоской волны с единичной амплитудой с углом $\theta_0 = 3\pi/4$, $\lambda = 5$, на цилиндре радиуса $a = 1$. Размер системы (556) был выбран равным $N = 128$. На Рис. 16 приводится график зависимости $|\nu(s)|$ от длины дуги.

На Рис. 17 мы сравниваем данные $|u_{sc}(r, \theta)|$ для $r = 2$ в зависимости от $\theta \in [0, 2\pi]$, полученные на основе интегрального уравнения (551) и точного решения (569). В дальней зоне рассеянного

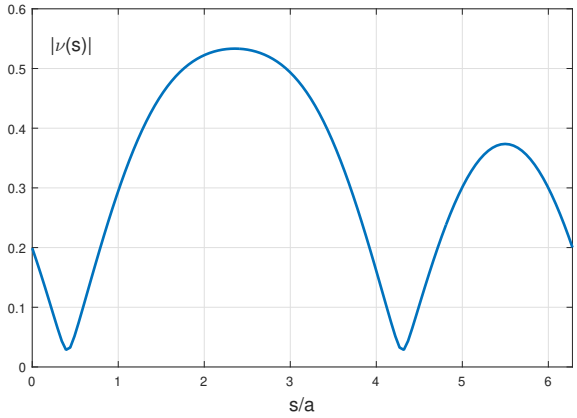


Рис. 22: Зависимость $|\nu(s)|$ от длины дуги, полученная на основе интегрального уравнения.

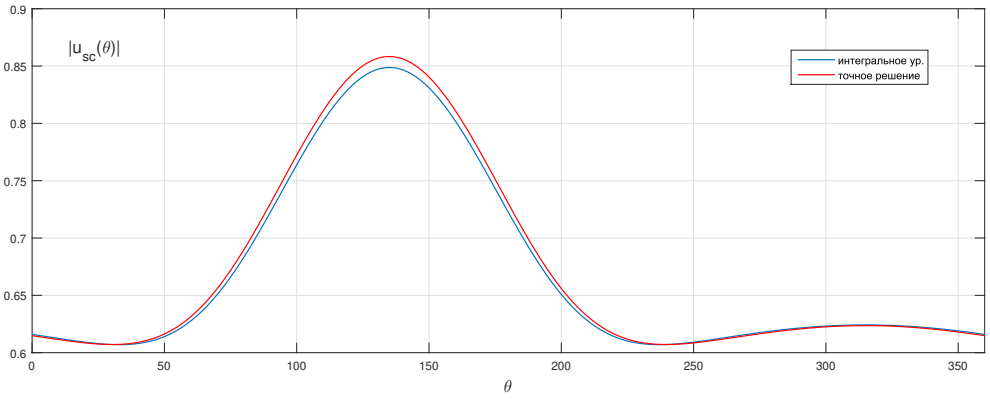


Рис. 23: Данные рассеянного поля на круглом цилиндре $|u_{sc}(r,\theta)|$ для $r = 2$ в зависимости от θ, полученные на основе интегрального уравнения и точного решения.

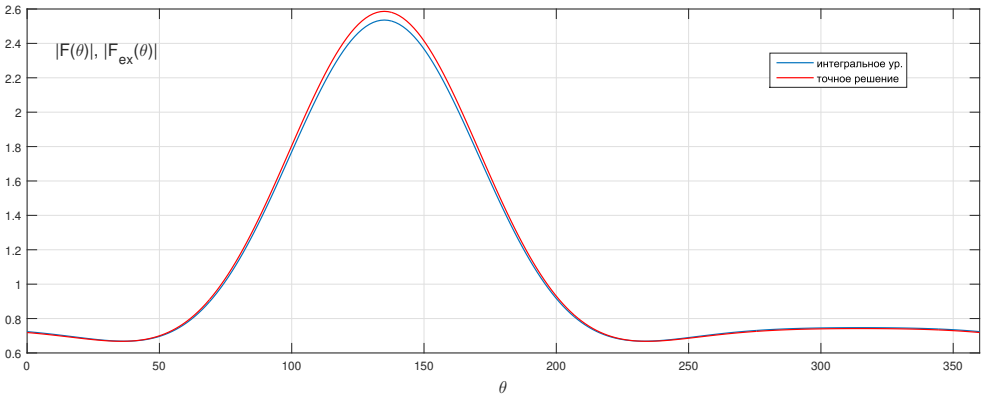


Рис. 24: Сравнение диаграмм рассеяния $|F(\theta)|$, $|F_{ex}(\theta)|$, рассеянного поля на круглом цилиндре в зависимости от θ, полученные на основе интегрального уравнения и точного решения.

