Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики

Лекции по курсу

Численные методы

Лекторы
А. В. Гулин, Н. И. Ионкин

Москва, 2013

Оглавление

Предисловие от авторов

Читателю предлагается курс лекций по численным методам, который авторы читали в течение десятков лет студентам III – IV курсов программистских кафедр факультета ВМК МГУ. Безусловно, программа и содержание курса неоднократно менялись как в связи с обновлением курса, так и в связи с преобразованиями учебных планов, происходившими в разные годы на факультете. Здесь представлен вариант курса, читаемого в последние годы.

Решение издать курс лекций обусловлено постоянными из года в год просьбами студентов, слушающих этот курс, оформить лекции в печатной и электронной версиях.

Данный курс лекций ориентирован на студентов, основной специализацией которых не является разработка и обоснование численных методов решения прикладных задач.

Тем не менее, одной из главных задач этого курса является обретение студентами навыка ориентирования в области численных методов, умения применять к решению прикладных задач основополагающие приемы построения и исследования вычислительных алгоритмов.

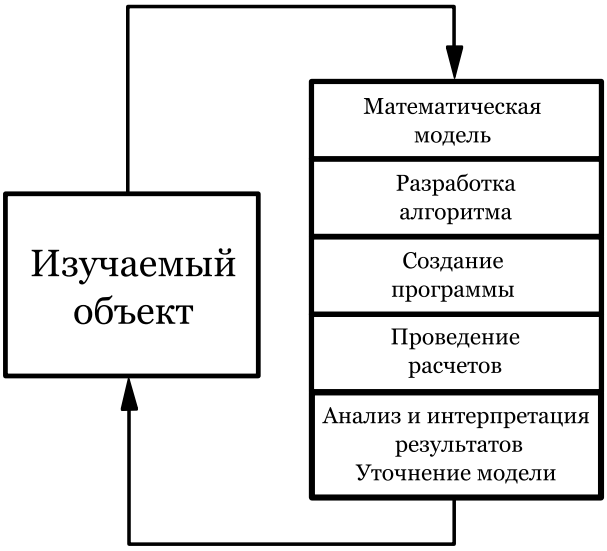
Авторы считают своим приятным долгом выразить благодарность студентам кафедры АСВК В. С. Алтухову, М. А. Казачук, М. В. Коростелевой, С. В. Селецкому, А. В. Фролову, и В. И. Шахуро, которые с энтузиазмом и творчески записали и оформили лекции курса.

Заслуженный профессор МГУ, А. В. Гулин
Заслуженный преподаватель МГУ, доцент Н. И. Ионкин

Введение

Настоящее пособие представляет собой конспект курса лекций по численным методам, читаемого студентам третьего курса факультета вычислительной математики и кибернетики МГУ им. М.В.Ломоносова.

Численные методы — это методы приближенного решения математических задач, сводящиеся к выполнению конечного числа элементарных операций над входными данными. В современном мире численные методы играют огромную роль не только как способ проведения и оптимизации расчетов для конкретных прикладных задач, но и как часть системы научного познания; они позволяют расширить наши представления об окружающем мире. Наглядно демонстрирует данный факт принцип колеса Самарского, использующийся при изучении объектов и явлений окружающего мира математическими методами.



Принцип колеса Самарского заключен в следующем: сначала по изучаемому объекту строится его математическая модель, которая отражает существенные в данной задаче свойства изучаемого объекта. Затем для построенной модели предлагается алгоритм решения поставленной задачи и приводится его формальное обоснование. По предложенному алгоритму создается программа для выполнения численных расчетов на ЭВМ, после чего уже производятся сами расчеты, анализ результатов выполнения алгоритма, их интерпретация и, возможно, уточнение модели. Получение новых данных расширяет существующие знания об изучаемом объекте, появляются новые задачи, и колесо Самарского замыкается.

В рамках данного курса численных методов рассматривается этап разработки алгоритма для некоторых классов математических моделей. Мы предполагаем, что каждая из рассматриваемых нами математических моделей построена корректно (рассмотрение решения задач для некорректных математических моделей выходит за рамки нашего курса).

Данный курс разделен на пять глав. В главе I рассматриваются прямые и итерационные численные методы решения систем линейных алгебраических уравнений, а также исследуются итерационные методы решения частичной и полной проблем собственных значений. В главе II представлены методы интерполирования и приближения функций. В главе III описаны методы решения нелинейных уравнений и систем нелинейных уравнений. На практике часто встречается задача численного решения дифференциальных уравнений, которой посвящены главы IV и V. Так, в главе IV приводятся описание и анализ разностных методов решения задач математической физики. А в заключительной, пятой, главе рассматриваются методы численного решения задач Коши для обыкновенных дифференциальных уравнений.

Свежую версию лекций можно скачать на странице github.com/shahurik/num-cmc.
Кроме того, по всем вопросам можно написать по адресу num-cmc@ya.ru.

Список обозначений

$\mathbb{N}$ — множество натуральных чисел: $\{1,2, \dots\}$;
$\mathbb{Z}$ — множество целых чисел;
$\mathbb{Z}_+$ — множество целых неотрицательных чисел;
$\mathbb{R}$ — множество вещественных чисел;
$\mathbb{R}_+$ — множество вещественных неотрицательных чисел;
$\mathbb{C}$ — множество комплексных чисел;
$f(x) = \mathrm{O}\big(g(x)\big)$ — функция f асимптотически ограничена сверху функцией g (с точностью до постоянного множителя);
$\boldsymbol{f}(x)$ — вектор-функция;
$[x]$ — целая часть числа x.

В следующих обозначениях m и n — натуральные числа.
$A\ (m \times n)$ — вещественная (если не сказано иное) матрица A, содержащая m строк и n столбцов;
$\mathbb{R}^{m\times n}$ — множество всех матриц размера $m \times n$ над полем вещественных чисел;
$\mathbb{C}^{m\times n}$ — множество всех матриц размера $m \times n$ над полем комплексных чисел.

Размер следующих матриц и вектора определяется по контексту.
θ — нулевой вектор-столбец;
E — единичная матрица;
$\mathbf{0}$ — нулевая матрица;

$\square$ — конец доказательства;
δ_{ij} — символ Кронекера:

$$\delta_{ij} = \begin{cases} 1 & \text{при } \ i = j, \\ 0 & \text{при } \ i \neq j. \end{cases}$$

Глава 1

Численные методы линейной алгебры

§1 Введение

Решение систем линейных уравнений

Рассмотрим матричное уравнение вида

$$Ax = f, \tag{1}$$

где $|A| \neq 0,\ A\ (m \times m),\ x = (x_1, x_2, \ldots, x_m)^T,\ f = (f_1, f_2, \ldots, f_m)^T$.
Так как матрица A невырождена, то решение системы (1) существует и единственно. Существуют две группы методов решения СЛАУ:

1. Прямые методы (методы Гаусса, Крамера, Холецкого и другие), позволяющие за конечное количество действий получить решение задачи. Эффективность методов этой группы оценивается по необходимому количеству умножений и делений. Несмотря на то, что эти методы часто называют точными, прямые методы таковыми не являются из-за ошибок округления при вычислении.

2. Итерационные методы (методы Якоби, Зейделя, Самарского и другие), в которых задается начальное приближение x^0 и итерационный процесс, по которому строится x^n — последовательность приближений, такая, что $\|x - x^n\| < \varepsilon$ ($\varepsilon > 0$ — точность приближения), и, следовательно, $\lim\limits_{n\to\infty} x^n = x$.

   Эффективность итерационного метода определяется числом итераций $n_0 = n_0(\varepsilon)$, необходимых для получения решения с заданной точностью ε.

Поиск собственных значений матрицы

Задача нахождения собственных значений матрицы $A\ (m \times m)$ состоит в решении уравнения

$$Ax = \lambda x,\ x \neq \theta. \tag{2}$$

Здесь λ — собственное значение, x — собственный вектор. Собственные значения находятся из уравнения $|A - \lambda E| = 0$, которое в общем случае представляет из себя многочлен степени n. Однако, как было доказано Абелем и Галуа, при $n \geqslant 5$ данное уравнение не имеет общего решения в радикалах. Таким образом, в общем виде задачу можно решить только вычислительными методами.

Рассматривают две проблемы поиска собственных значений:

1. Частичная проблема собственных значений — нахождение отдельных собственных значений (например, максимального и минимального по модулю).
2. Полная проблема собственных значений (для решения обычно используется метод QR–разложения матрицы A) — нахождение спектра (всех собственных значений) матрицы.

Нахождение обратной матрицы

Определение. *Матрица A^{-1} называется обратной к матрице A, если она удовлетворяет равенствам*

$$AA^{-1} = A^{-1}A = E.$$

Как мы помним из курса линейной алгебры, если нам известна матрица, обратная к матрице A, например, в задаче поиска решения системы линейных уравнений (1), то решение находится очень просто: $x = A^{-1}f$. В дальнейшем мы будем активно использовать понятие обратной матрицы не только в контексте прямого поиска решения, но и при исследовании на сходимость численных методов нахождения решений различных задач и оценке скорости их сходимости.

§2 Связь метода Гаусса с факторизацией матрицы

Рассмотрим матричное уравнение вида

$$Ax = f, \tag{1}$$

где $|A| \neq 0,\ A\ (m \times m),\ x = (x_1, x_2, \ldots, x_m)^T,\ f = (f_1, f_2, \ldots, f_m)^T$. Матрица A, вообще говоря, может быть матрицей с комплексными элементами.

Рассмотрим факторизацию (разложение в произведение) матрицы $A\ (m \times m)$

$$A = B \cdot C, \tag{2}$$

где B — нижнетреугольная матрица, а C — верхнетреугольная матрица с единицами на главной диагонали:

$$B = \begin{pmatrix} b_{11} & 0 & \cdots & 0 \\ b_{21} & b_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mm} \end{pmatrix}, \quad C = \begin{pmatrix} 1 & c_{12} & \cdots & c_{1m} \\ 0 & 1 & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Ясно, что не любую матрицу A можно представить в виде (2). В дальнейшем мы покажем, что нахождение элементов матриц B и C возможно при определенном ограничении на матрицу A. Запишем определение элемента матрицы A через произведение i-ой строки матрицы B и j-ого столбца матрицы C:

$$a_{ij} = \sum_{l=1}^{m} b_{il}c_{lj}.$$

Выделим j-ое слагаемое:

$$a_{ij} = \sum_{l=1}^{j-1} b_{il}c_{lj} + b_{ij}c_{jj} + \sum_{l=j+1}^{m} b_{il}c_{lj}.$$

Учитывая структуру матрицы C ($c_{lj} = 0, l > j, c_{jj} = 1$), получим

$$b_{ij} = a_{ij} - \sum_{l=1}^{j-1} b_{il}c_{lj}, \quad i \geqslant j. \tag{3}$$

Аналогично, в определении элемента матрицы A выделим i-ое слагаемое:

$$a_{ij} = \sum_{l=1}^{i-1} b_{il}c_{lj} + b_{ii}c_{ij} + \sum_{l=i+1}^{m} b_{il}c_{lj}.$$

Исходя из вида матрицы B ($b_{il} = 0, l > i$), получим

$$b_{ii}c_{ij} = a_{ij} - \sum_{l=1}^{i-1} b_{il}c_{lj}.$$

Предполагая, что $b_{ii} \neq 0$, поделим левую и правую части уравнения на b_{ii}:

$$c_{ij} = \frac{a_{ij} - \sum\limits_{l=1}^{i-1} b_{il}c_{lj}}{b_{ii}}, \quad i < j. \tag{4}$$

Несмотря на то, что уравнения (3) и (4) образуют нелинейную систему уравнений, элементы матриц B и C можно вычислить по явным формулам. Приведем алгоритм нахождения элементов матриц B и C.

1. $b_{11} = a_{11}$. Найдем элементы 1-й строки матрицы C:
$$c_{1j} = \frac{a_{1j}}{b_{11}}, \quad j = \overline{2, m}.$$
2. Рассмотрим элементы 1-ого столбца матрицы B:
$$b_{i1} = a_{i1}, \quad i = \overline{2, m}.$$
3. $b_{22} = a_{22} - b_{21}c_{12}$. Далее, аналогично 1-ому шагу, найдем элементы 2-ой строки матрицы C:
$$c_{2j} = \frac{a_{2j} - b_{21}c_{1j}}{b_{22}}, \quad j = \overline{3, m}.$$
4. Вычислим элементы 2-ого столбца матрицы B аналогично 2-ому шагу:
$$b_{i2} = a_{i2} - b_{i1}c_{12}, \quad i = \overline{3, m}.$$
5. Повторяя последовательно шаги алгоритма для столбцов матрицы B и строк матрицы C, найдем все элементы матриц B и C.

Утверждение. *Пусть все угловые миноры матрицы A отличны от нуля. Тогда представление матрицы A в виде (2) существует и единственно.*

Доказательство. Обозначим $|A_1| = a_{11} \neq 0$, $A_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \ldots, A_i = \begin{vmatrix} a_{11} & \ldots & a_{1i} \\ \vdots & \ddots & \vdots \\ a_{i1} & \ldots & a_{ii} \end{vmatrix}$, $i = \overline{1, m}$.

Поскольку $|A_i| \neq 0$, введем для определенности $|A_0| = 1$.

$$A_i = B_i \cdot C_i, \quad i = \overline{1, m}.$$

Подсчитаем значение определителя матрицы A_i, приняв во внимание вид матриц C_i и B_i и равенство $|C_i| = 1$:

$$|A_i| = |B_i||C_i| = \underbrace{b_{11}b_{22} \cdot \ldots \cdot b_{i-1,i-1}}_{|A_{i-1}|} b_{ii},$$

$$b_{ii} = \frac{|A_i|}{|A_{i-1}|} \neq 0, \quad i = \overline{1,m}.$$

Подставив b_{ii} в формулы (3) и (4), получим факторизацию матрицы A. Следовательно, факторизация матрицы A в виде (2) существует и определяется единственным образом. □

Задача. *Показать, что для вычисления элементов матриц B и C по формулам (3) и (4) требуется $\frac{m^3-m}{3}$ умножений и делений.*

Решение. Оценим необходимое количество операций для вычисления элементов b_{ij} по формуле (3). Для вычисления фиксированного b_{ij} потребуется $(j-1)$ умножение. Зафиксировав i и учитывая, что $i \geqslant j$, получим

$$\sum_{j=1}^{i}(j-1) = \frac{i(i-1)}{2}.$$

Далее, варьируя i от 1 до m, получим

$$\sum_{i=1}^{m}\frac{i(i-1)}{2} = \frac{1}{2}\left(\sum_{i=1}^{m}i^2 - \sum_{i=1}^{m}i\right) = \frac{1}{2}\left(\frac{m(m+1)(2m+1)}{6} - \frac{m(m+1)}{2}\right) = \frac{m(m-1)(m+1)}{6}.$$

Оценим необходимое количество операций для вычисления элементов c_{ij} по формуле (4). Для вычисления фиксированного c_{ij} потребуется $(i-1)$ умножение и одно деление. При фиксированном j получим

$$\sum_{i=1}^{j-1} i = \frac{j(j-1)}{2}.$$

Далее, варьируя j от 1 до m, получим аналогичную формулу:

$$\sum_{j=1}^{m}\frac{j(j-1)}{2} = \frac{m(m-1)(m+1)}{6}.$$

Сложив необходимое количество операций для вычисления b_{ij} и c_{ij}, получим искомый результат:

$$\frac{m(m-1)(m+1)}{6} + \frac{m(m-1)(m+1)}{6} = \frac{m^3-m}{3}.$$

□

Замечание. *Классическим методом решения СЛАУ вида (1) является метод Гаусса. Кратко вспомним, в чем он заключается:*

1. *Прямой ход. С помощью элементарных преобразований матрица $[A|f]$, получаемая приписыванием к матрице A вектор-столбца f правых частей системы уравнений (1), приводится к матрице $[A'|f']$, где A' — верхнетреугольная матрица с единицами на главной диагонали:*

$$[A|f] \to \ldots \to [A'|f'].$$

На этом этапе мы получили новую СЛАУ

$$A'y = f', \tag{5}$$

эквивалентную данной: ее решение совпадает с решением исходной задачи.

2. *Обратный ход метода Гаусса. Последовательно, начиная с последнего уравнения СЛАУ* (5) *и поднимаясь к первому, по явным формулам вычисляются все компоненты решения системы.*

Число действий, необходимое для преобразований в прямом ходе метода Гаусса равно $\frac{m^3-m}{3}$. *Подробный подсчет числа действий можно найти, например, в [8]. Заметим, что матрица* A', *к которой приводится матрица* A *в прямом ходе метода Гаусса, в точности совпадает с матрицей* C, *полученной в результате факторизации матрицы в виде* (2). *Таким образом факторизация матрицы в виде* (2) *требует такое же число действий, что и сведение матрицы* A *к* A' *в прямом ходе метода Гаусса.*

В матричном уравнении (1) подставим $A = BC$: $BCx = f$, обозначим $Cx = y$ и получим две системы уравнений с треугольными матрицами:

$$\begin{cases} By = f & (6) \\ Cx = y, \qquad y = (y_1, \ldots, y_m)^T. & (7) \end{cases}$$

Запишем i-ое уравнение системы (6):

$$b_{i1}y_1 + b_{i2}y_2 + \ldots + b_{ii}y_i = f_i, \quad i = \overline{1,m}.$$

Предполагая, что $b_{ii} \neq 0$, получим

$$y_i = \frac{f_i - \sum\limits_{l=1}^{i-1} b_{il}y_l}{b_{ii}}.$$

Для вычисления y_i требуется $(i-1)$ умножение и 1 деление — всего i операций. Учитывая, что i изменяется от 1 до m, получим, что для решения системы (6) требуется $1+2+\ldots+m = \frac{m(m+1)}{2}$ операций.

Замечание 1. *На вычисление новых правых частей, т.е. вектора* f', *в методе Гаусса уходит* $\frac{m(m+1)}{2}$ *действий. Как мы можем видеть, это число совпадает с количеством операций, необходимых для вычисления вектора* y *при решении системы* (6).

Аналогично, запишем i-ое уравнение системы (7):

$$x_i + c_{i,i+1}x_{i+1} + \ldots + c_{im}x_m = y_i,$$

$$x_i = y_i - \sum_{l=i+1}^{m} c_{il}x_l, \quad i = \overline{1,m}.$$

Для вычисления x_i требуется $(m-i)$ умножений. Изменяя i от 1 до m, получим, что для решения системы (7) требуется $(m-1) + (m-2) + \ldots + 2 + 1 = \frac{m(m-1)}{2}$ умножений.

Замечание 2. *Число операций, затрачиваемых на выполнение обратного хода метода Гаусса, равно* $\frac{m(m-1)}{2}$, *что совпадает с числом действий, требуемых для решения системы* (7).

В итоге получим, что для решения систем (6) и (7) требуется $\frac{m(m-1)}{2} + \frac{m(m+1)}{2} = m^2$ операций. Тогда все решение системы (1) с использованием факторизации матриц требует $\frac{m^3-m}{3} + m^2 = \frac{m^3+3m^2-m}{3}$ операций, что равно общему числу операций, необходимых для решения этой же системы методом Гаусса. Таким образом, решение системы (1) методом Гаусса эквивалентно по числу операций факторизации матрицы и решению двух систем уравнений.

Замечание 3. *Возникает вопрос о необходимости решения СЛАУ (1) именно с использованием факторизации вместо классического метода Гаусса. Выигрыш по числу операций обусловлен особенностями задач, встречающихся на практике: как правило, решаются целые серии задач с одной и той же матрицей A, которая описывает математическую модель изучаемого объекта или процесса, и с различными правыми частями f, которые соответствуют изменяющимся входным условиям. Таким образом, можно один раз факторизовать матрицу A, а затем для нахождения решения каждой задачи решать лишь СЛАУ вида (6) и (7) для каждого наблюдения.*

§3 Обращение матрицы методом Гаусса-Жордана

Рассмотрим задачу обращения (поиска обратной матрицы) невырожденной матрицы A ($m \times m$). Согласно критерию обратимости матрицы, для невырожденной матрицы всегда существует обратная. Введем обозначение: $A^{-1} = X = (x_{ij}),\ i, j = \overline{1, m}$. С учетом этого задача обращения матрицы состоит в решении системы

$$AX = E, \tag{1}$$

где A ($m \times m$), $|A| \neq 0$, или, если записать поэлементно:

$$\sum_{l=1}^{m} a_{il} x_{lj} = \delta_{ij}. \tag{2}$$

Можно приступить к решению последней системы методом Гаусса без учета структуры матрицы коэффициентов. Эта система имеет m^2 неизвестных переменных, число требуемых для решения операций будет пропорционально m^6. Покажем, что существует способ обращения матрицы, требующий ровно m^3 операций. Более того, в случае, если матрица A имеет специальную структуру (например, если матрица A — блочная или трехдиагональная), число операций уменьшится.

Сведем уравнение (2) к решению m систем линейных уравнений с матрицей A. Для этого введем вектор-столбец матрицы X: $X^{(j)} = (x_{1j}, x_{2j}, \ldots, x_{mj})^T$ и вектор-столбец правой части $\delta^{(j)} = (0, 0, \ldots, 0, 1, 0, \ldots, 0)^T$ с единицей на j-ой позиции. Теперь можем записать матричное уравнение (1) в виде m систем:

$$AX^{(j)} = \delta^{(j)}, \quad j = \overline{1, m}. \tag{3}$$

Факторизуем матрицу A в виде

$$A = B \cdot C. \tag{4}$$

Для этого требуется $\frac{m^3-m}{3}$ умножений и делений. Получаем две системы линейных уравнений:

$$\begin{cases} By^{(j)} = \delta^{(j)} & (5) \\ Cx^{(j)} = y^{(j)}. & (6) \end{cases}$$

При фиксированном j решение систем (5) и (6) требует число действий, равное m^2. Для решения m таких систем при $j = \overline{1,m}$ потребуется m^3 действий. Значит, в целом для обращения матрицы A необходимо $m^3 + \frac{m^3-m}{3} \sim \frac{4}{3}m^3$ операций. Покажем теперь, что это число операций можно уменьшить. Рассмотрим систему уравнений (5):

$$\begin{aligned} b_{11}y_1^{(j)} = 0 \quad &\Rightarrow \quad y_1^{(j)} = 0, \\ b_{21}y_1^{(j)} + b_{22}y_2^{(j)} = 0 \quad &\Rightarrow \quad y_2^{(j)} = 0, \\ b_{31}y_1^{(j)} + b_{32}y_2^{(j)} + b_{33}y_3^{(j)} = 0 \quad &\Rightarrow \quad y_3^{(j)} = 0, \\ &\dots \\ b_{j-1,1}y_1^{(j)} + \ldots + b_{j-1,j-1}y_{j-1}^{(j)} = 0 \quad &\Rightarrow \quad y_{j-1}^{(j)} = 0. \end{aligned}$$

Рассмотрим j-ое уравнение: $b_{jj}y_j^{(j)} = 1$. Предполагая, что $b_{jj} \neq 0$, получим:

$$y_j^{(j)} = \frac{1}{b_{jj}}. \tag{7}$$

Запишем уравнения системы при $i > j$

$$b_{ij}y_j^{(j)} + b_{i,j+1}y_{j+1}^{(j)} + \ldots + b_{ii}y_i^{(j)} = 0, \quad i = \overline{(j+1),m}, \tag{8}$$

и выразим из них y_i:

$$y_i^{(j)} = \frac{-\sum\limits_{l=j}^{i-1} b_{il}y_l^{(j)}}{b_{ii}}, \quad i = \overline{(j+1),m}. \tag{9}$$

Перейдем к подсчету числа операций, необходимых для решения систем уравнений (5) и (6). При фиксированных i и j в формуле (9) получаем $(i-j)$ умножений и одно деление в уравнении (7). Варьируя индекс i от 1 до m, при фиксированном j получаем

$$(m-j) + (m-j-1) + \ldots + 1 = \frac{(m-j)(m-j+1)}{2}$$

умножений и $(m-j+1)$ делений. Таким образом, число действий, необходимое для решения одной системы (5) равно

$$\frac{(m-j)(m-j+1)}{2} + \frac{2(m-j+1)}{2} = \frac{(m-j+1)(m-j+2)}{2}.$$

Общее число действий, необходимое для решения всех систем (5) равно

$$\sum_{j=1}^{m} \frac{(m-j+1)(m-j+2)}{2}. \tag{10}$$

Задача. *Показать, что сумма* (10) *равна* $\frac{m(m+1)(m+2)}{6}$.

Решение. Сделаем замену $k = m - j + 1$ в формуле (10):

$$\sum_{j=1}^{m} \frac{(m-j+1)(m-j+2)}{2} = \sum_{k=1}^{m} \frac{k(k+1)}{2}.$$

Преобразовав полученное выражение, получим искомый результат:

$$\frac{1}{2}\left(\sum_{k=1}^{m} k^2 + \sum_{k=1}^{m} k\right) = \frac{1}{2}\left(\frac{m(m+1)}{2} + \frac{m(m+1)(2m+1)}{6}\right) = \frac{m(m+1)(m+2)}{6}.$$

□

Аналогично получаем, что число операций для решения всех систем вида (6) равно $\frac{m^2(m-1)}{2}$. Просуммируем число операций для факторизации исходной матрицы и для решения систем (5) и (6) при $j = \overline{1,m}$:

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

Описанный выше метод обращения произвольной невырожденной матрицы называется методом Гаусса-Жордана. Отметим, что он является самым эффективным методом обращения невырожденных матриц произвольного вида.

§4 Метод квадратного корня

Определение. *Квадратная матрица A называется эрмитовой (самосопряженной), если ее элементы связаны соотношением $a_{ij} = \overline{a_{ji}}$ или $A = A^*$.*

Рассмотрим задачу

$$Ax = f, \tag{1}$$

где $A \in \mathbb{C}^{m\times m}$, $A = A^*$, $|A| \neq 0$, $x = (x_1, x_2, \ldots, x_m)^T$, $f = (f_1, f_2, \ldots, f_m)^T$, и один из прямых методов ее решения — метод квадратного корня (метод Холецкого).

Заметим, что хотя класс эрмитовых матриц с точки зрения линейной алгебры достаточно узок, на практике часто возникают модели, описываемые именно этим классом матриц. Поэтому с практической точки зрения такое ограничение на систему (1) вполне допустимо. Факторизуем эрмитову матрицу A в виде

$$A = S^*DS, \tag{2}$$

где матрица S — верхнетреугольная матрица с положительными элементами на главной диагонали, а D — диагональная матрица со значениями ± 1 на главной диагонали:

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ 0 & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_{mm} \end{pmatrix}, \ s_{ii} > 0, \qquad D = \begin{pmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{mm} \end{pmatrix}, \ d_{ii} = \pm 1.$$

Покажем, что факторизация (2) возможна на примере вещественной симметрической матрицы второго порядка:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = A^T, \quad a_{12} = a_{21}.$$

Матрицы S и D будем искать в виде

$$S = \begin{pmatrix} s_{11} & s_{12} \\ 0 & s_{22} \end{pmatrix}, \ s_{ii} > 0, \ i = 1, 2,$$

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

$$D = \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}, \ d_{ii} = \pm 1, \quad i = 1, 2.$$

Найдем матрицу DS:

$$DS = \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix} \begin{pmatrix} s_{11} & s_{12} \\ 0 & s_{22} \end{pmatrix} = \begin{pmatrix} d_{11}s_{11} & d_{11}s_{12} \\ 0 & d_{22}s_{22} \end{pmatrix}.$$

Домножим матрицу DS слева на S^T:

$$S^T DS = \begin{pmatrix} s_{11} & 0 \\ s_{12} & s_{22} \end{pmatrix} \begin{pmatrix} d_{11}s_{11} & d_{11}s_{12} \\ 0 & d_{22}s_{22} \end{pmatrix} = \begin{pmatrix} d_{11}s_{11}^2 & d_{11}s_{11}s_{12} \\ d_{11}s_{11}s_{12} & d_{11}s_{12}^2 + d_{22}s_{22}^2 \end{pmatrix}.$$

Приравняем элементы матриц A и $S^T DS$:

$$\begin{cases} a_{11} = d_{11}s_{11}^2 & (3) \\ a_{12} = d_{11}s_{11}s_{12} & (4) \\ a_{22} = d_{11}s_{12}^2 + d_{22}s_{22}^2. & (5) \end{cases}$$

Из неравенства $s_{11} > 0$ и из уравнения (3) следует, что

$$d_{11} = \operatorname{sgn} a_{11},\ s_{11} = \sqrt{|a_{11}|}.$$

Рассмотрим уравнение (4). Заметим, что $s_{11}d_{11} \neq 0$, и получим

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

Наконец, рассмотрим уравнение (5). Получим соотношение $s_{22}^2 d_{22} = a_{22} - d_{11}s_{12}^2$, правая часть которого известна. Следовательно,

$$d_{22} = \operatorname{sgn}(a_{22} - s_{12}^2 d_{11}),\ s_{22} = \sqrt{|a_{22} - s_{12}^2 d_{11}|}.$$

Таким образом, вещественную симметрическую матрицу второго порядка можно факторизовать в виде (2).

Рассмотрим теперь произвольную эрмитову матрицу A $(m \times m)$. Запишем уравнение для элементов матрицы DS:

$$(DS)_{ij} = \sum_{l=1}^{m} d_{il}s_{lj}, \quad i, j = \overline{1, m}.$$

Учитывая диагональную структуру матрицы D, получим:

$$(DS)_{ij} = d_{ii}s_{ij}.$$

Домножим матрицу DS слева на S^*:

$$a_{ij} = (S^* DS)_{ij} = \sum_{l=1}^{m} (S^*)_{il} d_{ll} s_{lj}, \quad i, j = \overline{1, m}.$$

Выделим i-ое слагаемое из последней суммы и учтем, что $(S^*)_{ij} = \overline{s}_{ji}$:

$$a_{ij} = \sum_{l=1}^{i-1} \overline{s}_{li} d_{ll} s_{lj} + \overline{s}_{ii} d_{ii} s_{ij} + \sum_{l=i+1}^{m} \overline{s}_{li} d_{ll} s_{lj}, \quad i, j = \overline{1, m}.$$

Третье слагаемое из равенства равно нулю в силу того, что матрица S^* является нижнетреугольной: $\overline{s}_{li} = 0,\ l > i$. Тогда получим:

$$a_{ij} = \sum_{l=1}^{i-1} \overline{s}_{li} d_{ll} s_{lj} + \overline{s}_{ii} d_{ii} s_{ij}, \quad i, j = \overline{1, m}. \tag{6}$$

Так как матрица A эрмитова, можем рассматривать это равенство только в случае $i \leqslant j$. При $i = j$ получим:

$$a_{ii} = \sum_{l=1}^{i-1} \overline{s}_{li} d_{ll} s_{li} + \overline{s}_{ii} d_{ii} s_{ii}, \quad i = \overline{1, m}.$$

Учтем, что $s_{ij}\overline{s}_{ij} = |s_{ij}|^2$:

$$a_{ii} = \sum_{l=1}^{i-1} d_{ll} |s_{li}|^2 + d_{ii} |s_{ii}|^2, \quad i = \overline{1, m},$$

$$d_{ii} |s_{ii}|^2 = a_{ii} - \sum_{l=1}^{i-1} |s_{li}|^2 d_{ll}, \quad i = \overline{1, m}.$$

Выразим d_{ii} и s_{ii}:

$$d_{ii} = \operatorname{sgn}(a_{ii} - \sum_{l=1}^{i-1} |s_{li}|^2 d_{ll}), \quad i = \overline{1, m}, \tag{7}$$

$$s_{ii} = \sqrt{\left| a_{ii} - \sum_{l=1}^{i-1} |s_{li}|^2 d_{ll} \right|}, \quad i = \overline{1, m}. \tag{8}$$

Рассмотрим случай $i \neq j$ $(i < j)$. В уравнении (6) выделим второе слагаемое:

$$\overline{s}_{ii} d_{ii} s_{ij} = a_{ij} - \sum_{l=1}^{i-1} \overline{s}_{li} d_{ll} s_{lj}, \quad i, j = \overline{1, m}.$$

В силу того, что s_{ii} — вещественные положительные числа, получим

$$s_{ii} d_{ii} s_{ij} = a_{ij} - \sum_{l=1}^{i-1} \overline{s}_{li} d_{ll} s_{lj}, \quad i, j = \overline{1, m}.$$

Так как $s_{ii} d_{ii} \neq 0$, то получим выражения для коэффициентов s_{ij}:

$$s_{ij} = \frac{a_{ij} - \sum\limits_{l=1}^{i-1} \overline{s}_{li} d_{ll} s_{lj}}{s_{ii} d_{ii}}, \quad i, j = \overline{1, m},\ i < j. \tag{9}$$

Таким образом, для вычисления элементов матриц в разложении (2) были получены явные формулы (7) – (9).

Метод квадратного корня позволяет примерно вдвое уменьшить количество операций, необходимых для решения системы (1), по сравнению с методом Гаусса — до $\sim \frac{m^3}{6}$ умножений и делений и m операций извлечения квадратного корня. Однако, метод справедлив только в случае, если матрица системы линейных уравнений эрмитова.

§5 Примеры и канонический вид итерационных методов решения СЛАУ

Рассмотрим матричное уравнение

$$Ax = f, \tag{1}$$

где $|A| \neq 0,\ A\ (m \times m),\ x = (x_1, x_2, \ldots, x_m)^T,\ f = (f_1, f_2, \ldots, f_m)^T$.
Распишем систему (1) покоординатно:

$$\sum_{j=1}^{m} a_{ij}x_j = f_i, \quad i = \overline{1, m}. \tag{2}$$

Выделим i-ое слагаемое в сумме:

$$\sum_{j=1}^{i-1} a_{ij}x_j + a_{ii}x_i + \sum_{j=i+1}^{m} a_{ij}x_j = f_i, \quad i = \overline{1, m}.$$

Предположим, что элементы главной диагонали матрицы A отличны от нуля: $a_{ii} \neq 0,\ i = \overline{1, m}$. Тогда уравнение (2) разрешимо относительно x_i:

$$x_i = \frac{f_i - \sum\limits_{j=1}^{i-1} a_{ij}x_j - \sum\limits_{j=i+1}^{m} a_{ij}x_j}{a_{ii}}, \quad i = \overline{1, m}.$$

Все итерационные методы основаны на построении последовательности векторов $x^n = (x_1^n, \ldots, x_m^n)$ такой, что $x^n \to x$ при $n \to \infty$, где x — точное решение матричного уравнения (1). Вектор x^n называется *n-ой итерацией метода*.

Отметим, что при выборе итерационного метода важно, чтобы метод был легко реализуем и сходился к решению достаточно быстро.

Определение. *Итерационный метод называется двухслойным, если для вычисления текущей итерации используются только элементы текущей и предыдущей итераций.*

Замечание. *Определенный выше итерационный метод также можно называть одношаговым.*

В этом случае для того, чтобы начать процесс построения последовательности x^n, необходимо задать начальное приближение x^0. Далее будем предполагать, что *начальное приближение уже задано*.

Рассмотрим в качестве примера два простейших из двухслойных итерационных методов: метод Якоби и метод Зейделя.

Метод Якоби

Метод Якоби является явным итерационным методом и задается уравнением

$$x_i^{n+1} = \frac{f_i - \sum\limits_{j=1}^{i-1} a_{ij}x_j^n - \sum\limits_{j=i+1}^{m} a_{ij}x_j^n}{a_{ii}}, \quad i = \overline{1, m},\ n \in \mathbb{Z}_+.$$

Забегая вперед, заметим, что метод Якоби является легко реализуемым, но при этом медленно сходящимся.

Метод Зейделя

Метод Зейделя, в отличие от метода Якоби, является неявным итерационным методом и задается уравнением

$$x_i^{n+1} = \frac{f_i - \sum\limits_{j=1}^{i-1} a_{ij} x_j^{n+1} - \sum\limits_{j=i+1}^{m} a_{ij} x_j^n}{a_{ii}}, \quad i = \overline{1,m},\ n \in \mathbb{Z}_+.$$

В правой части уравнения используются координаты $(n+1)$-ой итерации, поэтому метод Зейделя является неявным. Но если разумно организовать вычисления, то можно найти координаты $(n+1)$-ой итерации по явным формулам.
Рассмотрим метод Зейделя при $i = 1$:

$$x_1^{n+1} = \frac{f_1 - \sum\limits_{j=2}^{m} a_{1j} x_j^n}{a_{11}}, \quad n \in \mathbb{Z}_+.$$

Видно, что x_1^{n+1} находится по явной формуле. Рассмотрим вторую координату $(n+1)$-ой итерации:

$$x_2^{n+1} = \frac{f_2 - a_{21} x_1^{n+1} - \sum\limits_{j=3}^{m} a_{2j} x_j^n}{a_{22}}, \quad n \in \mathbb{Z}_+.$$

Так как координата x_1^{n+1} известна, то координату x_2^{n+1} можно найти по явной формуле. Продолжая вычисления, получим, что каждый элемент $(n+1)$-ой итерации можно найти по явным формулам от уже известных элементов. Заметим, что метод Зейделя прост в реализации, но медленно сходится.

Каноническая запись итерационных методов

Для исследования сходимости итерационных методов удобно записывать их в матричном виде. Представим матрицу A в виде

$$A = R_1 + D + R_2,$$

где R_1 — нижнетреугольная матрица с нулевой главной диагональю, D — диагональная матрица, R_2 — верхнетреугольная матрица с нулевой главной диагональю:

$$R_1 = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ a_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & 0 \end{pmatrix}, \quad D = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{mm} \end{pmatrix}, \quad R_2 = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1m} \\ 0 & 0 & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Перепишем матричное уравнение (1) в виде

$$(R_1 + D + R_2)x = f.$$

Оставим в левой части слагаемое с матрицей D, остальные слагаемые перенесем в правую часть уравнения:

$$Dx = f - R_1 x - R_2 x.$$

Предположим, что матрица D обратима ($a_{ii} \neq 0,\ i = \overline{1,m}$). Тогда получим:

$$x = D^{-1} f - D^{-1} R_1 x - D^{-1} R_2 x. \tag{3}$$

Запишем итерационные методы Якоби (МЯ) и Зейделя (МЗ) исходя из уравнения (3):

$$\begin{aligned} \text{МЯ}: &\quad x^{n+1} = D^{-1}f - D^{-1}R_1x^n - D^{-1}R_2x^n, \qquad n \in \mathbb{Z}_+, \\ \text{МЗ}: &\quad x^{n+1} = D^{-1}f - D^{-1}R_1x^{n+1} - D^{-1}R_2x^n, \quad n \in \mathbb{Z}_+. \end{aligned}$$

Рассмотрим эти два метода без обращения матрицы D:

$$\begin{aligned} \text{МЯ}: &\quad Dx^{n+1} + (R_1 + R_2)x^n = f, \quad n \in \mathbb{Z}_+, \\ \text{МЗ}: &\quad (D + R_1)x^{n+1} + R_2x^n = f, \quad n \in \mathbb{Z}_+. \end{aligned}$$

Перепишем эти соотношения в виде

$$\text{МЯ}: \quad D(x^{n+1} - x^n) + Ax^n = f, \qquad n \in \mathbb{Z}_+, \tag{4}$$

$$\text{МЗ}: \quad (D + R_1)(x^{n+1} - x^n) + Ax^n = f, \quad n \in \mathbb{Z}_+. \tag{5}$$

Из формул (4) и (5) видно, что если в каждом из методов последовательность итераций сходится, то она сходится к решению системы (1).

Мы видим, что один и тот же итерационный метод можно записать различными способами. Поэтому целесообразно ввести какую-то стандартную (каноническую) форму записи итерационных методов.

Определение. *Канонической формой записи двухслойного итерационного метода решения системы* (1) *называется его запись в виде*

$$B_{n+1}\frac{x^{n+1} - x^n}{\tau_{n+1}} + Ax^n = f, \tag{6}$$

где $n \in \mathbb{Z}_+$, начальное приближение x^0 задано, τ_{n+1} — положительное вещественное число, называемое итерационным параметром, B_{n+1} — некоторая обратимая матрица.

Определение. *Если в методе* (6) *параметр τ_{n+1} и матрица B_{n+1} не зависят от номера итерации ($B_{n+1} = B$, $\tau_{n+1} = \tau$), то такой метод называется стационарным, в противном случае — нестационарным.*

Определение. *Если $B_{n+1} = E$, то метод* (6) *называется явным, в противном случае — неявным.*

При рассмотрении итерационных методов обычно исследуют достаточные условия, при которых данный метод сходится, и оценивают скорость сходимости метода.

Рассмотрим далее еще несколько примеров итерационных методов: метод простой итерации, метод Ричардсона и попеременно-треугольный итерационный метод. В этих методах введение параметров τ и B позволяет увеличить скорость сходимости по сравнению с методами Якоби и Зейделя.

Метод простой итерации

Метод простой итерации (метод релаксации) определяется итерационной схемой вида

$$\frac{x^{n+1} - x^n}{\tau} + Ax^n = f, \ \tau > 0, \ n \in \mathbb{Z}_+, \ x^0 \text{ — задано.} \tag{7}$$

Метод Ричардсона

Метод Ричардсона определяется итерационной схемой вида

$$\frac{x^{n+1}-x^n}{\tau_{n+1}}+Ax^n=f,\ \tau>0,\ n\in\mathbb{Z}_+,\ x^0\text{ — задано.} \tag{8}$$

Замечание. *Для итерационных методов (7) и (8) в случае, когда матрица A является симметричной и положительно определенной, известен такой набор итерационных параметров (Чебышевский набор), при котором сходимость этих методов будет наиболее быстрая.*

Попеременно-треугольный итерационный метод (метод Самарского)

Представим матрицу A в виде

$$A=R_1+R_2,$$

где R_1 — нижнетреугольная матрица, R_2 — верхнетреугольная матрица:

$$R_1=\begin{pmatrix}0.5a_{11}&0&\cdots&0\\a_{21}&0.5a_{22}&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\a_{m1}&a_{m2}&\cdots&0.5a_{mm}\end{pmatrix},\quad R_2=\begin{pmatrix}0.5a_{11}&a_{12}&\cdots&a_{1m}\\0&0.5a_{22}&\cdots&a_{2m}\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&0.5a_{mm}\end{pmatrix}.$$

Итерационная схема попеременно-треугольного метода имеет вид

$$(E+\omega R_1)(E+\omega R_2)\frac{x^{n+1}-x^n}{\tau}+Ax^n=f,\ \tau>0,\ n\in\mathbb{Z}_+, \tag{9}$$

где $\tau>0,\ \omega>0$ — итерационные параметры, позволяющие, вообще говоря, ускорить процесс сходимости итерационного метода. Рассматриваемый метод формально является неявным, однако можно показать, что $(n+1)$-ая итерация выражается с помощью явных формул за три шага. Введем обозначения:

$$w^{n+1}=(E+\omega R_2)\frac{x^{n+1}-x^n}{\tau},$$

$$v^{n+1}=\frac{x^{n+1}-x^n}{\tau}.$$

Определение. *Вектор $r^n=f-Ax^n$ называется невязкой на n-ой итерации.*

В нашем случае невязка r^n известна. На первом шаге решим уравнение

$$(E+\omega R_1)w^{n+1}=r^n.$$

Заметим, что $(E+\omega R_1)$ — нижнетреугольная матрица. Нахождение вектора решения системы с нижнетреугольной матрицей осуществляется по явным формулам, начиная с первой компоненты вектора w^{n+1}. На втором шаге аналогично решим уравнение с верхнетреугольной матрицей $(E+\omega R_2)$:

$$(E+\omega R_2)v^{n+1}=w^{n+1}.$$

На третьем шаге найдем $(n+1)$-ую итерацию по формуле

$$x^{n+1}=x^n+\tau v^{n+1}.$$

Таким образом, несмотря на то, что метод Самарского является неявным, его реализация не представляет никакой трудности.