поля для сравнения диаграмм рассеяния на Рис. 18 мы приводим графики для $|F(\theta)|$ (см. (538)) и $|F_{ex}(\theta)|$ (см. (571)) в зависимости от $\theta \in [0, 2\pi]$, вычисленные на основе интегрального уравнения (551) и точного решения (569). На Рис. 17-18 видно хорошее совпадение между результатами, полученными на основе интегрального уравнения и точного решения.

12.3 Рассеяние плоской электромагнитной волны на тонком, конечной длины, идеально-проводящем вибраторе

Задача рассеяния плоской электромагнитной волны на тонком, конечной длины, идеально-проводящем проводе хорошо известна прежде всего в радиофизике, в теории антенн. Заметим, что в физике метаматериалов простейшие метаповерхности являются периодическими решетками из тонких проводов. Эта задача также весьма актуальна в терагерцовой ближнепольной спектроскопии.

Рассмотрим рассеяние плоской электромагнитной волны $\mathbf{E}_{in} = \mathbf{E}_0 e^{i\mathbf{k}_0 \cdot \mathbf{r}}$ на вертикальном, тонком, конечной длины $(-L < z < L)$, идеально-проводящем вибраторе (см. Рис. 19), где $\mathbf{E}_0$ есть постоянный вектор амплитуды волны, $\mathbf{k}_0$ - волновой вектор. Вибратор имеет форму цилиндра кругового сечения, где a - радиус круга сечения

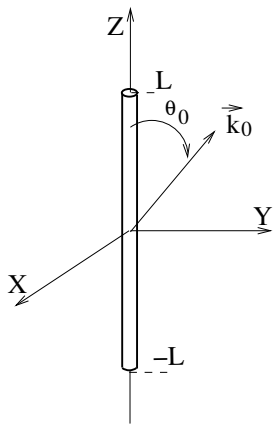


Рис. 25: Рассеяние плоской электромагнитной волны на тонком, конечной длины, идеально-проводящем вибраторе.

(см. Рис. 19). Мы построим решение этой задачи в длинноволновом приближении, то есть при условиях, что $a << \lambda$, $a << 2L$, $\lambda \sim L$, где λ - длина волны ($k = 2\pi/\lambda$). В этом случае влиянием торцов в формировании поверхностного тока можно пренебречь. Более того, оказывается что в первом приближении вектор бокового поверхностного тока имеет только продольную компоненту, которая зависит только от z координаты. Итак, в длинноволновом приближении индуцированный поверхностный ток $I(z)$ создает рассеянное поле, описываемое векторным потенциалом (см. [24])

$$A_z(x,y,z) = \frac{1}{c}\int\limits_{-L}^{L} I(z')\frac{ikR}{R}dz',$$

$$R = \sqrt{x^2+y^2+(z-z')^2+a^2}, \qquad A_x = A_y = 0, \quad (572)$$

где c - скорость света в вакууме. Электрическая компонента рассеянного поля находится по формуле (см. [24])

$$\mathbf{E}_{sc}(x,y,z) = \frac{i}{k}(\nabla \mathrm{div}\mathbf{A} + k^2\mathbf{A}). \quad (573)$$

Удовлетворим краевому условию на боковой цилиндрической поверхности вибратора - касательная компонента полного электрического поля равна нулю:

$$(\mathbf{E}_{in}(z) + \mathbf{E}_{sc}(z)) \cdot \mathbf{e}_z = 0, \qquad -L < z < L, \quad (574)$$

где $\mathbf{e}_z$ - орт оси Z. Так как

$$\mathbf{E}_{sc}(z) \cdot \mathbf{e}_z = \frac{i}{k}(\frac{d^2}{dz^2} + k^2)A_z, \tag{575}$$

и учитывая (572), то мы получаем из граничного условия интегральное уравнение Поклингтона (см. [25], [26])

$$\mathbf{E}_{in}(z) \cdot \mathbf{e}_z + \frac{i}{ck}(\frac{d^2}{dz^2} + k^2) \int\limits_{-L}^{L} I(z') \frac{e^{ikR}}{R} dz' = 0, \qquad -L < z < L, \tag{576}$$

$$R = \sqrt{(z - z')^2 + a^2},$$

с учетом краевых условий для тока

$$I(-L) = I(L) = 0. \tag{577}$$

Заметим, что это интегральное уравнение Фредгольма первого рода является квази-сингулярным ядром при $a \to 0$. Уравнение (576) для случая рассеяния плоской волны можно представить в виде

$$\frac{i}{ck}(\frac{d^2}{dz^2} + k^2) \int\limits_{-L}^{L} I(z') \frac{ikR}{R} dz' = E_0 \sin\theta_0 E^{ik\cos\theta_0 z}, \tag{578}$$

$$L < z < L, \qquad I(-L) = I(L) = 0.$$

Если обратить дифференциальный оператор $\frac{d^2}{dz^2} + k^2$ в левой части интегрального уравнения Поклингтона (578), то получим известное интегральное уравнение Галлена (см. [25], [26])

$$\frac{i}{ck} \int\limits_{-L}^{L} I(z') \frac{e^{ikR}}{R} dz' + C_1 \sin kz + C_2 \cos kz = \frac{E_0}{ik \sin\theta_0} E^{ik\cos\theta_0 z}, \tag{579}$$

$$-L < z < L$$

где неизвестные постоянные $C_{1,2}$ находятся из краевых условий

$$I(-L) = I(L) = 0.$$

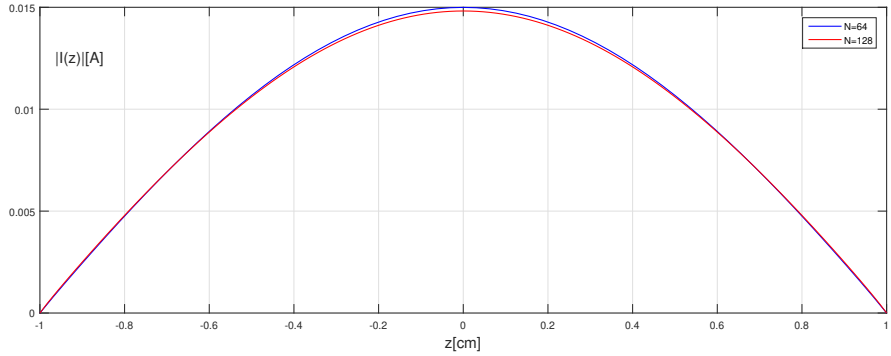


Рис. 26: Расчеты $|I(z)|$ для рассеяния плоской волны с $\lambda = 4cm$ с амплитудой $E_0 = 1V/cm$ для нормального падения $\theta_0 = \pi/2$ для вибратора с параметрами $L = 1cm$, $a = 0.001cm$, для $N = 64, 128$. Расчеты получены на основе применения метода интегрального уравнения Поклингтона. Случай полуволнового резонанса.