§6 Теоремы о сходимости итерационных методов

Рассмотрим матричное уравнение вида

$$Ax = f, \tag{1}$$

где $|A| \neq 0,\ A\ (m \times m),\ x = (x_1, x_2, \ldots, x_m)^T,\ f = (f_1, f_2, \ldots, f_m)^T$.
Рассмотрим также двухслойный стационарный метод решения уравнения (1):

$$B\frac{x^{n+1} - x^n}{\tau} + Ax^n = f, \tag{2}$$

где $n \in \mathbb{Z}_+$, начальное приближение x^0 задано, τ — положительное вещественное число, B — обратимая матрица размера $(m \times m)$.

Чтобы говорить о сходимости итерационного метода, необходимо ввести линейное пространство и определить в нем норму. Внимательный читатель может помнить из курса линейной алгебры, что в конечномерном пространстве все нормы эквивалентны. То есть найдутся такие константы, при помощи которых можно выразить одну норму через другую. Но при исследовании сходимости итерационных методов мы будем устремлять к нулю параметры этих методов, и если они будут участвовать в записях констант перехода от одной нормы к другой, то смысл таких оценок, вообще говоря, может сойти на нет. Поэтому всегда при рассмотрении сходимости итерационных методов мы будем указывать, в какой именно норме производится исследование.

Пусть H — линейное вещественное пространство размерности m:

$$\dim H = m.$$

Рассмотрим два произвольных вектора x и y из этого пространства:

$$x \in H, \quad x = (x_1, x_2, \ldots, x_m)^T,$$

$$y \in H, \quad y = (y_1, y_2, \ldots, y_m)^T.$$

Определим скалярное произведение двух векторов, заданных в ортонормированном базисе пространства H:

$$(x, y) = \sum_{i=1}^{m} x_i y_i.$$

Введем евклидову норму:

$$\|x\| = \sqrt{(x, x)} = \left(\sum_{i=1}^{m} x_i^2\right)^{\frac{1}{2}}.$$

Эту норму также часто называют среднеквадратичной нормой.

Далее будем считать, что понятия линейный оператор и матрица эквивалентны. Рассмотрим самосопряженный положительный линейный оператор $D = D^* > 0$.

Определение. *Линейный оператор D называется положительным (неотрицательным), если $(Dx, x) > 0\ \ \forall x \in H,\ x \neq \theta$ (соответственно $(Dx, x) \geqslant 0\ \ \forall x \in H$).*

Определение. *Скалярным произведением в смысле оператора D называется скалярное произведение, определяемое соотношением*

$$(x, y)_D = (Dx, y).$$

Определение. *Энергетической нормой, порождаемой линейным самосопряженным положительно определенным оператором D, называется норма, задаваемая соотношением*

$$\|x\|_D = \sqrt{(x,x)_D} = \sqrt{(Dx,x)}.$$

Задача. *Пусть $D = D^* > 0$. Доказать, что $\exists\ \delta > 0 :\ (Dx,x) \geqslant \delta(x,x) = \delta\|x\|^2$.*

Рассмотрим свойства положительного самосопряженного линейного оператора.
Если $D = D^* > 0$, то определены матрицы

$$D^{-1} = \left(D^{-1}\right)^* > 0,\quad D^{\frac{1}{2}} = \left(D^{\frac{1}{2}}\right)^* > 0,\quad D^{-\frac{1}{2}} = \left(D^{-\frac{1}{2}}\right)^* > 0.$$

Определение. *Погрешностью итерационного метода на n-ой итерации называется вектор*

$$v^n = x^n - x. \tag{3}$$

Определение. *Итерационный метод сходится в норме $\|\cdot\|$, если $\|v^n\| \to 0$ при $n \to \infty$.*

Выразим x^n из формулы (3) и подставим в уравнение (2). Получим однородное уравнение:

$$B\frac{v^{n+1} - v^n}{\tau} + Av^n = 0, \tag{4}$$

где $n \in \mathbb{Z}_+,\ v^0 = x^0 - x$.

Приступим к исследованию задачи (4). Выразим $(n+1)$-ую итерацию через n-ую с учетом того, что для матрицы B существует обратная. Домножим уравнение (4) на B^{-1} слева:

$$\frac{v^{n+1} - v^n}{\tau} + B^{-1}Av^n = 0.$$

Выразим из уравнения погрешность на $(n+1)$-ой итерации:

$$v^{n+1} = v^n - \tau B^{-1}Av^n = (E - \tau B^{-1}A)v^n = Sv^n.$$

Таким образом, мы получили матрицу S, которая связывает предыдущую итерацию с последующей:

$$S = E - \tau B^{-1}A. \tag{5}$$

Определение. *Матрица S из уравнения (5) называется матрицей перехода от n-ой итерации к $(n+1)$-ой.*

Теорема 1. *Итерационный метод (2) решения системы (1) сходится при любом начальном приближении тогда и только тогда, когда все собственные значения матрицы перехода S по модулю меньше единицы. (Без доказательства).*

Таким образом, сходимость итерационного метода (2) всецело зависит от свойств матрицы S, а именно, от ее спектра.

Заметим, что данная теорема практически неприменима, так как задача нахождения полного спектра матрицы S аналитически решается крайне редко.

Прежде чем приступить к рассмотрению вопроса о сходимости итерационного метода, заметим, что отныне и впредь будем считать, что линейное пространство H задано над полем $\mathbb{R}$ вещественных чисел.

Теорема 2 (теорема Самарского). *Пусть A — самосопряженный положительный оператор, τ — положительное вещественное число и выполнено матричное неравенство*

$$B - \frac{\tau}{2}A > 0. \tag{6}$$

Тогда итерационный метод (2) решения системы (1) сходится в среднеквадратичной норме при любом начальном приближении:

$$\|x^n - x\| = \sqrt{\sum_{j=1}^{m}\left(x_j^n - x_j\right)^2} \underset{n\to\infty}{\longrightarrow} 0, \quad \forall x^0.$$

Доказательство. Введем числовую последовательность $y_n = (Av^n, v^n) \geqslant 0$. Покажем, что $\{y_n\}$ — невозрастающая и ограниченная снизу последовательность. Для этого рассмотрим y_{n+1}:

$$y_{n+1} = (Av^{n+1}, v^{n+1}) = (ASv^n, Sv^n) = ((A - \tau AB^{-1}A)v^n, (E - \tau B^{-1}A)v^n). \tag{7}$$

Воспользуемся линейностью скалярного произведения и преобразуем правую часть равенства:

$$(Av^n, v^n) - \tau(Av^n, B^{-1}Av^n) - \tau(AB^{-1}Av^n, v^n) + \tau^2(AB^{-1}Av^n, B^{-1}Av^n). \tag{8}$$

В силу того, что оператор A — самосопряженный ($A = A^*$), получим

$$\left(AB^{-1}Av^n, v^n\right) = \left(B^{-1}Av^n, A^*v^n\right) = \left(Av^n, B^{-1}Av^n\right).$$

Преобразуем выражение (8):

$$y_n - \tau\left(2(Av^n, B^{-1}Av^n) - \tau(AB^{-1}Av^n, B^{-1}Av^n)\right) = y_n - 2\tau\left(\left(B - \frac{\tau}{2}A\right)B^{-1}Av^n, B^{-1}Av^n\right).$$

Подставив полученное выражение в равенство (7), получим тождество

$$\frac{y_{n+1} - y_n}{\tau} + 2\left(\left(B - \frac{\tau}{2}A\right)B^{-1}Av^n, B^{-1}Av^n\right) = 0, \tag{9}$$

в котором оператор $\left(B - \frac{\tau}{2}A\right)$ положителен по условию. Следовательно, второе слагаемое тождества неотрицательно. Отсюда следует, что $y_{n+1} \leqslant y_n$, что и означает монотонность последовательности $\{y_n\}$.

У невозрастающей последовательности $\{y_n\}$, все члены которой неотрицательны, по теореме Вейерштрасса существует предел y:

$$\lim_{n\to\infty} y_n = y.$$

Для дальнейшего доказательства нам понадобится свойство положительно определенного линейного оператора, которое мы сформулируем в виде задачи.

Задача. *Пусть H — вещественное линейное пространство, C — положительный линейный оператор в H. Доказать, что*

$$\exists\ \delta > 0 :\ (Cx, x) \geqslant \delta\|x\|^2, \quad \forall x \in H. \tag{10}$$

Воспользуемся свойством (10): существует константа $\delta > 0$ такая, что

$$\left(\left(B - \frac{\tau}{2}A\right) B^{-1}Av^n, B^{-1}Av^n\right) \geqslant \delta\|B^{-1}Av^n\|^2 \geqslant 0. \tag{11}$$

Введем вектор w^n:

$$w^n = B^{-1}Av^n. \tag{12}$$

Устремим n к бесконечности в равенстве (9):

$$\frac{y-y}{\tau} + 2\lim_{n\to\infty}\left(\left(B - \frac{\tau}{2}A\right) w^n, w^n\right) = 0.$$

Устремим теперь n к бесконечности в неравенстве (11) и примем во внимание полученное равенство:

$$0 \leqslant \lim_{n\to\infty} \delta\|w^n\|^2 \leqslant 0.$$

Получим, что

$$\lim_{n\to\infty}\|w_n\| = 0.$$

Выразим погрешность на n-ой итерации из уравнения (12):

$$v^n = A^{-1}Bw^n.$$

Так как норма произведения операторов не превосходит произведения их норм, а матрица $A^{-1}B$ не зависит от номера итерации, то получим, что погрешность стремится к нулю при n, стремящемся к бесконечности:

$$\|v^n\| \leqslant \|A^{-1}B\|\|w^n\| \underset{n\to\infty}{\longrightarrow} 0.$$

Следовательно,

$$\lim_{n\to\infty}\|v^n\| = \lim_{n\to\infty}\|x^n - x\| = 0.$$

Так как в ходе доказательства мы не использовали начальное приближение, то оно может быть произвольным. □

Следствие 1. *Пусть $A = A^* > 0$. Тогда метод Якоби сходится в среднеквадратичной норме при любом начальном приближении, если выполнено неравенство:*

$$2D > A,$$

где $A = R_1 + D + R_2$, $D = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{mm})$.

Доказательство. В методе Якоби $\tau = 1$, а $B = D$. По теореме Самарского метод сходится, если

$$B - \frac{\tau}{2}A > 0.$$

В нашем случае

$$D - \frac{1}{2}A > 0,$$

а это выполняется в силу условия $2D > A$. Следовательно, метод Якоби сходится в среднеквадратичной норме при любом начальном приближении. □

Следствие 2. *Пусть положительная самосопряженная матрица $A = A^* > 0$ является матрицей со строгим диагональным преобладанием:*

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

Тогда метод Якоби сходится в среднеквадратичной норме при любом начальном приближении x^0.

Доказательство. Рассмотрим квадратичную форму с матрицей A:

$$(Ax, x) = \sum_{i,j=1}^{m} a_{ij} x_i x_j \leqslant \sum_{i,j=1}^{m} |a_{ij}|\ |x_i|\ |x_j|. \tag{13}$$

Для дальнейшей оценки квадратичной формы (13) воспользуемся неравенством $ab \leqslant \frac{a^2+b^2}{2}$:

$$(Ax, x) \leqslant \frac{1}{2} \sum_{i,j=1}^{m} |a_{ij}|\ |x_i|^2 + \frac{1}{2} \sum_{i,j=1}^{m} |a_{ij}|\ |x_j|^2$$

Преобразуем правую часть неравенства с учетом того, что матрица A является самосопряженной ($|a_{ij}| = |a_{ji}|$):

$$\frac{1}{2} \sum_{i,j=1}^{m} |a_{ij}|\ |x_i|^2 + \frac{1}{2} \sum_{i,j=1}^{m} |a_{ij}|\ |x_i|^2 = \sum_{i,j=1}^{m} |a_{ij}|\ |x_i|^2.$$

Вынесем суммирование по индексу i и воспользуемся свойством диагонального преобладания матрицы A:

$$\sum_{i=1}^{m} |x_i|^2 \left(a_{ii} + \sum_{j=1, j\neq i}^{m} |a_{ij}| \right) < \sum_{i=1}^{m} 2x_i^2 a_{ii} = (2Dx, x),$$

где $D = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$. Таким образом, мы получили, что

$$(Ax, x) < (2Dx, x).$$

Из этого неравенства следует, что $2D > A$.

Следовательно, выполняется условие следствия 1, и итерационный метод Якоби сходится при любом начальном приближении. □

Задача. *Пусть $A = A^* > 0$. Доказать, что $a_{ii} > 0,\ i = \overline{1, m}$.*

Следствие 3. *Пусть $A = A^* > 0$. Тогда метод Зейделя сходится в среднеквадратичной норме при любом начальном приближении x^0.*

Доказательство. Из условия теоремы Самарского следует, что для сходимости метода Зейделя достаточно выполнения неравенства

$$B - \frac{\tau}{2} A > 0. \tag{14}$$

Представим матрицу A в виде $A = R_1 + D + R_2$. В канонической записи метода Зейделя $\tau = 1,\ B = R_1 + D$. Тогда достаточное условие (14) преобразуется к виду

$$D + R_1 - \frac{R_1 + D + R_2}{2} > 0.$$

И, следовательно,

$$D + R_1 - R_2 > 0. \tag{15}$$

Запишем это неравенство в виде

$$(Dx, x) + (R_1 x, x) - (R_2 x, x) > 0, \quad x \neq \theta.$$

Так как $A = A^*$, то $R_2^* = R_1$. Тогда

$$(R_2 x, x) = (x, R_2^* x) = (x, R_1 x) = (R_1 x, x).$$

Следовательно, неравенство (15) принимает вид

$$(Dx, x) > 0, \quad x \neq 0. \tag{16}$$

Если матрица самосопряженная и положительно определенная, то все ее диагональные элементы больше нуля (см. задачу). Следовательно, матрица D также является положительно определенной, откуда следует неравенство (16). □

Следствие 4. *Пусть* $A = A^* > 0$, $\gamma_2 = \max\limits_{1 \leqslant k \leqslant m} \lambda_k > 0$. *Если* $0 < \tau < \frac{2}{\gamma_2}$, *то метод простой итерации сходится в среднеквадратичной норме при любом начальном приближении* x^0.

Доказательство. Из условия теоремы Самарского следует, что для того, чтобы метод простой итерации сходился в среднеквадратичной норме при любом начальном приближении, достаточно выполнения неравенства

$$B - \frac{\tau}{2} A > 0. \tag{17}$$

В методе простой итерации $B = E$. Следовательно, условие (17) преобразуется к виду

$$E - \frac{\tau}{2} A > 0. \tag{18}$$

Неравенство (18) равносильно неравенству $\lambda^E - \frac{\tau}{2}\lambda^A > 0$, которое справедливо, если

$$1 - \frac{\tau}{2}\gamma_2 > 0.$$

Из положительности параметра τ следует, что для сходимости метода простой итерации достаточно выполнения условия

$$0 < \tau < \frac{2}{\gamma_2}.$$

□

§7 Оценка скорости сходимости итерационных методов

Рассмотрим матричное уравнение вида

$$Ax = f, \tag{1}$$

где $|A| \neq 0$, A $(m \times m)$, $x = (x_1, x_2, \ldots, x_m)^T$, $f = (f_1, f_2, \ldots, f_m)^T$ и двухслойный стационарный метод решения этого уравнения:

$$B\frac{x^{n+1} - x^n}{\tau} + Ax^n = f, \tag{2}$$

где $n \in \mathbb{Z}_+$, начальное приближение x^0 задано, τ — положительное вещественное число, B — обратимая матрица размера $(m \times m)$.

Введем погрешность $v^n = x_n - x$. Тогда из уравнения (2) получим однородную задачу:

$$B\frac{v^{n+1} - v^n}{\tau} + Av^n = 0, \quad n \in \mathbb{Z}_+, \quad v^0 = x^0 - x. \tag{3}$$

Предположим, что выполняется оценка

$$\|v^{n+1}\| \leqslant \rho\|v^n\|, \quad 0 < \rho < 1. \tag{4}$$

Тогда можно говорить о скорости сходимости итерационного метода (2) в зависимости от параметра ρ. Применив эту оценку n раз для v^n, получим:

$$\|v^n\| \leqslant \rho^n\|v^0\|. \tag{5}$$

При $0 < \rho < 1$ видно, что $\|v^n\| \underset{n\to\infty}{\longrightarrow} 0$. Заметим, что чем ближе параметр ρ к нулю, тем выше скорость сходимости метода (2). Кроме того, оценка (4) позволяет посчитать необходимое количество итераций для достижения заданной точности $\varepsilon > 0$:

$$\|x^n - x\| \leqslant \varepsilon\|x^0 - x\| \tag{6}$$

Из неравенств (5) и (6) получим

$$\rho^n \leqslant \varepsilon, \ \frac{1}{\rho^n} \geqslant \frac{1}{\varepsilon}.$$

Прологарифмируем обе части второго неравенства и выразим n:

$$n \geqslant \frac{\ln\frac{1}{\varepsilon}}{\ln\frac{1}{\rho}}.$$

Таким образом, для достижения заданной точности ε достаточно провести количество итераций, равное

$$n_0(\varepsilon) = \left[\frac{\ln\frac{1}{\varepsilon}}{\ln\frac{1}{\rho}}\right], \text{ где } [x] \text{ — целая часть числа } x.$$

Определение. *Величина* $\ln\dfrac{1}{\rho}$ *называется скоростью сходимости итерационного метода.*

Пусть H — вещественное линейное пространство размерности m. Введем в H скалярное произведение и среднеквадратичную норму:

$$(x, y) = \sum_{i=1}^{m} x_i y_i,$$

$$\|x\| = \sqrt{(x, x)}.$$

Пусть $D = D^* > 0$. Введем энергетическую норму, порождаемую оператором D:

$$\|x\|_D = \sqrt{(Dx, x)}.$$

В пространстве H существует ортонормированный базис $\{e_k\}$ из собственных векторов оператора D:

$$De_k = \lambda_k^D e_k, \ e_k \neq \theta, \ k = \overline{1, m},$$

$$(e_i, e_j) = \delta_{ij} = \begin{cases} 1 & \text{при } i = j, \\ 0 & \text{при } i \neq j, \end{cases} \quad i, j = \overline{1, m}.$$

Тогда любой вектор $x \in H$ можно однозначно разложить по этому базису:

$$x = \sum_{i=1}^{m} c_k e_k, \; c_k = (x, e_k).$$

Кроме того, в линейном пространстве с заданной в нем нормой и ортонормированным базисом выполняется равенство Парсеваля:

$$\|x\|^2 = \sum_{k=1}^{m} c_k^2, \quad x \in H. \tag{7}$$

Теорема 1 (об оценке скорости сходимости)**.** *Пусть* $A = A^* > 0, B = B^* > 0$. *Пусть также существует* ρ, $0 < \rho < 1$, *такое, что выполнено операторное неравенство:*

$$\frac{1-\rho}{\tau} B \leqslant A \leqslant \frac{1+\rho}{\tau} B. \tag{8}$$

Тогда для итерационного метода (2) *решения системы* (1) *справедлива оценка:*

$$\|v^{n+1}\|_B \leqslant \rho \|v^n\|_B, \quad n \in \mathbb{N}. \tag{9}$$

Доказательство. Так как $B = B^* > 0$, то существует матрица $B^{-\frac{1}{2}} = \left(B^{-\frac{1}{2}}\right)^*$. Домножим обе части уравнения (3) на $B^{-\frac{1}{2}}$ слева:

$$B^{\frac{1}{2}} \frac{v^{n+1} - v^n}{\tau} + B^{-\frac{1}{2}} A v^n = 0. \tag{10}$$

Введем вектор $z^n = B^{\frac{1}{2}} v^n$ и перепишем задачу (10) через вектор z^n:

$$\frac{z^{n+1} - z^n}{\tau} + B^{-\frac{1}{2}} A B^{-\frac{1}{2}} z^n = 0.$$

Выразим z^{n+1} через z^n:

$$z^{n+1} = z^n - \tau B^{-\frac{1}{2}} A B^{-\frac{1}{2}} z^n = S z^n.$$

Здесь матрица

$$S = E - \tau B^{-\frac{1}{2}} A B^{-\frac{1}{2}} \tag{11}$$

называется матрицей перехода от n-ой итерации к $(n+1)$-ой итерации вектора z.
В силу определения z^{n+1} и с учетом самосопряженности оператора B верно равенство

$$\|z^{n+1}\|^2 = (z^{n+1}, z^{n+1}) = (B^{\frac{1}{2}} v^{n+1}, B^{\frac{1}{2}} v^{n+1}) = (B v^{n+1}, v^{n+1}) = \|v^{n+1}\|_B^2.$$

Таким образом, чтобы доказать утверждение теоремы, достаточно получить оценку

$$\|z^{n+1}\| \leqslant \rho \|z^n\|.$$

Покажем, что S — самосопряженный оператор:

$$S^* = \left(E - \tau B^{-\frac{1}{2}} A B^{-\frac{1}{2}}\right)^* = E - \tau \left(B^{-\frac{1}{2}}\right)^* A^* \left(B^{-\frac{1}{2}}\right)^* = S.$$

Пусть s_k – собственные значения матрицы S. В силу самосопряженности матрицы S в линейном пространстве существует ортонормированный базис из собственных векторов оператора S:

$$Se_k = s_k e_k,\ e_k \neq \theta,\ k = \overline{1,m}. \tag{12}$$

Покажем, что все собственные значения s_k не превосходят по модулю ρ: $|s_k| \leqslant \rho$, $k = \overline{1,m}$.

Подставим выражение S из (11) в уравнение (12) и умножим слева обе части равенства на $B^{\frac{1}{2}}$:

$$\left(B^{\frac{1}{2}} - \tau A B^{-\frac{1}{2}}\right) e_k = s_k B^{\frac{1}{2}} e_k, \quad k = \overline{1,m}.$$

Введем вектор $y = B^{-\frac{1}{2}} e_k$ и перепишем это равенство в виде

$$(B - \tau A)y = s_k B y, \quad k = \overline{1,m}.$$

Отсюда следует равенство:

$$Ay = \frac{1 - s_k}{\tau} By.$$

Умножим левую и правую части этого равенства скалярно на вектор y:

$$(Ay, y) = \frac{1 - s_k}{\tau}(By, y).$$

Воспользуемся неравенством (8) из условия теоремы:

$$\frac{1-\rho}{\tau}(By, y) \leqslant \frac{1 - s_k}{\tau}(By, y) \leqslant \frac{1+\rho}{\tau}(By, y).$$

Из данных неравенств и неравенства $y \neq \theta$ следует

$$|s_k| \leqslant \rho,\ k = \overline{1,m}.$$

Разложим вектор z^n по ортонормированному базису $\{e_k\}$ из собственных векторов матрицы S:

$$z^n = \sum_{k=1}^{m} c_k^{(n)} e_k,\ c_k^{(n)} = (z^n, e_k).$$

Найдем разложение для z^{n+1}:

$$z^{n+1} = Sz^n = \sum_{k=1}^{m} c_k^{(n)} S e_k = \sum_{k=1}^{m} c_k^{(n)} s_k e_k.$$

Запишем равенство Парсеваля (7) для z^{n+1}:

$$\|z^{n+1}\|^2 = \sum_{k=1}^{m} \left(c_k^{(n)} s_k\right)^2.$$

В силу того, что спектр матрицы S по модулю не превосходит ρ, верно неравенство

$$\|z^{n+1}\|^2 \leqslant \rho^2 \sum_{k=1}^{m} \left(c_k^{(n)}\right)^2 = \rho^2 \|z^n\|^2.$$

Из этого неравенства следует оценка $\|z^{n+1}\| \leqslant \rho \|z^n\|$, которая, как мы показали выше, эквивалентна утверждению теоремы. □

Замечание. *Оценка* (9) *справедлива и в энергетической норме* $\|\cdot\|_A$.

Следствие 1. *Пусть* A, B *— самосопряженные положительно определенные операторы, и пусть существуют* $\gamma_2 > \gamma_1 > 0$, *для которых выполняется условие*

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B.$$

Тогда, если

$$\tau = \tau_0 = \frac{2}{\gamma_1 + \gamma_2},$$

то двухслойный итерационный метод решения системы уравнений сходится, и верна оценка

$$\|x^{n+1} - x\|_B \leqslant \rho \|x^n - x\|_B, \tag{13}$$

где $\rho = \frac{1-\xi}{1+\xi},\ \xi = \frac{\gamma_1}{\gamma_2}$.

Доказательство. Для того, чтобы воспользоваться теоремой 1, рассмотрим неравенство (8) из условия теоремы. Очевидно, что $\gamma_1 = \frac{1-\rho}{\tau}$ и $\gamma_2 = \frac{1+\rho}{\tau}$. Сложив эти равенства, получим

$$\gamma_1 + \gamma_2 = \frac{2}{\tau},\ \tau = \frac{2}{\gamma_1 + \gamma_2}.$$

Вычитая из второго равенства первое, получим

$$\gamma_2 - \gamma_1 = \frac{2\rho}{\tau} = \rho(\gamma_1 + \gamma_2),$$

$$\rho = \frac{\gamma_2 - \gamma_1}{\gamma_1 + \gamma_2} = \frac{1-\xi}{1+\xi},\ \xi = \frac{\gamma_1}{\gamma_2}.$$

Таким образом, оценка (13) выполнена с найденной выше константой ρ. □

Сформулируем следующее следствие для метода простой итерации:

$$\frac{x^{n+1} - x^n}{\tau} + Ax^n = f, \quad n \in \mathbb{Z}_+.$$

Следствие 2. *Пусть* A *— самосопряженный положительно определенный оператор, а* γ_1 *и* γ_2 *— его минимальное и максимальное собственные значения:*

$$\gamma_1 = \min_{1 \leqslant k \leqslant m} \lambda_k^A,\ \gamma_2 = \max_{1 \leqslant k \leqslant m} \lambda_k^A.$$

Кроме того, пусть $\tau = \frac{2}{\gamma_1+\gamma_2}$. *Тогда верна оценка*

$$\|x^{n+1} - x\| \leqslant \rho \|x^n - x\|,$$

где $\rho = \frac{1-\xi}{1+\xi},\ \xi = \frac{\gamma_1}{\gamma_2}$.

Доказательство следствия 2 очевидно.

§8 Исследование скорости сходимости ПТИМ

Рассмотрим матричное уравнение вида

$$Ax = f, \tag{1}$$

где $|A| \neq 0,\ A\ (m \times m),\ x = (x_1, x_2, \ldots, x_m)^T,\ f = (f_1, f_2, \ldots, f_m)^T$.
Представим матрицу A в виде

$$A = R_1 + R_2,$$

где R_1 — нижнетреугольная матрица, R_2 — верхнетреугольная матрица:

$$R_1 = \begin{pmatrix} 0.5a_{11} & 0 & \cdots & 0 \\ a_{21} & 0.5a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & 0.5a_{mm} \end{pmatrix}, \quad R_2 = \begin{pmatrix} 0.5a_{11} & a_{12} & \cdots & a_{1m} \\ 0 & 0.5a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0.5a_{mm} \end{pmatrix}.$$

Очевидно, что такое представление существует для произвольной матрицы A.

Запишем каноническую форму попеременно-треугольного итерационного метода (ПТИМ):

$$(E + \omega R_1)(E + \omega R_2)\frac{x^{n+1} - x^n}{\tau} + Ax^n = f, \quad \omega > 0, \quad \tau > 0, \quad n \in \mathbb{Z}_+.$$

Обозначим

$$B = (E + \omega R_1)(E + \omega R_2).$$

Теорема 1 (о сходимости ПТИМ)**.** *Пусть A — самосопряженный положительно определенный оператор и $\omega > \frac{\tau}{4}$. Тогда* ПТИМ *сходится в среднеквадратичной норме при любом начальном приближении x^0.*

Доказательство. Раскроем скобки в выражении для B, учитывая, что $R_1 = R_2^*$:

$$B = (E + \omega R_2^*)(E + \omega R_2) = E + \omega(R_2^* + R_2) + \omega^2 R_2^* R_2 = E + \omega A + \omega^2 R_2^* R_2. \tag{2}$$

Очевидно, что

$$B = (E - \omega R_2^*)(E - \omega R_2) + 2\omega A. \tag{3}$$

Кроме того,

$$((E - \omega R_2^*)(E - \omega R_2)x, x) = ((E - \omega R_2)x, (E - \omega R_2)x) \geqslant 0.$$

Тогда из уравнения (3) следует неравенство

$$B \geqslant 2\omega A. \tag{4}$$

Учитывая условие теоремы ($\omega > \frac{\tau}{4}$), получим, что $B > \frac{\tau}{2}A$ и ПТИМ сходится по теореме Самарского при любом начальном приближении x^0. □

Теорема 2 (о скорости сходимости ПТИМ)**.** *Пусть A — положительно определенный самосопряженный оператор и числа $\delta > 0,\ \Delta > 0$ таковы, что выполняются неравенства*

$$A \geqslant \delta E,\ R_2^* R_2 \leqslant \frac{\Delta}{4}A. \tag{5}$$

Положим

$$\omega = \frac{2}{\sqrt{\delta\Delta}},\ \tau = \frac{2}{\gamma_1 + \gamma_2},\ \gamma_1 = \frac{\sqrt{\delta}}{2}\left(\frac{\sqrt{\delta\Delta}}{\sqrt{\delta} + \sqrt{\Delta}}\right),\ \gamma_2 = \frac{\sqrt{\delta\Delta}}{4}.$$

Тогда ПТИМ *сходится и имеет место оценка*

$$\|x^{n+1} - x\|_B \leqslant \rho \|x^n - x\|_B,$$

где $\rho = \frac{1-\sqrt{\eta}}{1+3\sqrt{\eta}},\ \eta = \frac{\delta}{\Delta}$.

Доказательство. Покажем, что из неравенств (5) следует $\eta \leqslant 1$. Рассмотрим второе неравенство и воспользуемся определением сопряженного оператора:

$$R_2^* R_2 \leqslant \frac{\Delta}{4} A \ \Rightarrow\ (R_2^* R_2 x, x) = (R_2 x, R_2 x) = \|R_2 x\|^2 \leqslant \frac{\Delta}{4}(Ax, x). \tag{6}$$

Рассмотрим первое неравенство:

$$A \geqslant \delta E \ \Rightarrow\ (Ax, x) \geqslant \delta \|x\|^2.$$

Очевидно, что из представления $A = R_1 + R_2 = R_2^* + R_2$ следует равенство

$$(Ax, x) = (R_2^* x, x) + (R_2 x, x) = 2(R_2 x, x).$$

Тогда предположим, что x — ненулевой вектор, и получим

$$\delta \|x\|^2 \leqslant (Ax, x) = \frac{(Ax, x)^2}{(Ax, x)} = \frac{4(R_2 x, x)^2}{(Ax, x)}.$$

Воспользуемся неравенством Коши-Буняковского и неравенством (6):

$$\delta \|x\|^2 \leqslant \frac{4\|R_2 x\|^2 \|x\|^2}{(Ax, x)} \leqslant \frac{4\Delta (Ax, x)\|x\|^2}{4(Ax, x)} = \Delta \|x\|^2.$$

Таким образом, справедливо неравенство $\delta \leqslant \Delta$.

При доказательстве будем опираться на следствие 1 из теоремы об оценке скорости сходимости итерационного метода общего вида. Чтобы воспользоваться следствием 1 из теоремы об оценке скорости сходимости, найдем из условия теоремы γ_1 и γ_2 такие, что

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B. \tag{7}$$

Из неравенства (4) ($B \geqslant 2\omega A$), полученного в ходе доказательства теоремы о сходимости ПТИМ следует оценка $A \leqslant \frac{B}{2\omega}$. Тогда можно положить в неравенстве (7) $\gamma_2 = \frac{1}{2\omega}$.
Оценим выражение (2), воспользовавшись неравенствами (5):

$$B = E + \omega A + \omega^2 R_2^* R_2 \leqslant \frac{1}{\delta} A + \omega A + \frac{\Delta \omega^2}{4} A = \left(\frac{1}{\delta} + \omega + \frac{\Delta \omega^2}{4} \right) A.$$

Тогда положим в неравенстве (7) $\gamma_1 = \left(\frac{1}{\delta} + \omega + \frac{\Delta \omega^2}{4} \right)^{-1}$.

Для нахождения максимально возможной скорости сходимости будем минимизировать функцию $\rho(\omega)$ (как известно, чем меньше ρ, тем быстрее сходится метод):

$$\rho(\omega) = \frac{1 - \xi(\omega)}{1 + \xi(\omega)},\ \xi(\omega) = \frac{\gamma_1(\omega)}{\gamma_2(\omega)},$$

что эквивалентно минимизации функции $f(\omega)$:

$$f(\omega) = \frac{\gamma_2(\omega)}{\gamma_1(\omega)} = \frac{1}{2} \left(1 + \frac{1}{\omega \delta} + \frac{\Delta \omega}{4} \right) \to min.$$

Для нахождения экстремальных точек найдем производную $f(\omega)$ и приравняем ее к нулю:

$$f'(\omega) = \frac{1}{2}\left(\frac{\Delta}{4} - \frac{1}{\omega^2\delta}\right) = 0 \;\Rightarrow\; \omega = \omega_0 = \frac{2}{\sqrt{\delta\Delta}}.$$

Учтем, что $\omega > 0$, и проверим, что точка ω_0 доставляет минимум функции $f(\omega)$, найдя знак второй производной функции в этой точке:

$$f''(\omega) = \frac{1}{\delta\omega^3} > 0.$$

Подставим ω_0 в выражения для γ_1, γ_2, ρ:

$$\gamma_1 = \frac{1}{\frac{1}{\delta} + \frac{2}{\sqrt{\delta\Delta}} + \frac{\Delta}{4}\frac{4}{\delta\Delta}} = \frac{1}{\frac{2}{\delta} + \frac{2}{\sqrt{\delta\Delta}}} = \frac{\delta\sqrt{\Delta}}{2\sqrt{\Delta} + 2\sqrt{\delta}} = \frac{\sqrt{\delta}}{2}\left(\frac{\sqrt{\delta\Delta}}{\sqrt{\Delta} + \sqrt{\delta}}\right)$$

$$\gamma_2 = \frac{1}{2\omega_0} = \frac{\sqrt{\delta\Delta}}{4}$$

$$\xi(\omega) = \frac{\gamma_1(\omega)}{\gamma_2(\omega)} = \frac{4}{\sqrt{\delta\Delta}}\frac{\sqrt{\delta}}{2}\left(\frac{\sqrt{\delta\Delta}}{\sqrt{\Delta} + \sqrt{\delta}}\right) = \frac{2\sqrt{\delta}}{\sqrt{\Delta} + \sqrt{\delta}}$$

$$\left.\begin{aligned} 1 - \xi = 1 - \frac{2\sqrt{\delta}}{\sqrt{\Delta} + \sqrt{\delta}} = \frac{\sqrt{\Delta} - \sqrt{\delta}}{\sqrt{\Delta} + \sqrt{\delta}} \\ 1 + \xi = 1 + \frac{2\sqrt{\delta}}{\sqrt{\Delta} + \sqrt{\delta}} = \frac{\sqrt{\Delta} + 3\sqrt{\delta}}{\sqrt{\Delta} + \sqrt{\delta}} \end{aligned}\right\} \Rightarrow \rho = \frac{1 - \xi}{1 + \xi} = \frac{\sqrt{\Delta} - \sqrt{\delta}}{\sqrt{\Delta} + 3\sqrt{\delta}} = \frac{1 - \sqrt{\eta}}{1 + 3\sqrt{\eta}},\ \eta = \frac{\delta}{\Delta}\ (\Delta \neq 0).$$

Исходя из полученных соотношений и следствия 1, получаем оценку

$$\|x^{n+1} - x\|_B \leqslant \rho\|x^n - x\|_B.$$

Таким образом, теорема доказана. □

Покажем, что ПТИМ сходится на порядок быстрее метода простой итерации (МПИ), метода Зейделя (МЗ) и метода Якоби (МЯ).

Число итераций, необходимое для достижения заданной точности $\varepsilon > 0$ равно

$$n_0(\varepsilon) = \left[\frac{\ln\frac{1}{\varepsilon}}{\ln\frac{1}{\rho}}\right],$$

где $[x]$ означает целую часть числа x, а $\ln\frac{1}{\rho}$ — скорость сходимости итерационного метода. В практических задачах η часто является величиной порядка $\mathrm{O}\big(m^{-2}\big)$.

Оценим скорость сходимости ПТИМ:

$$\frac{1}{\rho} = \frac{1 + 3\sqrt{\eta}}{1 - \sqrt{\eta}} = \frac{(1 + 3\sqrt{\eta})(1 + \sqrt{\eta})}{1 - \eta} \approx 1 + 4\sqrt{\eta},$$

$$\ln\frac{1}{\rho} \approx \ln(1 + 4\sqrt{\eta}) = \mathrm{O}\big(m^{-1}\big),\ n_0(\varepsilon) = \mathrm{O}\big(m\big).$$

Оценим скорость сходимости МПИ:

$$\rho = \frac{1 - \xi}{1 + \xi} = \frac{1 - \eta}{1 + \eta},\ \frac{1}{\rho} = \frac{1 + \eta}{1 - \eta} = \frac{(1 + \eta)^2}{1 - \eta^2} \approx 1 + 2\eta,$$

$$\ln\frac{1}{\rho} \approx \ln(1 + 2\eta) = \mathrm{O}\big(m^{-2}\big),\ n_0(\varepsilon) = \mathrm{O}\big(m^2\big).$$

Таким образом, МПИ сходится на порядок медленнее, чем ПТИМ. МЯ и МЗ имеют тот же порядок сходимости, что и МПИ.

§9 Методы решения задач на собственные значения

Рассмотрим задачу поиска собственных значений, которая состоит в нахождении чисел λ и векторов x, удовлетворяющих уравнению

$$Ax = \lambda x,\ x \neq \theta,$$

где A — вещественная матрица порядка $(m \times m)$. λ называется *собственным значением* матрицы A, а x — соответствующим ему *собственным вектором*. У любой вещественной матрицы порядка $(m \times m)$ существует ровно m собственных значений, вообще говоря, комплексных.

Собственный вектор определяется с точностью до константы $C \neq 0$. В вычислительных методах собственные векторы обычно нормируют с условием $\|x_k\| = 1$, чтобы избежать быстрого накопления ошибок округления.

Задача поиска собственных значений эквивалентна задаче нахождения корней характеристического многочлена матрицы A:

$$|A - \lambda E| = a_m\lambda^m + a_{m-1}\lambda^{m-1} + \ldots + a_1\lambda + a_0 = 0,$$

где $a_i \in \mathbb{R}$, $i = \overline{0, m}$, $a_m \neq 0$. Это уравнение имеет общее решение в радикалах только при $m \leqslant 4$, в реальных же задачах m может быть порядка 10^5 или 10^6 и выше. Таким образом, при больших m задачу поиска собственных значений можно решить только численными методами.

Собственные значения необходимы для оценки скорости сходимости итерационных методов решения систем линейных уравнений. При этом обычно достаточно найти минимальное и максимальное по модулю собственные значения. Таким образом, различают два вида проблем, связанных с поиском собственных значений матрицы:

1. Частичная проблема собственных значений, которая заключается в нахождении некоторых собственных значений.

2. Полная проблема собственных значений, которая заключается в нахождении всего спектра матрицы.

Очевидно, что частичная проблема является более простой, чем полная проблема.

Степенной метод

Рассмотрим частичную проблему собственных значений. Будем искать собственный вектор по формуле

$$x^{n+1} = Ax^n, \quad n \in \mathbb{Z}_+,\ x^0 \text{ задано.} \tag{1}$$

Пусть $\{\lambda_k\}_{k=1}^m$ — собственные значения матрицы A, среди которых могут быть повторяющиеся. Упорядочим их по неубыванию модулей:

$$|\lambda_1| \leqslant |\lambda_2| \leqslant \ldots \leqslant |\lambda_m|.$$

Будем доказывать сходимость степенного метода при выполнении трех условий:

A) В пространстве $\mathbb{R}^m$ существует базис $\{e_k\}$ из собственных векторов матрицы A.

B) $\left|\frac{\lambda_{m-1}}{\lambda_m}\right| < 1$.

C) $x^0 = c_1e_1 + c_2e_2 + \ldots + c_me_m$, где $c_m \neq 0$.

Утверждение. *Пусть вещественная матрица A $(m\times m)$ такова, что выполнены условия A) – C). Тогда степенной метод для матрицы A сходится по направлению к собственному вектору, отвечающему максимальному по модулю собственному значению:*

$$x^n \underset{n\to\infty}{\longrightarrow} e_m.$$

Кроме того, для последовательности $\left\{\lambda_m^{(n)}\right\}$, заданной одной из формул

$$\lambda_m^{(n)} = \frac{x_i^{n+1}}{x_i^n},$$

$$\lambda_m^{(n)} = \frac{(Ax^n, x^n)}{(x^n, x^n)}$$

справедлива следующая оценка сходимости к λ_m:

$$\lambda_m^{(n)} - \lambda_m = \mathrm{O}\left(\left(\frac{\lambda_{m-1}}{\lambda_m}\right)^n\right).$$

Доказательство. Покажем, что при выполнении условий A) – C) степенной метод сходится к собственному вектору матрицы A, отвечающему максимальному по модулю собственному значению.