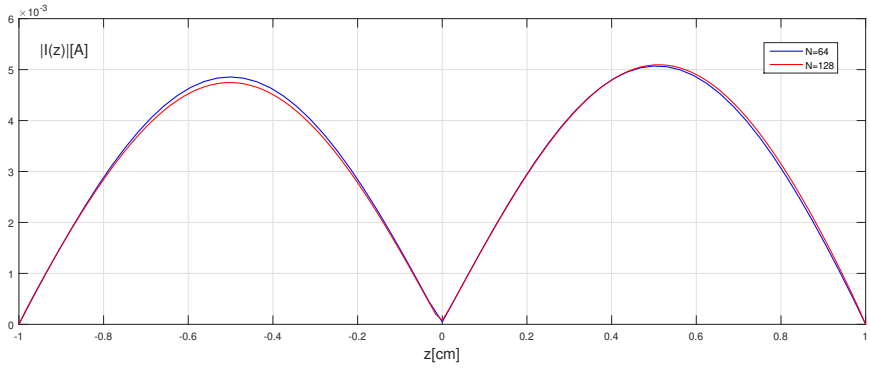


Рис. 27: Расчеты $|I(z)|$ для рассеяния плоской волны с $\lambda = 2cm$ с амплитудой $E_0 = 1V/cm$ для наклонного падения $\theta_0 = \pi/6$ для вибратора с параметрами $L = 1cm$, $a = 0.001cm$, для $N = 64, 128$. Расчеты получены на основе применения метода интегрального уравнения Поклингтона.

Это интегральное уравнение также является уравнением Фредгольма первого рода с квази-сингулярным ядром при $a \to 0$. Но степень сингулярности для уравнения Галлена меньше, чем для уравнения Поклингтона.

Воспользуемся простейшим вариантом метода моментов для интегральных уравнений (578) и (579), вычисляя интеграл с помощью формулы трапеций (кусочно-линейная аппроксимация решения) с равномерной сеткой с N узлами. На Рис. 20-21 представлены расчеты $|I(z)|$ для $-L < z < L$, полученные на основе применения метода интегрального уравнения Поклингтона (578), для рассеяния плоской волны с $\lambda = 4cm,\ 2cm$ и амплитудой $E_0 = 1V/cm$ при нормальном $\theta_0 = \pi/2$ и наклонном $\theta_0 = \pi/6$ падениях для вибратора с параметрами $L = 1cm$, $a = 0.001cm$, для $N = 64, 128$. Оба графика демонстрируют отличную сходимость метода интегрального уравнения Поклингтона.

Вопросы для самоконтроля к главе 12

1. Уравнения какого вида называются интегральным уравнением Фредгольма второго рода?

2. При каком условии метод последовательных приближений сходится к точному решению интегрального уравнения Фредгольма второго рода?

3. На основании каких причин интегральное уравнение Фредгольма второго рода с вырожденным ядром решается точно?

4. Каким образом задачу рассеяния плоской электромагнитной волны на бесконечном идеально-проводящем цилиндре можно свести к интегральному уравнению Фредгольма второго рода?

5. Что представляют собой потенциалы простого и двойного слоя? Каковы их основные свойства?

6. Каковы основные детали метода моментов численного решения интегрального уравнения Фредгольма второго рода?

7. Что представляет собой диаграмма рассеяния в задаче рассеяния плоской электромагнитной волны на бесконечном идеально-проводящем цилиндре?

8. Каким образом задачу рассеяния плоской электромагнитной волны на конечном идеально-проводящем тонком проводе можно

свести к интегральному уравнению Фредгольма первого рода?

13 Приложение: задачи для лабораторных работ

13.1 Решение систем линейных алгебраических уравнений методом последовательных исключений Гаусса. LU разложение матрицы системы

Решить систему линейных алгебраических уравнений методом последовательных исключений Гаусса.

$$Ax = f, \quad A = \begin{pmatrix} 2 & 1 & 3 \\ 11 & 7 & 5 \\ 9 & 8 & 4 \end{pmatrix}, \quad f = \begin{pmatrix} 10 \\ 2 \\ 6 \end{pmatrix}.$$

Получить LU разложение матрицы системы.

13.2 Решение систем линейных алгебраических уравнений итерационными методами Якоби, Зейделя, SOR методом

Решить систему линейных алгебраических уравнений итерационными методами Якоби, Зейделя и SOR методом

$$Ax = f, \quad A = \begin{pmatrix} 10 & 3 & 0 \\ 3 & 15 & 1 \\ 0 & 1 & 7 \end{pmatrix}, \quad f = \begin{pmatrix} 2 \\ 12 \\ 5 \end{pmatrix}.$$

с относительной погрешностью, удовлетворяющей

$$\frac{||\bar{x} - x^{(n)}||}{||\bar{x}||} \leq \epsilon,$$

где $\bar{x}$ есть точное решение, $\bar{x}^{(n)}$ - итерация с номером n, а ϵ задано (например, $\epsilon = 10^{-4}$). Сравнить количество итераций для всех трех методов.

13.3 Аппроксимация функций тригонометрическим многочленом и полиномами Чебышева

Сравним аппроксимации бесконечно дифференцируемой функции

$$f(x) = A_1 \cos \omega_1 x + A_2 \sin \omega_2 x$$

на отрезке $[-1, 1]$, полученные с помощью тригонометрического многочлена ряда Фурье

$$S_N^{(1)}(x) = \frac{a_0^{(1)}}{2} + \sum_{n=1}^{N} (a_n^{(1)} \cos(\pi n x) + b_n^{(1)} \sin(\pi n x)), \quad |x| \le 1, \quad (580)$$

$$a_n^{(1)} = \int_{-1}^{1} f(x) \cos(\pi n x) dx, \quad n = 0, 1, ..., N, \quad (581)$$

$$b_n^{(1)} = \int_{-1}^{1} f(x) \sin(\pi n x) dx, \quad n = 1, ..., N, \quad (582)$$

и аппроксимации чебышевскими многочленами

$$S_N^{(2)}(x) = \sum_{n=0}^{N} a_n^{(2)} T_n(x), \quad |x| \le 1, \quad T_n(x) = \cos(n \arccos x). \quad (583)$$

Для аппроксимации чебышевскими многочленами следует воспользоваться заменой переменной $x = \cos t$, $t \in [0, \pi]$, а именно:

$$f(\cos t) \sim S_N^{(2)}(\cos t) = \sum_{n=0}^{N} a_n^{(2)} \cos nt, \quad (584)$$

$$a_0^{(2)} = \frac{1}{\pi} \int_0^{\pi} f(\cos t) dt, \quad (585)$$

$$a_n^{(2)} = \frac{2}{\pi} \int_0^{\pi} f(\cos t) \cos nt, \quad n > 0. \quad (586)$$

Для вычисления чебышевских коэффициентов следует использовать алгоритм быстрого дискретного преобразования Фурье

$$a_0^{(2)} \sim \frac{1}{N} \sum_{m=0}^{N-1} f(\cos t_m), \qquad t_m = \frac{\pi m}{N}, \quad (587)$$

$$a_n^{(2)} \sim \frac{2}{N} \sum_{m=0}^{N-1} f(\cos t_m) \cos(n t_m), \quad n > 0. \tag{588}$$

13.4 Нахождение собственных значений матрицы с помощью метода Ньютона и собственных векторов с помощью итерационных методов Якоби, Зейделя, SOR методом

Вычислить собственные значения симметричной и положительно-определенной матрицы

$$A = \begin{pmatrix} 16 & 3 & 2 \\ 3 & 5 & 1 \\ 2 & 1 & 10 \end{pmatrix},$$

с помощью метода Ньютона. А затем вычислить соответствующие собственные векторы с помощью итерационных методов Якоби, Зейделя, а также SOR методом.

13.5 Определение диэлектрической проницаемости слоя по данным рассеяния плоской электромагнитной волны

Пусть через горизонтальный однородный слой диэектрика из диспергирующей среды с электродинамическими параметрами $\epsilon = \epsilon_1(\omega) + i\epsilon_2(\omega)$, $\mu = 1$ ($\omega = 2\pi f$ - частота) распространяется плоская электромагнитная волна (см. Рис. 1) вдоль оси Z снизу вверх. Пусть электрическая компонента поля направлена вдоль оси Y. Предполагается, что выше и ниже слоя находится ваккуум. Пусть временной сигнал $E_{tr}(t, z)$ прошедшей волны в точке наблюдения с координатой $z > 0$ задан (см. секцию 5.5. и формулу (242)). Решить обратную задачу определения функции $\epsilon(\omega)$. Для этого, используя формулу (243) и итерационную последовательность метода Ньютона для уравнения (246), следует восстановить функцию коэффициента преломления $n(\omega)$, а значит, и $\epsilon(\omega) = n^2(\omega)$, то есть

решить обратную задачу восстановления параметров диэлектрического слоя.