Из рекуррентной формулы (1) получим:

$$x^n = A^n x^0, \quad n \in \mathbb{N}.$$

Воспользуемся условиями A), C) и разложим n-ую итерацию по базису из собственных векторов $\{e_k\}$ матрицы A:

$$x^n = A^n x^0 = \sum_{k=1}^{m} c_k A^n e_k = \sum_{k=1}^{m} c_k \lambda_k^n e_k = c_m\lambda_m^n e_m + c_{m-1}\lambda_{m-1}^n e_{m-1} + \ldots + c_1\lambda_1^n e_1.$$

В силу условия C) $c_m \neq 0$. Кроме того, поскольку у матрицы A существует хотя бы одно ненулевое собственное значение, то максимальное по модулю из них гарантированно не равно нулю: $\lambda_m \neq 0$. Поделив равенство на $c_m\lambda_m^n$, получим:

$$\frac{x^n}{c_m\lambda_m^n} = e_m + \frac{c_{m-1}}{c_m}\left(\frac{\lambda_{m-1}}{\lambda_m}\right)^n e_{m-1} + \ldots + \frac{c_1}{c_m}\left(\frac{\lambda_1}{\lambda_m}\right)^n e_1.$$

Перейдя к пределу при $n \to \infty$ и учитывая условие B), получим, что x^n сходится по направлению к e_m:

$$\lim_{n\to\infty} x^n = e_m.$$

Рассмотрим два способа вычисления максимального по модулю собственного значения матрицы A. Первый способ состоит в вычислении отношения i-ых координат $(n+1)$-ой и n-ой итераций.

$$x_i^n = c_1\lambda_1^n e_1^{(i)} + \ldots + c_m\lambda_m^n e_m^{(i)}, \quad i = \overline{1,m},$$
$$x_i^{n+1} = c_1\lambda_1^{n+1} e_1^{(i)} + \ldots + c_m\lambda_m^{n+1} e_m^{(i)}, \quad i = \overline{1,m}.$$

Здесь $e_j^{(i)}$ — i-ая координата вектора e_j, $j = \overline{1,m}$.

$$\lambda_m^{(n)} = \frac{x_i^{n+1}}{x_i^n}, \tag{2}$$

$$\lambda_m^{(n)} = \frac{c_m\lambda_m^{n+1}e_m^{(i)} + c_{m-1}\lambda_{m-1}^{n+1}e_{m-1}^{(i)} + \ldots + c_1\lambda_1^{n+1}e_1^{(i)}}{c_m\lambda_m^{n}e_m^{(i)} + c_{m-1}\lambda_{m-1}^{n}e_{m-1}^{(i)} + \ldots + c_1\lambda_1^{n}e_1^{(i)}} =$$

$$= \frac{c_m\lambda_m^{n+1}e_m^{(i)}\left(1 + \frac{c_{m-1}}{c_m}\left(\frac{\lambda_{m-1}}{\lambda_m}\right)^{n+1}\frac{e_{m-1}^{(i)}}{e_m^{(i)}} + \ldots + \frac{c_1}{c_m}\left(\frac{\lambda_1}{\lambda_m}\right)^{n+1}\frac{e_1^{(i)}}{e_m^{(i)}}\right)}{c_m\lambda_m^{n}e_m^{(i)}\left(1 + \frac{c_{m-1}}{c_m}\left(\frac{\lambda_{m-1}}{\lambda_m}\right)^{n}\frac{e_{m-1}^{(i)}}{e_m^{(i)}} + \ldots + \frac{c_1}{c_m}\left(\frac{\lambda_1}{\lambda_m}\right)^{n}\frac{e_1^{(i)}}{e_m^{(i)}}\right)} = \lambda_m + \mathrm{O}\left(\left(\frac{\lambda_{m-1}}{\lambda_m}\right)^n\right).$$

Заметим, что начальное приближение x^0 — ненулевой вектор, и в силу этого вектор $x^n = A^n x^0$ имеет хотя бы одну ненулевую координату. Поэтому возможно деление на i-ую координату вектора x^n, где i — некоторое целое число от 1 до m.

Второй способ состоит в вычислении выражения

$$\lambda_m^{(n)} = \frac{(Ax^n, x^n)}{(x^n, x^n)} = \frac{(x^{n+1}, x^n)}{(x^n, x^n)}. \tag{3}$$

Пусть A — самосопряженная матрица. Тогда в пространстве $\mathbb{R}^{m\times m}$ существует ортонормированный базис $\{e_k\}$ из собственных векторов матрицы A:

$$(e_i, e_j) = \delta_{ij} = \begin{cases} 1 & \text{при } i = j, \\ 0 & \text{при } i \neq j, \end{cases} \quad i, j = \overline{1, m}.$$

Тогда выражение (3) можно преобразовать следующим образом:

$$\lambda_m^{(n)} = \frac{c_m^2\lambda_m^{2n+1} + c_{m-1}^2\lambda_{m-1}^{2n+1} + \ldots + c_1^2\lambda_1^{2n+1}}{c_m^2\lambda_m^{2n} + c_{m-1}^2\lambda_{m-1}^{2n} + \ldots + c_1^2\lambda_1^{2n}} =$$

$$= \frac{c_m^2\lambda_m^{2n+1}\left(1 + \left(\frac{c_{m-1}}{c_m}\right)^2\left(\frac{\lambda_{m-1}}{\lambda_m}\right)^{2n+1} + \ldots + \left(\frac{c_1}{c_m}\right)^2\left(\frac{\lambda_1}{\lambda_m}\right)^{2n+1}\right)}{c_m^2\lambda_m^{2n}\left(1 + \left(\frac{c_{m-1}}{c_m}\right)^2\left(\frac{\lambda_{m-1}}{\lambda_m}\right)^{2n} + \ldots + \left(\frac{c_1}{c_m}\right)^2\left(\frac{\lambda_1}{\lambda_m}\right)^{2n}\right)} = \lambda_m + \mathrm{O}\left(\left(\frac{\lambda_{m-1}}{\lambda_m}\right)^{2n}\right).$$

Заметим, что показатель степени равен $2n$, в отличие от заявленного в условии утверждения показателя, равного n. Таким образом, если матрица A — самосопряженная, то оценку сходимости из условия утверждения можно улучшить.

Рассмотрим теперь выражение (3) для произвольной матрицы A и воспользуемся условием A) сходимости степенного метода:

$$\lambda_m^{(n)} = \frac{\sum\limits_{i,j=1}^{m} c_i c_j \lambda_i^{n+1}\lambda_j^n (e_i, e_j)}{\sum\limits_{i,j=1}^{m} c_i c_j \lambda_i^{n}\lambda_j^n (e_i, e_j)} =$$

$$= \frac{\lambda_m^{2n+1}c_m^2(e_m, e_m) + \lambda_m^{n+1}\lambda_{m-1}^n c_{m-1}c_m(e_{m-1}, e_m) + \ldots + c_1^2\lambda_1^{2n+1}(e_1, e_1)}{\lambda_m^{2n}c_m^2(e_m, e_m) + \lambda_m^{n}\lambda_{m-1}^n c_{m-1}c_m(e_{m-1}, e_m) + \ldots + c_1^2\lambda_1^{2n}(e_1, e_1)} =$$

$$= \frac{\lambda_m^{2n+1}c_m^2(e_m, e_m)\left(1 + \frac{c_{m-1}}{c_m}\left(\frac{\lambda_{m-1}}{\lambda_m}\right)^{n+1}\frac{(e_{m-1}, e_{m-1})}{(e_m, e_m)} + \ldots + \left(\frac{c_1}{c_m}\right)^2\left(\frac{\lambda_1}{\lambda_m}\right)^{2n+1}\frac{(e_1, e_1)}{(e_m, e_m)}\right)}{\lambda_m^{2n}c_m^2(e_m, e_m)\left(1 + \frac{c_{m-1}}{c_m}\left(\frac{\lambda_{m-1}}{\lambda_m}\right)^{n}\frac{(e_{m-1}, e_{m-1})}{(e_m, e_m)} + \ldots + \left(\frac{c_1}{c_m}\right)^2\left(\frac{\lambda_1}{\lambda_m}\right)^{2n}\frac{(e_1, e_1)}{(e_m, e_m)}\right)},$$

$$\lambda_m^{(n)} = \lambda_m + \mathrm{O}\left(\left(\frac{\lambda_{m-1}}{\lambda_m}\right)^n\right).$$

Утверждение доказано. □

Замечание. *Пусть у вещественной матрицы A $(m \times m)$ существует комплексное собственное значение с ненулевой мнимой частью: $\lambda = \lambda_0 + i\lambda_1,\ \lambda_1 \neq 0$. Тогда соответствующий собственный вектор — комплексный и имеет ненулевую мнимую часть: $x = x_0 + ix_1,\ x_1 \neq \theta$, и начальное приближение x^0 вектора x в итерационном методе также должно быть комплексным с ненулевой мнимой частью.*

Доказательство. Подействуем на x оператором A:

$$A(x_0 + ix_1) = (\lambda_0 + i\lambda_1)(x_0 + ix_1).$$

Разделим вещественную и мнимую части уравнения:

$$\begin{cases} Ax_0 = \lambda_0 x_0 - \lambda_1 x_1 \\ Ax_1 = \lambda_0 x_1 + \lambda_1 x_0. \end{cases}$$

Предположим, что $x_1 = \theta$. Тогда из второго уравнения следует, что $x_0 = \theta$ и $x = \theta$. Однако x — собственный вектор и поэтому не может быть нулевым. Полученное противоречие завершает доказательство. □

Метод обратных итераций

Пусть матрица A — невырожденная. Рассмотрим следующую форму записи неявного итерационного метода:

$$Ax^{n+1} = x^n, \quad n \in \mathbb{Z}_+,\ x^0 \text{ задано.}$$

Умножим обе части равенства слева на A^{-1} и получим формулу степенного метода для матрицы A^{-1}:

$$x^{n+1} = A^{-1}x^n, \quad n \in \mathbb{Z}_+,\ x^0 \text{ задано.} \tag{4}$$

Из свойств обратной матрицы следует, что собственные значения невырожденной матрицы A и обратной к ней матрицы A^{-1} связаны соотношением

$$\lambda_k^{A^{-1}} = \frac{1}{\lambda_k^A},\ k = \overline{1, m}.$$

Заметим, что если собственные значения λ_k^A упорядочены по возрастанию модулей, то соответствующие им собственные значения $\lambda_k^{A^{-1}}$ будут упорядочены по убыванию модулей. В данном методе обозначим $\lambda_k = \lambda_k^A$, и пусть $\{\lambda_k\}$ упорядочены по возрастанию модулей.

Сформулируем три условия:

A) В пространстве $\mathbb{R}^m$ существует базис $\{e_k\}$ из собственных векторов матрицы A.

B) $\left|\frac{\lambda_1}{\lambda_2}\right| < 1$.

C) $x^0 = c_1e_1 + c_2e_2 + \ldots + c_me_m,\ c_1 \neq 0$.

Утверждение. *Пусть невырожденная вещественная матрица A $(m \times m)$ такова, что выполнены условия A) – C). Тогда метод обратных итераций для матрицы A^{-1} сходится по направлению к собственному вектору, отвечающему минимальному по модулю собственному значению:*

$$x^n \underset{n\to\infty}{\longrightarrow} e_1.$$

Доказательство. Разложим n-ую итерацию по базису $\{e_k\}$ из собственных векторов матрицы A:

$$x^n = A^{-n}x^0 = \sum_{k=1}^{m} c_k A^{-n} e_k = \sum_{k=1}^{m} c_k \lambda_k^{-n} e_k = c_1\lambda_1^{-n}e_1 + c_2\lambda_2^{-n}e_2 + \ldots + c_m\lambda_m^{-n}e_m.$$

В силу условия C) $c_1 \neq 0$. Кроме того, поскольку матрица A невырождена, $\lambda_1^n \neq 0$. Поделив равенство на $c_1\lambda_1^{-n}$, получим

$$\frac{x^n}{c_1\lambda_1^{-n}} = e_1 + \frac{c_2}{c_1}\left(\frac{\lambda_1}{\lambda_2}\right)^n e_2 + \ldots + \frac{c_m}{c_1}\left(\frac{\lambda_1}{\lambda_m}\right)^n e_m.$$

Перейдя к пределу при $n \to \infty$ и учитывая условие B), получим, что x^n сходится по направлению к e_1:

$$\lim_{n\to\infty} x^n = e_1.$$

□

Сформулируем утверждения о вычислении минимального собственного значения в виде задачи.

Задача. *Пусть выполнены условия A) – C) сходимости метода обратных итераций. Показать, что в случае произвольной матрицы A справедливы следующие оценки:*

$$\lambda_1 - \frac{x_i^n}{x_i^{n+1}} = \mathrm{O}\left(\left(\frac{\lambda_1}{\lambda_2}\right)^n\right),$$

$$\lambda_1 - \frac{(x^n, x^n)}{(x^{n+1}, x^n)} = \mathrm{O}\left(\left(\frac{\lambda_1}{\lambda_2}\right)^n\right).$$

Показать, что если матрица A — самосопряженная, то последнюю оценку можно улучшить:

$$\lambda_1 - \frac{(x^n, x^n)}{(x^{n+1}, x^n)} = \mathrm{O}\left(\left(\frac{\lambda_1}{\lambda_2}\right)^{2n}\right).$$

Метод обратных итераций со сдвигом

Рассмотрим итерационный метод, задаваемый формулой

$$(A - \alpha E)x^{n+1} = x^n, \quad n \in \mathbb{Z}_+,\ x^0 \text{ задано},$$

где α — такое вещественное число, что матрица $(A - \alpha E)$ невырождена. Домножим обе части равенства слева на $(A - \alpha E)^{-1}$ и получим формулу степенного метода с матрицей $(A - \alpha E)^{-1}$:

$$x^{n+1} = (A - \alpha E)^{-1}x^n. \tag{5}$$

Таким образом, метод обратных итераций эквивалентен степенному методу, записанному для матрицы $B = (A - \alpha E)^{-1}$. Следовательно, векторы x_n будут сходиться при $n \to \infty$ по направлению к такому собственному вектору e_r матрицы A, для которого величина

$$|\lambda_r - \alpha|^{-1} = \max_{1\leqslant k\leqslant m} |\lambda_k - \alpha|^{-1}.$$

Это означает, что если требуется найти собственный вектор e_r, отвечающий данному собственному значению λ_r, то надо задать число α, близкое к λ_r, и вычислить векторы x_n,

исходя из формулы (5).
Само собственное значение λ_r находится из выражения:

$$\lambda_r = \lim_{n\to\infty}\left(\alpha + \frac{x_n^{(i)}}{x_{n+1}^{(i)}}\right), \quad i = \overline{1,m}.$$

Следовательно, метод обратных итераций со сдвигом позволяет в принципе отыскать любое собственное значение матрицы A. Этот метод очень часто используют для нахождения и уточнения собственных векторов, если собственные значения уже известны.

§10 Приведение матрицы к верхней почти треугольной форме

Рассмотрим полную проблему собственных значений матрицы A $(m\times m)$. Идея QR-алгоритма, позволяющего решить эту проблему, состоит в использовании сохраняющих спектр преобразований для приведения матрицы A к более простому виду: верхней почти треугольной форме, и построении итерационного процесса, приводящего преобразованную матрицу к виду, в котором найти спектр матрицы достаточно легко — верхнетреугольной или диагональной форме.

Определение. *Матрица A имеет верхнюю почти треугольную форму* (ВПТФ)*, если ее можно записать в виде*

$$A = \begin{pmatrix} \times & \times & \times & \dots & \times & \times \\ \times & \times & \times & \dots & \times & \times \\ 0 & \times & \times & \dots & \times & \times \\ 0 & 0 & \times & \dots & \times & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \times & \times \end{pmatrix},$$

где символами $\times$ обозначены, вообще говоря, ненулевые элементы матрицы.

Определение. *Элементарным отражением, соответствующим вещественному вектор-столбцу $v = (v_1, v_2, \dots, v_m)^T$, называется преобразование, задаваемое матрицей*

$$H = E - 2\frac{vv^T}{\|v\|^2}. \tag{1}$$

Убедимся, что формула (1) задает матрицу порядка $(m \times m)$:

$$v^T v = v_1^2 + v_2^2 + .. + v_m^2 = \|v\|^2 \text{ — число},$$

$$vv^T = \begin{pmatrix} v_1^2 & v_1v_2 & \cdots & v_1v_m \\ v_2v_1 & v_2^2 & \cdots & v_2v_m \\ \vdots & \vdots & \ddots & \vdots \\ v_mv_1 & v_mv_2 & \cdots & v_m^2 \end{pmatrix} \text{ — симметрическая (эрмитова) матрица.}$$

Сформулируем свойства матрицы элементарного отражения:

1. H — симметричная матрица, $H = H^T$.
2. H — ортогональная матрица, $H^{-1} = H^T$.

   Для доказательства этого свойства рассмотрим произведение $H^T H$:

$$H^T H = H^2 = \left(E - 2\frac{vv^T}{\|v\|^2}\right)\left(E - 2\frac{vv^T}{\|v\|^2}\right) = E^2 - 4\frac{vv^T}{\|v\|^2} + 4\frac{v(v^Tv)v^T}{\|v\|^4} = E.$$

   Домножив полученное равенство на H^{-1} справа, получим требуемое утверждение.

Утверждение. *Пусть задан вещественный вектор-столбец* $x = (x_1, x_2, .., x_m)^T$. *Тогда можно выбрать вектор* v *так, чтобы было выполнено равенство*

$$Hx = (-\|x\|, 0, 0, .., 0)^T, \quad \|x\| = \sqrt{(x, x)},$$

где H *— элементарное отражение, соответствующее вектор-столбцу* v.

Доказательство. Будем искать вектор v в виде

$$v = x + \sigma z, \quad \sigma \in \mathbb{R}_+, \ z = (1, 0, .., 0)^T.$$

Подставим выражение для v в формулу (1):

$$Hx = x - 2x\frac{(x+\sigma z)(x+\sigma z)^T}{(x+\sigma z)^T(x+\sigma z)} = x - (x+\sigma z)\frac{2(x+\sigma z)^T x}{(x+\sigma z)^T(x+\sigma z)}. \tag{2}$$

Рассмотрим отдельно числитель и знаменатель дроби:

$$2(x+\sigma z)^T x = 2(\|x\|^2 + \sigma x_1),$$

$$(x+\sigma z)^T(x+\sigma z) = \|x\|^2 + \sigma x_1 + \sigma x_1 + \sigma^2.$$

Пусть $\sigma = \|x\|$. Тогда

$$\frac{2(x+\sigma z)^T x}{(x+\sigma z)^T(x+\sigma z)} = 1.$$

Подставив последнее выражение в равенство (2), получим искомое равенство:

$$Hx = x - x - \sigma z = (-\|x\|, 0, 0, \ldots, 0)^T.$$

□

Утверждение. *Любую вещественную матрицу* A $(m \times m)$ *можно привести к верхней почти треугольной форме с помощью преобразования подобия с ортогональной матрицей* Q*:*

$$C = Q^{-1}AQ = \begin{pmatrix} \times & \times & \times & \ldots & \times & \times \\ \times & \times & \times & \ldots & \times & \times \\ 0 & \times & \times & \ldots & \times & \times \\ 0 & 0 & \times & \ldots & \times & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \times & \times \end{pmatrix},$$

где $Q^T = Q^{-1}$.

Доказательство. Представим матрицу A в виде

$$A = \begin{pmatrix} a_{11} & y_{m-1} \\ x_{m-1} & A_{m-1} \end{pmatrix},$$

где $x_{m-1} = (a_{21}, a_{31}, .., a_{m1})^T$, $y_{m-1} = (a_{12}, a_{13}, .., a_{1m})$.

Согласно предыдущему утверждению, можно задать такое элементарное отражение с матрицей H_{m-1} порядка $((m-1) \times (m-1))$, что будет справедливо равенство

$$H_{m-1}x_{m-1} = -\sigma_1 z_{m-1} = (-\|x_{m-1}\|, 0, 0, .., 0)^T, \ z_{m-1} = \underbrace{(1, 0, \ldots, 0)}_{m-1}^T, \ \sigma_1 = \|x_{m-1}\|. \tag{3}$$

Соответствующий матрице H вещественный вектор v можно представить в виде

$$v = x_{m-1} + \sigma_1 z_{m-1}.$$

Из-за несовпадения размерностей мы не можем напрямую применить преобразование H_{m-1} к матрице A. Поэтому рассмотрим матрицу U_1 $(m \times m)$:

$$U_1 = \begin{pmatrix} 1 & \theta^T \\ \theta & H_{m-1} \end{pmatrix},\ \theta = (\underbrace{0, 0, \ldots, 0}_{m-1}).$$

В силу того, что матрица H_{m-1} симметричная и ортогональная, матрица U_1 также является симметричной и ортогональной. Вычислим матрицу $C_1 = U_1^{-1} A U_1$, полученную действием преобразования подобия U_1 на матрицу A:

$$U_1^{-1} A = \begin{pmatrix} 1 & \theta^T \\ \theta & H_{m-1} \end{pmatrix} \begin{pmatrix} a_{11} & y_{m-1} \\ x_{m-1} & A_{m-1} \end{pmatrix} = \begin{pmatrix} a_{11} & y_{m-1} \\ H_{m-1}x_{m-1} & H_{m-1}A_{m-1} \end{pmatrix},$$

$$U_1^{-1} A U_1 = \begin{pmatrix} a_{11} & y_{m-1} \\ H_{m-1}x_{m-1} & H_{m-1}A_{m-1} \end{pmatrix} \begin{pmatrix} 1 & \theta^T \\ \theta & H_{m-1} \end{pmatrix} = \begin{pmatrix} a_{11} & y_{m-1}H_{m-1} \\ H_{m-1}x_{m-1} & H_{m-1}A_{m-1}H_{m-1} \end{pmatrix}.$$

В силу равенства (3) матрица C_1 имеет следующий вид:

$$C_1 = U_1^{-1} A U_1 = \begin{pmatrix} \times & \times & \times & \ldots & \times & \times \\ \times & \times & \times & \ldots & \times & \times \\ 0 & \times & \times & \ldots & \times & \times \\ 0 & \times & \times & \ldots & \times & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \times & \times & \ldots & \times & \times \end{pmatrix}.$$

Введем вектор $x_{m-2} = (c_{32}^{(1)}, c_{42}^{(1)}, \ldots, c_{m2}^{(1)})^T$, где $c_{i2}^{(1)}$ — элемент матрицы C_1, стоящий в позиции (i, 2), $i = \overline{3, m}$. Воспользуемся предыдущим утверждением и построим матрицу H_{m-2}, удовлетворяющую равенству

$$H_{m-2} x_{m-2} = -\sigma_2 z_{m-2} = (-\|x_{m-2}\|, 0, \ldots, 0)^T,\ z_{m-2} = (\underbrace{1, 0, \ldots, 0}_{m-2})^T,\ \sigma_2 = \|x_{m-2}\|.$$

По аналогичным соображениям рассмотрим матрицу U_2 $(m \times m)$:

$$U_2 = \left(\begin{array}{cc|c} 1 & 0 & \\ 0 & 1 & \mathbf{0} \\ \hline & \mathbf{0} & H_{m-2} \end{array}\right).$$

Матрица U_2 ортогональна и симметрична. Матрица $C_2 = U_2^{-1} C_1 U_2$ имеет следующий вид:

$$C_2 = U_2^{-1} C_1 U_2 = \begin{pmatrix} \times & \times & \times & \ldots & \times & \times \\ \times & \times & \times & \ldots & \times & \times \\ 0 & \times & \times & \ldots & \times & \times \\ 0 & 0 & \times & \ldots & \times & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \times & \ldots & \times & \times \end{pmatrix} = U_2^{-1} U_1^{-1} A U_1 U_2.$$

Через $(m-2)$ шага получим матрицу C, имеющую ВПТФ:

$$C = U_{m-2}^{-1}U_{m-3}^{-1}\dots U_2^{-1}U_1^{-1}AU_1U_2\dots U_{m-3}U_{m-2} = \begin{pmatrix} \times & \times & \times & \dots & \times & \times \\ \times & \times & \times & \dots & \times & \times \\ 0 & \times & \times & \dots & \times & \times \\ 0 & 0 & \times & \dots & \times & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \times & \times \end{pmatrix}.$$

Определим матрицу $Q = U_1U_2\dots U_{m-2}$. Покажем, что Q — ортогональная матрица:

$$Q^T = (U_1U_2\dots U_{m-2})^T = U_{m-2}^TU_{m-3}^T\dots U_1^T = U_{m-2}^{-1}\dots U_1^{-1} = (U_1U_2\dots U_{m-2})^{-1} = Q^{-1}.$$

Таким образом, произвольную матрицу A можно привести к матрице C с ВПТФ с помощью преобразования подобия, задаваемого ортогональной матрицей Q:

$$C = Q^{-1}AQ,\ c_{ij} = 0 \text{ при } i \geqslant j+2.$$

□

Замечание 1. *Преобразование подобия сохраняет спектр матрицы:* $\lambda_k^C = \lambda_k^A,\ k = \overline{1,m}$.

Доказательство. Рассмотрим ненулевой собственный вектор x_k матрицы A, отвечающий собственному значению λ_k^A:

$$Ax_k = \lambda_k^Ax_k,\ x_k \neq \theta.$$

Домножим обе части равенства на матрицу Q^{-1} слева:

$$Q^{-1}Ax_k = \lambda_k^AQ^{-1}x_k.$$

Обозначим $y_k = Q^{-1}x_k$. Отсюда $x_k = Qy_k$. Тогда справедливо равенство

$$\underbrace{Q^{-1}AQ}_{C}y_k = \lambda_k^Ay_k.$$

Таким образом, y_k является собственным вектором матрицы C, и выполнено требуемое равенство $\lambda_k^C = \lambda_k^A$. Доказательство в обратную сторону очевидно. □

Замечание 2. *Если A — симметричная матрица, то C также является симметричной матрицей:*

$$A = A^T \ \Rightarrow\ C = C^T.$$

Доказательство. $C = Q^{-1}AQ$. Запишем и преобразуем выражение для C^T:

$$C^T = (Q^{-1}AQ)^T = Q^TA^T\left(Q^{-1}\right)^T = Q^TA^TQ = Q^{-1}AQ = C.$$

□

§11 Понятие о QR-алгоритме решения полной проблемы собственных значений

Утверждение. *Произвольная матрица A $(m \times m)$ может быть представлена в виде:*

$$A = QR, \tag{1}$$

где Q — ортогональная матрица, а R — матрица, имеющая верхнюю треугольную форму (ВТФ).

Доказательство. Возьмем вектор $x = (a_{11}, a_{21}, \ldots, a_{m1})^T$ — первый столбец матрицы A. Рассмотрим вектор

$$v = x + \|x\|z, \quad z = (\underbrace{1, 0, \ldots, 0}_{m})^T.$$

и построим матрицу

$$H_1 = E - 2vv^T\|v\|^2.$$

По доказанному выше

$$H_1 x = (-\|x\|, 0, 0, \ldots, 0)^T.$$

Тогда матрица $A_1 = H_1 A$ будет иметь следующий вид:

$$A_1 = H_1 A = \begin{pmatrix} \times & \times & \times & \ldots & \times \\ 0 & \times & \times & \ldots & \times \\ 0 & \times & \times & \ldots & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \times & \times & \ldots & \times \end{pmatrix}.$$

Пусть теперь $x = \left(a_{22}^{(1)}, a_{32}^{(1)}, \ldots, a_{m2}^{(1)}\right)$. По вектору x однозначно определяется элементарное отражение с матрицей H $((m-1) \times (m-1))$, удовлетворяющей равенству

$$Hx = (-\|x\|, 0, \ldots, 0)^T.$$

Пусть $H_2 = \begin{pmatrix} 1 & \theta^T \\ \theta & H \end{pmatrix}$. Тогда матрица $A_2 = H_2 A_1$ имеет следующий вид:

$$A_2 = H_2 H_1 A = \begin{pmatrix} \times & \times & \times & \ldots & \times \\ 0 & \times & \times & \ldots & \times \\ 0 & 0 & \times & \ldots & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \times & \ldots & \times \end{pmatrix}.$$

После $(m-1)$ шага получим матрицу $R = H_{m-1}H_{m-2}\ldots H_2 H_1 A$, имеющую в ВТФ:

$$R = H_{m-1}H_{m-2}\ldots H_2 H_1 A = \begin{pmatrix} \times & \times & \times & \ldots & \times \\ 0 & \times & \times & \ldots & \times \\ 0 & 0 & \times & \ldots & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \times \end{pmatrix}.$$

Введем матрицу $Q = H_1 H_2 \ldots H_{m-1}$. Покажем, что матрица Q ортогональная, воспользовавшись свойством ортогональности элементарного отражения:

$$Q^{-1} = H_{m-1}^{-1}\ldots H_2^{-1}H_1^{-1} = H_{m-1}^T \ldots H_2^T H_1^T = (H_1 H_2 \ldots H_{m-1})^T = Q^T.$$

Таким образом, справедливо разложение (1) матрицы A. В силу того, что в ходе преобразований на матрицу A не накладывались ограничения, разложение справедливо для произвольной матрицы. □

Замечание. *Количество операций, необходимых для вычисления QR-разложения матрицы A, зависит от вида матрицы A. Для произвольной матрицы количество операций можно оценить величиной порядка m^3, для матрицы в* ВПТФ *— порядка m^2, для трехдиагональной матрицы — порядка m.*

Рассмотрим оптимальную версию алгоритма. Приведем матрицу A к матрице A_0, имеющую ВПТФ, и вычислим QR-разложение матрицы A_0:

$$A_0 = Q_0 R_0,$$

где Q_0 — ортогональная, а R_0 — верхнетреугольная матрица. Обозначим матрицу

$$A_1 = R_0 Q_0.$$

Покажем, что спектры матриц A_0 и A_1 совпадают. Из определения матриц A_0 и A_1 получим

$$R_0 = Q_0^{-1} A_0,$$
$$A_1 = Q_0^{-1} A_0 Q_0.$$

Матрица A_1 подобна матрице A_0, и из этого следует, что спектры матриц равны.

На следующем шаге вычислим QR-разложение матрицы $A_1 = Q_1 R_1$ и обозначим матрицу $A_2 = R_1 Q_1$. Аналогичным образом продолжая вычисления, на k-ом шаге вычислим QR-разложение матрицы $A_k = Q_k R_k$ и обозначим $A_{k+1} = R_k Q_k$. Справедливо следующее утверждение, которое мы приводим без доказательства ввиду его сложности. Доказательство можно посмотреть в [9] и [10].

Утверждение. *Если все собственные значения матрицы A вещественны, то последовательность матриц $\{A_k\}$ сходится к матрице, имеющей* ВТФ*:*

$$A_k \underset{k\to\infty}{\longrightarrow} \begin{pmatrix} \lambda_1 & \times & \dots & \times \\ 0 & \lambda_2 & \dots & \times \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_m \end{pmatrix}.$$

Если же матрица имеет комплексную пару собственных значений $\lambda_0 \pm i\lambda_1$, то ей на главной диагонали предельной матрицы будет соответствовать клетка размера 2×2:

$$A_k \underset{k\to\infty}{\longrightarrow} \begin{pmatrix} \times & & & & & \times \\ & \times & & & & \\ & & \lambda_0 & \lambda_1 & & \\ & & -\lambda_1 & \lambda_0 & & \\ & & & & \ddots & \\ \mathbf{0} & & & & & \times \end{pmatrix}.$$

Замечание 1. *Итерационный процесс останавливается, когда все элементы ниже главной диагонали, либо ниже побочной (в случае комплексно-сопряженных собственных значений) матрицы A_n при некотором n становятся равными нулю. Однако следует заметить, что в данном случае под нулем мы понимаем либо машинный ноль, либо число, меньшее некоторой заданной величины — необходимой точности вычисления.*

Замечание 2. *QR-алгоритм применим к произвольной матрице A.*

Замечание 3. *QR-алгоритм является очень затратным по необходимому количеству операций и объему памяти, используемому для хранения промежуточных матриц.*

§12 Предварительное преобразование матрицы к ВПТФ. Неухудшение ВПТФ при QR-алгоритме

Лемма 1. *Пусть $C = BA$, где B имеет* ВТФ, *а A имеет* ВПТФ. *Тогда C имеет* ВПТФ.

Доказательство. Выпишем элемент матрицы C по определению произведения матриц:

$$c_{ij} = \sum_{\alpha=1}^{m} b_{i\alpha} a_{\alpha j},\ i, j = \overline{1, m}.$$

Учтем, что $b_{i\alpha} = 0$ при $\alpha < i$ и $a_{\alpha j} = 0$ при $\alpha > j + 1$:

$$c_{ij} = \sum_{\alpha=i}^{m} b_{i\alpha} a_{\alpha j} = \sum_{\alpha=i}^{j+1} b_{i\alpha} a_{\alpha j},\ i, j = \overline{1, m}.$$

При $i > j + 1$ получим, что $c_{ij} = 0$. Таким образом, C имеет ВПТФ и лемма доказана. □

Аналогичным образом доказывается следующая лемма (ее непосредственное доказательство предоставляется читателю).

Лемма 2. *Пусть $C = BA$, где B — матрица с* ВПТФ, *а A — матрица с* ВТФ. *Тогда C — матрица с* ВПТФ.

Рассмотрим применение QR-алгоритма для матрицы A. Приведем матрицу A к верхней почти треугольной матрице A_0. Запишем QR-разложение матрицы A_0:

$$A_0 = Q_0 R_0.$$

Поскольку R_0 и R_0^{-1} — матрицы, имеющие ВТФ, то матрица Q_0, определяемая выражением

$$Q_0 = A_0 R_0^{-1},$$

в силу леммы 2 имеет ВПТФ. Матрица $A_1 = R_0 Q_0$ в силу леммы 1 также имеет ВПТФ. Таким образом, леммы 1 и 2 гарантируют на каждом шаге QR-алгоритма неухудшение ВПТФ матрицы $A_k,\ k \in \mathbb{Z}_+$.

Глава 2

Интерполирование и приближение функций

§1 Постановка задачи интерполирования

Рассмотрим некоторый технологический процесс, характеризуемый множеством параметров. Разместим в среде протекания процесса конечное число датчиков, позволяющих получать точные значения параметров процесса в ограниченном числе точек среды. Для получения исчерпывающей информации о протекании процесса необходимо уметь оценивать значения параметров процесса в точках, в которых нет возможности их измерить.

Под интерполированием (точное определение будет дано ниже) понимается процесс поиска промежуточных значений величины по имеющемуся дискретному набору известных значений. В вычислительной математике интерполирование обычно рассматривается в рамках задачи вычисления промежуточных значений функций, например, при вычислении значений специальных функций, являющихся решениями дифференциальных уравнений специального вида (функции Бесселя, Ханкеля и другие). Как правило, значения функций такого рода задаются таблицами, шаг которых может оказаться слишком большим для конкретной задачи. В таком случае используют интерполирование для получения значений функции с заданной точностью.

Интерполирование функций используется при исследовании сходимости разностных методов решения дифференциальных задач. При исследовании сходимости необходимо уметь сравнивать сеточные и непрерывные функции. Эту задачу можно решить двумя методами. Первый метод состоит в проецировании непрерывной функции на сетку и последующем сравнении сеточных функций. Второй способ состоит в восстановлении непрерывной функции по сеточной с помощью интерполирования и последующем сравнении непрерывных функций.

Постановка задачи. *Рассмотрим вещественную функцию*

$$f(x), \quad x \in [a, b] \subset \mathbb{R}$$

и произвольным образом заданное разбиение области определения этой функции:

$$a \leqslant x_0 < x_1 < x_2 < \ldots < x_n \leqslant b.$$

Точки $\{x_i\}_{i=0}^n$ *называются узловыми точками функции* $f(x)$*. В этих точках задано значение функции:*

$$f(x_i) = f_i, \quad i = \overline{0, n}.$$

Задача интерполирования состоит в нахождении значений функции $f(x)$ *на всем отрезке [a,b] по ее значениям в узловых точках.*

Замечание. *Далее будем считать термины «интерполирование функции» и «приближение функции» синонимами.*

Заметим, что в постановке задачи интерполирования не указан конкретный метод построения приближенных значений функции $f(x)$. В силу этого задача допускает сколь угодно много решений. В этой главе рассматривается задача приближения заданной функции вещественными полиномами:

$$P_n(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n, \quad a_i \in \mathbb{R}, \quad \sum_{i=0}^{n} a_i^2 \neq 0.$$

Определение. *Вещественный полином n-ой степени $P_n(x)$ называется интерполяционным полиномом для функции $f(x)$, построенным по узлам $\{x_i\}_{i=0}^n$, если его значения в узловых точках совпадают со значениями функции в этих точках:*

$$P_n(x_i) = f_i, \quad i = \overline{0, n}. \tag{1}$$

Утверждение. *Для любой функции $f(x)$ существует единственный интерполяционный полином степени n, построенный по $(n+1)$-му узлу.*

Доказательство. Распишем систему (1) покоординатно:

$$\begin{cases} a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_n x_0^n = f_0 \\ a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_n x_1^n = f_1 \\ \ldots \\ a_0 + a_1 x_n + a_2 x_n^2 + \ldots + a_n x_n^n = f_n \end{cases}. \tag{2}$$

Получим систему линейных уравнений с $(n+1)$-им уравнением относительно коэффициентов полинома $P_n(x)$ с матрицей

$$A = \begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{pmatrix}.$$

Определитель матрицы A — это определитель Вандермонда $(n+1)$-ого порядка:

$$|A| = \prod_{1 \leqslant j < i \leqslant n} (x_i - x_j).$$

Поскольку все узлы различны, матрица A невырождена: $|A| \neq 0$.

Из невырожденности матрицы A следует существование и единственность решения системы (2). Таким образом, для любой функции $f(x)$ существует интерполяционный полином $P_n(x)$, и его коэффициенты однозначно определяются по значениям функции в $(n+1)$-ой узловой точке. □

Замечание. *Помимо интерполирования иногда решают задачу экстраполирования — прогнозирования поведения функции за пределами отрезка. Задача экстраполирования имеет большую погрешность, чем задача интерполирования.*

§2 Интерполяционная формула Лагранжа

Рассмотрим вещественную функцию

$$f(x), \quad x \in [a,b] \subset \mathbb{R},$$

заданную в узловых точках произвольного разбиения отрезка $[a,b]$:

$$a \leqslant x_0 < x_1 < x_2 < \ldots < x_n \leqslant b,$$

$$f(x_i) = f_i, \quad i = \overline{0,n}.$$

Определение. *Интерполяционный полином для функции $f(x)$, заданный формулой*

$$L_n(x) = \sum_{k=0}^{n} c_k(x) f(x_k), \quad k = \overline{0,n}, \tag{1}$$

где $c_k(x)$ — полином степени n, называется интерполяционным полиномом в форме Лагранжа.

Из определения интерполяционного полинома следует, что

$$L_n(x_i) = f(x_i) = f_i, \quad i = \overline{0,n}.$$

Из этих равенств следуют условия

$$c_k(x_l) = \delta_{kl}, \quad k,l = \overline{0,n}. \tag{2}$$

Будем искать полиномы $c_k(x)$ с учетом этих условий.

Рассмотрим полином $(n+1)$-ой степени вида

$$\omega(x) = \prod_{i=0}^{n} (x - x_i).$$

Вынесем за скобку множитель $(x - x_k)$:

$$\omega(x) = (x - x_k) \left(\prod_{\substack{i=0 \\ i \neq k}}^{n} (x - x_i) \right),$$

продифференцируем по x:

$$\omega'(x) = (x - x_k) \left(\prod_{\substack{i=0 \\ i \neq k}}^{n} (x - x_i) \right)' + \left(\prod_{\substack{i=0 \\ i \neq k}}^{n} (x - x_i) \right)$$

и подставим в полученное выражение $x = x_k$:

$$\omega'(x_k) = \left(\prod_{\substack{i=0 \\ i \neq k}}^{n} (x_k - x_i) \right), \quad k = \overline{0,n}.$$

Искомые полиномы $c_k(x)$ можно представить следующим образом:

$$c_k(x) = \frac{\omega(x)}{(x - x_k)\omega'(x_k)}, \quad k = \overline{0, n}. \tag{3}$$

Заметим, что условия (2) для полиномов $c_k(x)$ выполнены. Учитывая равенства (1) и (3), запишем интерполяционный полином в форме Лагранжа:

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

Оценим точность приближения функции $f(x)$ интерполяционным полиномом в форме Лагранжа.

Определение. *Пусть $L_n(x)$ — интерполяционный полином в форме Лагранжа для функции $f(x)$. Тогда функция*

$$\psi_{L_n}(x) = f(x) - L_n(x) \tag{4}$$

называется погрешностью интерполирования функции $f(x)$ интерполяционным полиномом $L_n(x)$.

Пусть существует $(n+1)$-ая производная функции $f(x)$ на отрезке $[a, b]$. Тогда

$$\psi_{L_n}(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega(x), \quad \text{где } \xi \in [a, b]. \tag{5}$$

Обычно оценку погрешности аппроксимации (5) записывают в виде

$$|\psi_{L_n}(x)| \leqslant \frac{M_{n+1}}{(n+1)!} |\omega(x)|, \quad \text{где } M_{n+1} = \sup_{x \in [a,b]} \left| f^{(n+1)}(x) \right|. \tag{6}$$

Замечание 1. *Вывод формул (5) и (6) в данном курсе не рассматривается, его можно найти в [1].*

Замечание 2. *Если исходная функция является полиномом степени, не превышающей n, то интерполяционный полином в форме Лагранжа приближает ее точно.*

Замечание 3. *Наличие в оценке погрешности (6) быстро убывающего множителя $\frac{1}{(n+1)!}$ вовсе не гарантирует сходимость интерполяционного полинома в форме Лагранжа к заданной функции при увеличении числа узлов в разбиении. Более того, начальное разбиение может быть выбрано так, что мы вовсе не получим сходимости. Поэтому на практике лучше разбивать область определения функции на меньшие отрезки, на каждом из которых приближать функцию полиномом невысокой степени, и потом «сшивать» полученные приближения в одну функцию, определенную уже на всем отрезке.*

§3 Разделенные разности

Рассмотрим вещественную функцию

$$f(x), \quad x \in [a, b] \subset \mathbb{R},$$

заданную в узловых точках произвольного разбиения отрезка $[a, b]$:

$$a \leqslant x_0 < x_1 < x_2 < \ldots < x_n \leqslant b,$$

$$f(x_i) = f_i, \quad i = \overline{0, n}.$$

Определение. *Разделенной разностью первого порядка, построенной по несовпадающим узлам x_i и x_j, называется отношение*

$$f(x_i, x_j) = \frac{f(x_j) - f(x_i)}{x_j - x_i}, \quad 0 \leqslant i, j \leqslant n. \tag{1}$$

Обычно мы будем рассматривать разделенные разности, составленные по соседним узлам. Например,

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

Замечание. *Отношение (1) является дискретным аналогом первой производной.*

Определение. *Разделенной разностью второго порядка, построенной по несовпадающим узлам $x_0,\ x_1,\ x_2$, называется отношение*

$$f(x_0, x_1, x_2) = \frac{f(x_1, x_2) - f(x_0, x_1)}{x_2 - x_0}. \tag{2}$$

Замечание. *Отношение (2) является дискретным аналогом второй производной.*

Определение. *Пусть даны $f(x_j, .., x_{j+k})$ и $f(x_{j+1}, ...x_{j+k+1})$ — разделенные разности k-ого порядка по соответствующим узлам, где $0 \leqslant j, k \leqslant n$. Тогда разделенной разностью $(k+1)$-ого порядка, построенной по несовпадающим узлам $x_j,\ x_{j+1},\ \ldots,\ x_{j+k+1}$, называется отношение*

$$f(x_j, x_{j+1}, \ldots, x_{j+k+1}) = \frac{f(x_{j+1}, x_{j+2}, \ldots, x_{j+k+1}) - f(x_j, x_{j+1}, \ldots, x_{j+k})}{x_{j+k+1} - x_j}. \tag{3}$$

Замечание. *Отношение (3) является дискретным аналогом $(k+1)$-ой производной.*

Введем следующие обозначения:

$$\omega(x) = \prod_{i=0}^{n} (x - x_i) = \omega_{0,n}(x),$$

$$\omega_{\alpha,\beta}(x) = \prod_{i=\alpha}^{\beta} (x - x_i), \quad \alpha = 0, 1, \ldots, \beta,\ \beta = \overline{0, n}.$$

Очевидно, что

$$\omega'_{0,n}(x_i) = \prod_{\substack{j=0 \\ i \neq j}}^{n} (x_i - x_j),\ \omega'_{\alpha,\beta}(x_i) = \prod_{\substack{j=\alpha \\ i \neq j}}^{\beta} (x_i - x_j),\ i = \alpha, \alpha+1, \ldots, \beta.$$

Покажем, что разделенная разность произвольного порядка выражается через значения функции $f(x)$ в узлах $\{x_i\}_{i=0}^n$.