13.6 Нахождение минимума целевой функции, заданной в области

Вычислить минимум целевой функции

$$z = f(x, y) = \frac{1}{2}Ax^2 + Bxy + \frac{1}{2}Cy^2 - Dx - Ey + F,$$

в области $x^2 + y^2 \leq R$ аналитическим способом, с помощью метода градиентного спуска и метода Ньютона.

13.7 Вычисление двойных интегралов методом Монте-Карло

Вычислить двойной интеграл

$$I = \int_{x^2+y^2 \leq 1} (Ax^2 + By^2 + C)dxdy, \quad A > 0, \quad B > 0, \quad C > 0,$$

методом Монте-Карло двумя способами (см. секцию 7.3). Первый способ основан на применении теоремы о среднем, второй способ использует представление двойного интеграла как объема вертикального цилиндра, ограниченного сверху поверхностью $z = f(x, y) = Ax^2 + By^2 + C$. Сравнить результаты вычислений.

13.8 Решения начальной задачи Коши для линейного ОДУ первого порядка

Исследовать численные решения начальной задачи Коши для линейного ОДУ первого порядка методам Рунге-Кутта четвертого порядка и многошаговым методом Адамса-Бэшфорта-Молтона. Исследовать абсолютную и относительную погрешности вычислений для случая $y' = y, \quad y(0) = 0, \quad x \in [0, 5]$, и $y' = -y, \quad y(0) = 1, \quad x \in [0, 5]$.

13.9 Решения начальной задачи Коши для линейного ОДУ второго порядка с помощью линейного интегрального уравнения Вольтерра

Исследовать численные решения начальной задачи Коши для линейного ОДУ второго порядка методом последовательных приближений соответствующего линейного интегрального уравнения Вольтерра. Построить численное решение задачи $u'' + 3u' + 2u = cos(x)$, $u(0) = 1,\ u'(0) = 0$, для отрезка $[0, 10]$. Сравнить численное решение с точным.

13.10 Решения краевой задачи для линейного ОДУ второго порядка

Построить численное решение краевой задачи для линейного ОДУ второго порядка $y'' + k^2 q(x)y = f(x)$ на отрезке $[a, b]$ с краевыми условиями $y(a) = A,\ y(b) = B$ с помощью метода конечных разностей. Параметры k, a, b, A, B, и функции $q(x),\ f(x)$ заданы. Проверить численное решение задачи с помощью строгого решения в случае $q(x) \equiv 1$.

13.11 Одномерный фотонный кристалл для линейного ОДУ второго порядка

Исследовать задачу для линейного ОДУ второго порядка с периодическим коэффициентом:

$$y'' + (\frac{\omega}{c})^2 q(x)y = 0, \qquad q(x + d) = q(x), \tag{589}$$

где ω есть частота, c - скорость света (звука), а периодическая функция $q(x)$ с периодом d есть квадрат коэффициента преломления (индекса рефракции)

$$q(x) = \cos(\frac{2\pi}{d}x) + q_0, \quad q_0 > 1. \tag{590}$$

Построить методом конечных разностей дисперсионную диаграмму для заданного индекса рефракции с заданнымми параметрами, например, для $c = 1$, $q_0 = 2$, $d = 1$.

13.12 Спектральная задача для радиального уравнения Шредингера, описывающего движение электрона в кулоновском центральном поле

Для уравнения Шредингера для радиальной волновой функции

$$\frac{d^2R}{dr^2} + \left(2E + \frac{2Z}{r} - \frac{l(l+1)}{r^2}\right)R = 0 \tag{591}$$

с безразмерными величинами r и E рассмотреть приближенную спектральную задачу на дискретный спектр $U_{min} < \{E_{kl}\} < 0$, $k = 0, 1, 2, ...$, на отрезке $r \in [0, r_{max}]$, с учетом краевых условий $R(0) = R(r_{max}) = 0$, где значение r_{max} следует выбрать достаточно большим. Построить численное решение задачи методом конечных разностей для разных $Z = 2, 3, 4, ...$ Вычислить функции плотности вероятности $|R_{kl}(r)|^2$ электрона для значений $k = 0, 1, 2$, $l = 0, 1$. Сравнить полученные приближенные значения для спектра энергии $\{E_{kl}\}$ с точными для $k = 0, 1, 2$ и $l = 0, 1$. Проанализировать стабильность метода.