Утверждение. *Разделенная разность k-ого порядка представима в виде*

$$f(x_0, x_1, \ldots, x_k) = \sum_{i=0}^{k} \frac{f(x_i)}{\omega'_{0,k}(x_i)}. \tag{4}$$

Доказательство. Воспользуемся методом математической индукции.
Пусть $k = 1$. Тогда

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

Таким образом утверждение выполнено при k=1. Пусть теперь утверждение верно для некоторого $k = l$. Докажем его для $k = l + 1$.

Следующие соотношения вытекают из предположения индукции:

$$f(x_0, x_1, \ldots, x_l) = \sum_{i=0}^{l} \frac{f(x_i)}{\omega'_{0,l}(x_i)}, \tag{5}$$

$$f(x_1, x_2, \ldots, x_{l+1}) = \sum_{i=1}^{l+1} \frac{f(x_i)}{\omega'_{1,l+1}(x_i)}. \tag{6}$$

Запишем разделенную разность $(l + 1)$-ого порядка:

$$f(x_0, x_1, \ldots, x_{l+1}) = \frac{f(x_1, x_2, \ldots, x_{l+1}) - f(x_0, x_1, \ldots, x_l)}{x_{l+1} - x_0}. \tag{7}$$

Подставим выражения (5) и (6) в уравнение (7) и вынесем общий множитель за скобку:

$$f(x_0, x_1, \ldots, x_{l+1}) = \frac{1}{x_{l+1} - x_0} \left(\sum_{i=1}^{l+1} \frac{f(x_i)}{\omega'_{1,l+1}(x_i)} - \sum_{i=0}^{l} \frac{f(x_i)}{\omega'_{0,l}(x_i)} \right).$$

Вынесем за скобку $(l + 1)$-ое слагаемое первой суммы и нулевое слагаемое второй:

$$\begin{aligned} f(x_0, x_1, \ldots, x_{l+1}) = \frac{f(x_0)}{(x_0 - x_{l+1})\omega'_{0,l}(x_0)} + \frac{f(x_{l+1})}{(x_{l+1} - x_0)\omega'_{1,l+1}(x_{l+1})} + \\ + \frac{1}{x_{l+1} - x_0} \left(\sum_{i=1}^{l} f(x_i) \left(\frac{1}{\omega'_{1,l+1}(x_i)} - \frac{1}{\omega'_{0,l}(x_i)} \right) \right). \end{aligned} \tag{8}$$

Рассмотрим отдельно некоторые элементы этого равенства. Заметим, что:

$$(x_0 - x_{l+1})\omega'_{0,l}(x_0) = \omega'_{0,l+1}(x_0),$$

$$(x_{l+1} - x_0)\omega'_{1,l+1}(x_{l+1}) = \omega'_{0,l+1}(x_{l+1}),$$

$$\frac{1}{x_{l+1} - x_0} \left(\frac{1}{\omega'_{1,l+1}(x_i)} - \frac{1}{\omega'_{0,l}(x_i)} \right) =$$

$$= \frac{1}{x_{l+1} - x_0} \left(\frac{x_i - x_0}{\omega'_{1,l+1}(x_i)(x_i - x_0)} - \frac{x_i - x_{l+1}}{\omega'_{0,l}(x_i)(x_i - x_{l+1})} \right) = \frac{1}{\omega'_{0,l+1}(x_i)}.$$

Подставив полученные выражения в равенство (8), получим:

$$f(x_0, x_1, \ldots, x_{l+1}) = \frac{f(x_0)}{\omega'_{0,l+1}(x_0)} + \frac{f(x_{l+1})}{\omega'_{0,l+1}(x_{l+1})} + \sum_{i=1}^{l} \frac{f(x_i)}{\omega'_{0,l+1}(x_i)} = \sum_{i=0}^{l+1} \frac{f(x_i)}{\omega'_{0,l+1}(x_i)}.$$

Утверждение для $k = l + 1$ доказано, и в силу индукции справедлива формула (4). □

Утверждение. *Значение функции $f(x)$ в произвольном узле x_i, $i = \overline{0,n}$ можно выразить через значение функции в узле x_0 и разделенные разности до порядка i включительно.*

Доказательство. Пусть $k = 1$. Запишем разделенную разность первого порядка:

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

Домножим обе части уравнения на $(x_1 - x_0) \neq 0$:

$$(x_1 - x_0)f(x_0, x_1) = f(x_1) - f(x_0).$$

Следовательно,

$$f(x_1) = f(x_0) + (x_1 - x_0)f(x_0, x_1).$$

Докажем утверждение для $k = 2$. Аналогично предыдущему случаю запишем разделенную разность 2-ого порядка и домножим обе части равенства на $(x_2 - x_0)(x_2 - x_1) \neq 0$:

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

Введем обозначения:

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

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

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

Следовательно,

$$(x_2 - x_0)(x_2 - x_1)f(x_0, x_1, x_2) = \alpha + \beta + f(x_2) =$$

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

$$= f(x_2) - f(x_0) - (x_2 - x_0)f(x_0, x_1).$$

Выразив из последнего выражения $f(x_2)$, получим:

$$f(x_2) = f(x_0) + (x_2 - x_0)f(x_0, x_1) + (x_2 - x_0)(x_2 - x_1)f(x_0, x_1, x_2).$$

Переход от $k = l$ к $k = l + 1$ для произвольного $l \in N$ производится по аналогии с рассмотренным переходом от $k = 1$ к $k = 2$, но здесь не приводится, так как сопровождается более громоздкими выкладками. Далее мы иногда будем пользоваться таким приемом, чтобы избегать громоздкости выкладок.

Обобщив полученные результаты, запишем формулу для $f(x_n)$:

$$\begin{aligned} f(x_n) = f(x_0) + (x_n - x_0)f(x_0, x_1) + (x_n - x_0)(x_n - x_1)f(x_0, x_1, x_2) + \\ + \ldots + (x_n - x_0)(x_n - x_1)\ldots(x_n - x_{n-1})f(x_0, x_1, \ldots, x_n). \end{aligned} \tag{9}$$

□

Замечание. *Формула (9) является дискретным аналогом формулы Тейлора.*

§4 Интерполяционная формула Ньютона

Рассмотрим вещественную функцию

$$f(x), \quad x \in [a,b] \subset \mathbb{R},$$

заданную в узловых точках произвольного разбиения отрезка $[a,b]$:

$$a \leqslant x_0 < x_1 < x_2 < \ldots < x_n \leqslant b,$$

$$f(x_i) = f_i, \quad i = \overline{0,n}.$$

Воспользуемся результатами предыдущего параграфа и запишем формулу для $f(x_n)$:

$$\begin{gathered} f(x_n) = f(x_0) + (x_n - x_0)f(x_0,x_1) + (x_n - x_0)(x_n - x_1)f(x_0,x_1,x_2) + \\ + \ldots + (x_n - x_0)(x_n - x_1)\ldots(x_n - x_{n-1})f(x_0,x_1,\ldots,x_n). \end{gathered} \tag{1}$$

Подставив в эту формулу x вместо x_n, получим полином степени n от x:

$$f(x) = f(x_0) + (x - x_0)f(x_0,x_1) + (x - x_0)(x - x_1)f(x_0,x_1,x_2) +$$

$$+ \ldots + (x - x_0)(x - x_1)\ldots(x - x_{n-1})f(x_0,x_1,\ldots,x_n).$$

Обозначим полученный полином как $N_n(x)$:

$$\begin{gathered} N_n(x) = f(x_0) + (x - x_0)f(x_0,x_1) + (x - x_0)(x - x_1)f(x_0,x_1,x_2) + \\ + \ldots + (x - x_0)(x - x_1)\ldots(x - x_{n-1})f(x_0,x_1,\ldots,x_n). \end{gathered} \tag{2}$$

Утверждение. *Полином (2) интерполирует функцию $f(x)$.*

Доказательство. Для доказательства утверждения достаточно показать, что

$$N_n(x_i) = f(x_i), \quad i = \overline{0,n}.$$

Подставив в формулу (2) x_i вместо x, получим:

$$\begin{gathered} N_n(x_i) = f(x_0) + (x_i - x_0)f(x_0,x_1) + (x_i - x_0)(x_i - x_1)f(x_0,x_1,x_2) + \\ + \ldots + (x_i - x_0)(x_i - x_1)\ldots(x_i - x_{n-1})f(x_0,x_1,\ldots,x_n). \end{gathered} \tag{3}$$

В равенстве (3) все слагаемые, начиная с i-ого, содержат множитель $(x_i - x_i)$, тождественно равный нулю. Тогда получим

$$\begin{gathered} N_n(x_i) = f(x_0) + (x_i - x_0)f(x_0,x_1) + (x_i - x_0)(x_i - x_1)f(x_0,x_1,x_2) + \\ + \ldots + (x_i - x_0)(x_i - x_1)\ldots(x_i - x_{i-1})f(x_0,x_1,\ldots,x_i) = f(x_i), \quad i = \overline{0,n}, \end{gathered}$$

что и требовалось доказать. □

Определение. *Интерполяционный полином, задаваемый формулой (2), называется интерполяционным полиномом Ньютона.*

Замечание 1. *Интерполяционный полином Ньютона тождественно совпадает с интерполяционным полиномом в форме Лагранжа.*

Доказательство. Этот факт следует из доказанного в первом параграфе утверждения, что для любой функции $f(x)$ существует единственный интерполяционный полином, построенный по $(n+1)$ узлу. То есть интерполяцоннный полином Ньютона и интерполяцоннный полином в форме Лагранжа являются различными вариантами записи одного и того же интерполяционного полинома. □

Замечание 2. *Так как интерполяционный полином Ньютона тождественно совпадает с интерполяционным полиномом в форме Лагранжа, он имеет такую же погрешность:*

$$\left|\psi_{N_n(x)}(x)\right| \leqslant \frac{M_{n+1}}{(n+1)!}\left|\omega(x)\right|, \quad \text{где } M_{n+1} = \sup_{x\in[a,b]}\left|f^{(n+1)}(x)\right|.$$

Замечание 3. *Аналогично случаю с интерполяционным полиномом Лагранжа, если исходная функция является полиномом степени, не превышающей n, то интерполяционный полином Ньютона приближает ее точно.*

Замечание 4. *Выбор формы записи интерполяционного полинома функции $f(x)$ зависит от особенностей каждой конкретной задачи. Например, если узлы зафиксированы и их число постоянно, а искомая функция меняется, то удобно использовать интерполяционный полином в форме Лагранжа. Если же появляется необходимость в добавлении или удалении узлов при условии сохранения функции, то удобно использовать интерполяционный полином в форме Ньютона.*

§5 Интерполирование с кратными узлами. Полином Эрмита

Рассмотрим вещественную функцию

$$f(x), \quad x \in [a,b] \subset \mathbb{R},$$

заданную в узловых точках произвольного разбиения отрезка $[a,b]$:

$$a \leqslant x_0 < x_1 < x_2 < \ldots < x_m \leqslant b,$$

$$f(x_i) = f_i, \quad i = \overline{0,m}.$$

Пусть, кроме того, в узле x_k заданы значения всех производных функции $f(x)$ до порядка $(a_k - 1)$, $k = \overline{0,m}$. Натуральное число a_k называется кратностью соответствующего узла x_k.

Постановка задачи. *Необходимо построить полином $H_n(x)$ степени n, удовлетворяющий условию:*

$$H_n^{(i)}(x_k) = f^{(i)}(x_k), \quad i = \overline{0,(a_k-1)},\ k = \overline{0,m}.$$

Определение. *Полином $H_n(x)$ называется интерполяционным полиномом Эрмита.*

Будем искать интерполяционный полином $H_n(x)$ в виде

$$H_n(x) = \sum_{k=0}^{m}\sum_{i=0}^{a_{k-1}} c_{k,i}(x) f^{(i)}(x_k), \quad i = \overline{0,(a_k-1)},\ k = \overline{0,m},$$

где $c_{k,i}(x)$ - полиномы степени n.

Сформулируем условие, при котором можно найти интерполяционный полиномом Эрмита.

Утверждение. *Если сумма кратностей узлов функции f(x) равна $(n+1)$:*

$$\sum_{k=0}^{m} a_k = n+1,$$

то существует, причем единственный, интерполяционный полином Эрмита степени n для функции $f(x)$.

Рассмотрение задачи построения интерполяционного полинома Эрмита в общей постановке, которую мы привели выше, выходит за рамки нашего курса. Интересующиеся могут обратится к [1], мы же далее будем рассматривать частный случай: построение интерполяционного полинома Эрмита для функции $f(x)$ по трем узлам, один из которых имеет кратность.

Построение полинома Эрмита по трем узлам

Рассмотрим функцию $f(x)$, определенную вместе со своей первой производной на отрезке $[a, b]$. Построим для функции $f(x)$ интерполяционный полином Эрмита $H_3(x)$ по трем узлам x_0, x_1 и x_2: $a \leqslant x_0 < x_1 < x_2 \leqslant b$, где узел x_1 — кратный.

По определению интерполяционного полинома Эрмита для $H_3(x)$ должны выполняться следующие равенства:

$$H_3(x_0) = f(x_0),\ H_3(x_1) = f(x_1),\ H_3(x_2) = f(x_2),\ H_3'(x_1) = f'(x_1). \tag{1}$$

Будем искать полином Эрмита $H_3(x)$ в виде

$$H_3(x) = c_0(x)f(x_0) + c_1(x)f(x_1) + c_2(x)f(x_2) + b_1(x)f'(x_1), \tag{2}$$

где $b_1(x)$ и $c_i(x),\ i = \overline{0,2}$ - полиномы третьей степени.
Равенства (1) и (2) позволяют сформулировать условия нахождения коэффициентов $b_1(x)$ и $c_i(x),\ i = \overline{0,2}$:

$$\begin{aligned}
&c_0(x_0) = 1, \quad c_1(x_0) = 0, \quad c_2(x_0) = 0, \quad b_1(x_0) = 0,\\
&c_0(x_1) = 0, \quad c_1(x_1) = 1, \quad c_2(x_1) = 0, \quad b_1(x_1) = 0,\\
&c_0(x_2) = 0, \quad c_1(x_2) = 0, \quad c_2(x_2) = 1, \quad b_1(x_2) = 0,\\
&c_0'(x_1) = 0, \quad c_1'(x_1) = 0, \quad c_2'(x_1) = 0, \quad b_1'(x_1) = 1.
\end{aligned}$$

Воспользуемся этими условиями и получим коэффициенты интерполяционного полинома (2) в явном виде.

Из условий $c_0(x_1) = 0$, $c_0(x_2) = 0$ и $c_0'(x_1) = 0$ следует, что узлы x_1 и x_2 являются корнями полинома $c_0(x)$ двойной и единичной кратности соответственно. Поэтому коэффициент $c_0(x)$ будем искать в виде

$$c_0(x) = k(x - x_1)^2(x - x_2), \quad \text{где } k \in \mathbb{R}.$$

Для нахождения k воспользуемся условием $c_0(x_0) = 1$:

$$c_0(x_0) = k(x_0 - x_1)^2(x_0 - x_2) = 1.$$

Поделим это равенство на $(x_0 - x_1)^2(x_0 - x_2)$ (мы можем это сделать, так как узлы x_0, x_1, x_2 различны):

$$k = \frac{1}{(x_0 - x_1)^2(x_0 - x_2)}.$$

Замечание. *В дальнейшем при делении на множители, содержащие разности узлов, мы не будем оговаривать неравенство нулю этих множителей, считая это очевидным.*

Запишем представление для $c_0(x)$ с учетом выражения для коэффициента k:

$$c_0(x) = \frac{(x - x_1)^2(x - x_2)}{(x_0 - x_1)^2(x_0 - x_2)}.$$

Очевидно, что коэффициент $c_2(x)$ имеет аналогичную структуру с двукратным корнем x_1 и однократным корнем x_2:

$$c_2(x) = \frac{(x-x_1)^2(x-x_0)}{(x_2-x_1)^2(x_2-x_0)}.$$

Рассмотрим коэффициент $b_1(x)$, для которого точки x_0, x_1, x_2 являются однократными корнями. Тогда

$$b_1(x) = k_1(x-x_0)(x-x_1)(x-x_2),$$

$$b_1'(x) = k_1((x-x_1)(x-x_2) + (x-x_0)(x-x_2) + (x-x_0)(x-x_1)).$$

Для нахождения k_1 воспользуемся условием $b_1'(x_1) = 1$:

$$b_1'(x_1) = k_1(x_1-x_0)(x_1-x_2) = 1.$$

Получаем выражение для k_1:

$$k_1 = \frac{1}{(x_1-x_0)(x_1-x_2)}.$$

Тогда $b_1(x)$ принимает вид

$$b_1(x) = \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_1-x_0)(x_1-x_2)}.$$

Из условий $c_1(x_0) = 0$, $c_1(x_2) = 0$ следует, что коэффициент $c_1(x)$ обращается в ноль в точках x_0 и x_2. Будем искать его в виде

$$c_1(x) = (x-x_0)(x-x_2)(ax+b), \quad \text{где } a, b \in \mathbb{R}.$$

Так как $c_1(x_1) = 1$, то получаем, что

$$c_1(x_1) = (x_1-x_0)(x_1-x_2)(ax_1+b) = 1.$$

Перепишем равенство относительно $(ax_1 + b)$:

$$ax_1 + b = \frac{1}{(x_1-x_0)(x_1-x_2)}. \tag{3}$$

Для нахождения коэффициента a вычислим производную $c_1'(x)$ в точке x_1:

$$c_1'(x) = a(x-x_0)(x-x_2) + (ax+b)(2x-x_0-x_2).$$

Значит,

$$c_1'(x_1) = a(x_1-x_0)(x_1-x_2) + (ax_1+b)(2x_1-x_0-x_2).$$

Подставив вместо $(ax_1 + b)$ равенство (3), получим представление для коэффициента a:

$$a = -\frac{(2x_1-x_0-x_2)}{(x_1-x_0)^2(x_1-x_2)^2}.$$

Выразим из равенства (3) коэффициент b:

$$b = \frac{1}{(x_1-x_0)(x_1-x_2)} - ax_1 = \frac{1}{(x_1-x_0)(x_1-x_2)} + x_1\frac{(2x_1-x_0-x_2)}{(x_1-x_0)^2(x_1-x_2)^2}.$$

Тогда коэффициент $c_1(x)$ принимает вид:

$$c_1(x) = (x-x_0)(x-x_2)\left(-\frac{(2x_1-x_0-x_2)}{(x_1-x_0)^2(x_1-x_2)^2}x + \frac{1}{(x_1-x_0)(x_1-x_2)} + x_1\frac{(2x_1-x_0-x_2)}{(x_1-x_0)^2(x_1-x_2)^2}\right).$$

Упростив последнее выражение, получим

$$c_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}\left(1 - \frac{(x-x_1)(2x_1-x_0-x_2)}{(x_1-x_0)(x_1-x_2)}\right).$$

Итак, мы нашли все необходимые коэффициенты для построения полинома Эрмита $H_3(x)$.

Замечание. *Заметим, что из-за появления кратных узлов сложность вычисления коэффициентов полинома Эрмита значительно возросла. Если для интерполяционных полиномов в форме Лагранжа и в форме Ньютона существуют единые формулы для вычисления всех коэффициентов, то для полинома Эрмита необходимо вычислять коэффициенты для разных узлов по-разному.*

Оценка погрешности для $H_3(x)$

Зафиксируем $x \in (x_0, x_2) \subset \mathbb{R}$: $x \neq x_1$. Введем функцию $g(s)$:

$$g(s) = f(s) - H_3(s) - k\omega(s), \quad s \in [x_0, x_2],$$

где $\omega(s) = (s-x_0)(s-x_1)^2(s-x_2)$, а k — некая зависящая от x постоянная.
Выберем константу k так, чтобы $g(x) = 0$. Тогда

$$f(x) - H_3(x) - k\omega(x) = 0,$$

$$k = \frac{f(x) - H_3(x)}{\omega(x)}.$$

Введем погрешность для полинома Эрмита $H_3(x)$:

$$\psi_{H_3}(x) = f(x) - H_3(x).$$

Пусть для любого $x \in [x_0, x_2]$ существует $f^{(4)}(x)$. Функция $g(s)$ имеет не менее четырех нулей: три — в узлах x_0, x_1, x_2, а четвертый — в точке x (мы подобрали коэффициент k таким образом, чтобы x был корнем). Воспользуемся теоремой Ролля. Так как $g(s)$ имеет не менее 4-ех нулей, то $g'(s)$ имеет не менее 3-ех нулей на отрезке $[x_0, x_2]$. Так как узел x_1 является кратным узлом для интерполяционного полинома Эрмита $H_3(x)$, то точка x_1 является нулем $g'(s)$: $g'(x_1) = 0$. Следовательно, первая производная имеет не меньше четырех нулей. Тогда вторая производная имеет не менее трех нулей, а третья — не менее двух. Тогда существует точка ξ такая, что

$$g^{(4)}(\xi) = 0 = \left(f^{(4)}(s) - 4!\,k\right)\Big|_{s=\xi} = f^{(4)}(\xi) - 4!\frac{f(x) - H_3(x)}{\omega(x)}.$$

Тогда получим следующее выражение для погрешности:

$$\psi_{H_3}(\xi) = f(\xi) - H_3(\xi) = \frac{f^{(4)}(\xi)}{4!}\omega(\xi).$$

Обозначим

$$M_4 = \sup_{x\in[x_0,x_2]}\left|f^{(4)}(x)\right|.$$

Следовательно,

$$|\psi_{H_3}(x)| \leqslant \frac{M_4}{4!}|\omega(x)|,$$

где $\omega(x) = (x-x_0)(x-x_1)^2(x-x_2)$.

Замечание 1. *В общем случае погрешность интерполяционного полинома Эрмита степени n, $n \in \mathbb{N}$, для функции $f(x)$ имеет вид*

$$\psi_{H_n}(x) = \frac{f^{(n+1)}(x)}{(n+1)!}(x-x_0)^{a_0}(x-x_1)^{a_1}\dots(x-x_m)^{a_m}, \quad a_0+a_1+\dots+a_m = n+1,$$

где $\{x_i\}_{i=0}^m$ — разбиение области определения функции $f(x)$, $m \in \mathbb{N}$, и функция $f(x)$ должна быть $(n+1)$ раз дифференцируема на своей области определения.

Замечание 2. *Интерполяционный полином Эрмита дает более гладкое приближение, чем ранее рассмотренные интерполяционные полиномы в форме Лагранжа и в форме Ньютона.*

Задача. *Показать, что интерполяционный полином Эрмита $H_3(x)$ можно получить из интерполяционного полинома Лагранжа $L_3(x)$ с помощью предельного перехода.*

Решение. Пусть x_0, x_1, x_2 — узловые точки функции $f(x)$ на отрезке $[x_0, x_2]$. Добавим фиктивный узел $x_3 \in [x_0, x_2]$, $x_3 \neq x_i$, $i = \overline{0,2}$. Построим полином в форме Лагранжа по этим четырем узлам:

$$\begin{aligned} L_3(x) = {} & \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}f(x_3) + \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}f(x_1) + \\ & + \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)}f(x_0) + \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)}f(x_2). \end{aligned} \tag{4}$$

Покажем, что $\lim\limits_{x_3 \to x_1} L_3(x) = H_3(x)$.

При стремлении x_3 к x_1, коэффициент при $f(x_0)$ в формуле (4) примет вид:

$$\frac{(x-x_1)^2(x-x_2)}{(x_0-x_1)^2(x_0-x_2)} = c_0(x).$$

Аналогично получим, что выражение коэффициента при $f(x_2)$ совпадает с коэффициентом $c_2(x)$ из интерполяционного полинома Эрмита (2) при $x_3 \to x_1$.

Рассмотрим два оставшихся коэффициента: обозначим через $\alpha(x_3)$ первые два слагаемых суммы (4). $\alpha(x_3)$ можно представить в виде

$$\alpha(x_3) = \frac{\beta(x_3)}{x_3 - x_1},$$

$$\beta(x_3) = \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_2)}f(x_3) - \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)}f(x_1).$$

При переходе к пределу функции $\alpha(x_3)$ при $x_3 \to x_1$ возникает неопределенность вида $\left[\frac{0}{0}\right]$. Для ее устранения воспользуемся правилом Лопиталя и получим:

$$\lim_{x_3 \to x_1} \alpha(x_3) = \lim_{x_3 \to x_1} \frac{\beta'(x_3)}{(x_3 - x_1)'} = \lim_{x_3 \to x_1} \beta'(x_3).$$

Так как $\beta'(x_3)$ уже не содержит неопределенности при $x_3 \to x_1$, то

$$\lim_{x_3 \to x_1} \beta'(x_3) = \beta'(x_1).$$

После проведения всех необходимых вычислений получим, что

$$\beta'(x_1) = \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_1-x_0)(x_1-x_2)}f'(x_1) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}\left(1 - \frac{(x-x_1)(2x_1-x_0-x_2)}{(x_1-x_0)(x_1-x_2)}\right)f(x_1).$$

Видно, что при $f'(x_1)$ и $f(x_1)$ мы получили выражения, в точности совпадающие с коэффициентами $b_1(x)$ и $c_1(x)$ из формулы для интерполяционного полинома Эрмита $H_3(x)$ (2).

□

§6 Использование интерполяционного полинома Эрмита $H_3(x)$ для оценки погрешности квадратурной формулы Симпсона

Рассмотрим задачу приближенного вычисления определенного интеграла

$$I = \int_a^b f(x)dx \tag{1}$$

от интегрируемой по Риману на отрезке $[a, b] \subset \mathbb{R}$ функции $f(x)$.

Построим разбиение отрезка $[a, b]$:

$$a \leqslant x_0 < x_1 < \ldots < x_n \leqslant b, \quad \text{где } n \in \mathbb{N},$$

так, чтобы выполнялось условие

$$x_i - x_{i-1} = h, \quad i = \overline{1, n},$$

где h — некоторая константа, задающая шаг разбиения, причем $hn = b-a$. Отрезки $[x_{i-1}, x_i]$, $i = \overline{1, n}$, называются частичными сегментами.

Будем искать интеграл I в виде суммы интегралов по всем частичным сегментам:

$$I = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)dx. \tag{2}$$

Для вычисления интеграла на всем отрезке достаточно построить приближение интеграла на i-ом частичном сегменте $[x_{i-1}, x_i]$ для $i = \overline{1, n}$.

Замечание. *Часто формулы для приближенного вычисления определенного интеграла называют квадратурными.*

Запишем формулу Симпсона для i-ого частичного сегмента функции $f(x)$, $i = \overline{1, n}$:

$$\int_{x_{i-1}}^{x_i} f(x)dx \approx \frac{h}{6}\left(f(x_{i-1}) + 4f(x_{i-\frac{1}{2}}) + f(x_i)\right), \tag{3}$$

где $x_{i-\frac{1}{2}} = \frac{x_i + x_{i-1}}{2}$ — полуцелая точка.

Утверждение. *Квадратурная формула Симпсона (3) является точной для любого полинома степени не выше трех.*

Доказательство. Приведем доказательство данного утверждения для i-ого частичного сегмента, $i = \overline{1, n}$.
Пусть

$$f(x) = a_0 + a_1x + a_2x^2 + a_3x^3 = L_2(x) + a_3x^3, \quad a_3 \neq 0.$$

Квадратурная формула Симпсона (3) точна для $L_2(x)$, так как по построению задает приближение функций параболами, то есть полиномами второй степени. Покажем, что формула Симпсона точна для функции x^3. Для этого вычислим интеграл $\int_{x_{i-1}}^{x_i} x^3dx$ по формуле

Ньютона-Лейбница:

$$\int\limits_{x_{i-1}}^{x_i} x^3 dx = \frac{x_i^4 - x_{i-1}^4}{4} = \frac{(x_i^2 - x_{i-1}^2)(x_i^2 + x_{i-1}^2)}{4} = \\ = \frac{(x_i - x_{i-1})(x_i + x_{i-1})(x_i^2 + x_{i-1}^2)}{4} = \frac{h}{4}(x_i + x_{i-1})(x_i^2 + x_{i-1}^2) \tag{4}$$

и по квадратурной формуле Симпсона:

$$\int\limits_{x_{i-1}}^{x_i} x^3 dx = \frac{h}{6}(x_{i-1}^3 + 4x_{i-\frac{1}{2}}^3 + x_i^3) = \frac{h}{6}\left((x_{i-1} + x_i)(x_{i-1}^2 - x_i x_{i-1} + x_i^2) + 4\left(\frac{x_i + x_{i-1}}{2}\right)^3\right) =$$

$$= \frac{h}{6}\left((x_{i-1} + x_i)(x_{i-1}^2 - x_i x_{i-1} + x_i^2) + \frac{(x_i + x_{i-1})(x_i^2 + 2x_i x_{i-1} + x_{i-1}^2)}{2}\right) =$$

$$= \frac{h}{6}(x_i + x_{i-1})\left(\frac{2x_{i-1}^2 - 2x_i x_{i-1} + 2x_i^2 + x_i^2 + 2x_i x_{i-1} + x_{i-1}^2}{2}\right) =$$

$$= \frac{h}{12}(x_i + x_{i-1})3(x_{i-1}^2 + x_i^2) = \frac{h}{4}(x_i + x_{i-1})(x_i^2 + x_{i-1}^2).$$

Полученные выражения для интеграла от функции x^3 совпадают, значит, формула Симпсона точна для полиномов третьей степени. □

Перейдем к оценке погрешности квадратурной формулы Симпсона (3), для чего воспользуемся интерполяционным полиномом Эрмита $H_3(x)$, рассмотренным в предыдущем параграфе.

Если для оценки погрешности квадратурной формулы Симпсона мы воспользуемся выражением для погрешности интерполяционного полинома Лагранжа второй степени, то получим сильно завышенную оценку. Правильная оценка получается при использовании полинома Эрмита $H_3(x)$.

Зафиксируем узлы x_{i-1}, $x_{i-\frac{1}{2}}$ и x_i и построим на этих узлах интерполяционный полином Эрмита $H_{3,i}(x)$ для функции $f(x)$. Ранее в §5 было доказано, что такой полином существует, единственен и удовлетворяет следующим условиям:

$$H_{3,i}(x_{i-1}) = f(x_{i-1}),$$

$$H_{3,i}(x_{i-\frac{1}{2}}) = f(x_{i-\frac{1}{2}}),$$

$$H_{3,i}(x_i) = f(x_i),$$

$$H'_{3,i}(x_{i-\frac{1}{2}}) = f'(x_{i-\frac{1}{2}}).$$

Запишем погрешность для полинома $H_{3,i}(x)$:

$$\psi_{H_{3,i}}(x) = \frac{f^{(4)}(\xi)}{4!}(x - x_{i-1})(x - x_{i-\frac{1}{2}})^2(x - x_i), \quad \xi \in [x_{i-1}, x_i]. \tag{5}$$

Введем обозначение:

$$\Psi_i(f) = \int\limits_{x_{i-1}}^{x_i} \psi_{H_{3,i}}(x) dx. \tag{6}$$

Представим исходную функцию $f(x)$ в виде $f(x) = H_{3,i}(x) + \psi_{H_{3,i}(x)}$. Тогда

$$\int\limits_{x_{i-1}}^{x_i} f(x)dx = \int\limits_{x_{i-1}}^{x_i} H_{3,i}(x)dx + \int\limits_{x_{i-1}}^{x_i} \psi_{H_{3,i}}(x)dx. \tag{7}$$

Так как формула Симпсона (3) точна для полиномов третьей степени, то мы можем заменить интеграл $\int\limits_{x_{i-1}}^{x_i} H_{3,i}(x)dx$ на соответствующую ему правую часть формулы (3):

$$\int\limits_{x_{i-1}}^{x_i} H_{3,i}(x)dx = \frac{h}{6}\left(H_{3,i}(x_{i-1}) + 4H_{3,i}(x_{i-\frac{1}{2}}) + H_{3,i}(x_i)\right).$$

Тогда

$$\int\limits_{x_{i-1}}^{x_i} f(x)dx = \frac{h}{6}\left(H_{3,i}(x_{i-1}) + 4H_{3,i}(x_{i-\frac{1}{2}}) + H_{3,i}(x_i)\right) + \int\limits_{x_{i-1}}^{x_i} \psi_{H_{3,i}}(x)dx =$$

$$= \frac{h}{6}\left(f(x_{i-1}) + 4f(x_{i-\frac{1}{2}}) + f(x_i)\right) + \Psi_i(f).$$

Следовательно,

$$\Psi_i(f) = \int\limits_{x_{i-1}}^{x_i} f(x)dx - \frac{h}{6}\left(f(x_{i-1}) + 4f(x_{i-\frac{1}{2}}) + f(x_i)\right). \tag{8}$$

Таким образом мы получаем, что $\Psi_i(f)$ задает погрешность формулы Симпсона (3) на i-ом частичном сегменте.

Так как выполнены равенства (5) и (6), то погрешность (8) можно оценить следующим образом:

$$|\Psi_i(f)| \leqslant \int\limits_{x_{i-1}}^{x_i} |\psi_{H_{3,i}}(x)|dx \leqslant \int\limits_{x_{i-1}}^{x_i} \frac{M_4}{4!}(x - x_{i-1})(x - x_{i-\frac{1}{2}})^2(x_i - x)dx,$$

$$M_4 = \sup_{x\in[x_{i-1},x_i]} \left|f^{(4)}(x)\right|.$$

Задача. *Показать, что*

$$\int\limits_{x_{i-1}}^{x_i} (x - x_{i-1})(x - x_{i-\frac{1}{2}})^2(x_i - x)dx = \frac{h^5}{120}.$$

Решение. Произведем замену в подынтегральном выражении: $x = x_{i-1} + th,\ t \in [0, 1]$. Тогда $dx = hdt$ и $x - x_{i-1} = th,\ x_i - x = h(1 - t),\ (x - x_{i-\frac{1}{2}})^2 = h^2\left(t - \frac{1}{2}\right)^2$, и мы получаем, что

$$\int\limits_{x_{i-1}}^{x_i} (x - x_{i-1})(x - x_{i-\frac{1}{2}})^2(x_i - x)dx =$$

$$= h^5 \int\limits_0^1 t\ \left(t - \frac{1}{2}\right)^2 (1 - t)dt = h^5 \int\limits_0^1 \left(2t^3 - \frac{5}{4}t^2 - t^4 + \frac{1}{4}t\right)dt = \frac{h^5}{120}.$$

□

Таким образом, погрешность формулы Симпсона (3) на i-ом частичном сегменте имеет пятый порядок точности:

$$|\Psi_i(f)| \leqslant \frac{M_4}{4!}\frac{h^5}{120}. \tag{9}$$

Оценим погрешность приближения интеграла I (1) на всем отрезке $[a,b]$, учитывая представление этого интеграла в виде суммы ингералов по всем частичным сегментам (2) и воспользовавшись формулой Симпсона (3):

$$|\Psi_n(f)| = \left|\int_a^b f(x)dx - \sum_{i=0}^{n}\frac{h}{6}\left(f(x_{i-1}) + 4f(x_{i-\frac{1}{2}}) + f(x_i)\right)\right| \leqslant \left|\sum_{i=0}^{n}\Psi_i(f)\right|.$$

Мы выбирали разбиение отрезка $[a,b]$ так, что $nh = b-a$, поэтому с учетом оценки (9) получим, что

$$|\Psi_n(f)| \leqslant \left(\frac{h}{2}\right)^4 \frac{M_4(b-a)}{180}.$$

Следовательно, квадратурная формула Симпсона на всем отрезке $[a,b]$ имеет четвертый порядок точности.

§7 Наилучшее среднеквадратичное приближение функции

Рассмотрим гильбертово пространство L_2 — линейное пространство функций, интегрируемых с квадратом:

$$\int_a^b f^2(x)dx < \infty.$$

Заметим, что здесь рассматривается интегрирование любого типа, не только интегрирование по Риману.

Введем скалярное произведение в пространстве L_2:

$$\forall f,g \in L_2 \quad (f,g) = \int_a^b f(x)g(x)dx.$$

Теперь введем норму в пространстве L_2:

$$\|f\|_{L_2} = \|f\| = \sqrt{(f,f)} = \left(\int_a^b f^2(x)dx\right)^{\frac{1}{2}}.$$

Определение. *Пусть дана система $(n+1)$ линейно независимых функций в пространстве L_2 $\{\varphi_i(x)\}_{i=0}^n$. Многочлен $\varphi(x)$ вида*

$$\varphi(x) = c_0\varphi_0(x) + c_1\varphi_1(x) + \ldots + c_n\varphi_n(x) = \sum_{k=0}^{n} c_k\varphi_k(x), \text{ где } c_k \in \mathbb{R}, k = \overline{1,n},$$

называется обобщенным многочленом.

Так как коэффициенты обобщенного многочлена задаются произвольным образом, то, варьируя их значения, можно получить бесконечно много различных обобщенных многочленов.

Определение. *Пусть $f(x) \in L_2$ и дана система из $(n+1)$ линейно независимых функций*

$$\varphi_i(x) \in L_2, \quad i = \overline{0,n}.$$

Обобщенный многочлен $\overline{\varphi}(x)$, имеющий минимальное отклонение по норме от функции $f(x)$:

$$\|f(x) - \overline{\varphi}(x)\| = \min_{\varphi(x)} \|f(x) - \varphi(x)\| = \min_{\varphi(x)} \left(\int_a^b (f(x) - \varphi(x))^2 \, dx \right)^{\frac{1}{2}},$$

называется наилучшим среднеквадратичным приближением функции $f(x)$ по системе функций $\{\varphi_i(x)\}_{i=0}^n$.

Утверждение. *Наилучшее среднеквадратичное приближение функции $f(x)$ по системе функций $\{\varphi_i(x)\}_{i=0}^n$ существует и единственно.*

Доказательство. Вначале рассмотрим доказательство для частного случая: выберем систему функций, состоящую из одной функции $\varphi_0(x) \in L_2$. Тогда обобщенный многочлен имеет вид

$$\varphi(x) = c_0 \varphi_0(x).$$

Рассмотрим задачу для функции $f(x)$: среди всех обобщенных многочленов найдем тот, который минимизирует функционал

$$F(c_0) = \int_a^b (f(x) - c_0\varphi_0(x))^2 dx.$$

Преобразуем это выражение:

$$F(c_0) = \int_a^b f^2(x)dx - 2c_0 \int_a^b f(x)\varphi_0(x)dx + c_0^2 \int_a^b \varphi_0^2(x)dx = (f,f) - 2c_0(f,\varphi_0) + c_0^2(\varphi_0,\varphi_0).$$

Мы получили квадратичную функцию относительно c_0. Найдем ее экстремум:

$$F'(c_0) = 0,$$

$$c_0(\varphi_0,\varphi_0) = (f,\varphi_0).$$

Тогда коэффициент $\overline{c_0}$, доставляющий минимум функционалу $F(c_0)$, равен:

$$\overline{c_0} = \frac{(f,\varphi_0)}{(\varphi_0,\varphi_0)} = \frac{\int_a^b f(x)\varphi_0(x)dx}{\int_a^b \varphi_0^2(x)dx}. \tag{1}$$

Получим наилучшее среднеквадратичное приближение $\overline{\varphi}(x)$ для функции $f(x)$:

$$\overline{\varphi}(x) = \overline{c_0}\varphi_0(x) = \frac{(f,\varphi_0)}{(\varphi_0,\varphi_0)}\varphi_0. \tag{2}$$

Заметим, что при $\varphi_0(x) = 1$, из выражений (1) и (2) можно получить выражение для среднего значения интеграла:

$$\overline{\varphi}(x) = \frac{\int_a^b f(x)dx}{(b-a)},$$

которое и является наилучшим среднеквадратичным приближением в этом случае.

Разумеется, увеличивая число n базисных функций $\varphi_i(x)$, мы вправе ожидать увеличения точности приближения. Покажем, как строится наилучшее среднеквадратичное приближение в случае произвольного n. Пусть $\{\varphi_i(x)\}_{i=0}^n$ — система линейно независимых функций, $\varphi_i(x) \in L_2[a,b]$. Обозначим обобщенный многочлен через

$$\varphi(x) = \sum_{k=0}^{n} c_k \varphi_k(x), \quad \text{где } c_k \in \mathbb{R}$$

и рассмотрим функционал

$$F(c_0, c_1, \ldots, c_n) = \int_a^b (f(x) - \varphi(x))^2 dx = \int_a^b (f(x) - \sum_{k=0}^{n} c_k \varphi_k(x))^2 dx.$$

Преобразуем это равенство:

$$F(c_0, c_1, \ldots, c_n) = \int_a^b f^2(x)dx - 2\sum_{k=0}^{n} c_k \int_a^b f(x)\varphi_k(x)dx + \sum_{k=0}^{n} c_k \sum_{l=0}^{n} c_l \int_a^b \varphi_k(x)\varphi_l(x)dx =$$

$$= (f,f) - 2\sum_{k=0}^{n} c_k (f, \varphi_k) + \sum_{k=0}^{n} c_k \sum_{l=0}^{n} c_l (\varphi_k, \varphi_l).$$

Минимум функционала $F(c_0, c_1, \ldots, c_n)$ достигается в точке, в которой все частные производные первого порядка обращаются в ноль:

$$\frac{\partial F(c_0, \ldots, c_n)}{\partial c_k} = 0, \quad k = \overline{0,n}.$$

Получаем систему уравнений относительно коэффициентов $c_l, l = \overline{0,n}$:

$$\sum_{l=0}^{n} c_l (\varphi_k, \varphi_l) = (f, \varphi_k), \quad k = \overline{0,n}.$$

Запишем эту систему более подробно:

$$\begin{cases} c_0(\varphi_0,\varphi_0) + c_1(\varphi_0,\varphi_1) + \ldots + c_n(\varphi_0,\varphi_n) = (f,\varphi_0) \\ c_1(\varphi_1,\varphi_0) + c_1(\varphi_1,\varphi_1) + \ldots + c_n(\varphi_1,\varphi_n) = (f,\varphi_1) \\ \ldots \\ c_0(\varphi_n,\varphi_0) + c_1(\varphi_n,\varphi_1) + \ldots + c_n(\varphi_n,\varphi_n) = (f,\varphi_n). \end{cases} \tag{3}$$

Выпишем матрицу коэффициентов системы:

$$\begin{pmatrix} (\varphi_0,\varphi_0) & (\varphi_0,\varphi_1) & \ldots & (\varphi_0,\varphi_n) \\ (\varphi_1,\varphi_0) & (\varphi_1,\varphi_1) & \ldots & (\varphi_1,\varphi_n) \\ \vdots & \vdots & \ddots & \vdots \\ (\varphi_n,\varphi_0) & (\varphi_n,\varphi_1) & \ldots & (\varphi_n,\varphi_n) \end{pmatrix} = G(\varphi_0, \ldots, \varphi_n).$$

Полученная матрица является матрицей Грама системы функций $\{\varphi_i(x)\}_{i=0}^n$. Так как $\{\varphi_i(x)\}_{i=0}^n$ — система линейно независимых функций, то определитель матрицы Грама положителен:

$$|G(\varphi_0, \ldots, \varphi_n)| > 0.$$

Следовательно система линейных уравнений (3) имеет единственное решение $(\overline{c_0}, \overline{c_1}, \ldots, \overline{c_n})^T$. Тогда наилучшее среднеквадратичное приближение для функции $f(x)$ существует и определено единственным образом:

$$\overline{\varphi}(x) = \sum_{i=0}^{n} \overline{c_i}\varphi_i(x).$$

□

Замечание 1. *Можно заметить, что чем больше базисных функций мы вводим, тем точнее среднеквадратичное приближение заданной функции. В пределе мы переходим в базис всего пространства и получаем точное разложение заданной функции по базису. Однако следует помнить, что при увеличении числа базисных функций увеличивается и размер соответствующей матрицы Грама, а определитель этой матрицы приближается к нулю. Это создает определенные проблемы при решении задач на практике, связанные с увеличением влияния ошибок округления.*

Замечание 2. *Заметим, что если исходная система функций $\{\varphi_i(x)\}_{i=0}^n$ — ортогональная, то матрица Грама этой системы — диагональная, что значительно упрощает нахождение среднеквадратичного приближения заданной функции.*

Замечание 3. *Если $\{\varphi_i(x)\}_{i=0}^n$ — ортонормированная система функций в пространстве L_2, то соответствующая этой системе матрица Грама является единичной, и решение системы (3) имеет вид*

$$\overline{c}_k = (f, \varphi_k), \quad k = \overline{0, n}, \tag{4}$$

где $\overline{c}_k$ — коэффициенты обобщенного многочлена, реализующего наилучшее среднеквадратичное приближение функции $f(x)$. Коэффициенты такого вида называются коэффициентами Фурье функции $f(x)$.