13.13 Решение краевой задачи для уравнения Пуассона для прямоугольной области

Решить краевую задачу для уравнения Пуассона

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y), \quad (x, y) \in \Omega, \qquad u|_{\partial\Omega} = \phi, \quad (x, y) \in \partial\Omega, \tag{592}$$

используя метод конечных разностей, где область Ω есть прямоугольник со сторонами $x \in [0, a]$ и $y \in [0, b]$. Используя следующие граничные условия

$$\phi(x, 0) = 0, \quad \phi(x, b) = \sin(x)/\sin(a), \quad x \in [0, a],$$

$$\phi(0,y)=0, \quad \phi(a,y)=\sinh(y)/\sinh(b), \quad y\in[0,b], \tag{593}$$

решите задачу с помощью итерационного метода и методом сведения задачи к системе линейных алгебраических уравнений для $f(x,y)=xy$ и в случае $f(x,y)=0$. Сравните оба результата. Получите точное решение в случае $f(x,y)=0$ и сравните его с полученными численными результатами.

13.14 Решение начально-краевой задачи для уравнения теплопроводности методом конечных разностей

Решите начально-краевую задачу для уравнения теплопроводности методом конечных разностей

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2}, \qquad u(x,0)=\sin(\pi x), \qquad 0\le x\le 1,$$

$$u(0,t)=u(1,t)=0, \qquad t\ge 0.$$

Решите задачу численно тремя методами: явная схема, неявная схема и метода Кранка-Николсона с параметром $\theta=0$ (явная схема), $\theta=1$ (неявная схема), $\theta=1/2$, для разных соотношений между шагами по переменным x и t. Постройте точное решение и сравните его с численными расчетами. Проанализируйте стабильность трех методов.

13.15 Решение начально-краевой задачи для волнового уравнения методом конечных разностей

Решите начально-краевую задачу для волнового уравнения методом конечных разностей

$$\frac{\partial^2 u}{\partial t^2}=\frac{\partial^2 u}{\partial x^2}, \qquad u(x,0)=\sin(\pi x), \qquad u_t(x,0)=\sin(\pi x),$$

$$0\le x\le 1, \qquad u(0,t)=u(1,t)=0, \qquad t\ge 0.$$

Примените метод конечных разностей с явной схемой для разных соотношений между шагами по переменным x и t. Постройте точное решение и сравните его с численными расчетами. Проанализируйте стабильность трех методов.

13.16 Решение задачи рассеяния плоской электромагнитной волны на бесконечном идеально-проводящем цилиндре методом граничных интегральных уравнений

Решить задачу рассеяния плоской электромагнитной волны в случае TM-поляризации (задача Дирихле) на бесконечном идеально-проводящем цилиндре методом граничного интегрального уравнения (552) для кругового цилиндра радиуса (см. секцию 12.2.). Пусть рассеивается плоская волна с единичной амплитудой с углом падения $\theta_0 = 3\pi/4$, $\lambda = 5$, на цилиндре радиуса $a = 1$. Построить точное решение в виде ряда, аналогичного (569). Сравнить численное решение с точным. Построить графики зависимостей, описанных в секции 12.2.

13.17 Решение задачи рассеяния плоской электромагнитной волны на на тонком, конечной длины, идеально-проводящем вибраторе методом интегрального уравнения Галлена

Решить задачу рассеяния плоской электромагнитной волны на на тонком, конечной длины, идеально-проводящем вибраторе методом интегрального уравнения Галлена (579) (см. секцию 12.3.). Пусть рассеивается плоская волна с единичной амплитудой и углом падения $\theta_0 = \pi/2,\ \pi/6$, $\lambda = 4, 2cm$, на цилиндре длинной $2L = 1cm$, радиуса $a = 0.001cm$. Построить графики зависимостей, описанных в секции 12.3.