Замечание 4. *Рассмотрим систему линейно независимых функций*

$$\varphi_k(x) = x^k, \quad k = \overline{0, n}.$$

Введем в пространстве скалярное произведение следующим образом:

$$\int_{\alpha}^{\beta} \rho(x)\varphi_k(x)\varphi_l(x)dx = (\varphi_k, \varphi_l),$$

где $\rho(x) > 0$ — весовая функция. Если определенным образом выбирать границы α и β и весовую функцию, то можно построить систему ортогональных полиномов (например, полиномы Якоби, Лежандра, Чебышева).

Утверждение. *Если $\{\varphi_i(x)\}_{i=0}^n$ — ортонормированная система функций, то для этой системы функций выполняется неравенство Бесселя:*

$$\sum_{k=0}^{n} c_k^2 \leqslant \|f\|^2,$$

где c_k — коэффициенты обобщенного многочлена, реализующего наилучшее среднеквадратичное приближение функции $f(x)$.

Доказательство. Действительно, если система функций $\{\varphi_i(x)\}_{i=0}^n$ ортонормирована, то выполнено замечание 3. Обозначим $\overline{c}_k = c_k$ и вычислим отклонение от наилучшего среднеквадратичного приближения:

$$\int_a^b (f(x) - \sum_{k=0}^n c_k\varphi_k(x))^2 dx = (f,f) - 2\sum_{k=0}^n c_k(f,\varphi_k) + \sum_{k=0}^n c_k^2 = (f,f) - \sum_{k=0}^n c_k^2 \geqslant 0.$$

Следовательно неравенство Бесселя выполнено. □

Замечание 5. *Если $\{\varphi_i(x)\}_{i=0}^\infty$ — ортонормированный базис, то выполняется равенство Парсеваля:*

$$\sum_{k=0}^\infty c_k^2 = \|f\|^2.$$

Замечание 6. *В процессе построения наилучшего среднеквадратичного приближения возникает следующий ряд вопросов:*

1. *Как решать системы линейных уравнений высокого порядка?*
2. *Как вычислять интегралы для поиска скалярных произвдений функций для построения системы (3)?*
3. *Как производить суммирование с коэффициентами Фурье?*

На первый из этих вопросов мы ответили в главе I, второго коснулись в §6, рассмотрение остальных вопросов выходит за рамки нашего курса.

§8 Наилучшее среднеквадратичное приближение функций, заданных таблично

Пусть H — линейное пространство функций, заданных таблично, то есть элементы $f \in H$ — функции, заданные в узлах $a \leqslant x_0 < x_1 < \ldots < x_N \leqslant b$, $N \in \mathbb{N}$:

$$f(x_i) = f_i, \quad i = \overline{0,N}.$$

Введем скалярное произведение в пространстве H:

$$\forall f,g \in H \quad (f,g) = \sum_{i=0}^N f_i g_i.$$

Введем соответствующую норму — эта норма является аналогом среднеквадратичной нормы в пространстве функций, определенных на всем отрезке $[a,b]$:

$$\forall f \in H \quad \|f\| = \sqrt{(f,f)} = \left(\sum_{i=0}^N f_i^2\right)^{\frac{1}{2}}.$$

В предыдущем параграфе предполагалось, что функция $f(x)$ задана аналитически. Здесь функция задана таблично, то есть известны только ее значения $f_i = f(x_i)$ в конечном числе точек x_i, $i = \overline{0,N}$.

Мы хотим приблизить функцию $f(x)$ некоторой функцией, заданной аналитически. Один из способов приближения мы уже знаем — это интерполяция по данным значениям

$f_0, f_1, \ldots, f_N$. Однако при больших N такой способ приближения трудоемок и может даже дать неверное представление о поведении функции. Одним из распространенных способов приближения функций, заданных таблично, является способ, основанный на минимизации среднеквадратичной погрешности.

Как и в предыдущем параграфе, предположим, что задана система базисных функций $\{\varphi_i(x)\}_{i=0}^n$ (например, $\varphi_i(x) = x^i,\ i = \overline{0,n}$). Можем считать, что функции $\varphi_i(x)$ заданы только в точках $x_j,\ j = \overline{0,N}$. Задача состоит в подборе коэффициентов c_k, для которых величина отклонения

$$\left\| f - \sum_{k=0}^{n} c_k \varphi_k \right\| = \left(\sum_{i=0}^{N} \left(f_i - \sum_{k=0}^{n} c_k \varphi_k(x_i) \right)^2 \right)^{\frac{1}{2}}$$

являлась бы минимальной. Эта задача является дискретным аналогом задачи о минимизации функционала $F(c_0, c_1, \ldots, c_n)$, рассмотренной в предыдущем параграфе, и решается аналогичным образом.

Введем функционал

$$F(c_0, c_1, \ldots, c_n) = \left\| f - \sum_{k=0}^{n} c_k \varphi_k \right\|^2.$$

Этот функционал имеет тот же вид, что и аналогичный функционал для функций гильбертового пространства, рассмотренный в предыдущем параграфе.

Запишем систему линейных уравнений для поиска коэффициентов $\{c_k\}_{k=0}^n$, на которых функционал $F(c_0, c_1, \ldots, c_n)$ достигает своего минимума:

$$\frac{\partial F}{\partial c_k} = 0, \quad k = \overline{0,n},$$

$$\sum_{l=0}^{n} c_l(\varphi_k, \varphi_l) = (f, \varphi_k), \quad k = \overline{0,n}.$$

Вид полученной системы аналогичен виду системы, которую мы рассматривали в предыдущем параграфе, следовательно, для рассматриваемой системы сохраняется свойство существования и единственности решения — $c_k,\ k = \overline{0,n}$.

Значит, для построения наилучшего среднеквадратичного приближения функции с помощью некоторой системы функций достаточно знать значения этой функции лишь в некоторых точках интересующего отрезка.

Глава 3

Численное решение нелинейных уравнений и систем нелинейных уравнений

§1 Введение

Рассмотрим задачу поиска корней нелинейного уравнения: нелинейные уравнения, вообще говоря, не имеют аналитического решения, поэтому для поиска решения используют вычислительные методы, хотя такое решение является лишь приближенным.

Заметим, что принципиальное отличие численных методов решения нелинейных уравнений от численных методов решения систем линейных уравнений состоит в необходимости специально выбирать для конкретного итерационного метода начальное приближение, так как от этого выбора зависит сходимость рассматриваемых итерационных методов решения нелинейных уравнений.

Постановка задачи. *Рассмотрим функцию* $f(x),\ x \in \mathbb{R}$, *и уравнение*

$$f(x) = 0. \tag{1}$$

Пусть x^* *— корень уравнения, и определена его окрестность радиуса* a*, не содержащая других корней уравнения:*

$$U_a(x^*) = \{x : |x - x^*| < a\},$$

причем заданная функция $f(x)$ *определена на этой окрестности. Будем считать, что начальное приближение* $x^0 \in U_a(x^*)$ *задано. Тогда для нахождения численного решения уравнения в рассматриваемой окрестности необходимо построить последовательность* $\{x^n\}$*, сходящуюся к корню* x^* *уравнения* (1)*:*

$$\lim_{n\to\infty} f(x^n) = f(x^*) = 0.$$

Численное решение нелинейных уравнений можно разбить на два этапа:

1. Локализация корня, т.е. определение окрестности $U_a(x^*)$.
2. Задание итерационного процесса — построение последовательности $\{x^n\}$, сходящейся к корню уравнения.

Пусть $f(x)$ — непрерывная функция, заданная на отрезке $[a, b]$. Рассмотрим два приема локализации вещественного корня (известно, что уравнение (1) может иметь и комплексные корни, но в данном курсе мы не будем ими заниматься).

Первый прием

Пусть задано разбиение сегмента $[a, b]$:

$$a \leqslant x_0 < x_1 < x_2 < \ldots < x_n \leqslant b,$$

и если для некоторого $i = \overline{1, n}$ выполняется условие

$$f(x_{i-1})f(x_i) < 0, \tag{2}$$

то на интервале (x_{i-1}, x_i) существует по крайней мере один корень уравнения (1) или число корней на этом интервале нечетно. Если же выполняется условие

$$f(x_{i-1})f(x_i) > 0, \quad i = \overline{1, n},$$

то на каждом из интервалов (x_{i-1}, x_i) либо нет корней уравнения (1), либо их число четно.

В случае выполнения условия (2) интервал (x_{i-1}, x_i) вновь разбивается на частичные интервалы, и для частичных интервалов повторяется описанная выше процедура, которая в итоге позволит найти промежуток меньшей длины, содержащий корень.

Второй прием

Более регулярным способом отделения действительных корней является метод бисекции (деления пополам).

Предположим, что на сегменте (a, b) расположен лишь один корень x_* уравнения (1). Тогда $f(a)$ и $f(b)$ имеют различные знаки. Пусть для определенности $f(a) > 0$, $f(b) < 0$.

Положим

$$x_0 = \frac{a + b}{2}$$

и рассмотрим значения функции $f(x)$ в этой точке.
Если $f(x_0) < 0$, то значение искомого корня x_* лежит в интервале (a, x_0), если же $f(x_0) > 0$, то $x_* \in (x_0, b)$. Далее из этих двух интервалов (a, x_0) и (x_0, b) выбираем тот, на границе которого функция $f(x)$ имеет различные знаки.
Затем находим точку x_1 — середину выбранного интервала, вычисляем $f(x_1)$ и повторяем указанный выше алгоритм.

В результате получаем последовательность интервалов, содержащих искомый корень x_*, причем каждый последующий интервал имеет длину в 2 раза меньшую, чем предыдущий. Процесс заканчивается, когда длина вновь полученного интервала станет меньше заданного числа $\varepsilon > 0$.

Замечание. *Как правило рассматриваемая функция $f(x)$ имеет больше одного корня, и задача состоит в поиске всех корней уравнения (1) на области определения функции $f(x)$. Тогда можно поступать следующим образом: пусть мы нашли один из корней этого уравнения, причем этот корень имеет единичную кратность. Тогда для поиска других корней рассматриваемого уравнения осуществим переход к функции $g(x)$ вида*

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

Очевидно, что уравнение $g(x) = 0$ имеет на единицу меньше корней, чем уравнение (1), и все корни этого уравнения являются также корнями уравнения (1). Поэтому после решения данного уравнения получаем корни исходного уравнения, отличные от уже найденных. Таким образом мы сможем найти по крайней мере все некратные корни уравнения (1).

Круг вопросов, которые мы рассматриваем в связи с решением одного нелинейного уравнения, переносится и на поиск решения системы нелинейных уравнений. Рассмотрим нелинейную систему уравнений

$$f_i(x_1, x_2, \ldots, x_m) = 0, \quad i = \overline{1, m}. \tag{3}$$

Введем векторы $x = (x_1, x_2, \ldots, x_m)^T$, $f = (f_1, f_2, \ldots, f_m)^T$. Тогда система уравнений (3) запишется в векторной форме, как

$$f(x) = \theta.$$

Последнее уравнение удобно рассматривать как операторное уравнение в m-мерном пространстве $\mathbb{R}^m$. При этом отображение

$$f\colon \mathbb{R}^m \longrightarrow \mathbb{R}^m$$

представляет собой нелинейное отображение пространства $\mathbb{R}^m$ в себя, и рассуждения о методах решения нелинейных систем проводится аналогично одномерному случаю.

§2 Метод простой итерации

Рассмотрим функцию $f(x)$, $x \in \mathbb{R}$ и уравнение

$$f(x) = 0. \tag{1}$$

Пусть x^* — корень этого уравнения, и определена его окрестность радиуса a, не содержащая других корней рассматриваемого уравнения:

$$U_a(x^*) = \{x : |x - x^*| < a\},$$

причем заданная функция $f(x)$ определена на этой окрестности.

Будем считать, что начальное приближение $x_0 \in U_a(x^*)$ задано. Рассмотрим итерационные методы, задаваемые общей формулой

$$x^{n+1} = S(x^n), \quad n \in \mathbb{Z}_+ \tag{2}$$

с некоторой функцией $S(x)$, определенной на $U_a(x^*)$. Пусть функция $S(x)$ имеет вид

$$S(x) = x + r(x)f(x), \ S(x^*) = x^*, \tag{3}$$

где $r(x)$ — функция, не обращающаяся в ноль в окрестности $U_a(x^*)$, то есть $sgn(r(x)) \neq 0$, $x \in U_a(x^*)$.

Определение. *Итерационный метод, описываемый формулой (2) с функцией $S(x)$ вида (3), называется методом простой итерации.*

Определение. *Функция $S(x)$ называется Липшиц-непрерывной при $x \in U_a(x^*)$ с константой $q > 0$, если для любых точек $x_1, x_2 \in U_a(x^*)$ выполнено неравенство*

$$|S(x_1) - S(x_2)| \leqslant q|x_1 - x_2|.$$

Утверждение. *Пусть функция $S(x)$ Липшиц-непрерывна с константой $q \in (0, 1)$ в некоторой окрестности $U_a(x^*)$, и пусть задано начальное приближение $x_0 \in U_a(x^*)$. Тогда метод простой итерации (2) сходится со скоростью геометрической прогрессии со знаменателем q.*

Доказательство. Докажем с помощью метода математической индукции, что $x^k \in U_a(x^*)$ при $k \in \mathbb{Z}_+$.

Справедливость утверждения $x^0 \in U_a(x^*)$ следует из условия. Пусть требуемое условие верно при $k = n$. Рассмотрим $(n+1)$-ую итерацию:

$$x^{n+1} = S(x^n)$$

и оценим $|x^{n+1} - x^*|$, учитывая, что функция $S(x)$ Липшиц-непрерывна:

$$|x^{n+1} - x^*| = |S(x^n) - S(x^*)| \leqslant q|x^n - x^*|. \tag{4}$$

Из условия $q \in (0, 1)$ следует неравенство

$$|x^{n+1} - x^*| \leqslant q|x^n - x^*| < a.$$

Таким образом, $x^{n+1} \in U_a(x^*)$.

Докажем сходимость метода простой итерации. Используя оценку (4) как рекуррентную, получим:

$$|x^n - x^*| \leqslant q^n|x^0 - x^*|. \tag{5}$$

Из условия $q \in (0, 1)$ следует, что

$$\lim_{n\to\infty} q^n = 0.$$

Тогда в силу неравенства (5) и неотрицательности модуля выполнено равенство

$$\lim_{n\to\infty} |x^n - x^*| = 0.$$

Следовательно, метод простой итерации сходится со скоростью геометрической прогрессии со знаменателем q. □

Замечание. *Если функция $S(x)$ непрерывно дифференцируема, то в качестве q можно взять максимальное значение $|S'(x)|$, и сходимость будет иметь место, если*

$$q = \max_{x\in U_a(x^*)} \left|S'(x)\right| < 1.$$

Рассмотрим итерационный метод, записанный уравнением:

$$\frac{x^{n+1} - x^n}{\tau} + f(x^n) = 0, \quad \tau \in \mathbb{R}_+,\ \tau > 0,\ n \in \mathbb{Z}_+,\ x^0 \in U_a(x^*). \tag{6}$$

Выразим из этого равенства x^{n+1}:

$$x^{n+1} = x^n - \tau f(x_n).$$

Этот метод является методом простой итерации вида (2) с функцией $S(x)$, имеющий вид

$$S(x) = x - \tau f(x).$$

Получим оценку параметра τ, которая будет гарантировать сходимость метода простой итерации вида (6), то есть обеспечивать выполнение условий замечания к доказанному выше утверждению.

Пусть окрестность $U_a(x^*)$ выбрана таким образом, чтобы в ней выполнялось условие $|S'(x)| < 1$. В предположении об ограниченности функции $f'(x)$ вычислим точную верхнюю грань M ее модуля:

$$M = \sup_{x\in U_a(x^*)} \left|f'(x)\right|.$$

Продифференцируем функцию $S(x)$:

$$S'(x) = 1 - \tau f'(x).$$

Пусть для определенности $f'(x) > 0,\ x \in U_a(x^*)$. Потребовав, чтобы выполнялось условие $|S'(x)| < 1$, получим оценку для τ:

$$|1 - \tau M| < 1, \quad 0 < \tau < \frac{2}{M}.$$

Таким образом, если для поиска корня x^* применяется итерационный метод, записанный в виде (6), то значение параметра τ следует выбирать из интервала $\left(0, \frac{2}{M}\right)$.

Метод Эйткена ускорения сходимости итерационного метода

Предположим, что существует число A, не зависящее от n и такое, что

$$x^n - x^* \approx Aq^n, \quad n \in \mathbb{Z}_+, \quad A \in \mathbb{R}.$$

Запишем оценки для трех последовательных итераций:

$$x^{n-1} - x^* \approx Aq^{n-1}, \quad x^n - x^* \approx Aq^n, \quad x^{n+1} - x^* \approx Aq^{n+1}, \tag{7}$$

Выразим Aq^{n+1} через итерации $x^{n-1},\ x^n,\ x^{n+1}$. Для этого рассмотрим равенства

$$(x^{n+1} - x^n)^2 = A^2q^{2n}(q-1)^2,$$

$$x^{n+1} - 2x^n + x^{n-1} = Aq^{n-1}(q-1)^2,$$

получающиеся из выражений (7). Разделим первое равенство на второе:

$$\frac{(x^{n+1} - x^n)^2}{x^{n+1} - 2x^n + x^{n-1}} = Aq^{n+1}.$$

Подставим полученное выражение для Aq^{n+1} в оценку (7) для корня x^* и $(n+1)$-ой итерации x^{n+1} и получим представление для корня x^*:

$$x^* \approx x^{n+1} - \frac{(x^{n+1} - x^n)^2}{x^{n+1} - 2x^n + x^{n-1}}.$$

Метод Эйткена позволяет ускорить сходимость метода простой итерации. Идея метода заключается в том, что после вычисления $x^{n-1},\ x^n,\ x^{n+1}$ производится пересчет по формуле

$$x'_{n+1} = x^{n+1} - \frac{(x^{n+1} - x^n)^2}{(x^{n+1} - 2x^n + x^{n-1})},$$

и значение x'_{n+1} берется в качестве нового приближения.

§3 Метод Ньютона и метод секущих

Рассмотрим функцию $f(x),\ x \in \mathbb{R}$ и уравнение

$$f(x) = 0. \tag{1}$$

Пусть x^* — корень этого уравнения, и определена его окрестность радиуса a, не содержащая других корней уравнения:

$$U_a(x^*) = \{x : |x - x^*| < a\},$$

причем заданная функция $f(x)$ определена на этой окрестности.

Будем считать, что начальное приближение $x^0 \in U_a(x^*)$ задано. Пусть в $U_a(x^*)$ существует и не обращается в ноль непрерывная первая производная функции $f(x)$:

$$f'(x) \neq 0, \quad x \in U_a(x^*).$$

Разложим $f(x^*)$ по формуле Тейлора в малой окрестности точки $x \in U_a(x^*)$:

$$f(x^*) = f(x) + (x^* - x)f'(x) + \dots$$

и отбросим в этом разложении величины, имеющие второй и выше порядок малости по $(x^* - x)$.
Заменив x^* на x^{n+1} и x на x^n, получим уравнение

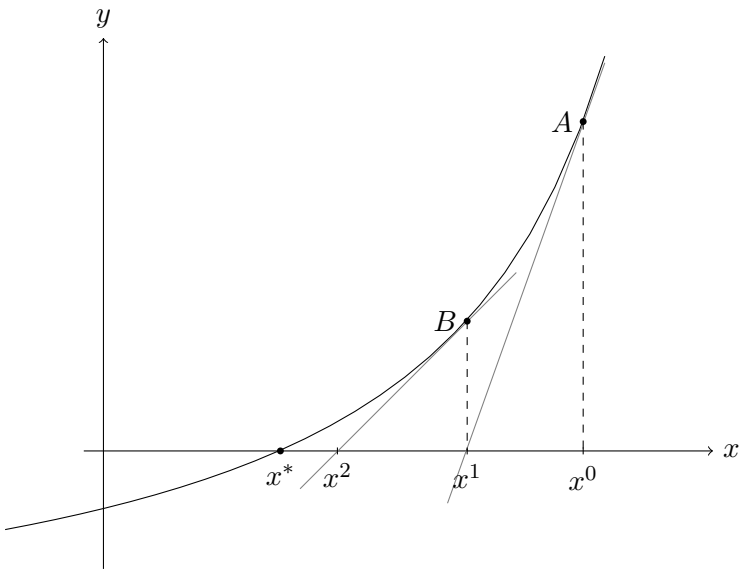
$$f(x^n) + (x^{n+1} - x^n)f'(x^n) = 0, \quad n \in \mathbb{Z}_+.$$

Учитывая, что $f'(x^n) \neq 0$, и разрешив последнее уравнение относительно x^{n+1}, имеем:

$$x^{n+1} = x^n - \frac{f(x^n)}{f'(x^n)}, \quad n \in \mathbb{Z}_+. \tag{2}$$

Определение. *Итерационный процесс поиска корня уравнения (1), задаваемый формулой (2), называется итерационным методом Ньютона.*

Дадим геометрическую интерпретацию метода Ньютона. Рассмотрим точку $A\,(x^0, f(x^0))$. Определим первую итерацию x^1 рассматриваемого процесса как абсциссу точки пересечения с осью Ox касательной к функции $f(x)$ в точке A. Аналогично получаем значение x^2 как точку пересечения с осью Ox касательной к функции $f(x)$ в точке $B\,(x^1, f(x^1))$. Продолжая таким образом, на n-ом шаге получаем значение x^n, приближающее корень x^* уравнения (1) с заданной точностью.



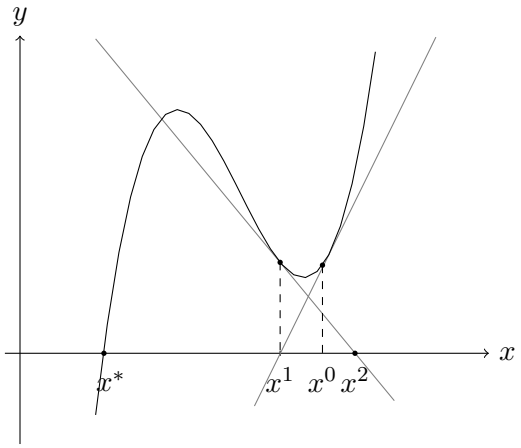
Выпишем уравнение касательной к функции $f(x)$ в точке x^n:

$$y - f(x^n) = f'(x^n)(x - x^n).$$

Очевидно, что значение x^{n+1}, найденное по формуле (2), представляет собой абсциссу точки пересечения с осью x касательной к кривой $y = f(x)$, проведенной через точку $(x^n, f(x^n))$.

Замечание. *Итерационный метод Ньютона часто называют методом касательных.*

Если не выполнено условие неравенства нулю производной функции $f(x)$ в области $U_a(x^*)$, то метод Ньютона может расходиться. На графике показан пример такого случая.



Замечание 1. *Метод Ньютона является вычислительно сложным, поскольку на каждой итерации проводится вычисление значений производной функции $f(x)$, что является, вообще говоря, неустойчивым процессом.*

Замечание 2. *При решении задач на практике часто рассматривается модифицированный метод Ньютона, задаваемый формулой*

$$x^{n+1} = x^n - \frac{f(x^n)}{f'(x^0)}, \quad n \in \mathbb{Z}_+.$$

Преимущество этого метода перед классическим методом заключается в том, что в нем не требуется вычислять значения функции $f'(x)$ на каждой итерации. Однако при этом модифицированный метод Ньютона сходится медленнее классического метода Ньютона.

Метод Ньютона для нелинейных систем уравнений

Рассмотрим систему двух нелинейных уравнений:

$$\begin{cases} f_1(x_1, x_2) = 0 \\ f_2(x_1, x_2) = 0 \end{cases}. \tag{3}$$

Пусть точка (x_1^*, x_2^*) — решение этой системы. Разложим значение функции $f_1(x_1^*, x_2^*)$ по формуле Тейлора в малой окрестности точки (x_1, x_2), лежащей в окрестности решения:

$$f_1(x_1^*, x_2^*) = f_1(x_1, x_2) + (x_1^* - x_1)\frac{\partial f_1(x_1, x_2)}{\partial x_1} + (x_2^* - x_2)\frac{\partial f_1(x_1, x_2)}{\partial x_2} + \dots$$

Заменим в этом разложении x_i на x_i^n, x_i^* на x_i^{n+1}, $i = 1, 2$ и учтем, что (x_1^*, x_2^*) — решение первого уравнения системы (3):

$$f_1(x_1^n, x_2^n) + (x_1^{n+1} - x_1^n)\frac{\partial f_1(x_1^n, x_2^n)}{\partial x_1^n} + (x_2^{n+1} - x_2^n)\frac{\partial f_1(x_1^n, x_2^n)}{\partial x_2^n} = 0. \tag{4}$$

Аналогичным образом разложив функцию $f_2(x_1^*, x_2^*)$ по формуле Тейлора и произведя такую же замену переменных, получим

$$f_2(x_1^n, x_2^n) + (x_1^{n+1} - x_1^n)\frac{\partial f_2(x_1^n, x_2^n)}{\partial x_1^n} + (x_2^{n+1} - x_2^n)\frac{\partial f_2(x_1^n, x_2^n)}{\partial x_2^n} = 0. \tag{5}$$

Введем векторы

$$f = (f_1, f_2)^T,\ x = (x_1, x_2)^T$$

и матрицу Якоби системы (3) — матрицу из частных производных функций $f_1(x)$ и $f_2(x)$:

$$J(x) = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1}(x) & \dfrac{\partial f_1}{\partial x_2}(x) \\ \dfrac{\partial f_2}{\partial x_1}(x) & \dfrac{\partial f_2}{\partial x_2}(x) \end{pmatrix}. \tag{6}$$

Перепишем уравнения (4) и (5) в матричном виде:

$$f(x^n) + J(x^n)(x^{n+1} - x^n) = \theta. \tag{7}$$

Пусть матрица Якоби невырождена. Выразим $(n+1)$-ую итерацию через n-ую:

$$x^{n+1} = x^n - J^{-1}(x^n)f(x^n),\quad n \in \mathbb{Z}_+. \tag{8}$$

Заметим, что нахождение матрицы J не является простой процедурой, так как нахождение производных является, вообще говоря, неустойчивым процессом.

Замечание. *При поиске значения каждой следующей итерации x^{n+1} необходимо сначала решить следующую систему:*

$$J(x^n)v^n = -f(x^n),\quad n \in \mathbb{Z}_+,$$

где $v^n = x^{n+1} - x^n$. Теперь значение x^{n+1} получается из найденного v^n: $x^{n+1} = x^n + v^n$.

Теперь перейдем к рассмотрению системы из $m > 2$ нелинейных уравнений

$$\begin{cases} f_1(x_1, x_2, \ldots, x_n) = 0 \\ f_2(x_1, x_2, \ldots, x_n) = 0 \\ \ldots \\ f_m(x_1, x_2, \ldots, x_n) = 0 \end{cases}. \tag{9}$$

Введем векторы

$$f = (f_1, f_2, \ldots, f_m)^T,\ x = (x_1, x_2, \ldots, x_m)^T$$

и матрицу Якоби системы (9):

$$J = (f_{ij}),\ f_{ij} = \frac{\partial f_i}{\partial x_j},\quad i, j = \overline{1, m}.$$

Запишем схему итерационного метода Ньютона, используя матрицу Якоби:

$$x^{n+1} = x^n - J^{-1}(x^n)f(x^n),\quad n \in \mathbb{Z}_+.$$

Заметим, что вычислять матрицу J на каждом шаге достаточно трудоемко.

Замечание. *Аналогично одномерному случаю можно рассматривать модифицированный метод Ньютона для решения нелинейных систем:*

$$x^{n+1} = x^n - J^{-1}(x^0)f(x^n),\quad n \in \mathbb{Z}_+.$$

Реализация модифицированного метода Ньютона проще классического варианта, но скорость сходимости при данном подходе меньше.

Метод секущих

Ранее мы рассматривали одношаговые методы решения нелинейных уравнений — метод простых итераций и итерационный метод Ньютона. Рассмотрим многошаговый итерационный метод — метод секущих.

Запишем итерационный метод Ньютона для решения уравнения (1):

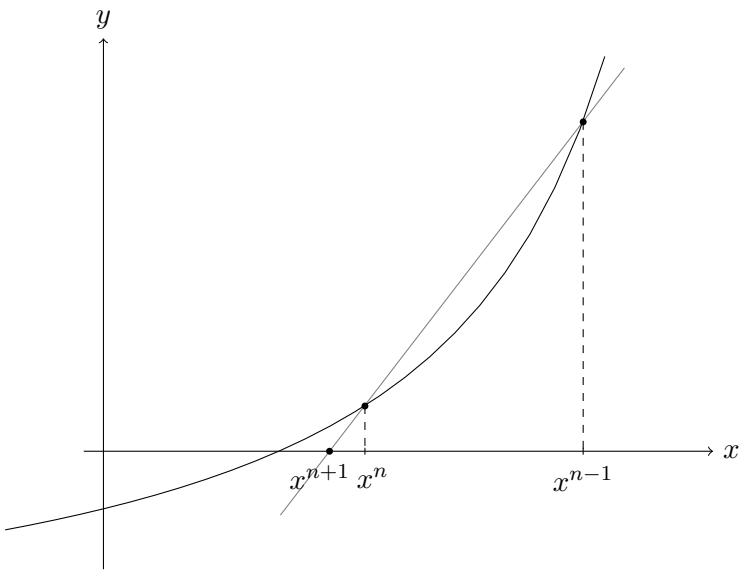
$$x^{n+1} = x^n - \frac{f(x^n)}{f'(x^n)}, \quad n \in \mathbb{Z}_+, \ x^0 \in U_a(x_0). \tag{10}$$

Заменим производную $f'(x^n)$ на соответствующий дискретный аналог $\frac{f(x^n)-f(x^{n-1})}{x^n-x^{n-1}}$ и подставим это отношение в уравнение (10).
Получим итерационный метод

$$x^{n+1} = x^n - \frac{(x^n - x^{n-1})f(x^n)}{f(x^n) - f(x^{n-1})}, \quad n \in \mathbb{N}, \quad x^0, x^1 \text{ заданы.} \tag{11}$$

Определение. *Итерационный процесс (11) задает двухшаговый метод решения нелинейных уравнений, называемый методом секущих.*

Рассмотрим геометрическую интерпретацию метода секущих.



Через точки $(x^{n-1}, f(x^{n-1}))$, $(x^n, f(x^n))$ проводится секущая. За новое значение x^{n+1} принимается абсцисса точки пересечения секущей и оси Ox. Иначе говоря, на отрезке $[x^{n-1}, x^n]$ функция $f(x)$ интерполируется полиномом первой степени, и за очередное приближение x^{n+1} принимается корень этого полинома.

§4 Сходимость метода Ньютона. Оценка скорости сходимости

Рассмотрим функцию $f(x)$, $x \in \mathbb{R}$ и уравнение

$$f(x) = 0. \tag{1}$$

Пусть x^* — корень этого уравнения, и определена его окрестность радиуса a, не содержащая других корней уравнения:

$$U_a(x^*) = \{x : |x - x^*| < a\},$$

причем заданная функция $f(x)$ определена на этой окрестности.

Будем считать, что начальное приближение $x^0 \in U_a(x^*)$ задано. Запишем формулу итерационного метода Ньютона решения уравнения (1):

$$x^{n+1} = x^n - \frac{f(x^n)}{f'(x^n)}, \quad n \in \mathbb{Z}_+, \ x^0 \in U_a(x^*).$$

Будем рассматривать итерационный метод Ньютона как метод простой итерации с функцией

$$S(x) = x - \frac{f(x)}{f'(x)}.$$

При изучении сходимости метода простой итерации было замечено, что, если $|S'(x)| < 1$ при $x \in U_a(x^*)$, то он сходится. Предполагая, что функция $f(x)$ дифференцируема достаточное количество раз, продифференцируем функцию $S(x)$:

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

Так как x^* — корень уравнения (1), то $f(x^*) = 0$, и, следовательно, $S'(x^*) = 0$, и по непрерывности функции $S'(x)$, $x \in U_a(x^*)$, $|S'(x)| < 1$ метод сходится.

Введем погрешность приближенного решения:

$$z^n = x^n - x^*.$$

Покажем, что связь между z^n и z^{n+1} квадратичная. Рассмотрим выражение для z^{n+1}:

$$z^{n+1} = x^{n+1} - x^* = S(z^n + x^*) - S(x^*). \tag{2}$$

Разложим $S(z^n + x^*)$ по формуле Тейлора и учтем, что $S'(x^*) = 0$:

$$z^{n+1} = S(x^*) + S'(x^*)z^n + \frac{1}{2}S''(\tilde{x}^n)\,(z^n)^2 - S(x^*) = \frac{1}{2}S''(\tilde{x}^n)(z^n)^2, \tag{3}$$

$$\tilde{x}^n = x^n + \theta z^n, \ \theta \in \mathbb{R}, \ |\theta| < 1.$$

Замечание. *Пусть функция $f(x)$ трижды непрерывно дифференцируема в окрестности $U_a(x^*)$. Тогда*

$$S''(x) = \left(\frac{f(x)f''(x)}{(f'(x))^2}\right)'.$$

Пусть существует постоянная $M > 0$ такая, что для любого $x \in U_a(x^*)$ выполняется неравенство

$$M \geqslant \frac{1}{2}\left|S''(x)\right|. \tag{4}$$

Из этого неравенства и уравнения (3) следует оценка

$$|z^{n+1}| \leqslant M|(z^n)^2|.$$

Домножим это неравенство на M и обозначим $v^n = M|z^n|$. Тогда получим, что

$$v^{n+1} \leqslant (v^n)^2. \tag{5}$$

Отсюда следует, что $v^n \leqslant (v^0)^{2^n}$, значит,

$$M|z^n| \leqslant \left(M\left|z^0\right|\right)^{2^n},$$

$$|z^n| \leqslant \frac{1}{M}\left(M\left|z^0\right|\right)^{2^n}.$$

Введем обозначение $q = M|z_0|$. Если $0 < q < 1$, то последовательность $\{z^n\}_{n=0}^{\infty}$ стремится к нулю:

$$z^n \underset{n\to\infty}{\longrightarrow} 0,$$

и итерационный метод Ньютона сходится. Условие на q $(0 < q < 1)$ будет выполнено, если $0 < |z^0| < \frac{1}{M}$, то есть $|x^0 - x^*| \leqslant \frac{1}{M}$.

Таким образом, мы доказали следующую теорему.

Теорема 1. *Пусть существует такая константа $M > 0$, для которой выполнена оценка*

$$\frac{1}{2}\left|S''(x)\right| \leqslant M, \quad x \in U_a(x^*).$$

Тогда если начальное приближение x^0 выбрать в соответствии с условием

$$|x^0 - x^*| \leqslant \frac{1}{M},$$

то итерационный метод Ньютона сходится, и имеет место оценка:

$$|x^n - x^*| < \frac{1}{M}\left(M|x^0 - x^*|\right)^{2^n}.$$

Замечание 1. *Если итерационный метод Ньютона сходится, то достаточно быстро.*

Замечание 2. *Из условий теоремы следует, что начальное приближение нужно выбирать достаточно близко к точному решению рассматриваемого уравнения.*

Замечание 3. *Другие рассмотренные нами методы (модифицированный метод Ньютона и метод секущих) обладают, по крайней мере, линейной сходимостью. Это следует из того, что если их записать в виде $x^{n+1} = S(x^n)$, то $S(x^*) = x^*$ и $S'(x^*) \neq 0$. Например, для модифицированного метода Ньютона $S'(x^*) = 1 - \frac{f'(x^*)}{f'(x^0)}$, и чем ближе взять x^0 к x^*, тем быстрее будет сходимость.*

Глава 4

Разностные методы решения задач математической физики

§1 Введение

Эта глава посвящена решению задач математической физики с помощью численных методов. Численные методы позволяют находить решение произвольной дифференциальной задачи, в то время как аналитические подходы разработаны лишь для некоторых классов задач и, как правило, используют целый ряд допущений. К примеру, мы будем рассматривать уравнение теплопроводности, которое является аналитически неразрешимым, если область задания уравнения определена произвольным образом, или уравнение содержит переменные коэффициенты. Разностные схемы позволят нам находить решение уравнения теплопроводности и в таких сложных случаях.

Постановка задачи. *Рассмотрим классическую формулировку первой краевой задачи для уравнения теплопроводности в области* $G = \{(x,t) \ : \ x \in (0,1), \ t \in (0,T]\}$ *для некоторого* $T > 0$. *Для простоты возьмем коэффициент при производной искомой функции в правой части уравнения равным единице.*

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2} + f(x,t), \quad (x,t) \in G. \tag{1}$$

Выпишем краевые условия первого рода:

$$\begin{cases} u(0,t) = \mu_1(t) \\ u(1,t) = \mu_2(t), \end{cases} \quad t \in [0,T], \tag{2}$$

и начальное условие:

$$u(x,0) = u_0(x), \quad x \in [0,1]. \tag{3}$$

Заметим, что мы рассматриваем только те задачи, для которых существует классическое решение, то есть решение задачи существует, единственно и удовлетворяет условиям:

1. *Решение обладает достаточной гладкостью, то есть функция* $u(x,t)$ *непрерывна в замкнутой области* $\overline{G} = \{(x,t) \ : \ x \in [0,1], \ t \in [0,T]\}$, *непрерывно дифференцируема один раз по* t *и два раза по* x *внутри области* G.

2. $u(x,t)$ *удовлетворяет внутри области* G *уравнению* (1), *на границе — условию* (2) *и условию* (3) *в начальный момент времени.*

Кроме того, условия на границе (2) *и в начальный момент времени должны быть согласованы:* $\mu_1(0) = u_0(0)$ *и* $\mu_2(0) = u_0(1)$.

Из курса «Уравнения математической физики» известно, что в такой постановке существует единственное решение $u(x,t)$, которое непрерывно зависит от правой части уравнения $f(x,t)$, начального условия $u_0(x)$ и краевых условий (2).

Чтобы решить эту задачу численно, поставим ей в соответствие разностную схему, то есть дискретный аналог рассматриваемого уравнения и дополнительных условий. Таким образом мы сведем непрерывную задачу к конечной системе линейных уравнений, которые уже можно решать с использованием вычислительных машин.

Сначала введем в рассматриваемой области G равномерную по переменным x и t сетку.

Определение. *Сеткой в заданной области называется совокупность конечного числа точек, принадлежащих данной области. Эти точки называются узлами сетки.*

В частности, равномерная сетка размера $(N-1) \times M$, $N, M \in \mathbb{N}$ в рассматриваемой области G вводится так:

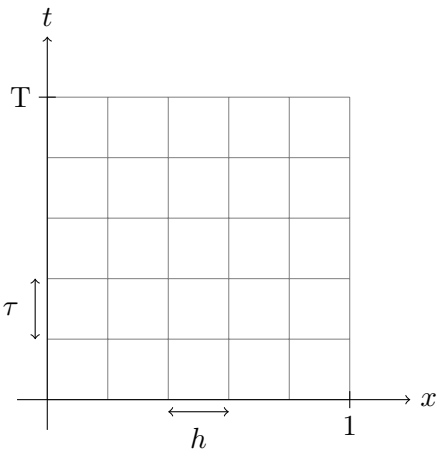
$$\omega_h = \left\{ x_i = ih,\ i = \overline{1,(N-1)} \right\}, \quad \omega_\tau = \left\{ t_j = j\tau,\ j = \overline{1,M} \right\},$$

$$h = \frac{1}{N} > 0,\ \tau = \frac{T}{M} > 0.$$

Величину h назовем шагом по переменной x, величину τ — шагом по времени.
Тогда множество точек

$$\omega_{\tau h} = \omega_\tau \times \omega_h \subset G$$

задает равномерную сетку с шагом h по переменной x и шагом τ по времени в области G. Эта сетка изображена на рисунке.



Аналогичным образом введем равномерную сетку размера $(N+1) \times (M+1)$ на замыкании области G с теми же размерами шагов h и τ по переменной x и по времени соответственно. Эту сетку задает множество точек

$$\overline{\omega}_{\tau h} = \overline{\omega}_\tau \times \overline{\omega}_h \subset \overline{G} = \{(x,t)\ :\ x \in [0,1],\ t \in [0,T]\},$$

где

$$\overline{\omega}_h = \left\{ x_i = ih,\ i = \overline{0,N} \right\}, \quad \overline{\omega}_\tau = \left\{ t_j = j\tau,\ j = \overline{0,M} \right\}.$$

В дальнейшем везде, где мы рассматриваем уравнение теплопроводности, будем использовать введенные сетки, если не указано иное.

Замечание. *В общем случае сетки могут иметь более сложную структуру, например, использовать переменный шаг, который зависит от расположения конкретной пары узлов, или для многомерной области иметь более сложную структуру расположения узлов относительно друг друга (в рассматриваемом примере равномерная сетка является прямоугольной). В последнее время часто используются сетки, автоматически подстраивающиеся под решение конкретной задачи.*

Определение. *Совокупность всех узлов в фиксированный момент времени t_n называется слоем. Слой, для которого $t_n = 0$, будем называть нулевым слоем, в котором задано начальное приближение.*

§2 Явная разностная схема. Погрешность, сходимость, устойчивость

Рассмотрим уравнение теплопроводности с краевыми условиями первого рода:

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2} + f(x,t), \quad (x,t) \in G = \{(x,t) \ : \ x \in (0,1), \ t \in (0,T]\}, \tag{1}$$

$$\begin{cases} u(0,t) = \mu_1(t) \\ u(1,t) = \mu_2(t), \end{cases} \quad t \in [0,T], \tag{2}$$

$$u(x,0) = u_0(x), \quad x \in [0,1] \tag{3}$$

и построим для него разностную схему.

Воспользуемся сетками $\omega_{\tau h}$ и $\overline{\omega}_{\tau h}$, введенными в первом параграфе данной главы на множествах G и $\overline{G}$ соответственно.

Определение. *Сеточной функцией называется функция дискретного аргумента на заданной сетке, то есть такая функция определена только в узлах данной сетки.*

Поставим в соответствие непрерывным функциям $u(x,t)$ и $f(x,t)$ их дискретные аналоги. Введем обозначения для $(x_i, t_n) \in \omega_{\tau h}$:

$$f_i^n = f(x_i, t_n),$$

$$u_i^n = u(x_i, t_n).$$

Обозначим численное решение задачи через

$$y(x_i, t_n) = y_i^n, \quad (x_i, t_n) \in \overline{\omega}_{\tau h}.$$

$y(x_i, t_n)$ является сеточной функцией, заданной на сетке $\overline{\omega_{\tau h}}$.

Поставим в соответствие производным функции $u(x,t)$ их дискретные аналоги для функции $y(x_i, t_n)$:

$$\frac{\partial u(x_i,t_n)}{\partial t} \approx \frac{y_i^{n+1} - y_i^n}{\tau},$$

$$\frac{\partial^2 u(x_i,t_n)}{\partial x^2} \approx \frac{y_{i-1}^n - 2y_i^n + y_{i+1}^n}{h^2}.$$

Получаем дискретный аналог уравнения (1):

$$\frac{y_i^{n+1} - y_i^n}{\tau} = \frac{y_{i-1}^n - 2y_i^n + y_{i+1}^n}{h^2} + f(x_i, t_n), \quad (x_i, t_n) \in \omega_{\tau h}. \tag{4}$$

Запишем дискретные аналоги краевых условий первого рода (2) и начального условия (3):

$$\begin{cases} y_0^{n+1} = \mu_1(t_{n+1}) \\ y_N^{n+1} = \mu_2(t_{n+1}), \end{cases} \quad t_{n+1} \in \overline{\omega}_\tau, \tag{5}$$

$$y_i^0 = u_0(x_i), \quad x_i \in \overline{\omega}_h. \tag{6}$$

Определение. *Дискретным аналогом задачи* (1) – (3)*, или ее разностной схемой, называется система линейных уравнений* (4) – (6)*.*

Замечание 1. *В первой краевой задаче численные значения решения y_0^{n+1} и y_N^{n+1} равны значениям функций $\mu_1(t)$ и $\mu_2(t)$ соответственно при $t = t_{n+1}$ (хотя это и не обязательно). В случае краевых условий иного типа, аппроксимация краевых условий должна быть согласована по порядку погрешности с порядком аппроксимации уравнения.*

Замечание 2. *Заметим, что в уравнении* (4) *значения функции $f(x,t)$ необязательно брать именно в узлах рассматриваемой сетки, можно использовать значения этой функции с некоторой «поправкой». Что именно имеется в виду под «поправкой», будет рассмотрено далее, а так же будет показано, что выбор значений функции $f(x,t)$ для разностной схемы, использующих такую «поправку», позволит получить более высокий порядок погрешности аппроксимации, а стало быть и более точное решение исходного уравнения.*

Замечание 3. *Качество и скорость решения численной задачи* (4) – (6) *во многом зависит от выбора числа узлов сетки $\omega_{\tau h}$: чем меньше узлов в сетке, тем меньше уравнений содержится в системе, тем проще и быстрее ее решать, но и приближение решения исходной задачи в этом случае будет более грубым и наоборот.*

При изучении разностных схем возникают следующие вопросы:

1. Погрешность аппроксимации на решении (невязка).

   Каждой задаче может быть сопоставлено бесконечное число разностных схем, оценка погрешности аппроксимации позволяет их сравнивать. Разностная схема должна аппроксимировать исходную дифференциальную задачу. Если же аппроксимация отсутствует, то не будет сходимости решения численной задачи к решению исходной задачи, и рассмотрение такой разностной схемы не имеет смысла.

2. Существование и единственность решения разностной задачи.

   Построенная разностная задача должна быть корректной, то есть должно существовать единственное решение. В ряде случаев доказательство существования и единственности решения является нетривиальной задачей.

3. Алгоритм нахождения разностного решения.

   В разностных схемах матрица системы линейных уравнений как правило содержит большое количество нулей. Для таких систем существуют более эффективные алгоритмы решения, чем универсальный метод Гаусса, например для систем с трехдиагональной матрицей разумно использовать метод прогонки.

4. Сходимость разностной схемы.

   Необходимо изучить условия, при которых решение данной разностной схемы сходится к точному решению исходной задачи с любой наперед заданной точностью.

Напомним, что сходимость рассматривается для каждой нормы, введенной на пространстве сеточных функций, независимо (то есть из сходимости в некоторой норме конечномерного пространства, вообще говоря, не следует сходимость в другой норме этого пространства).

5. Устойчивость разностной схемы.

   Устойчивость в данном контексте является чисто внутренним свойством разностных схем: разностная схема называется устойчивой в норме $\|\cdot\|$, если выполнена априорная оценка

   $$\|y\| \leqslant M\|f\|,$$

   где $M > 0$ — константа, не зависящая от шагов сетки.

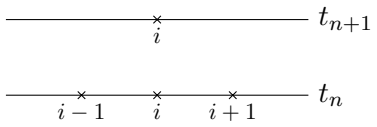
Для построения разностной схемы, обладающей хорошими свойствами, необходимо изучить все пять вопросов.

Замечание. *Вопросы сходимости и устойчивости разностной схемы являются ключевыми, однако обычно достаточно рассмотреть только один из этих двух вопросов: в конце курса будет доказано, что из устойчивости разностной схемы следует ее сходимость к решению исходной задачи при условии, что разностная схема аппроксимирует исходную задачу.*

Определение. *Совокупность узлов, которые участвуют в записи разностной схемы, называют шаблоном.*

Вернемся к изучению явной разностной схемы.

В рассматриваемой разностной схеме использован четырехточечный шаблон, схематично изображенный на рисунке.



Для построенной разностной схемы решение на $(n+1)$-ом слое находится явно, поэтому и рассматриваемая разностная схема называется явной:

$$y_i^{n+1} = y_i^n + \frac{\tau}{h^2}(y_{i-1}^n - 2y_i^n + y_{i+1}^n) + \tau f_i^n, \quad i = \overline{1,(N-1)},$$

$$\begin{cases} y_0^{n+1} = \mu_1(t_{n+1}) \\ y_N^{n+1} = \mu_2(t_{n+1}), \end{cases} \quad t_{n+1} \in \overline{\omega}_\tau,$$

$$y_i^0 = u_0(x_i), \quad i = \overline{0,N}.$$

Выведенные явные формулы нахождения решения позволяют утверждать, что решение разностной схемы (4) – (6) существует и единственно, значит, мы получили ответ на вопрос (2).

Перейдем к исследованию оставшихся вопросов. Как мы уже упоминали в главе «Интерполирование и приближение функций», существует два подхода к измерению близости точного решения задачи (1) – (3) (непрерывной функции) и численного решения задачи (4) – (6) (сеточной функции):

1. Спроектировать непрерывную функцию $u(x,t)$ на дискретное пространство и измерять близость функций $u(x,t)$ и y_i^n в норме дискретного пространства.

2. С помощью интерполирования восполнить функцию y_i^n до непрерывной и сравнивать рассматриваемые функции в пространстве непрерывных функций.

В этом курсе мы будем пользоваться первым подходом.

Определение. *Сеточная функция вида*

$$z_i^n = z(x_i, t_n) = y_i^n - u_i^n, \quad (x_i, t_n) \in \overline{\omega}_{\tau h} \tag{7}$$

называется погрешностью решения разностной схемы (4) – (6).

Выразим $y_i^n = z_i^n + u_i^n$ и подставим это выражение в разностную схему. Получим систему уравнений для z_i^n, аналогичную разностной схеме, но с нулевыми краевыми условиями и нулевой начальной функцией:

$$\frac{z_i^{n+1} - z_i^n}{\tau} = \frac{z_{i-1}^n - 2z_i^n + z_{i+1}^n}{h^2} + \psi_i^n, \quad (x_i, t_n) \in \omega_{\tau h}, \tag{8}$$

$$z_0^{n+1} = z_N^{n+1} = 0, \quad t_{n+1} \in \overline{\omega}_\tau, \tag{9}$$

$$z_i^0 = 0, \quad x_i \in \overline{\omega}_h. \tag{10}$$

Определение. *Сеточная функция, задаваемая равенством*

$$\psi_i^n = \frac{u_{i-1}^n - 2u_i^n + u_{i+1}^n}{h^2} - \frac{u_i^{n+1} - u_i^n}{\tau} + f_i^n, \tag{11}$$

называется погрешностью аппроксимации разностной схемы (4) – (6) *на решении исходной задачи.*

Задача. *Доказать, что* $\psi_i^n = \mathrm{O}\big(\tau + h^2\big)$.

Решение. Здесь и далее $(x_i, t_n) \in \overline{\omega}_{\tau h},\ i = \overline{0, N},\ n = \overline{0, M}$.

Разложим $u(x_i, t_{n+1})$ в узле (x_i, t_n) по формуле Тейлора:

$$u(x_i, t_{n+1}) = u_i^{n+1} = u(x_i, t_n) + u_t'(x_i, t_n)\tau + O(\tau^2).$$

Разложим $u(x_{i+1}, t_n)$ в узле (x_i, t_n) по формуле Тейлора:

$$u(x_{i+1}, t_n) = u_{i+1}^n = u(x_i, t_n) + u_x'(x_i, t_n)h + \frac{1}{2}u_{xx}''(x_i, t_n)h^2 + \frac{1}{6}u_{xxx}'''(x_i, t_n)h^3 + O(h^4).$$

Разложим $u(x_{i-1}, t_n)$ в узле (x_i, t_n) по формуле Тейлора (далее всюду при использовании формулы Тейлора мы будем предполагать, что разлагаемая функция обладает нужной гладкостью, то есть имеет непрерывные производные до соответствующего по ходу разложения порядка):

$$u(x_{i-1}, t_n) = u_{i-1}^n = u(x_i, t_n) - u_x'(x_i, t_n)h + \frac{1}{2}u_{xx}''(x_i, t_n)h^2 - \frac{1}{6}u_{xxx}'''(x_i, t_n)h^3 + O(h^4).$$

Полученные разложения подставим в формулу (11) и после приведения подобных слагаемых получим оценку $\psi_i^n = \mathrm{O}\big(\tau + h^2\big)$. □

Введем норму в пространстве сеточных функцих на n-ом слое, $n = \overline{0, M}$:

$$\|z^n\|_C = \max_{0 \leqslant i \leqslant N} |z_i^n|.$$

Мы рассматриваем решение разностной задачи по слоям, поэтому нет необходимости вводить норму как максимум модуля для всех слоев.

Теорема. *Пусть функция $u(x,t)$ обладает достаточной гладкостью (четыре раза дифференцируема по x и два раза по t). Тогда для сходимости решения разностной схемы (4) – (6) к решению исходной задачи (1) – (3) в норме $\|\cdot\|_C$ необходимо и достаточно, чтобы выполнялось условие:*

$$\gamma = \frac{\tau}{h^2} \leqslant 0.5.$$

Кроме того, выполняется оценка:

$$\left\|z^{n+1} - u^{n+1}\right\|_C \leqslant M_1\left(\tau + h^2\right),$$

где $M_1 > 0$ — константа, не зависящая от τ и h.

Доказательство. Докажем, что выполнения условия теоремы достаточно для сходимости разностной схемы к решению исходной задачи.

Запишем выражение для z_i^{n+1} в виде

$$z_i^{n+1} = (1-2\gamma)\, z_i^n + \gamma\left(z_{i-1}^n + z_{i+1}^n\right) + \tau\psi_i^n$$

и ограничим левую часть равенства по модулю с учетом того, что $1 - 2\gamma > 0$, так как выполнено условие теоремы

$$\left|z_i^{n+1}\right| \leqslant (1-2\gamma)\left|z_i^n\right| + \gamma\left(\left|z_{i-1}^n\right| + \left|z_{i+1}^n\right|\right) + \tau\left|\psi_i^n\right|.$$

Перейдем в правой части неравенства от модулей слагаемых к нормам соответствующих векторов. При таком переходе правая часть неравенства может только увеличиться:

$$\left|z_i^{n+1}\right| \leqslant (1-2\gamma)\left\|z^n\right\|_C + 2\gamma\|z^n\|_C + \tau\|\psi^n\|_C.$$

Полученное неравенство верно для всех $i = \overline{0,N}$, а значит, оно выполнено и для максимального из $\left|z_i^{n+1}\right|$. Следовательно, можно заменить левую часть неравенства на норму $\left\|z^{n+1}\right\|_C$, и, с учетом приведения подобных слагаемых, получить

$$\|z^{n+1}\|_C \leqslant \|z^n\|_C + \tau\|\psi^n\|_C.$$

Получили рекуррентную оценку для нормы $\left\|z^{n+1}\right\|_C$. Раскроем ее:

$$\left\|z^{n+1}\right\|_C \leqslant \left\|z^0\right\|_C + \tau\sum_{k=0}^{n}\left\|\psi^k\right\|_C. \tag{12}$$

Так как

$$\left\|\psi^k\right\|_C \leqslant M\left(\tau + h^2\right),$$

где $M > 0$ — константа, не зависящая от τ и h,

$$\left\|z^0\right\|_C = 0,$$

$$\sum_{k=0}^{n}\tau = t_n \leqslant T,$$

то получим окончательную оценку:

$$\left\|z^{n+1}\right\|_C \leqslant M_1\left(\tau + h^2\right), \quad M_1 = TM.$$

Из данной оценки следует, что при $\tau \to 0,\ h \to 0$

$$\left\|z^{n+1}\right\|_C = \left\|y^{n+1} - u^{n+1}\right\|_C \to 0.$$

Следовательно, решение разностной схемы сходится к решению исходной задачи.

Перед тем, как доказать необходимость, докажем устойчивость разностной схемы. Рассмотрим разностную схему (4) – (6) с нулевыми краевыми условиями, получим задачу, совпадающую с рассмотренной задачей (8) – (10). После проведения оценок, аналогичных показанным выше, получим

$$\left\|y^{n+1}\right\|_C \leqslant \left\|u_0\right\|_C + \sum_{k=0}^{n} \tau \left\|f^k\right\|_C.$$

Эта априорная оценка означает устойчивость решения разностной схемы по начальным условиям и правой части уравнения. Окончательная оценка имеет вид

$$\left\|y^{n+1}\right\|_C \leqslant \left\|u_0\right\|_C + M_1 \left\|f^n\right\|_C,$$

где константа M_1 не зависит от τ и h.

Перейдем к доказательству необходимости выполнения условия теоремы для сходимости разностной схемы. Рассмотрим однородное уравнение относительно y_i^n:

$$\frac{y_i^{n+1} - y_i^n}{\tau} = \frac{y_{i-1}^n - 2y_i^n + y_{n+1}^n}{h^2},$$

где $n = \overline{0, (M-1)},\ i = \overline{1, (N-1)}$.

Покажем, что при нарушении условия теоремы появятся неограниченные возрастающие гармоники — функции вида

$$y_j^n = q^n e^{ijh\varphi}, \quad \text{где } i^2 = -1,\ \varphi,\ h \in \mathbb{R},\ q \in \mathbb{C}. \tag{13}$$

Предположим, что $\gamma > 0.5$. Подставим выражение (13) в рассматриваемое относительно y_i^n однородное уравнение и выразим q:

$$q = 1 + \gamma\left(e^{ih\varphi} - 2 + e^{-ih\varphi}\right) = 1 + 2\gamma\left(\cos h\varphi - 1\right) = 1 - 4\gamma \sin^2 \frac{h\varphi}{2}.$$

Так как, по предположению, $\gamma > 0.5$, то

$$1 - 4\gamma \sin^2 \frac{h\varphi}{2} < -1,$$

и $|q| > 1$. Тогда y_i^n неограниченно возрастает при $n \to \infty$, и о сходимости говорить не приходится.

Следовательно, если условие теоремы нарушено, то решение разностной схемы не будет сходиться к решению исходной задачи. □

Замечание 4. *Разностные схемы могут сходиться условно (и быть условно устойчивыми) и абсолютно. Условная сходимость определяется наличием ограничений на шаги сетки любого характера, для абсолютной сходимости требуется, чтобы какие-либо ограничения отсутствовали.*

Замечание 5. *Важно помнить, что сходимость и устойчивость разностной схемы рассматриваются в каждой норме отдельно. В данном параграфе доказана сходимость и устойчивость решений разностной схемы (4)–(6) по норме $\|\cdot\|_C$, которая является достаточно сильной нормой, а значит, обеспечивает более точную оценку, по сравнению, например, со среднеквадратичной нормой.*

§3 Чисто неявная разностная схема (схема с опережением). Погрешность, устойчивость, сходимость

Рассмотрим уравнение теплопроводности с краевыми условиями первого рода:

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2} + f(x,t), \quad (x,t) \in G = \{(x,t) \ : \ x \in (0,1), \ t \in (0,T]\}, \tag{1}$$

$$\begin{cases} u(0,t) = \mu_1(t) \\ u(1,t) = \mu_2(t), \end{cases} \quad t \in [0,T], \tag{2}$$

$$u(x,0) = u_0(x), \quad x \in [0,1]. \tag{3}$$

Воспользуемся сетками $\omega_{\tau h}$ и $\overline{\omega}_{\tau h}$, введенными в первом параграфе данной главы, на множествах G и $\overline{G}$ соответственно.

Поставим в соответствие задаче (1) – (3) следующую разностную схему:

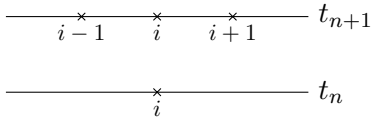
$$\frac{y_i^{n+1} - y_i^n}{\tau} = \frac{y_{i-1}^{n+1} - 2y_i^n + y_{i+1}^{n+1}}{h^2} + f(x_i, t_{n+1}), \quad (x_i,t_n), (x_i,t_{n+1}) \in \omega_{\tau h}, \tag{4}$$

$$\begin{cases} y_0^{n+1} = \mu_1(t_{n+1}) \\ y_N^{n+1} = \mu_2(t_{n+1}), \end{cases} \quad t_{n+1} \in \overline{\omega}_\tau, \tag{5}$$

$$y_i^0 = u_0(x_i), \quad x_i \in \overline{\omega}_h, \tag{6}$$

где $y_i^n = y(x_i,t_n)$ — искомое численное решение в точке $(x_i,t_n) \in \overline{\omega}_{\tau h}$.

В рассматриваемой разностной схеме использован четырехточечный шаблон вида



Как мы видим, разностная схема является неявной, а это значит, что для получения решения на $(n+1)$-ом слое необходимо решить трехточечное уравнение. Таким образом найти решение на $(n+1)$-ом слое «в лоб» не получится. В связи с этим возникает вопрос о разрешимости разностной задачи. Покажем, что эта задача имеет единственное решение, и укажем алгоритм его нахождения. Выразим y_i^{n+1} из уравнения (4):

$$y_i^{n+1} = y_i^n + \gamma \left(y_{i-1}^{n+1} - 2y_i^{n+1} + y_{i+1}^{n+1} \right) + \tau f_i^{n+1},$$

где $\gamma = \frac{\tau}{h^2}$, $(x_i,t_n), (x_i,t_{n+1}) \in \omega_{\tau h}$.

Перенесем слагаемые, относящиеся к $(n+1)$-ому слою, в левую часть уравнения и получим следующую систему уравнений относительно неизвестных $\{y_i^{n+1}\}_{i=1}^{N-1}$:

$$\begin{cases} -\gamma y_{i-1}^{n+1} + (1+2\gamma)\, y_i^{n+1} - \gamma y_{i+1}^{n+1} = y_i^n + \tau f_i^{n+1}, \quad i = \overline{1,(N-1)}, \\ y_0^{n+1} = \mu_1^{n+1}, \quad y_N^{n+1} = \mu_2^{n+1}. \end{cases}$$

Эта система имеет трехдиагональную матрицу порядка $(N-1)$:

$$A = \begin{pmatrix} 1+2\gamma & -\gamma & 0 & \dots & 0 & 0 \\ -\gamma & 1+2\gamma & -\gamma & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1+2\gamma & -\gamma \\ 0 & 0 & 0 & \dots & -\gamma & 1+2\gamma \end{pmatrix},$$

обладающую строгим диагональным преобладанием:

$$a_{ii} > \sum_{j=1}^{N} |a_{ij}|, \quad i = \overline{1,(N-1)}.$$

Матрицы со строгим диагональным преобладанием обладают свойством невырожденности, поэтому $|A| \neq 0$, и решение задачи (4) – (6) существует и единственно. Так как матрица A — трехдиагональная, разумно использовать метод прогонки для нахождения решения системы. Этот метод является разновидностью метода Гаусса, адаптированной для матриц специального вида, и, в отличие от классического метода Гаусса, имеет сложность $\mathrm{O}(N)$. Кроме того, так как рассматриваемая матрица обладает строгим диагональным преобладанием, метод прогонки будет устойчивым, а значит, ошибки округления нарастать не будут.

Введем сеточную функцию погрешности решения разностной схемы, равную разности приближенного и точного решений:

$$z_i^n = z(x_i, t_n) = y_i^n - u_i^n,$$

где $u_i^n = u(x_i, t_n), (x_i, t_n) \in \overline{\omega}_{\tau h}$.
Выразив из последнего соотношения y_i^n и подставив это выражение в разностную схему, с учетом линейности уравнения (4) получим уравнение для z_i^n с нулевыми краевыми и начальным условиями:

$$\frac{z_i^{n+1} - z_i^n}{\tau} = \frac{z_{i-1}^{n+1} - 2z_i^{n+1} + z_{i+1}^{n+1}}{h^2} + \psi_i^n, \quad (x_i, t_n), (x_i, t_{n+1}) \in \omega_{\tau h}, \tag{7}$$

$$\begin{cases} z_0^{n+1} = 0 \\ z_N^{n+1} = 0, \end{cases} \quad t_{n+1} \in \overline{\omega}_\tau, \tag{8}$$

$$z_i^0 = 0, \quad x_i \in \overline{\omega}_h, \tag{9}$$

где ψ_i^n — погрешность аппроксимации на решении y_i^n:

$$\psi_i^n = \psi(x_i, t_n) = -\frac{u_i^{n+1} - u_i^n}{\tau} + \frac{u_{i-1}^{n+1} - 2u_i^{n+1} + u_{i+1}^{n+1}}{h^2} + f(x_i, t_{n+1}), \tag{10}$$

где $(x_i, t_n), (x_i, t_{n+1}) \in \omega_{\tau h}$.

Задача. *Доказать, что*

$$\psi_i^n = \mathrm{O}\big(\tau + h^2\big). \tag{11}$$

Для оценки погрешности z_i^n воспользуемся нормой $\|\cdot\|_C$ в пространстве сеточных функций на слое, которую мы ввели в предыдущем параграфе.

Теорема. *Пусть функция $u(x,t)$ имеет достаточную гладкость (четыре раза дифференцируема по x и два раза по t). Тогда чисто неявная разностная схема сходится к решению исходной задачи в норме $\|\cdot\|_C$ с первым порядком точности по τ и вторым порядком точности по h.*

Доказательство. Пусть $x_{i_0} \in \overline{\omega}_h$ — узел, на котором достигается максимум погрешности на $(n+1)$-ом слое:

$$\big|z_{i_0}^{n+1}\big| = \max_{0 \leqslant i \leqslant N} \big|z_i^{n+1}\big| = \big\|z^{n+1}\big\|_C.$$

Заметим, что такой узел всегда существует, так как в противном случае $z^{n+1} = \theta$, и дальнейшие рассуждения не имеют смысла.

Для доказательства теоремы воспользуемся принципом максимума. Запишем уравнение (7) относительно узла x_{i_0}:

$$(1+2\gamma)\, z_{i_0}^{n+1} = z_{i_0}^{n} + \gamma\left(z_{i_0-1}^{n+1} + z_{i_0+1}^{n+1}\right) + \tau\psi_{i_0}^{n}, \quad \gamma = \frac{\tau}{h^2} > 0.$$

Оценим левую часть равенства по модулю с учетом того, что $(1+2\gamma) > 0$:

$$(1+2\gamma)\left|z_{i_0}^{n+1}\right| \leqslant \left|z_{i_0}^{n}\right| + \gamma\left(\left|z_{i_0-1}^{n+1}\right| + \left|z_{i_0+1}^{n+1}\right|\right) + \tau\left|\psi_{i_0}^{n}\right|.$$

Перейдем в правой части неравенства от модулей слагаемых к нормам соответствующих функций. При таком переходе правая часть неравенства может только увеличиться:

$$(1+2\gamma)\left|z_{i_0}^{n+1}\right| \leqslant \left\|z^{n}\right\|_C + 2\gamma\left\|z^{n+1}\right\|_C + \tau\left\|\psi^{n}\right\|_C.$$

Так как по предположению $\left|z_{i_0}^{n+1}\right| = \left\|z^{n+1}\right\|_C$, то полученное неравенство имеет вид

$$(1+2\gamma)\left\|z^{n+1}\right\|_C \leqslant \left\|z^{n}\right\|_C + 2\gamma\left\|z^{n+1}\right\|_C + \tau\left\|\psi^{n}\right\|_C.$$

Отсюда следует, что

$$\left\|z^{n+1}\right\|_C \leqslant \left\|z^{n}\right\|_C + \tau\left\|\psi^{n}\right\|_C.$$

Раскроем рекуррентное соотношение:

$$\left\|z^{n+1}\right\|_C \leqslant \left\|z^{0}\right\|_C + \sum_{k=0}^{n}\tau\left\|\psi^{k}\right\|_C.$$

$\left\|z^{0}\right\|_C = 0$, так как начальная погрешность равна нулю, значит

$$\left\|z^{n+1}\right\|_C \leqslant \sum_{k=0}^{n}\tau\left\|\psi^{k}\right\|_C.$$

Из (11) следует, что

$$\left\|\psi^{k}\right\|_C \leqslant M\left(\tau + h^2\right),$$

где $M > 0$ — константа, не зависящая от τ и h, и

$$\sum_{k=0}^{n}\tau = t_n \leqslant T.$$

Таким образом получим окончательную оценку:

$$\left\|z^{n+1}\right\|_C \leqslant M_1\left(\tau + h^2\right),$$

где $M_1 = TM > 0$ — константа, не зависящая от τ и h. Устремив τ и h к нулю, получим:

$$\lim_{\substack{\tau\to 0\\ h\to 0}}\left\|y^{n+1} - u^{n+1}\right\|_C = 0.$$

Равенство предела разности нулю означает, что решение разностной схемы сходится к решению исходной задачи с первым порядком точности по τ и вторым порядком точности по h. □

Замечание. *Если в разностной задаче (4) – (6) взять нулевые краевые условия*

$$y_0^{n+1} = y_N^{n+1} = 0,$$

то для y_i^n можно вывести оценку, аналогичную полученной выше:

$$\left\|y^{n+1}\right\|_C \leqslant \left\|u^{0}\right\|_C + \tau\sum_{k=0}^{N}\left\|f^{k}\right\|_C.$$

Эта оценка означает, что решение разностной схемы устойчиво по начальному условию и по правой части уравнения.

§4 Симметричная разностная схема (схема Кранка-Никольсона)

Рассмотрим уравнение теплопроводности с краевыми условиями первого рода:

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2} + f(x,t), \quad (x,t) \in G = \{(x,t) \ : \ x \in (0,1), \ t \in (0,T]\}, \tag{1}$$

$$\begin{cases} u(0,t) = \mu_1(t) \\ u(1,t) = \mu_2(t), \end{cases} \quad t \in [0,T], \tag{2}$$

$$u(x,0) = u_0(x), \quad x \in [0,1]. \tag{3}$$

Воспользуемся сетками $\omega_{\tau h}$ и $\overline{\omega}_{\tau h}$, введенными в первом параграфе данной главы на множествах G и $\overline{G}$ соответственно.

Введем вторую разностную производную для дискретной функции $y_i^n = y(x_i, t_n)$, определенной на множестве $\overline{\omega}_{\tau h}$:

$$y^n_{\overline{x}x,i} = \frac{y^n_{i-1} - 2y^n_i + y^n_{i+1}}{h^2}.$$

Эта производная является дискретным аналогом второй производной по x функции $u(x,t)$.

Поставим в соответствие уравнению (1) его дискретный аналог в виде

$$\frac{y_i^{n+1} - y_i^n}{\tau} = \frac{y^{n+1}_{\overline{x}x,i} + y^n_{\overline{x}x,i}}{2} + f(x_i, t_{n+\frac{1}{2}}), \tag{4}$$

где $(x_i, t_{n+\frac{1}{2}}) = \left(x_i, t_n + \frac{\tau}{2}\right) \in \omega_{\tau h}$.

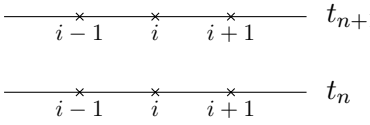
Определение. *Слой $t_{n+\frac{1}{2}} = t_n + \frac{\tau}{2}$ называется полуцелым слоем.*

Добавим краевые и начальное условия:

$$\begin{cases} y_0^{n+1} = \mu_1(t_{n+1}) \\ y_N^{n+1} = \mu_2(t_{n+1}), \end{cases} \quad t_{n+1} \in \overline{\omega}_\tau, \tag{5}$$

$$y_i^0 = u_0(x_i), \quad x_i \in \overline{\omega}_h. \tag{6}$$

В рассматриваемой разностной схеме использован шеститочечный шаблон вида



Заметим, что данная схема похожа на ту, которую мы рассматривали в предыдущем параграфе, в частности, матрица системы, соответствующей этой схеме, является трехдиагональной со строгим диагональным преобладанием. Это значит, что решение разностной схемы (4) – (6) такой задачи существует, единственно и находится с помощью метода прогонки.

Введем погрешность решения разностной схемы:

$$z_i^n = z(x_i, t_n) = y_i^n - u_i^n,$$

где $u_i^n = u(x_i, t_n),\ (x_i, t_n) \in \overline{\omega}_{\tau h}$.
Выразив y_i^n из этого выражения и подставив его в уравнение (4), получим задачу относительно z_i^n:

$$\frac{z_i^{n+1} - z_i^n}{\tau} = \frac{z_{\overline{x}x,i}^{n+1} + z_{\overline{x}x,i}^n}{2} + \psi_i^n, \quad (x_i, t_n), (x_i, t_{n+1}) \in \omega_{\tau h}, \tag{7}$$

$$\begin{cases} z_0^{n+1} = 0 \\ z_N^{n+1} = 0, \end{cases} \quad t_{n+1} \in \overline{\omega}_\tau, \tag{8}$$

$$z_i^0 = 0, \quad x_i \in \overline{\omega}_h, \tag{9}$$

где ψ_i^n — погрешность аппроксимации на решении исходной задачи (1)–(3):

$$\psi_i^n = \psi(x_i, t_n) = -\frac{u_i^{n+1} - u_i^n}{\tau} + \frac{u_{\overline{x}x,i}^{n+1} + u_{\overline{x}x,i}^n}{2} + f(x_i, t_{n+\frac{1}{2}}), \quad (x_i, t_n), (x_i, t_{n+1}) \in \omega_{\tau h}. \tag{10}$$

Задача. *Доказать, что*

$$\psi_i^n = \mathrm{O}\big(\tau^2 + h^2\big). \tag{11}$$

Переходим к изучению вопросов сходимости и устойчивости разностной задачи (4)–(6).

Рассмотрим вещественное пространство H_{N-1} сеточных функций w, заданных на одномерной сетке ω_h, содержащей $(N-1)$ узел и обращающихся в нуль на границе ($w_0 = w_N = 0$).

Значение функции $w \in H_{N-1}$ в i-ом узле сетки, $i = \overline{1, (N-1)}$, обозначим через w_i. Заметим, что

$$\dim H_{N-1} = N - 1.$$

Введем скалярное произведение в пространстве H_{N-1}:

$$(z, v) = \sum_{i=1}^{N-1} z_i v_i h, \quad z, v \in H_{N-1}.$$

Введем норму в пространстве H_{N-1}:

$$\|z\|_{L_2(\omega_h)} = \|z\|_{L_2} = \left(\sum_{i=1}^{N-1} z_i^2 h \right)^{\frac{1}{2}}, \quad z \in H_{N-1}. \tag{12}$$

Заметим, что если взять значения сеточной функции z_i^n, рассматриваемой на сетке $\omega_{\tau h}$, принадлежащие одному слою, пусть n-ому, то эти значения образуют функцию z^n, принадлежащую пространству H_{N-1}. Тогда, если будет верна оценка

$$\left\| z^{n+1} \right\|_{L_2} \leqslant M \left(\tau^2 + h^2 \right),$$

где константа M не зависит от τ и h, то это будет означать сходимость рассматриваемой разностной схемы к решению исходной задачи в норме L_2 с соответствующими порядками точности по τ и h.

Наряду с вещественным пространством H_{N-1} будем рассматривать гильбертово пространство L_2 — линейное пространство функций, интегрируемых с квадратом на интервале $(0, 1)$:

$$\int_0^1 f^2(x) dx < \infty.$$

Введем скалярное произведение в пространстве L_2:

$$(f,g) = \int\limits_0^1 f(x)g(x)dx, \quad f(x), g(x) \in L_2.$$

Теперь введем норму в пространстве L_2:

$$\|f\|_{L_2} = \left(\int\limits_0^1 f^2(x)dx \right)^{\frac{1}{2}}.$$

Задача на собственные значения

Рассмотрим задачу на собственные значения (задачу Штурма-Лиувилля) для функции $u(x) \in L_2$, обладающей достаточной гладкостью:

$$\begin{cases} \dfrac{d^2u}{dx^2} + \lambda u(x) = 0, \quad x \in (0,1), \\ u(0) = u(1) = 0, \end{cases} \tag{13}$$

причем $u(x) \not\equiv 0$.

Решениями данной задачи являются собственные значения λ_k и собственные функции $u_k(x)$:

$$\lambda_k = \pi^2 k^2, \quad k \in \mathbb{N},$$

$$0 < \lambda_1 < \lambda_2 < \ldots < \lambda_n < \ldots,$$

$$u_k(x) = c\sin(\pi k x), \quad c = const \neq 0.$$

Одним из свойств собственных функций задачи Штурма-Лиувилля является тот факт, что эти функции образуют ортогональный базис пространства L_2.

Положим $c = \sqrt{2}$ и получим:

$$u_k(x) = \sqrt{2}\sin(\pi k x).$$

Тогда функции $\{u_k(x)\}_{k=1}^{\infty}$ образуют ортонормированный базис в пространстве L_2:

$$(u_k, u_l) = \delta_{kl}.$$

Значит, произвольную функцию $f(x) \in L_2$ можно разложить по базису $\{u_k(x)\}_{k=1}^{\infty}$:

$$f(x) = \sum_{k=1}^{\infty} f_k u_k(x),$$

где коэффициенты $f_k = (f, u_k)$ называются коэффициентами Фурье. Тогда справедливо равенство Парсеваля:

$$\|f\|_{L_2}^2 = \sum_{k=1}^{\infty} f_k^2.$$

Рассмотрим теперь разностный аналог задачи Штурма-Лиувилля для сеточной функции $y \in H_{N-1}$:

$$\begin{cases} y_{\overline{x}x_i} + \lambda y(x_i) = 0, \quad x_i \in w_h, \quad i = \overline{1,(N-1)}, \\ y_0 = y_N = 0. \end{cases} \tag{14}$$

причем $y(x) \not\equiv 0$. Будем искать собственные функции в виде

$$y(x_i) = \sin(\alpha x_i), \quad \alpha \in \mathbb{R}, \quad i = \overline{1,(N-1)}.$$

Распишем уравнение (14) подробнее:

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + \lambda y_i = 0$$

и перенесем слагаемые, содержащие y_i, в правую часть:

$$y_{i+1} + y_{i-1} = \left(2 - h^2\lambda\right) y_i, \quad i = \overline{1,(N-1)}.$$

Очевидно, что

$$y_{i+1} + y_{i-1} = y(x_i + h) + y(x_i - h) = \sin\alpha(x_i + h) + \sin\alpha(x_i - h) = 2\sin(\alpha x_i)\cos(\alpha h).$$

Следовательно,

$$2\sin(\alpha x_i)\cos(\alpha h) = (2 - h^2\lambda)\sin\alpha x_i.$$

$\sin(\alpha x_i) \neq 0$, так как собственные функции не могут быть нулевыми, значит

$$\frac{\lambda h^2}{2} = 1 - \cos\alpha h = 2\sin^2\left(\frac{\alpha h}{2}\right).$$

Отсюда следует, что

$$\lambda = \frac{4}{h^2}\sin^2\left(\frac{\alpha h}{2}\right).$$

Для того, чтобы найти α, воспользуемся краевым условием для y:

$$y_N = \sin\alpha = 0,$$

откуда следует, что $\alpha_k = \pi k$, $k \in \mathbb{N}$. Тогда собственные значения λ_k равны

$$\lambda_k = \frac{4}{h^2}\sin^2\left(\frac{\pi k h}{2}\right), \quad k = \overline{1,(N-1)},$$

а соответствующие им собственные функции имеют вид

$$y_k = C\sin(\pi k x_i), \quad k \in \overline{1,(N-1)}.$$

Система функций $y_k(x_i)$, $k = \overline{1,(N-1)}$ ортогональна, а если положить $C = \sqrt{2}$, то совокупность сеточных функций $y_k(x_i) = \sqrt{2}\sin(\pi k x_i)$ образует ортонормированный (в смысле скалярного произведения (12)) базис пространства H_{N-1}. Следовательно, любая сеточная функция $f(x_i)$, $i = \overline{1,(N-1)}$, однозначно разложима по базису $\{y_k\}_1^{N-1}$, то есть

$$f(x_i) = \sum_{k=1}^{N-1} c_k y_k(x_i),$$

где $c_k = (f, y_k)$, $k = \overline{1,(N-1)}$ — коэффициенты Фурье. Имеет место равенство Парсеваля:

$$\|f\|^2_{L_2(\omega_h)} = \sum_{k=1}^{N-1} c_k^2. \tag{15}$$

Воспользуемся рассмотренной задачей Штурма-Лиувилля для доказательства следующей теоремы.

Теорема. *Пусть функция $u(x,t)$, являющаяся решением задачи для уравнения теплопроводности (1)–(3), имеет достаточную гладкость. Тогда симметричная разностная схема (4)–(6) сходится к решению исходной задачи со вторым порядком по τ и вторым порядком по h в L_2-норме пространства сеточных функций.*

Доказательство. Обратимся к рассмотрению задачи (7)–(9) для погрешности решения разностной схемы z_i^n:

$$\frac{z_i^{n+1}-z_i^n}{\tau}=\frac{z_{\overline{x}x,i}^{n+1}+z_{\overline{x}x,i}^n}{2}+\psi_i^n,\quad (x_i,t_n),(x_i,t_{n+1})\in\omega_{\tau h},\tag{16}$$

$$\begin{cases} z_0^{n+1}=0 \\ z_N^{n+1}=0, \end{cases}\quad t_{n+1}\in\overline{\omega}_\tau,\tag{17}$$

$$z_i^0=0,\quad x_i\in\overline{\omega}_h,\tag{18}$$

где ψ_i^n — погрешность аппроксимации на решении задачи (1)–(3), $\psi_i^n=\mathrm{O}\big(\tau^2+h^2\big)$:

$$\psi_i^n=\psi(x_i,t_n)=-\frac{u_i^{n+1}-u_i^n}{\tau}+\frac{u_{\overline{x}x,i}^{n+1}+u_{\overline{x}x,i}^n}{2}+f(x_i,t_{n+\frac12}),\quad (x_i,t_n),(x_i,t_{n+1})\in\omega_{\tau h}.\tag{19}$$

Будем искать погрешность z_i^n в виде

$$z_i^n=\sum_{k=1}^{N-1}c_k(t_n)\mu_k(x_i),\quad x_i\in\overline{w}_h,\tag{20}$$

где $c_k(t_n), k=\overline{1,(N-1)}$ — дискретные функции только аргумента t_n, $\mu_k,\ k=\overline{1,(N-1)}$ — собственные функции задачи, зависящие только от $x_i\in\omega_h$.:

$$\mu_{\overline{x}x,i}+\lambda\mu(x_i)=0,\quad i=\overline{1,(N-1)},\tag{21}$$

$$\mu_0=\mu_N=0.$$

Задача (21) была рассмотрена нами выше.

Функции μ_k имеют вид

$$\mu_k(x_i)=\sqrt{2}\sin(\pi kx),\quad i,k=\overline{1,(N-1)}$$

и образуют ортонормированный базис в H_{N-1}. Этим функциям соответствуют собственные значения λ_k, равные

$$\lambda_k=\frac{4}{h^2}\sin^2\left(\frac{\pi kh}{2}\right),\quad k=\overline{1,(N-1)},$$

Так как функции $\{\mu_k\}_{k=1}^{N-1}$ образуют ортонормированный базис пространства H_{N-1}, то любой элемент пространства H_{N-1} можно разложить по этим функциям, следовательно, представление (20) корректно.

Разложим по базисным функциям погрешность аппроксимации ψ_i^n на решении:

$$\psi_i^n=\sum_{k=1}^{N}\psi^{(k)}(t_n)\mu_k(x_i),\tag{22}$$

где $\psi^{(k)}(t_n)$ — дискретные функции только аргумента t_n.
Подставим выражения (20) и (22) в уравнение (7):

$$\frac{\sum_{k=1}^{N-1}(c_k(t_{n+1})-c_k(t_n))}{\tau}\mu_k(x_i)=\frac{1}{2}\sum_{k=1}^{N-1}\left(c_k(t_{n+1})+c_k(t_n)\right)(\mu_k)_{\overline{x}x,i}+\sum_{k=1}^{N-1}\psi^{(k)}(t_n)\mu_k(x_i).$$

Принимая во внимание уравнение (21), получаем

$$\sum_{k=1}^{N-1}\left(\frac{c_k(t_{n+1})-c_k(t_n)}{\tau}+\frac{\lambda_k}{2}(c_k(t_{n+1})+c_k(t_n))\right)\mu_k(x_i)=\sum_{k=1}^{N-1}\psi^{(k)}(t_n)\mu_k(x_i).$$

Так как $\{\mu_k\}_{k=1}^{N-1}$ — система линейно независимых функций, то полученное равенство выполняется тогда и только тогда, когда коэффициенты при соответствующих функциях $\mu_k(x_i),\ k=\overline{1,(N-1)}$ равны:

$$\frac{c_k(t_{n+1})-c_k(t_n)}{\tau}+\frac{\lambda_k}{2}(c_k(t_{n+1})+c_k(t_n))=\psi^{(k)}(t_n),\quad k=\overline{1,(N-1)}.$$

Разрешим это уравнение относительно $(n+1)$-ого слоя, домножив обе части на $\tau\neq 0$ и сгруппировав слагаемые с $c_k(t_{n+1})$ и $c_k(t_n)$:

$$\left(1+\frac{\tau\lambda_k}{2}\right)c_k(t_{n+1})=\left(1-\frac{\tau\lambda_k}{2}\right)c_k(t_n)+\tau\psi^{(k)}(t_n).$$

Учитывая, что $\left(1+\frac{\tau\lambda_k}{2}\right)\neq 0$, получаем

$$c_k(t_{n+1})=\frac{1-\frac{\tau\lambda_k}{2}}{1+\frac{\tau\lambda_k}{2}}c_k(t_n)+\frac{\tau}{1+\frac{\tau\lambda_k}{2}}\psi^{(k)}(t_n). \tag{23}$$

Обозначим $q_k=\dfrac{1-\frac{\tau\lambda_k}{2}}{1+\frac{\tau\lambda_k}{2}}$.