Список литературы

[1] Самарский А.А., Михайлов А.П. Математическое моделирование. М.:Физматлит. 1989. -320с.

[2] Самарский А.А., Гулин А.В. Численные методы. М.:Наука, 1989. -430с.

[3] Самарский А.А. Введение в численные методы. М.:Наука, 1987. -286с.

[4] Бахвалов Н.С., Жидков Н.П., Кобельков Г.Н. Численные методы. М.:Лаборатории Базовых Знаний, 2002. -632с.

[5] Калиткин Н.Н. Численные методы. М.:Наука, 1978. -512с.

[6] Владимиров В.С. Уравнения математической физики. М.:Наука, 1987. -512с.

[7] Тихонов А.Н., Самарский А.А. Уравнения математической физики. М.:Наука, 1977. -735с.

[8] Фихтенгольц Г.М. Курс дифференциального и интегрального исчисления. М.:ФИЗМАТЛИТ, 2001. т.1 - 616с.; т.2 - 810с.; т.3 - 662с.

[9] Смирнов В.И. Курс высшей математики. т.2 2008, - 848с.; т.3(2) 2010 - 816с.; т.4(1,2) 1981 - 551с.

[10] Burden R. and Faires J. Numerical Analysis. Brooks Cole. 2001. -896с.

[11] Kincaid D. and Cheney W. Numerical Analysis. Brooks Cole. 1996. -789с.

[12] Чивилихин С.А. Вычислительные методы в технологиях программирования. Элементы теории и практикум. СПб: СПбГУ ИТМО, 2008. -108с.

[13] Мирошниченко Г.П., Петрашень А.Г. Численные методы (учебное пособие). СПб: СПбГУ ИТМО, 2007 .- 119с. http://books.ifmo.ru/file/pdf/194.pdf

[14] Borovikov V. A. Uniform Stationary Phase Method. The Institution of Electrical Engineers, London, UK. 1994. -233c.

[15] Бреховских Л.М., Волны в слоистых средах. Издательство Москва: Наука, 1973. -343 с.

[16] Займан Дж. Принципы теории твердого тела. Издательство: М.: Мир, 1966. -478 с.

[17] Галишникова Т. Н., Ильинский А. С. Численные методы в задачах дифракции. Издательство Московского Университета. 1987. -208с.

[18] Ландау Л. Д., Лифшиц Е. М. Квантовая механика. Теоретическая физика, том 3. 4-е изд., испр. -М.: Наука. Гл. ред. физ.-мат. лит., 1989. - 768 с.

[19] Давыдов А. С. Квантовая механика. СПб.: БХВ-Петербург, 2011. — 704 с.

[20] Бабич В. М., Лялинов М. А., Грикуров В. Э. Метод Зоммерфельда-Малюжинца в задачах дифракции. Издательство Санкт-Петербургского Государственного Университета. 2003. -101с.

[21] Градштейн И. С., Рыжик И. М. Таблицы интегралов, сумм, рядов и произведений. Государственное издательство физико-математической литературы. Москва. 1963. -1108с.

[22] Абрамовиц М., Стиган И. (ред.). Справочник по специальным функциям с формулами, графиками и математическими таблицами. М.: Наука, 1979. — 832 с.

[23] Бронштейн И.Н., Семендяев К.А. Справочник по математике для инженеров и учащихся втузов. М.: Наука, 1986. — 544 с.

[24] Вайнштейн Л. А. Электромагнитные волны. М.: Радиосвязь, 1988. — 440 с.

[25] Митра Р. Вычислительные методы в электродинамике. М.: Мир, 1977. — 485 с.

[26] Сазонов Д. М. Антенны и устройства СВЧ. М.: Высшая школа, 1988. — 431 с.

Залипаев Виктор Васильевич
Гулевич Дмитрий Романович

Численные методы в физике и технике

Учебное пособие

В авторской редакции
Редакционно-издательский отдел Университета ИТМО
Зав. РИО Н.Ф. Гусарова
Подписано к печати
Заказ №
Тираж
Отпечатано на ризографе

Редакционно-издательский отдел
Университета ИТМО
197101, Санкт-Петербург, Кронверский пр., 49