Задача. *Показать, что*

$$|q_k|=\left|\frac{1-\frac{\tau\lambda_k}{2}}{1+\frac{\tau\lambda_k}{2}}\right|\leqslant 1.$$

Решение. Нужно показать, что $-1\leqslant q_k\leqslant 1$ или

$$-1\leqslant\frac{1-0.5\tau\lambda_k}{1+0.5\tau\lambda_k}\leqslant 1.$$

Неравенство

$$\frac{1-0.5\tau\lambda_k}{1+0.5\tau\lambda_k}\leqslant 1$$

очевидно в силу того, что $\tau>0$, $\lambda_k>0$. Рассмотрим теперь неравенство

$$\frac{1-0.5\tau\lambda_k}{1+0.5\tau\lambda_k}\geqslant -1$$

или

$$-1-0.5\tau\lambda_k\leqslant 1-0.5\tau\lambda_k,$$

Которое, как легко заметить, выполнено всегда. □

Подставим выражение (23) в разложение (20):

$$z_i^{n+1} = \sum_{k=1}^{N-1} c_k(t_{n+1})\mu_k(x_i) = \sum_{k=1}^{N-1} q_k c_k(t_n)\mu_k(x_i) + \sum_{k=1}^{N-1} \frac{\tau}{1+\frac{\tau\lambda_k}{2}}\psi^{(k)}(t_n)\mu_k(x_i).$$

Обозначим первую сумму через V, а вторую через W. Применим правило треугольника для оценки нормы погрешности z^{n+1} через нормы этих величин:

$$\left\|z^{n+1}\right\|_{L_2} \leqslant \left\|V\right\|_{L_2} + \left\|W\right\|_{L_2}. \tag{24}$$

Оценим квадрат нормы V, воспользовавшись результатом рассмотренной выше задачи и равенством Парсеваля:

$$\left\|V\right\|_{L_2}^2 = \sum_{k=1}^{N-1} q_k^2 c_k^2(t_n) \leqslant \sum_{k=1}^{N-1} c_k^2(t_n) = \left\|z^n\right\|_{L_2}^2.$$

Аналогичным образом поступим с W с учетом того, что $1+\frac{\tau\lambda_k}{2} > 1$:

$$\left\|W\right\|_{L_2}^2 = \sum_{k=1}^{N-1} \left(\frac{\tau}{1+\frac{\tau\lambda_k}{2}}\right)^2 \left(\psi^{(k)}(t_n)\right)^2 \leqslant \tau^2 \sum_{k=1}^{N-1} \left(\psi^{(k)}(t_n)\right)^2 = \tau^2 \left\|\psi^n\right\|_{L_2}^2.$$

Тогда неравенство (24) примет вид

$$\left\|z^{n+1}\right\|_{L_2} \leqslant \left\|z^n\right\|_{L_2} + \tau\left\|\psi^n\right\|_{L_2}.$$

Рассматривая полученную оценку, как рекуррентную, легко получим:

$$\left\|z^{n+1}\right\|_{L_2} \leqslant \left\|z^0\right\|_{L_2} + \sum_{j=1}^{n} \tau\left\|\psi(t_j)\right\|_{L_2}. \tag{25}$$

Учитывая, что $\left\|z^0\right\|_{L_2} = 0$ как норма начального приближения, а также используя оценку нормы погрешности аппроксимации, которая следует из оценки (11),

$$\left\|\psi^k\right\|_{L_2} \leqslant M\left(\tau^2 + h^2\right),$$

где $M > 0$ — константа, не зависящая от τ и h,

$$\sum_{k=0}^{n} \tau = t_n \leqslant T.$$

Обозначим $M_1 = TM > 0$ и получим окончательную оценку:

$$\left\|z^{n+1}\right\|_{L_2} \leqslant M_1\left(\tau^2 + h^2\right).$$

□

Замечание. *Если в разностной задаче (4) – (6) взять нулевые краевые условия*

$$y_0^{n+1} = y_N^{n+1} = 0,$$

то для y_i^n можно вывести априорную оценку, аналогичную полученной выше оценке (25):

$$\|y^{n+1}\|_{L_2} \leqslant \|u_0\|_{L_2} + \tau\sum_{j=0}^{n}\|f(t_j)\|_{L_2}.$$

Эта оценка означает, что решение разностной схемы устойчиво в норме L_2 по начальному условию u_0 и правой части уравнения.

§5 Разностные схемы с весами. Погрешность аппроксимации на решении

Рассмотрим уравнение теплопроводности с краевыми условиями первого рода:

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2} + f(x,t), \quad (x,t) \in G = \{(x,t) \mid x \in (0,1),\ t \in (0,T]\}, \tag{1}$$

$$\begin{cases} u(0,t) = \mu_1(t) \\ u(1,t) = \mu_2(t), \end{cases} \quad t \in [0,T], \tag{2}$$

$$u(x,0) = u_0(x), \quad x \in [0,1]. \tag{3}$$

Воспользуемся сетками $\omega_{\tau h}$ и $\overline{\omega}_{\tau h}$, введенными в первом параграфе данной главы, на множествах G и $\overline{G}$ соответственно.

Поставим в соответствие задаче (1) – (3) семейство разностных схем (в зависимости от параметра σ):

$$\frac{y_i^{n+1} - y_i^n}{\tau} = \sigma y_{\overline{x}x,i}^{n+1} + (1-\sigma)\, y_{\overline{x}x,i}^n + \varphi_i^n, \quad (x_i, t_n) \in \omega_{\tau h}, \tag{4}$$

где φ_i^n - некоторая аппроксимация правой части, но не точное значение функции $f(x,t)$ в соответствующем узле, $\sigma \in \mathbb{R}$ — весовой множитель.

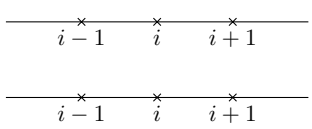
Замечание 1. *На практике обычно рассматривают параметр $\sigma \in [0,1]$, но данное условие не является обязательным.*

Добавим краевые и начальное условия:

$$\begin{cases} y_0^{n+1} = \mu_1(t_{n+1}) \\ y_N^{n+1} = \mu_2(t_{n+1}), \end{cases} \quad t_{n+1} \in \overline{\omega}_\tau, \tag{5}$$

$$y_i^0 = u_0(x_i), \quad x_i \in \overline{\omega}_h. \tag{6}$$

В рассматриваемой разностной схеме использован шеститочечный шаблон вида



При определенных значениях параметра σ мы получим разностные схемы, которые рассматривались в предыдущих параграфах:

1. При $\sigma = 0$, $\varphi_i^n = f_i^n$ получаем явную разностную схему.
2. При $\sigma = 1$, $\varphi_i^n = f(x_i, t_{n+1})$ получаем чисто неявную разностную схему.
3. При $\sigma = 0.5$, $\varphi_i^n = f(x_i, t_{n+\frac{1}{2}})$ получаем симметричную разностную схему.

Замечание. *Среди всех разностных схем семейства (4) явной является только схема с $\sigma = 0$, все остальные — неявные.*

Введем погрешность решения разностной схемы:

$$z_i^n = z(x_i, t_n) = y_i^n - u_i^n,$$

где $u_i^n = u(x_i, t_n),\ (x_i, t_n) \in \overline{\omega}_{\tau h}$.

Выразив y_i^n из этого выражения и подставив его в уравнение (4), получим задачу относительно z_i^n:

$$\frac{z_i^{n+1} - z_i^n}{\tau} = \sigma z_{\overline{x}x,i}^{n+1} + (1-\sigma) z_{\overline{x}x,i}^n + \psi_i^n, \quad (x_i, t_n), (x_i, t_{n+1}) \in \omega_{\tau h}, \tag{7}$$

$$\begin{cases} z_0^{n+1} = 0 \\ z_N^{n+1} = 0, \end{cases} \quad t_{n+1} \in \overline{\omega}_\tau, \tag{8}$$

$$z_i^0 = 0, \quad x_i \in \overline{\omega}_h, \tag{9}$$

где ψ_i^n — погрешность аппроксимации на решении задачи (1)–(3):

$$\psi_i^n = \sigma u_{\overline{x}x,i}^{n+1} + (1-\sigma) u_{\overline{x}x,i}^n - \frac{u_i^{n+1} - u_i^n}{\tau} + \varphi_i^n. \tag{10}$$

Далее считаем, что $i = \overline{1, N-1}, n = \overline{1, M-1}$.

Пусть решение $u(x,t)$ задачи (1)–(3) имеет достаточную гладкость (функция $u(x,t)$ шесть раз дифференцируема по x и три раза дифференцируема по t). Обозначим $u_t' = \dot{u} = \frac{\partial u}{\partial t},\ u_x' = u' = \frac{\partial u}{\partial x}$. Разложим значения $u_{i+1} = u(x_{i+1}, t)$ и $u_{i-1} = u(x_{i-1}, t)$ в ряд Тейлора в точке (x_i, t):

$$u_{i+1} = u_i + h u_i' + \frac{h^2}{2} u_i'' + \frac{h^3}{6} u_i''' + \frac{h^4}{24} u_i^{(4)} + \dots,$$

$$u_{i-1} = u_i - h u_i' + \frac{h^2}{2} u_i'' - \frac{h^3}{6} u_i''' + \frac{h^4}{24} u_i^{(4)} + \dots$$

Разложим в ряд Тейлора в точке $(x_i, t_{n+\frac{1}{2}})$ значения функции $u(x_i, t)$ на $(n+1)$-ом и n-ом слоях:

$$u_i^{n+1} = u_i(t_{n+\frac{1}{2}}) + \frac{\tau}{2}\dot{u}_i(t_{n+\frac{1}{2}}) + \frac{\tau^2}{8}\ddot{u}_i(t_{n+\frac{1}{2}}) + \frac{\tau^3}{48}\dddot{u}_i(t_{n+\frac{1}{2}}) + \dots,$$

$$u_i^n = u_i(t_{n+\frac{1}{2}}) - \frac{\tau}{2}\dot{u}_i(t_{n+\frac{1}{2}}) + \frac{\tau^2}{8}\ddot{u}_i(t_{n+\frac{1}{2}}) - \frac{\tau^3}{48}\dddot{u}_i(t_{n+\frac{1}{2}}) + \dots,$$

Воспользовавшись записанными выше разложениями, получим следующее выражение для второй дискретной производной:

$$u_{\overline{x}x,i} = \frac{u_{i+1} + u_{i-1} - 2u_i}{h^2} = u_i'' + \frac{h^2}{12} u_i^{(4)} + \mathrm{O}\big(h^4\big). \tag{11}$$

Вычтем выражение для u_i^n из выражения для u_i^{n+1}, разделим результат на $\tau \neq 0$ и получим:

$$\frac{u_i^{n+1} - u_i^n}{\tau} = \dot{u}_i(t_{n+\frac{1}{2}}) + \mathrm{O}\big(\tau^2\big). \tag{12}$$

Подставим выражения (11) и (12) в уравнение (10):

$$\begin{aligned} \psi_i^n = \sigma \left(u_i'' + \frac{\tau}{2}\dot{u}_i'' + \frac{h^2}{12} u_i^{(4)} + \mathrm{O}\big(\tau h^2\big) \right) + \\ + (1-\sigma) \left(u_i'' - \frac{\tau}{2}\dot{u}_i'' + \frac{h^2}{12} u_i^{(4)} + \mathrm{O}\big(\tau h^2\big) \right) - \dot{u}_i + \varphi_i^n + \mathrm{O}\big(\tau^2 + h^4\big). \end{aligned} \tag{13}$$

Воспользуемся неравенством, связывающим среднее арифметическое и среднее геометрическое чисел τ^2 и h^4:

$$\tau h^2 \leqslant \frac{\tau^2 + h^4}{2}.$$

Следовательно, $\mathrm{O}\big(\tau h^2\big) = \mathrm{O}\big(\tau^2 + h^4\big)$.

Сгруппируем слагаемые в уравнении (13) следующим образом:

$$\begin{aligned} \psi_i^n = u_i'' - \dot{u}_i + \varphi_i^n + \tau(\sigma - 0.5)\dot{u}_i'' + \frac{h^2}{12}u_i^{(4)} + \mathrm{O}\big(\tau^2 + h^4\big) = \\ = \underbrace{u_i'' - \dot{u}_i + f_i(t_{n+\frac{1}{2}})}_{0} + \varphi_i^n - f_i(t_{n+\frac{1}{2}}) + \tau(\sigma - 0.5)\dot{u}_i'' + \frac{h^2}{12}u_i^{(4)} + \mathrm{O}\big(\tau^2 + h^4\big). \end{aligned} \tag{14}$$

Для получения четвертого порядка по h для погрешности аппроксимации на решении необходимо исключить из уравнения (14) члены порядка h^2, то есть слагаемое $\dfrac{h^2}{12}u_i^{(4)}$. Рассмотрим уравнение:

$$u_i'' = \dot{u}_i - f_i.$$

Продифференцируем это равенство два раза по x и получим выражение для $u_i^{(4)}$:

$$u_i^{(4)} = \dot{u}_i'' - f_i''.$$

Подставим это выражение в равенство (14):

$$\psi_i^n = \varphi_i^n - f_i(t_{n+\frac{1}{2}}) + \left((\sigma - 0.5)\tau + \frac{h^2}{12}\right)\dot{u}_i'' - \frac{h^2}{12}f_i''(t_{n+\frac{1}{2}}) + \mathrm{O}\big(\tau^2 + h^4\big).$$

Выберем σ так, чтобы коэффициент $\left((\sigma - 0.5)\tau + \dfrac{h^2}{12}\right)$ обратился в нуль:

$$\sigma_* = \frac{1}{2} - \frac{h^2}{12\tau}.$$

Теперь если положить

$$\sigma = \sigma_*, \quad \varphi_i^n = f_i(t_{n+\frac{1}{2}}) + \frac{h^2}{12}f_i''(t_{n+\frac{1}{2}}),$$

то погрешность аппроксимации на решении задачи (1) – (3) будет иметь порядок $\mathrm{O}\big(\tau^2 + h^4\big)$.

Определение. *Разностная схема* (4) – (6) *при*

$$\sigma = \frac{1}{2} - \frac{h^2}{12\tau}, \quad \varphi_i^n = f_i(t_{n+\frac{1}{2}}) + \frac{h^2}{12}f_i''(t_{n+\frac{1}{2}})$$

называется разностной схемой повышенного порядка точности.

Замечание. *Если*

$$\begin{aligned} &\sigma = 0, \quad \varphi_i^n = f_i(t_{n+\frac{1}{2}}) + \mathrm{O}\big(h^2\big), \ \textit{то}\ \psi_i^n = \mathrm{O}\big(\tau + h^2\big), \\ &\sigma = 1, \quad \varphi_i^n = f_i(t_{n+\frac{1}{2}}) + \mathrm{O}\big(h^2\big), \ \textit{то}\ \psi_i^n = \mathrm{O}\big(\tau + h^2\big), \\ &\sigma = 0.5, \quad \varphi_i^n = f_i(t_{n+\frac{1}{2}}) + \mathrm{O}\big(\tau^2 + h^2\big), \ \textit{то}\ \psi_i^n = \mathrm{O}\big(\tau^2 + h^2\big). \end{aligned}$$

При всех остальных σ погрешность аппроксимации ψ_i^n имеет порядок $\mathrm{O}\big(\tau + h^2\big)$.

§6 Разностная схема для уравнения Пуассона. Первая краевая задача

Рассмотрим первую краевую задачу для уравнения Пуассона:

$$\begin{cases} \dfrac{\partial^2 u}{\partial x_1^2} + \dfrac{\partial^2 u}{\partial x_2^2} = f(x_1, x_2), \ (x_1, x_2) \in G, \\ u(x_1, x_2)|_\Gamma = \mu(x_1, x_2), \end{cases} \tag{1}$$

где G — прямоугольная область:

$$G = \{(x_1, x_2): \ x_1 \in \mathbb{R}, \ 0 < x_1 < l_1; \ x_2 \in \mathbb{R}, \ 0 < x_2 < l_2\},$$

а Γ — граница этой области.

Решением первой краевой задачи называется функция $u(x_1, x_2)$, удовлетворяющая системе уравнений (1), для которой выполнены следующие условия:

$$u(x_1, x_2) \in C\left(\overline{G}\right), \ \overline{G} = G \cup \Gamma, \quad u(x_1, x_2) \in C^2(G).$$

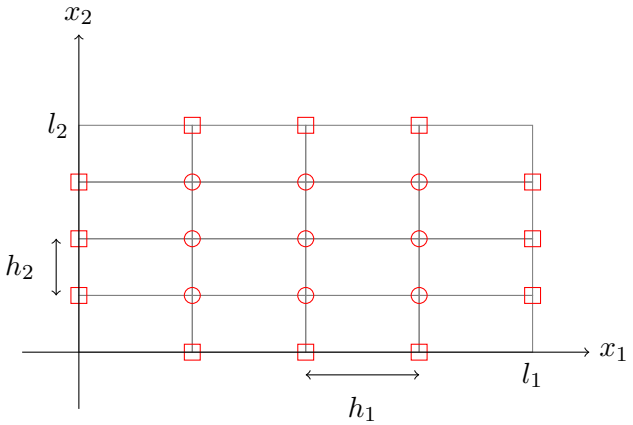
Введем на области G сетку с шагами $h_1 = \frac{l_1}{N_1}$ и $h_2 = \frac{l_2}{N_2}$, где $N_1, N_2 \in \mathbb{N}$ — размеры сетки (узлы этой сетки обозначены на рисунке окружностями):

$$\omega_h = \left\{ \left(x_1^{(i)}, x_2^{(j)}\right) : x_1^{(i)} = ih_1, \ x_2^{(j)} = jh_2 \right\}, \quad i = \overline{1, (N_1 - 1)}, \ j = \overline{1, (N_2 - 1)}.$$

Добавим к этой сетке узлы на границе Γ (обозначены на рисунке квадратами):

$$\Gamma_h = \{x_{0,j}\}_{j=1}^{N_2-1} \cup \{x_{N_1,j}\}_{j=1}^{N_2-1} \cup \{x_{i,0}\}_{i=1}^{N_1-1} \cup \{x_{i,N_2}\}_{i=1}^{N_1-1}.$$

Обозначим $\overline{\omega}_h = \omega_h \cup \Gamma_h$.



Пусть y_{ij} — сеточная функция, определенная на сетке ω_h. Определим для этой функции разностные производные второго порядка по x_1 и второго порядка по x_2 в узле $x_{ij} \in \omega_h$:

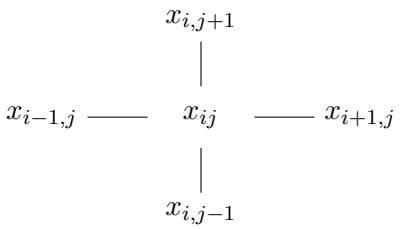
$$y_{\overline{x}_1 x_1, ij} = \frac{y_{i+1,j} - 2y_{ij} + y_{i-1,j}}{h_1^2},$$

$$y_{\overline{x}_2 x_2, ij} = \frac{y_{i,j+1} - 2y_{ij} + y_{i,j-1}}{h_2^2}$$

и поставим в соответствие задаче (1) разностную схему

$$\begin{cases} y_{\overline{x}_1 x_1, ij} + y_{\overline{x}_2 x_2, ij} = f_{ij}, \quad x_{ij} = \left(x_1^{(i)}, x_2^{(j)}\right) \in \omega_h, \\ y_{ij}|_{\Gamma_h} = \mu_{ij}, \end{cases} \tag{2}$$

где f_{ij}, μ_{ij} — значения функций $f(x_1, x_1)$ и $\mu(x_1, x_2)$ в узлах $x_{ij} \in \omega_h$. Этой разностной схеме соответствует пятиточечный шаблон типа «крест»:



Введем погрешность решения численной задачи:

$$z_{ij} = y_{ij} - u\left(x_1^{(i)}, x_2^{(j)}\right) = y_{ij} - u_{ij}.$$

Погрешность z_{ij} удовлетворяет следующей разностной схеме:

$$\begin{cases} z_{\overline{x}_1 x_1, ij} + z_{\overline{x}_2 x_2, ij} = -\psi_{ij}, \quad x_{ij} = \left(x_1^{(i)}, x_2^{(j)}\right) \in \omega_h, \\ z_{ij}|_{\Gamma_h} = 0. \end{cases}$$

где ψ_{ij} — погрешность аппроксимации на решении исходного уравнения (1):

$$\psi_{ij} = -f_{ij} + u_{\overline{x}_1 x_1, ij} + u_{\overline{x}_2 x_2, ij}.$$

Задача. *Показать, что справедлива следующая оценка погрешности аппроксимации на решении исходной задачи (1):*

$$\psi_{ij} = \mathrm{O}\big(h_1^2 + h_2^2\big).$$

§7 Разрешимость разностной задачи. Сходимость разностной задачи Дирихле

Продолжаем рассматривать задачу Дирихле

$$\begin{cases} \dfrac{\partial^2 u}{\partial x_1^2} + \dfrac{\partial^2 u}{\partial x_2^2} = f(x_1, x_2), \ (x_1, x_2) \in G, \\ u(x_1, x_2)|_{\Gamma} = \mu(x_1, x_2) \end{cases} \tag{1}$$

Запишем разностную схему (2) из §6 в виде:

$$\begin{cases} \dfrac{y_{i-1,j} - 2y_{ij} + y_{i+1,j}}{h_1^2} + \dfrac{y_{i,j-1} - 2y_{ij} + y_{i,j+1}}{h_2^2} = f_{ij}, \quad i = \overline{1, (N_1 - 1)}, \ j = \overline{1, (N_2 - 1)}, \\ y_{ij}|_{\Gamma_h} = \mu_{ij}. \end{cases}$$

Напомним, что f_{ij}, μ_{ij} — значения непрерывных функций $f(x_1, x_1)$ и $\mu(x_1, x_2)$ в узлах сетки ω_h. Разрешим эту схему относительно центрального узла x_{ij}:

$$\begin{cases} \left(\dfrac{2}{h_1^2} + \dfrac{2}{h_2^2}\right) y_{ij} = \dfrac{y_{i-1,j} + y_{i+1,j}}{h_1^2} + \dfrac{y_{i,j-1} + y_{i,j+1}}{h_2^2} - f_{ij}, \quad i = \overline{1, (N_1 - 1)}, \ j = \overline{1, (N_2 - 1)}, \\ y_{ij}|_{\Gamma_h} = \mu_{ij}. \end{cases} \tag{2}$$

Для того, чтобы эта система имела решение при любых значениях функций $f(x_1, x_2)$ и $\mu(x_1, x_2)$, необходимо и достаточно, чтобы однородная система линейных уравнений имела только тривиальное решение.

Пусть H_{N_1-1,N_2-1} — пространство сеточных функций, определенных на сетке ω_h и обращающихся в нуль на границе Γ_h. Введем норму в этом пространстве:

$$\|v\|_C = \max_{\substack{1\leqslant i\leqslant N_1-1\\ 1\leqslant j\leqslant N_2-1}} |v_{ij}|,\ v \in H_{N_1-1,N_2-1}.$$

Теорема 1. *Однородная система линейных уравнений*

$$\begin{cases} \left(\dfrac{2}{h_1^2}+\dfrac{2}{h_2^2}\right) v_{ij} = \dfrac{v_{i-1,j}+v_{i+1,j}}{h_1^2} + \dfrac{v_{i,j-1}+v_{i,j+1}}{h_2^2}, \quad i=\overline{1,(N_1-1)},\ j=\overline{1,(N_2-1)}, \\ v_{ij}|_{\Gamma_h} = 0 \end{cases}$$

имеет единственное решение, и оно является тривиальным:

$$v_{ij} = 0,\ x_{ij} \in \overline{\omega}_h.$$

Доказательство. Будем проводить доказательство методом от противного. Пусть существует узел $x_{ij} \in \omega_h$, в котором достигается ненулевое значение функции: $v_{ij} \neq 0$. Тогда найдется узел x_{i_0,j_0}, для которого выполнены два условия:

A) $v_{i_0,j_0} = \|v\|_C = \max\limits_{\substack{1\leqslant i\leqslant N_1-1\\ 1\leqslant j\leqslant N_2-1}} |v_{ij}|$.

B) Хотя бы для одного из оставшихся узлов шаблона выполнено условие

$$|v_{ij}| < |v_{i_0,j_0}|, \quad i \in \{i_0-1,\, i_0+1\},\ j \in \{j_0-1,\, j_0+1\}.$$

Такой узел существует, поскольку в противном случае значения во всех узлах совпадут и будут равны нулю, так как функция обращается в ноль на границе Γ_h.

Рассмотрим уравнение системы в узле x_{i_0,j_0}:

$$\left(\frac{2}{h_1^2}+\frac{2}{h_2^2}\right) v_{i_0,j_0} = \frac{v_{i_0-1,j_0}+v_{i_0+1,j_0}}{h_1^2} + \frac{v_{i_0,j_0-1}+v_{i_0,j_0+1}}{h_2^2}$$

и оценим его по модулю:

$$\left(\frac{2}{h_1^2}+\frac{2}{h_2^2}\right) |v_{i_0,j_0}| \leqslant \frac{|v_{i_0-1,j_0}|+|v_{i_0+1,j_0}|}{h_1^2} + \frac{|v_{i_0,j_0-1}|+|v_{i_0,j_0+1}|}{h_2^2}.$$

Значения функции v из правой части неравенства не превосходят v_{i_0,j_0} в силу условия A) и, кроме того, в силу условия B) хотя бы одно из значений функции строго меньше v_{i_0,j_0}. Таким образом, справедлива оценка

$$\left(\frac{2}{h_1^2}+\frac{2}{h_2^2}\right) |v_{i_0,j_0}| < \left(\frac{2}{h_1^2}+\frac{2}{h_2^2}\right) \|v\|_C.$$

Получили противоречие: $\|v\|_C < \|v\|_C$. Следовательно, предположение о существовании хотя бы одного ненулевого значения функции v неверно, и $v \equiv 0$. □

Следствие. *Разностная задача*

$$\begin{cases} y_{\overline{x}_1 x_1,ij} + y_{\overline{x}_2 x_2,ij} = f_{ij}, \quad x_{ij} = \left(x_1^{(i)}, x_2^{(j)}\right) \in \omega_h, \\ y_{ij}|_{\Gamma_h} = \mu_{ij} \end{cases}$$

имеет единственное решение при любых значениях f_{ij} и μ_{ij}, $x_{ij} \in \omega_h$.

Введем разностный оператор

$$L_h v_{ij} = \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right) v_{ij} - \frac{v_{i-1,j} + v_{i+1,j}}{h_1^2} - \frac{v_{i,j-1} + v_{i,j+1}}{h_2^2}, \quad x_{ij} \in \omega_h$$

и запишем разностную схему для погрешности $z_{ij} = y_{ij} - u_{ij}$ решения задачи (2) с помощью этого оператора:

$$\begin{cases} L_h z_{ij} = \psi_{ij}, \ x_{ij} \in \omega_h, \\ z_{ij}|_{\Gamma_h} = 0, \end{cases} \tag{3}$$

где ψ_{ij} погрешность аппроксимации на решении задачи (1) вида

$$\psi_{ij} = -f_{ij} + u_{\overline{x}_1 x_1, ij} + u_{\overline{x}_2 x_2, ij}.$$

Рассмотрим вопрос сходимости разностной схемы. Сходимость означает наличие оценки

$$\|z\|_C \leqslant M\left(h_1^2 + h_2^2\right),$$

где M — константа, не зависящая от h_1 и h_2. Такая оценка означает, что разностная схема имеет второй порядок точности по h_1 и h_2.

Лемма (принцип максимума)**.** *Пусть для сеточной функции v, определенной на сетке ω_h, выполнены неравенства*

$$v_{ij} \geqslant 0, \ x_{ij} \in \Gamma_h,$$

$$L_h v_{ij} \geqslant 0, \ x_{ij} \in \omega_h.$$

Тогда справедливо следующее неравенство:

$$v_{ij} \geqslant 0, \ x_{ij} \in \overline{\omega}_h.$$

Доказательство. Проведем доказательство методом от противного. Пусть существует узел $x_{ij} \in \omega_h$, в котором функция v отрицательна: $v_{ij} < 0$. Тогда найдется узел x_{i_0,j_0}, для которого выполнены два условия:

A) $v_{i_0,j_0} = \min\limits_{\substack{1 \leqslant i \leqslant N_1 - 1 \\ 1 \leqslant j \leqslant N_2 - 1}} v_{ij}$.

B) Хотя бы для одного из оставшихся узлов шаблона выполнено условие

$$|v_{ij}| > |v_{i_0,j_0}|, \quad i \in \{i_0 - 1, \, i_0 + 1\}, \ j \in \{j_0 - 1, \, j_0 + 1\}.$$

Такой узел существует, так как в противном случае $v \equiv 0$ и лемма доказана. Рассмотрим действие оператора L_h на значение функции v_{i_0,j_0}:

$$L_h v_{i_0,j_0} = \frac{v_{i_0,j_0} - v_{i_0-1,j_0}}{h_1^2} + \frac{v_{i_0,j_0} - v_{i_0+1,j_0}}{h_1^2} + \frac{v_{i_0,j_0} - v_{i_0,j_0-1}}{h_2^2} + \frac{v_{i_0,j_0} - v_{i_0,j_0+1}}{h_2^2}.$$

Все слагаемые в правой части этого равенства неположительны, и, кроме того, хотя бы одно из слагаемых в силу условия B) строго отрицательно. Таким образом,

$$L_h v_{i_0,j_0} < 0.$$

Это неравенство противоречит условию леммы, следовательно, предположение о существовании хотя бы одного узла, в котором функция v отрицательна, неверно. □

Следствие. *Рассмотрим две разностные задачи*

$$L_h y_{ij} = \varphi_{ij},\ x_{ij} \in \omega_h,\ y_{ij}|_{\Gamma_h} \text{— заданы},$$

$$L_h Y_{ij} = \Phi_{ij},\ x_{ij} \in \omega_h,\ Y_{ij}|_{\Gamma_h} \text{— заданы}.$$

Если выполнены неравенства

$$|y_{ij}| \leqslant Y_{ij},\ x_{ij} \in \Gamma_h,$$

$$|\varphi_{ij}| \leqslant \Phi_{ij},\ x_{ij} \in \omega_h,$$

то справедливо следующее неравенство:

$$|y_{ij}| \leqslant Y_{ij},\ x_{ij} \in \overline{\omega}_h.$$

Доказательство. Рассмотрим сеточные функции v и w, определенные на сетке $\overline{\omega}_h$:

$$v_{ij} = Y_{ij} - y_{ij},$$

$$w_{ij} = Y_{ij} + y_{ij}.$$

Воспользовавшись неравенствами из условия и принципом максимума, получим следующие оценки:

$$L_h v_{ij} = \Phi_{ij} - \varphi_{ij} \geqslant 0,\ x_{ij} \in \omega_h,\ v_{ij}|_{\Gamma_h} \geqslant 0,$$

$$v_{ij} \geqslant 0,\ x_{ij} \in \overline{\omega}_h.$$

$$L_h w_{ij} \geqslant 0,\ x_{ij} \in \omega_h,\ w_{ij}|_{\Gamma_h} \geqslant 0,$$

$$w_{ij} \geqslant 0,\ x_{ij} \in \overline{\omega}_h.$$

Из неотрицательности функций v и w следует искомая оценка для модуля функции y. □

Теорема 2. *Пусть решение исходной дифференциальной задачи четыре раза непрерывно дифференцируемо в $\overline{G}$:*

$$u(x_1, x_2) \in C^4\left(\overline{G}\right).$$

Тогда решение разностной задачи сходится к решению исходной задачи, и справедлива оценка

$$\left\| y_{ij} - u\left(x_1^{(i)}, x_2^{(j)}\right) \right\|_C \leqslant M\left(h_1^2 + h_2^2\right),$$

где $M > 0$ — константа, не зависящая от шагов сетки h_1 и h_2.

Доказательство. Запишем следующую разностную схему:

$$\begin{cases} L_h Y_{ij} = 4k,\ x_{ij} \in \omega_h,\ k > 0, \\ Y_{ij} \geqslant 0,\ x_{ij} \in \Gamma_h, \end{cases} \tag{4}$$

где $Y_{ij} = k\left(l_1^2 + l_2^2 - \left(x_1^{(i)}\right)^2 - \left(x_2^{(j)}\right)^2\right)$ — мажоранта. Напомним, что константы l_1 и l_2 используются в определении области G:

$$G = \{(x_1, x_2):\ x_1 \in \mathbb{R},\ 0 < x_1 < l_1;\ x_2 \in \mathbb{R},\ 0 < x_2 < l_2\}.$$

Задача. *Показать, что выполнено равенство $L_h Y_{ij} = 4k$.*

Далее рассмотрим две задачи: задачу для погрешности разностной схемы (3) и для мажоранты (4). Выберем k таким образом, чтобы было выполнено равенство

$$4k = \|\psi\|_C$$

и заметим, что справедливы следующие неравенства, вытекающие из доказанного следствия:

$$|z_{ij}| \leqslant Y_{ij},\ x_{ij} \in \Gamma_h,$$

$$|\psi_{ij}| \leqslant \|\psi\|_C = 4k,\ x_{ij} \in \omega_h.$$

Таким образом, во всех узлах сетки $\overline{\omega}_h$ выполняется оценка:

$$|z_{ij}| \leqslant Y_{ij},\ x_{ij} \in \overline{\omega}_h,\ 0 \leqslant Y_{ij} \leqslant \frac{l_1^2 + l_2^2}{4}\|\psi\|_C,$$

$$\|z\|_C \leqslant \underbrace{\frac{l_1^2 + l_2^2}{4}}_{M_1}\|\psi\|_C. \tag{5}$$

Так как

$$\|\psi\|_C \leqslant M_2\left(h_1^2 + h_2^2\right),$$

где M_2 — константа, не зависящая от шагов сетки h_1 и h_2. Тогда с учетом (5), получаем искомую оценку

$$\|z\|_C \leqslant M\left(h_1^2 + h_2^2\right),$$

где $M = M_1 M_2$ — константа, не зависящая от шагов сетки h_1 и h_2. □

Замечание. *Если в разностной схеме (2) заменить краевое условие нулевым, то получим задачу, аналогичную разностной задаче (3) для погрешности разностной схемы. Тогда для решения y разностной схемы (2) можно вывести оценку, аналогичную оценке (5) для погрешности z:*

$$\|y\|_C \leqslant M_1\|f\|_C = \frac{l_1^2 + l_2^2}{4}\|f\|_C.$$

Эта неравенство означает, что решение разностной схемы устойчиво.

§8 Методы решения разностной задачи Дирихле

Рассмотрим разностную задачу Дирихле:

$$\begin{cases} \left(\dfrac{2}{h_1^2} + \dfrac{2}{h_2^2}\right) y_{ij} = \dfrac{y_{i-1,j} + y_{i+1,j}}{h_1^2} + \dfrac{y_{i,j-1} + y_{i,j+1}}{h_2^2} - f_{ij}, \quad i = \overline{1,(N_1-1)},\ j = \overline{1,(N_2-1)}, \\ y_{ij}|_{\Gamma_h} = \mu_{ij}. \end{cases} \tag{1}$$

Для нахождения решения этой разностной схемы нужно решить СЛАУ с матрицей порядка $(N_1 - 1) \times (N_2 - 1)$. Заметим, что матрица системы разрежена, то есть среди элементов этой матрицы содержится большое количество нулей. Очевидно, что использование классического метода Гаусса для решения такой системы не будет оптимальным. Существуют значительно более эффективные как прямые, так и итерационные методы решения системы (1). Рассмотрим несколько итерационных методов, решающих поставленную задачу: методы Якоби, Зейделя и попеременно-треугольный итерационный метод.

Метод Якоби

Итерационный процесс задается схемой

$$\begin{cases} \left(\dfrac{2}{h_1^2}+\dfrac{2}{h_2^2}\right) y_{ij}^{(s+1)} = \dfrac{y_{i-1,j}^{(s)}+y_{i+1,j}^{(s)}}{h_1^2} + \dfrac{y_{i,j-1}^{(s)}+y_{i,j+1}^{(s)}}{h_2^2} - f_{ij}, \quad i=\overline{1,(N_1-1)},\ j=\overline{1,(N_2-1)}, \\ y_{ij}^{(s+1)}\Big|_{\Gamma_h} = \mu_{ij}, \end{cases}$$

где $s \in \mathbb{Z}_+$, $y_{ij}^{(0)}$ — задано.

Пусть шаги сетки равны: $h_1 = h_2 = h$. Тогда количество итераций n_0, необходимое для получения решения методом Якоби с заданной точностью $\varepsilon > 0$, обратно пропорционально квадрату шага h:

$$n_0(\varepsilon) \approx \mathrm{O}\big(h^{-2}\big).$$

Метод Зейделя

Метод Зейделя имеет такую же скорость сходимости, что и метод Якоби. Он задается схемой

$$\begin{cases} \left(\dfrac{2}{h_1^2}+\dfrac{2}{h_2^2}\right) y_{ij}^{(s+1)} = \dfrac{y_{i-1,j}^{(s+1)}+y_{i+1,j}^{(s)}}{h_1^2} + \dfrac{y_{i,j-1}^{(s+1)}+y_{i,j+1}^{(s)}}{h_2^2} - f_{ij}, \quad i=\overline{1,(N_1-1)},\ j=\overline{1,(N_2-1)}, \\ y_{ij}^{(s+1)}\Big|_{\Gamma_h} = \mu_{ij}, \end{cases}$$

где $s \in \mathbb{Z}_+$, $y_{ij}^{(0)}$ — задано.

Убедимся что решение этой разностной схемы, вообще говоря, неявной, можно найти по явным формулам:

1. Найдем $y_{11}^{(s+1)}$ из выражения

   $$\left(\frac{2}{h_1^2}+\frac{2}{h_2^2}\right) y_{11}^{(s+1)} = \frac{y_{2,1}^{(s)}}{h_1^2} + \frac{y_{1,2}^{(s)}}{h_2^2} - f_{11}.$$

   Видно, что $y_{11}^{(s+1)}$ находится по явной формуле.

2. Затем находим $y_{1j}^{(s+1)}$, $j=\overline{1,(N_2-1)}$:

   $$\left(\frac{2}{h_1^2}+\frac{2}{h_2^2}\right) y_{1j}^{(s+1)} = \frac{y_{2,j}^{(s)}}{h_1^2} + \frac{y_{1,j-1}^{(s+1)}+y_{1,j+1}^{(s)}}{h_2^2} - f_{1j},$$

   вычисляя значения $y_{1,j}^{(s+1)}$ последовательно, начиная с $j=2$.

3. Аналогично находим $y_{2j}^{(s+1)}$, $j=\overline{1,(N_2-1)}$, используя уже вычисленные на предыдущем шаге значения $y_{1j}^{(s+1)}$, $j=\overline{1,(N_2-1)}$.

4. Продолжая таким же образом вычисления, получим все значения $y_{ij}^{(s+1)}$, начиная с узла x_{11} и заканчивая узлом x_{N_1-1,N_2-1}, по явным формулам.

Попеременно-треугольный итерационный метод

Запишем систему уравнений (1) в виде

$$Ay = \varphi,$$

обозначив за A матрицу системы (1), за φ — ее правую часть; y — численное решение. Нетрудно проверить, что матрица A — самосопряженная и положительно определенная:

$$A = A^* > 0.$$

Как и в §8 главы 1, представим матрицу A в виде суммы

$$A = R_1 + R_2,$$

где матрица R_1 имеет нижнетреугольную форму, а матрица R_2 — верхнетреугольную. Запишем попеременно-треугольный итерационный метод:

$$(E + \omega R_1)(E + \omega R_2)\frac{y^{(s+1)} - y^{(s)}}{\tau} + Ay^{(s)} = \varphi, \quad x \in \omega_h,\ s \in \mathbb{Z}_+,\ y^{(0)} \text{ — задано},$$

где действие матриц R_1 и R_2 на сеточную функцию y определяется равенствами

$$(R_1 y)_{ij} = \frac{y_{ij} - y_{i-1,j}}{h_1^2} + \frac{y_{ij} - y_{i,j-1}}{h_2^2},$$

$$(R_2 y)_{ij} = \frac{y_{ij} - y_{i+1,j}}{h_1^2} + \frac{y_{ij} - y_{i,j+1}}{h_2^2}.$$

Заметим, воспользовавшись результатами, полученными в §8 главы 1, что этот итерационный метод сходится при $\omega > \frac{\tau}{4}$, а каждую следующую итерацию данного метода, как и в рассмотренных ранее методах, можно вычислить по явным формулам.

В заключение отметим, что при $h_1 = h_2 = h$ количество итераций n_0, необходимое для получения решения с заданной точностью $\varepsilon > 0$ обратно пропорционально шагу h:

$$n_0(\varepsilon) = \mathrm{O}\big(h^{-1}\big).$$

Таким образом, попеременно-треугольный итерационный метод на порядок быстрее методов Якоби и Зейделя и тем самым является эффективным методом, широко применяемым в настоящее время при численном решении разностной задачи Дирихле.

§9 Основные понятия теории разностных схем: аппроксимация, устойчивость, сходимость

Рассмотрим задачу:

$$L(u(x)) = f(x), \quad x \in G, \tag{1}$$

где L – линейный дифференциальный оператор, G – произвольная, в общем случае многомерная, область.

Произвольную линейную дифференциальную задачу, в том числе, содержащую краевые и начальные условия, можно привести к виду (1).

Мы рассматриваем алгоритмы решения только для корректно поставленных задач. Задача (1) называется корректно поставленной, если

1. Решение $u(x)$ рассматриваемой задачи существует и единственно.

2. Решение $u(x)$ непрерывно зависит от входных данных, к которым относится правая часть уравнения, краевые и начальные условия.

Построим в области G сетку G_h с некоторым обобщенным шагом $h = \max\limits_i h_i$ — максимальным среди всех используемых на сетке G_h шагов h_i. Заметим, что при $h \to 0$ число узлов в данной сетке стремится к бесконечности.

Поставим в соответствие непрерывным функциям $u(x)$ и $f(x)$ их дискретные аналоги $y_h(x)$ и $\varphi_h(x), x \in G_h$ соответственно. Пусть также оператор L_h является дискретным аналогом оператора L. Тогда поставим в соответствие исходной дифференциальной задаче ее дискретный аналог:

$$L_h y_h(x) = \varphi_h(x), \quad x \in G_h. \tag{2}$$

Замечание. *Заметим, что дискретных аналогов каждого дифференциального уравнения вида* (1) *существует бесконечно много, поэтому при выборе такого аналога необходимо руководствоваться потребностями конкретной задачи и выбирать приближение оптимальным для данного случая образом.*

Определение. *Дискретная задача вида* (2) *называется разностной схемой для уравнения* (1).

Замечание. *Заметим, что разностная схема является, фактически, системой линейных алгебраических уравнений, и при малых значениях шага h является системой высоких порядков.*

При рассмотрении приближений заданных в непрерывной области задач их дискретными аналогами встает вопрос о том, каким образом измерять близость решений обеих задач. Существует два способа измерения близости решений уравнений (1) и (2).

Введем два линейных пространства: B_0 – пространство непрерывных функций, которое удовлетворяет уравнению (1), и B_h – пространство дискретных функций, удовлетворяющих уравнению (2). Близость функций из пространств B_0 и B_h можно измерять:

1. В норме пространства B_h.

   Пусть P_h – оператор проектирования пространства B_0 на пространство B_h. Тогда близость функций $u(x)$ и $y_h(x)$ будем измерять в норме $\|\cdot\|_h$ пространства B_h. Если

   $$\|P_h(u(x)) - y_h(x)\|_h \to 0,$$

   то в этой норме $y_h(x)$ сходится к $u(x)$.

2. В норме пространства B_0 с помощью восстановления дискретной функции $y_h(x)$ до непрерывной функции пространства B_0.

Замечание. *При рассмотрении норм в пространствах B_0 и B_h необходимо учесть, что нормы в этих пространствах должны быть согласованными, то есть должен существовать предел*

$$\lim_{h \to 0} \|P_h(u(x))\|_h = \|u\|_0, \tag{3}$$

где $\|\cdot\|_0$ — норма в пространстве B_0. Как будет показано после доказательства теоремы Филиппова, согласованность норм гарантирует сходимость решения разностной схемы именно к решению исходного уравнения — если согласованность не выполнена, мы можем получить сходящуюся разностную схему, но сходится она будет не к решению исходного уравнения.

Пример. Рассмотрим область $G = \{x : 0 \leqslant x \leqslant 1\}$ и зададим на этой области равномерную сетку с шагом h и числом узлов, равным $N \in \mathbb{N}$:

$$G_h = \{x_i = ih, \quad i = \overline{0, N}, \quad hN = 1\}.$$

Введем нормы в пространствах B_0 и B_h. В пространстве B_0:

$$\|u\|_0 = \|u\|_c = \max_{0 \leqslant x \leqslant 1} |u(x)|, \quad u(x) \in B_0.$$

В пространстве B_h:

$$\forall y_h = (y_0, y_1, \ldots, y_n) \in B_h \quad \|y_h\|_h = \|y_h\|_c = \max_{0 \leqslant i \leqslant n} |y_i|.$$

Для введенных таким образом норм выполняется условие согласованности.

Рассмотрим еще несколько примеров:

1. Пусть B_0 — пространство функций, интегрируемых с квадратом. Введем норму

$$\|u\|_0 = \left(\int_0^1 u^2(x) dx \right)^{\frac{1}{2}} = \|u\|_{L_2}, \quad u(x) \in B_0.$$

Тогда рассмотрим в дискретном пространстве B_h следующую норму:

$$\|y_h\|_h = \left(\sum_{i=0}^{N} y_i^2 h \right)^{\frac{1}{2}} = \|y_h\|_{L_2(G_h)}, \quad y_h \in B_h.$$

Для этих норм выполнено условие согласованности.

2. Введем норму в дискретном пространстве B_h

$$\|y_h\|_h = \left(\sum_{i=0}^{N} y_i^2 \right)^{\frac{1}{2}}, \quad y_h \in B_h.$$

Докажем, что эта норма не согласована ни с одной нормой пространства B_0.

Пусть $u(x) \in B_0 : u(x) \equiv 1$. Тогда

$$\|P_h(u(x))\|_h = \left(\sum_{i=0}^{N} 1 \right)^{\frac{1}{2}} = \sqrt[2]{N+1}.$$

Следовательно, при $h \to 0$

$$\lim_{h \to 0} \|P_h(u(x))\|_h = \infty.$$

Очевидно, что $\|u\|_0$ не может равняться бесконечности, так как норма должна быть конкретным числом. Значит норма $\|\cdot\|_h$ не согласована ни с одной нормой пространства B_0.

Рассмотрим подробнее оператор проектирования P_h пространства B_0 на пространство B_h. Ранее в этом параграфе мы предполагали, что этот оператор задан следующим образом:

$$(P_h(u))(x_i) = u(x_i), \quad x_i \in G_h, \quad u(x) \in B_0.$$

Однако оператор проектирования можно ввести бесконечным числом способов. Например, оператор проектирования

$$(P_h(u))(x_i) = \frac{1}{h}\int\limits_{x_i-0.5h}^{x_i+0.5h} u(x)dx, \quad u(x) \in B_0,$$

задает среднее значение функции $u(x)$ в узле x_i, $i = \overline{1, N}$. Значения оператора в граничных точках области G_h определяются следующим образом:

$$(P_h(u))(x_0) = \frac{1}{0.5h}\int\limits_{0}^{0.5} u(x)dx,$$

$$(P_h(u))(x_N) = \frac{1}{0.5h}\int\limits_{1-0.5h}^{1} u(x)dx.$$

Перейдем к основным понятиям теории разностных схем.

Определение. *Сеточная функция*

$$z_h(x) = y_h(x) - u_h(x) = y_h(x) - P_h(u(x)), \quad y_h(x) \in B_h, u(x) \in B_0, \tag{4}$$

называется погрешностью разностной схемы (2).

Подставим выражение $y_h(x) = z_h(x) + u_h(x)$ в разностную схему (2) и получим разностную схему относительно погрешности $z_h(x)$:

$$L_h z_h(x) = \psi_h(x), \quad x \in G_h, \tag{5}$$

где

$$\psi_h(x) = \varphi_h(x) - L_h u_h(x). \tag{6}$$

Определение. *Сеточная функция, задаваемая соотношением* (6), *называется погрешностью аппроксимации на решении исходной задачи* (1).

Замечание. *Погрешность аппроксимации можно представить в виде суммы погрешности приближения оператора и погрешности приближения правой части.*

Определение. *Разностная схема* (2) *аппроксимирует исходную задачу, если*

$$\lim_{h\to 0}\|\psi_h\|_h = 0.$$

Определение. *Разностная схема* (2) *имеет k-ый порядок аппроксимации, если существуют положительные константы M_1 и k, не зависящие от шага h, такие, что*

$$\|\psi_h\|_h \leqslant M_1 h^k$$

для достаточно малых h.

Определение. *Разностная задача* (2) *называется корректно поставленной, если при достаточно малых h выполнено:*

1. *При любых погрешностях аппроксимации ψ_h и при любых правых частях φ_h решение задачи* (2) *$y_h(x)$ существует и единственно.*

2. *Существует константа M_2, не зависящая от шага h, для которой выполнена априорная оценка:*

$$\|\psi_h\|_h \leqslant M_2\|\varphi_h\|_h.$$

Это оценка означает устойчивость в норме $\|\cdot\|_h$ решения разностной схемы по правой части уравнения.

Замечание 1. *Свойства существования и единственности решения задачи определяют существование оператора L_h^{-1}.*

Замечание 2. *Второе условие корректности постановки задачи означает равномерную ограниченность по h оператора L_h^{-1}.*

Определение. *Решение разностной задачи (2) сходится к решению исходной дифференциальной задачи (1), если*

$$\lim_{h\to 0}\|z_h\|_h = \lim_{h\to 0}\|y_h - u_h\|_h = 0.$$

Определение. *Разностная схема (2) имеет k-ый порядок точности, если существует константа M_3, не зависящая от шага h, такая, что*

$$\|z_h\|_h \leqslant M_3 h^k.$$

Теорема 1. *(Филиппова). Пусть исходная задача (1) и разностная схема (2) поставлены корректно, и пусть разностная схема аппроксимирует исходную задачу. Тогда решение разностной задачи сходится к решению исходной задачи, причем порядок точности разностной схемы совпадает с порядком аппроксимации.*

Доказательство. Рассмотрим задачу для погрешности разностной схемы (5). Так как по условию разностная схема корректна, то выполнено условие

$$\|y_h\|_h \leqslant M_2\|\varphi_h\|_h,$$

где константа M_2 не зависит от h. Заметим, что задача для погрешности $z_h \in B_h$ тоже является корректно поставленной, значит, для погрешности выполнена аналогичная оценка:

$$\|z_h\|_h \leqslant M_2\|\psi_h\|_h.$$

Так как разностная схема аппроксимирует исходную задачу, то

$$\lim_{h\to 0}\|\psi_h\|_h = 0.$$

Пусть разностная схема имеет k-ый порядок точности:

$$\|\psi_h\|_h \leqslant M_1 h^k,$$

где константа M_1 не зависит от h. Тогда аналогичная оценка справедлива и для погрешности z_h:

$$\|z_h\|_h \leqslant M_3 h^k,$$

где $M_3 = M_2M_1 > 0$, не зависит от h. Это и означает, что разностная схема (2) сходится с k-ым порядком точности по h. □

Замечание. *При доказательстве теоремы условие согласованности норм (3) не использовалось. Это условие нужно для того, чтобы гарантировать единственность предельной функции.*

Покажем, что возможно отсутствие сходимости решения задачи (2) к исходной задаче (1), если нормы пространств B_0 и B_h не согласованы. Пусть

$$\lim_{h\to 0} \|y_h - u_h\|_h = 0.$$

Мы хотим выяснить, существует ли какая-либо другая функция $v(x) \in B_0$, для которой

$$\lim_{h\to 0} \|y_h - v_h\|_h = 0.$$

Покажем, что если норма $\|\cdot\|_0$ пространства B_0 согласована с нормой $\|\cdot\|_h$ пространства B_h, то функция $v(x)$, если существует, то тождественно совпадает с функцией $u(x)$. Оценим норму $\|u_h - v_h\|_h$:

$$\|u_h - v_h\|_h = \|y_h - v_h - (y_h - u_h)\|_h \leqslant \|y_h - u_h\|_h + \|y_h - v_h\|_h.$$

Так как

$$\lim_{h\to 0} \|y_h - u_h\|_h = 0,$$

$$\lim_{h\to 0} \|y_h - v_h\|_h = 0,$$

то

$$\lim_{h\to 0} \|u_h - v_h\|_h = 0.$$

В случае согласованности норм отсюда следует, что

$$\|u(x) - v(x)\|_0 = 0,$$

а это значит, что функции $u(x)$ и $v(x)$ совпадают: $u(x) \equiv v(x)$, и решение разностной схемы (2) y_h однозначно определяет решение исходной задачи (1).

Этот факт гарантируется согласованностью норм пространств B_0 и B_h. Что произойдет в случае отсутствия такой согласованности, заранее предсказать невозможно.

Глава 5

Методы решения обыкновенных дифференциальных уравнений и систем ОДУ

§1 Постановка задачи Коши и примеры численных методов решения задачи Коши

Рассмотрим задачу Коши для системы нелинейных обыкновенных дифференциальных уравнений первого порядка:

$$\begin{cases} \dfrac{d\boldsymbol{u}}{dt} = \boldsymbol{f}(t, \boldsymbol{u}(t)), \quad t > 0, \\ \boldsymbol{u}(0) = u_0, \end{cases} \tag{1}$$

где $\boldsymbol{u}(t) = (u_1(t), u_2(t), \ldots, u_m(t))^T$, $\boldsymbol{f}(t, \boldsymbol{u}(t)) = (f_1(t, \boldsymbol{u}(t)), \ldots, f_m(t, \boldsymbol{u}(t))^T$. Считаем, что задача поставлена корректно, то есть решение $\boldsymbol{u}(t)$ существует, единственно и обладает достаточной гладкостью в области $t \geqslant 0$.

Введем норму в пространстве вектор-функций:

$$\|\boldsymbol{u}(t)\| = \sqrt{u_1^2(t) + u_2^2(t) + \ldots + u_m^2(t)}.$$

Замечание 1. *Существование и единственность решения задачи Коши (1) можно гарантировать лишь локально в окрестности точки $(0, u_0)$, поэтому, вообще говоря, мы рассматриваем данную задачу в области $(t, \boldsymbol{u}(t)) \in G = \{0 \leqslant t \leqslant a, \|\boldsymbol{u}(t) - u_0\| \leqslant b,\ a, b \in \mathbb{R}\}$, при этом необходимо, чтобы вектор-функция $\boldsymbol{f}(t, u)$ удовлетворяла условию Липшица по переменной и в этой области с некоторой константой $L > 0$. В дальнейшем мы не будем акцентировать внимание на этом моменте.*

Замечание 2. *Нелинейность задачи Коши (1) порождена только правой частью $\boldsymbol{f}(t, \boldsymbol{u}(t))$. В таких случаях нелинейность называют слабой.*

Замечание 3. *В этом и последующих параграфах мы для простоты будем рассматривать задачу Коши (1) для одного уравнения и одной искомой функции, если не указано иное. Рассуждения для произвольного количества уравнений в системе ОДУ (1) производятся чаще всего аналогично одномерному случаю. Таким образом, рассматривается следующая задача Коши (с соответствующими ограничениями на функции $f(t, u)$ и $u(t)$,*

требуемыми для существования и единственности решения задачи):

$$\begin{cases} \dfrac{du}{dt} = f(t, u(t)), \quad t > 0, \\ u(0) = u_0, \end{cases} \tag{2}$$

Проинтегрируем уравнение (2), учтем начальное условие $u(0) = u_0$ и получим интегральное уравнение относительно искомой функции $u(t)$:

$$u(t) = u_0 + \int\limits_0^t f(t, u(x))dx.$$

На этом представлении основан метод Пикара приближенного решения задачи (2):

$$u_{n+1}(t) = u_0 + \int\limits_0^t f(t, u_n(x))dx, \quad n \in \mathbb{Z}_+. \tag{3}$$

Этот наиболее очевидный итерационный метод численного решения задачи Коши (2) обладает рядом существенных недостатков, которые не позволяют использовать данный метод на практике: на каждой итерации необходимо вычисление интеграла, взятие которого аналитическими методами практически невозможно, кроме того, итерационный метод, записанный в форме (3) (метод Пикара) сходится медленно. Поэтому на практике для решения ОДУ и систем ОДУ обычно применяют другие численные методы. Эти методы можно разделить на две группы:

1. Методы Рунге–Кутта.
2. Многошаговые разностные методы (эти методы мы рассмотрим подробнее в следующих параграфах).

Рассмотрим несколько примеров методов решения задачи (2). Введем равномерную сетку в области $t \geqslant 0$ с шагом $\tau > 0$:

$$\omega_\tau = \{t_n = n\tau,\ \tau > 0,\ n \in \mathbb{Z}_+\}.$$

Узлы t_n этой сетки будем иногда называть *слоями*.

Рассмотрим сеточную функцию $y_n = y(t_n)$, заданную на сетке ω_τ. Пусть значения этой функции в узлах сетки y_n приближают значения $u_n = u(t_n)$. Обозначим $f_n = f(t_n, y_n)$.

Пример 1. Пожалуй, наиболее простым методом решения задачи (2) является разностная схема (метод) Эйлера. Несмотря на всю простоту схемы, метод Эйлера часто используется на практике.

Замечание. *В данном примере мы рассмотрим только явную схему Эйлера, но нужно помнить, что существуют и неявный аналог этой схемы.*

$$\begin{cases} \dfrac{y_{n+1} - y_n}{\tau} = f(t_n, y_n), \quad t_n \in w_\tau \\ y_0 = u_0, \quad n \in \mathbb{Z}_+. \end{cases} \tag{4}$$

Эта схема является явной, так как значение численного решения в каждой следующей точке $t_{n+1}, n \in \mathbb{Z}_+$ находится по явной формуле:

$$y_{n+1} = y_n + \tau f_n, \quad n \in \mathbb{Z}_+.$$

Введем погрешность разностной схемы (4):

$$z_n = y_n - u_n, n \in \mathbb{Z}_+.$$

Если мы получим оценку $\|z_n\| \leqslant M\tau$, где константа M не зависит от τ, то будем говорить, что решение разностной схемы Эйлера сходится к решению исходного уравнения (2) с первым порядком точности по τ.

Запишем теперь погрешность аппроксимации разностной схемы (4) на решении исходной задачи (2):

$$\psi_n = -\frac{u_{n+1} - u_n}{\tau} + f(t_n, u_n). \tag{5}$$

Разложим u_{n+1} в ряд Тейлора в узле t_n:

$$u_{n+1} = u_n + \tau u_n + \mathrm{O}\big(\tau^2\big).$$

Тогда

$$\frac{u_{n+1} - u_n}{\tau} = u'_n + \mathrm{O}\big(\tau\big).$$

Подставим последнее выражение в уравнение (5):

$$\psi_n = -u'_n + f(t_n, u_n) + \mathrm{O}\big(\tau\big).$$

Воспользовавшись тем, что $-u'_n + f(t_n, u_n) = 0$, так как выполнено исходное уравнение (2), окончательно получаем:

$$\psi_n = \mathrm{O}\big(\tau\big).$$

Эта оценка означает, что разностная схема (4) аппроксимирует исходную задачу с первым порядком точности по τ. В дальнейшем мы покажем, что рассмотренная разностная схема будет сходиться к решению задачи (2) с первым порядком по τ.

Пример 2. Рассмотрим теперь двухэтапную разностную схему Рунге–Кутта (схему «предиктор–корректор»). В данной разностной схеме вводятся дополнительные точки, так называемые полуцелые слои:

$$t_{n+\frac{1}{2}} = t_n + 0.5\tau, n \in \mathbb{Z}_+.$$

Нахождение численного решения данной разностной схемы в каждой следующей точке t_{n+1} производится в два этапа:

$$t_n \longrightarrow t_{n+\frac{1}{2}} \longrightarrow t_{n+1}.$$

Выполним первый этап («предиктор») по схеме Эйлера:

$$\frac{y_{n+\frac{1}{2}} - y_n}{0.5\tau} = f(t_n, y_n). \tag{6}$$

Рассмотрим второй этап («корректор»):

$$\frac{y_{n+1} - y_n}{\tau} = f(t_{n+\frac{1}{2}}, y_{n+\frac{1}{2}}), \tag{7}$$

где $y_0 = u_0$, $n \in \mathbb{Z}_+$. Отсюда следует, что

$$y_{n+1} = y_n + \tau f(t_{n+\frac{1}{2}}, y_n + 0.5\tau f(t_n, y_n)). \tag{8}$$

Далее будет показано, что эта двухэтапная разностная схема имеет второй порядок точности по τ.

Оценка погрешности общего двухэтапного метода Рунге–Кутта.

Рассмотрим общий вид двухэтапного метода Рунге–Кутта:

$$\begin{cases} \dfrac{y_{n+1}-y_n}{\tau} = \sigma_1 K_1 + \sigma_2 K_2, \quad n \in \mathbb{Z}_+ \\ y_0 = u_0, \\ K_1 = f(t_n, y_n), \quad K_2 = f(t_n + a_2\tau, y_n + b_{21}\tau f(t_n, y_n)), \end{cases} \tag{9}$$

где $\sigma_1, \sigma_2, a_2, b_{21} \in \mathbb{R}$ — некоторые числа, от выбора которых зависит как погрешность аппроксимации, так и точность численного решения.

Подставим значения K_1 и K_2 в первое уравнение системы (9):

$$\frac{y_{n+1}-y_n}{\tau} = \sigma_1 f(t_n, y_n) + \sigma_2 f(t_n + a_2\tau, y_n + b_{21}\tau f(t_n, y_n)).$$

Тогда можем записать погрешность аппроксимации разностной схемы (9) на решении задачи (2):

$$\psi_n = -\frac{u_{n+1}-u_n}{\tau} + \sigma_1 f(t_n, u_n) + \sigma_2 f\left(t_n + a_2\tau, u_n + b_{21}\tau f(t_n, u_n)\right). \tag{10}$$

Разложим u_{n+1} в ряд Тейлора в окрестности точки (t_n, u_n):

$$\frac{u_{n+1}-u_n}{\tau} = u'_n + \frac{\tau}{2}u''_n + \mathrm{O}(\tau^2).$$

Далее разложим $f(t_n + a_2\tau, u_n + b_{21}\tau f_n)$ в окрестности точки (t_n, u_n):

$$f(t_n + a_2\tau, u_n + b_{21}\tau f(t_n, u_n)) = f(t_n, u_n) + a_2\tau\frac{\partial f_n}{\partial t} + b_{21}\tau f_n\frac{\partial f_n}{\partial u} + \mathrm{O}(\tau^2).$$

Заметим, что

$$u'' = \frac{d}{dt}\left(\frac{du}{dt}\right) = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial u}.$$

Тогда погрешность аппроксимации ψ_n принимает вид:

$$\begin{aligned} \psi_n = &- u'_n - 0.5\tau\left(\frac{\partial f_n}{\partial t} + \frac{\partial f_n}{\partial u}\right) + \mathrm{O}(\tau^2) + \sigma_1 f(t_n, u_n) + \\ &+\sigma_2\left(f(t_n, u_n) + \tau a_2\frac{\partial f_n}{\partial t} + \tau b_{21}\frac{\partial f_n}{\partial u}\right) + \mathrm{O}(\tau^2). \end{aligned}$$

Сгруппируем слагаемые следующим образом:

$$\psi_n = -u'_n + (\sigma_1 + \sigma_2)f(t_n, u_n) + \tau\left((a_2\sigma_2 - 0.5)\frac{\partial f_n}{\partial t} + (b_{21}\sigma_2 - 0.5)\frac{\partial f_n}{\partial u}\right) + \mathrm{O}(\tau^2).$$

Чтобы получить оценку погрешности аппроксимации ψ_n со вторым порядком по τ, необходимо избавиться от слагаемых, содержащих τ в первой степени. Для этого потребуем выполнение следующих условий:

1. $\sigma_1 + \sigma_2 = 1$ (это условие называется условием аппроксимации).
2. $\sigma_2 a_2 = \sigma_2 b_{21} = 0.5$.

Тогда погрешность аппроксимации этого метода имеет второй порядок малости по τ:

$$\psi_n = \mathrm{O}\big(\tau^2\big).$$

В записи общего метода Рунге–Кутта используется большое количество параметров, что обеспечивает широту класса описываемых этим методом разностных схем. Однако в двухэтапном методе Рунге–Кутта не имеет смысла пользоваться двумя параметрами σ_1 и σ_2, так наилучшая оценка погрешности метода достигается при $\sigma_1 + \sigma_2 = 1$, поэтому, как правило, в двухэтапном методе Рунге–Кутта выбирают один параметр $\sigma = \sigma_2$, тогда $\sigma_1 = 1 - \sigma$. Если положить $a = a_2 = b_{12}$, то двухэтапный метод Рунге–Кутта запишется, как однопараметрическое по σ семейство разностных схем вида:

$$\frac{y_{n+1} - y_n}{\tau} = (1 - \sigma)K_1 + \sigma K_2,$$

где $K_1 = f(t_n, y_n), \quad K_2 = f(t_n + a\tau, y_n + a\tau f(t_n, y_n))$.

Пример. Рассмотрим примеры разностных схем, являющихся частными случаями общего двухэтапного метода Рунге–Кутта.

1. При $\sigma = 1, \quad a = a_2 = 0.5, \quad b = b_{21} = 0.5$ мы получим схему Рунге–Кутта «предиктор–корректор» (8), которую мы уже рассматривали. Точность этой схемы равна $\mathrm{O}\big(\tau^2\big)$.

2. Если положить $\sigma = 0.5, \quad a = 1, \quad b = 1$, то мы получим симметричную разностную схему:
$$\begin{cases} \dfrac{y_{n+1} - y_n}{\tau} = 0.5\left(f(t_n, y_n) + f(t_{n+1}, y_{n+1})\right), \quad n \in \mathbb{Z}_+ \\ y_0 = u_0. \end{cases} \tag{11}$$
Эта разностная схема является очень эффективной, имеет второй порядок точности по τ и является наилучшей для интегрирования жестких систем ОДУ (с понятием жестких систем мы познакомимся в одном из следующих параграфов).

Оценка точности на примере двухэтапного метода Рунге–Кутта.

Выпишем еще раз разностную схему, описывающую общий двухэтапный метод Рунге–Кутта:

$$\begin{cases} \dfrac{y_{n+1} - y_n}{\tau} = (1 - \sigma)f(t_n, y_n) + \sigma f(t_n + a\tau, y_n + a\tau f(t_n, y_n)), \quad n \in \mathbb{Z}_+ \\ y_0 = u_0. \end{cases} \tag{12}$$

Введем погрешность разностной схемы (12):

$$z_n = y_n - u_n, n \in \mathbb{Z}.$$

Подставим выражение для погрешности в разностную схему (12) и получим задачу для нахождения функции z_n:

$$\begin{cases} \dfrac{z_{n+1} - z_n}{\tau} = (1 - \sigma)f_n + \sigma f(t_n + a\tau, y_n + a\tau f_n) - \dfrac{u_{n+1} - u_n}{\tau}, \quad n \in \mathbb{Z}_+ \\ z_0 = 0, \end{cases} \tag{13}$$

где $f_n = f(t_n, y_n)$.

Для доказательства сходимости решения разностной схемы (12) к решению исходной задачи Коши (2) достаточно показать, что

$$\lim_{n\to 0} |z_n| = 0.$$

Покажем, что $|z_n| \leqslant M|\psi_n|$, $n \in \mathbb{Z}_+$, где константа M не зависит от шага τ, ψ_n — погрешность аппроксимации на решении исходной задачи (2):

$$\psi_n = -\frac{u_{n+1}-u_n}{\tau} + (1-\sigma)f(t_n,u_n) + \sigma f(t_n + a\tau, u_n + a\tau f(t_n,u_n)).$$

Перепишем задачу (13) в эквивалентном виде, сформировав погрешность аппроксимации путем добавления недостающих слагаемых:

$$\begin{aligned}\frac{z_{n+1}-z_n}{\tau} = -\frac{u_{n+1}-u_n}{\tau} + (1-\sigma)f(t_n,u_n) + \sigma f(t_n + a\tau, u_n + a\sigma f(t_n,u_n)) + \\ +(1-\sigma)(f(t_n,y_n) - f(t_n,u_n)) + \\ +\sigma\big(f(t_n + a\tau, y_n + a\tau f(t_n,y_n)) - f(t_n + a\tau, u_n + a\tau f(t_n,u_n))\big) = \\ = \psi_n + \varphi_n^{(1)} + \varphi_n^{(2)},\end{aligned}$$

где

$$\varphi_n^{(1)} = (1-\sigma)(f(t_n,y_n) - f(t_n,u_n)),$$
$$\varphi_n^{(2)} = \sigma\big(f(t_n + a\tau, y_n + a\tau f(t_n,y_n)) - f(t_n + a\tau, u_n + a\tau f(t_n,u_n))\big).$$

Пусть функция $f(t,u)$ удовлетворяет условию Липшица по второму аргументу с константой $L > 0$:

$$|f(t,v) - f(t,u)| \leqslant L|u-v|, \quad (t,u),(t,v) \in G.$$

Замечание. *Требование липшицевости функции $f(t,u)$ естественно, так как является условием того, что решение исходной задачи (2) существует и единственно.*

Как правило, на практике выбирают $0 \leqslant \sigma \leqslant 1$, $a \geqslant 0$. Воспользуемся этими условиями и оценим выражения $\varphi_n^{(1)}$ и $\varphi_n^{(2)}$:

$$|\varphi_n^{(1)}| = (1-\sigma)|f(t_n,y_n) - f(t_n,u_n)| \leqslant (1-\sigma)L|y_n - u_n| = (1-\sigma)L|z_n|,$$

$$\begin{aligned}|\varphi_n^{(2)}| \leqslant \sigma L|y_n + a\tau f(t_n,y_n) - u_n - a\tau f(t_n,u_n)| \leqslant \\ \leqslant \sigma L\big(|y_n - u_n| + a\tau|f(t_n,y_n) - f(t_n,u_n)|\big) \leqslant \sigma L(|z_n| + a\tau L|z_n|) = \sigma L(1 + a\tau L)|z_n|.\end{aligned}$$

Пусть $\sigma a \leqslant 0.5$. Оценим сумму $|\varphi_n^{(1)}| + |\varphi_n^{(2)}|$:

$$|\varphi_n^{(1)}| + |\varphi_n^{(2)}| \leqslant (1-\sigma)L|z_n| + \sigma L(1 + a\tau L)|z_n| = L|z_n| + \sigma a\tau L^2|z_n| \leqslant L(1 + 0.5\tau L)|z_n|.$$

Приступим к получению оценки точности. Так как $f(t_n,u_n) = \psi_n + \varphi_n^{(1)} + \varphi_n^{(2)}$, то получаем

$$|z_{n+1}| \leqslant |z_n| + \tau|\psi_n| + \tau(|\varphi_n^{(1)}| + |\varphi_n^{(2)}|) = (1 + \tau L + 0.5\tau^2L^2)|z_n| + \tau|\psi_n|.$$

Заметим, что слагаемые в сумме $(1+\tau L+0.5\tau^2L^2)$ являются первыми членами разложения функции $e^{\tau L}$ по формуле Тейлора по переменной τ в окрестности нуля. Следовательно,

$$(1 + \tau L + 0.5\tau^2L^2) \leqslant e^{\tau L}.$$

Тогда

$$|z_{n+1}| \leqslant e^{\tau L}|z_n| + \tau|\psi_n|.$$

Введем обозначение $\rho = e^{\tau L}$. Тогда

$$|z_{n+1}| \leqslant \rho|z_n| + \tau|\psi_n|, \quad n \in \mathbb{Z}_+. \tag{14}$$

Раскроем полученное рекуррентное соотношение:

$$|z_{n+1}| \leqslant \rho^{n+1}|z_0| + \tau \sum_{j=0}^{n} \rho^{n-j}|\psi_j|.$$

Так как $z_0 = 0$, то получаем:

$$|z_{n+1}| \leqslant \max_{0 \leqslant j \leqslant n} |\psi_j| t_n e^{t_n L}. \tag{15}$$

Учтем, что $t_n \leqslant T$, тогда:

$$|z_{n+1}| \leqslant M \max_{0 \leqslant j \leqslant n} |\psi_j|,$$

где константа $M = Te^{TL} > 0$ не зависит от τ. Заметим, что

$$\lim_{\tau \to 0} |z_{n+1}| = 0,$$

так как $|\psi_j| \leqslant M_1(\tau^2)$ по доказанному выше. Тогда при достаточно малых τ получаем:

$$|z_{n+1}| = \mathrm{O}\big(\tau^2\big).$$

Это означает, что рассматриваемый общий двухэтапный метод Рунге–Кутта при выполнении соответствующих условий имеет квадратичную точность по τ, совпадающую с оценкой погрешности аппроксимации на решении исходного уравнения (2).

§2 Общий m-этапный метод Рунге–Кутта

Рассмотрим задачу Коши для нелинейного обыкновенного дифференциального уравнения первого порядка:

$$\begin{cases} \dfrac{du}{dt} = f(t, u(t)), \quad t > 0 \\ u(0) = u_0, \end{cases} \tag{1}$$

где функции $u(t)$ и $f(t, u)$ обладают достаточной гладкостью в соответствующих областях. Считаем, решение $u(t)$ существует и единственно.

Введем равномерную сетку в области $t \geqslant 0$ с шагом $\tau > 0$:

$$\omega_\tau = \{t_n = n\tau,\ \tau > 0,\ n \in \mathbb{Z}_+\}.$$

Рассмотрим сеточную функцию $y_n = y(t_n)$, заданную на сетке ω_τ. Пусть значения этой функции в узлах сетки y_n приближают значения $u_n = u(t_n)$. Обозначим $f_n = f(t_n, y_n)$.

Общая идея m-этапного метода Рунге–Кутта заключается в том, что для вычисления значения приближенного решения в каждой следующей точке t_{n+1} вводятся m дополнительных этапов. Промежуточные значения на каждом шаге $n \in \mathbb{Z}_+$ вычисляются по следующим формулам:

$$\begin{aligned} K_1 &= f(t_n, y_n), \\ K_2 &= f(t_n + a_2\tau, y_n + b_{21}\tau K_1), \\ K_3 &= f(t_n + a_3\tau, y_n + b_{31}\tau K_1 + b_{32}\tau K_2), \\ &\dots \\ K_m &= f(t_n + a_m\tau, y_n + b_{m1}\tau K_1 + b_{m2}\tau K_2 + \ldots + b_{mm-1}\tau K_{m-1}). \end{aligned}$$

При этом разностная схема для исходной задачи (1) имеет вид

$$\begin{cases} \dfrac{y_{n+1}-y_n}{\tau} = \sigma_1 K_1 + \sigma_2 K_2 + \ldots + \sigma_m K_m \\ y_0 = u_0, \quad n \in \mathbb{Z}_+, \end{cases} \tag{2}$$

где $\sigma_1, \sigma_2, \ldots, \sigma_m \in \mathbb{R}$.

Будем также считать, что выполнено следующее условие аппроксимации, без которого рассмотрение метода не имеет смысла:

$$\sum_{i=1}^{m} \sigma_i = 1.$$

Замечание. *Заметим, что формулы m-этапного метода Рунге–Кутта достаточно громоздки. Это является одной из причин того, что на практике редко используются методы Рунге–Кутта для $m > 4$.*

Приведем примеры трех- и четырех- этапных методов Рунге–Кутта, имеющих третий и четвертый порядок точности соответственно.

Пример 1. $m = 3$:

$$\frac{y_{n+1}-y_n}{\tau} = \frac{1}{6}(K_1 + 4K_2 + K_3),$$

где

$$\begin{aligned} K_1 &= f(t_n, y_n), \\ K_2 &= f(t_n + 0.5\tau, y_n + 0.5\tau K_1), \\ K_3 &= f(t_n + \tau, y_n - \tau K_1 + 2\tau K_2). \end{aligned}$$

Данная схема имеет третий порядок точности по τ.

Пример 2. $m = 4$:

$$\frac{y_{n+1}-y_n}{\tau} = \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4),$$

где

$$\begin{aligned} K_1 &= f(t_n, y_n), \\ K_2 &= f(t_n + 0.5\tau, y_n + 0.5\tau K_1), \\ K_3 &= f(t_n + 0.5\tau, y_n + 0.5\tau K_2), \\ K_4 &= f(t_n + \tau, y_n + \tau K_3). \end{aligned}$$

Данная схема имеет четвертый порядок точности по τ.

§3 Многошаговые разностные методы

Рассмотрим задачу Коши для нелинейного обыкновенного дифференциального уравнения первого порядка:

$$\begin{cases} \dfrac{du}{dt} = f(t, u(t)), \quad t > 0, \\ u(0) = u_0, \end{cases} \tag{1}$$

где функции $u(t)$ и $f(t, u)$ обладают достаточной гладкостью. Считаем, что решение $u(t)$ существует и единственно.

Введем равномерную сетку в области $t \geqslant 0$ с шагом $\tau > 0$:

$$\omega_\tau = \{t_n = n\tau,\ \tau > 0,\ n \in \mathbb{Z}_+\}.$$

Рассмотрим сеточную функцию $y_n = y(t_n)$, заданную на сетке ω_τ. Пусть значения этой функции в узлах сетки y_n приближают значения $u_n = u(t_n)$. Обозначим $f_n = f(t_n, y_n)$.

Определение. *Линейным m-шаговым разностным методом решения задачи (1) называется разностная схема вида*

$$\sum_{k=0}^{m} \frac{a_k}{\tau} y_{n-k} = \sum_{k=0}^{m} b_k f_{n-k}, \tag{2}$$

где $m \in \mathbb{N}$, $\tau > 0$ – шаг сетки ω_τ, $a_k, b_k \in \mathbb{R}$, $k = \overline{0, m}$, причем $a_0 \neq 0, b_m \neq 0$.

Замечание. *Уравнение (2) следует рассматривать как рекуррентное соотношение, выражающее новое значение $y_n = y(t_n)$ через найденные ранее значения $y_{n-1}, y_{n-2}, \ldots, y_{n-m}$. Уравнение (2) определено для $n = m, m+1, \ldots$ и требует для начала расчета задания m начальных значений $y_0, y_1, \ldots, y_{m-1}$. Значение $y_0 = u(0)$ определяется исходной задачей (1), а величины $y_1, \ldots, y_{m-1}$ можно вычислить с помощью других методов, например, с помощью рассмотренного выше метода Рунге–Кутта. В дальнейшем мы будем предполагать, что величины $y_0, y_1, \ldots, y_{m-1}$ уже заданы.*

Если в разностной схеме (2) $b_0 = 0$, то рассматриваемый метод называется явным, и искомое значение y_n выражается явным образом через предыдущие:

$$\frac{a_0}{\tau} y_n = \sum_{k=1}^{m} b_k f_{n-k} - \sum_{k=1}^{m} \frac{a_k}{\tau} y_{n-k}.$$

Если $b_0 \neq 0$, то метод называется неявным, и для нахождения y_n приходится решать нелинейное уравнение

$$\frac{a_0}{\tau} y_n - b_0 f(t_n, y_n) = F(y_{n-1}, \ldots, y_{n-m}),$$

где

$$F(y_{n-1}, \ldots, y_{n-m}) = \sum_{k=1}^{m} (b_k f_{n-k} - \frac{a_k}{\tau} y_{n-k}).$$

Обычно это уравнение решают итерационным методом Ньютона, выбирая начальное приближение равным y_{n-1} (этот метод мы рассматривали в §3 главы 3).

Заметим, что коэффициенты уравнения (2) определены с точностью до множителя. Для определенности будем считать, что выполнено условие

$$\sum_{k=0}^{m} b_k = 1.$$

Это означает, что правая часть разностного уравнения (2) аппроксимирует правую часть дифференциального уравнения (1).

Определение. *Погрешностью аппроксимации разностной схемы (2) на решении исходной задачи (1) называется сеточная функция*

$$\psi_n = -\sum_{k=0}^{m} \frac{a_k}{\tau} u_{n-k} + \sum_{k=0}^{m} b_k f(t_{n-k}, u_{n-k}), \tag{3}$$

заданная на сетке ω_τ, где $u_n = u(t_n)$ — решение исходной задачи (1).

Выясним вопрос о порядке погрешности аппроксимации при $\tau \to 0$ в зависимости от выбора коэффициентов $a_k, b_k,\ k = \overline{0,m}$. Будем предполагать в дальнейшем, что все рассматриваемые функции обладают необходимой гладкостью. Разложим u_{n-k} по формуле Тейлора в точке t_n:

$$u_{n-k} = u(t_n - k\tau) = \sum_{l=0}^{p} \frac{(-k\tau)^l}{l!} u^{(l)}(t_n) + \mathrm{O}\big(\tau^{p+1}\big).$$

Разложим правую часть исходного дифференциального уравнения в этой же точке:

$$f_{n-k} = f(t_n - k\tau) = u'(n - k\tau) = \sum_{l=0}^{p-1} \frac{(-k\tau)^l}{l!} u^{(l+1)}(t_n) + \mathrm{O}\big(\tau^{p}\big).$$

Подставим эти разложения в выражение (3) и получим

$$\psi_n = -\sum_{k=0}^{m} \frac{a_k}{\tau} \sum_{l=0}^{p} \frac{(-k\tau)^l}{l!} u^{(l)}(t_n) + \sum_{k=0}^{m} b_k \sum_{l=0}^{p-1} \frac{(-k\tau)^l}{l!} u^{(l+1)}(t_n) + \mathrm{O}\big(\tau^{p}\big).$$

Передвинем на единицу индекс суммирования l во втором слагаемом, а также домножим и поделим на l выражение, стоящее под знаком суммирования:

$$\psi_n = -\sum_{l=0}^{p} \sum_{k=0}^{m} \frac{a_k}{\tau} \frac{(-k\tau)^l}{l!} u^{(l)}(t_n) + \sum_{l=1}^{p} \sum_{k=0}^{m} b_k l \frac{(-k\tau)^{l-1}}{l(l-1)!} u^{(l)}(t_n) + \mathrm{O}\big(\tau^{p}\big).$$

Объединив две суммы под общим знаком суммирования (для этого необходимо выписать отдельно нулевое слагаемое первой суммы), получим

$$\psi_n = -\sum_{k=0}^{m} \frac{a_k}{\tau} u(t_n) + \sum_{l=1}^{p} \left(-\sum_{k=1}^{m} \frac{a_k}{\tau} \frac{(-k\tau)^l}{l!} u^{(l)}(t_n) + \sum_{k=1}^{m} l b_k \frac{(-k\tau)^{l-1}}{l!} u^{(l)}(t_n) \right) + \mathrm{O}\big(\tau^{p}\big).$$

После очевидных преобразований получаем:

$$\psi_n = -\sum_{k=0}^{m} \frac{a_k}{\tau} u(t_n) - \sum_{l=1}^{p} \sum_{k=0}^{m} \frac{(-k\tau)^{l-1}}{l!} u^{(l)}(t_n)(k a_k + l b_k) + \mathrm{O}\big(\tau^{p}\big).$$

Отсюда видно, что погрешность аппроксимации (3) имеет порядок p, если выполнены следующие условия:

$$\sum_{k=0}^{m} a_k = 0,$$

$$\sum_{k=0}^{m} k^{l-1}(k a_k + l b_k) = 0, \quad l = 1, 2, \ldots, p.$$

Вместе с условием нормировки

$$\sum_{k=0}^{m} b_k = 1$$

эти условия образуют систему из $(p+2)$-ух линейных алгебраических уравнений относительно $2(m+1)$ неизвестных $a_0, a_1, \ldots, a_m, b_0, b_1, \ldots, b_m$.

Полученную систему можно несколько упросить. Рассмотрим последние условия при $l = 1$:

$$\sum_{k=0}^{m}(ka_k + b_k) = 0,$$

$$\sum_{k=0}^{m} ka_k = -\sum_{k=0}^{m} b_k = -1,$$

то есть

$$\sum_{k=0}^{m} ka_k = -1.$$

Окончательно получаем следующую систему уравнений:

$$\begin{cases} \sum\limits_{k=1}^{m} ka_k = -1, \\ \sum\limits_{k=0}^{m} k^{l-1}(ka_k + lb_k) = 0, \quad l = \overline{2,p}, \end{cases} \tag{4}$$

в которой коэффициенты a_0, b_0 вычисляются по формулам

$$a_0 = -\sum_{k=1}^{m} a_k, \quad b_0 = 1 - \sum_{k=1}^{m} b_k.$$

Таким образом, мы уменьшили число уравнений в системе до p и число неизвестных до $2m$. Чтобы система не была переопределенной (в таких системах число уравнений больше числа неизвестных) необходимо выполнение условия $p \leqslant 2m$.

Таким образом наибольший возможный порядок аппроксимации неявных m-шаговых разностных методов равен $2m$, явных — $(2m-1)$, так как в явных методах $b_0 = 0$, и число неизвестных в системе (4) меньше на единицу по сравнению с системой, записанной для неявного метода.

Замечание 1. *Если убрать последние n уравнений системы (4), $n = \overline{1,(p-1)}$, то получим условия, обеспечивающие порядок погрешность аппроксимации $\mathrm{O}\big(\tau^{p-n}\big)$.*

Замечание 2. *В практике вычислений наибольшее распространение получили методы Адамса, которые представляют собой частный случай многошаговых методов (2), когда производная $u'(t)$ в исходном уравнении аппроксимируется по двум крайним точкам t_0 и t_n, то есть $a_0 = 1,\ a_1 = -1,\ a_k = 0,\ k = \overline{2,m}$:*

$$\frac{y_n - y_{n-1}}{\tau} = \sum_{k=0}^{m} b_k f_{n-k}.$$

Замечание 3. *Разностные схемы вида (2), обладающие наивысшими порядками аппроксимации на решении исходного уравнения, неустойчивы и не могут быть использованы на практике. Максимальный порядок аппроксимации устойчивого неявного m-шагового метода не превосходит $(m+1)$, если m нечетно, и не превосходит $(m+2)$, если m четно. Порядок аппроксимации устойчивых явных схем не превосходит m. Подробнее понятие устойчивости m-шагового разностного метода мы рассмотрим в следующем параграфе.*

В завершение рассмотрим достоинства и недостатки многошаговых разностных методов по сравнению с методом Рунге–Кутта.

Достоинства:

1. Формулы многошаговых методов значительно проще.
2. Многошаговые методы позволяют достигать большей точности.

Недостатки:

1. В многошаговых методах необходимо хранить в памяти большее число элементов — значения нескольких предыдущих шагов вместо одного.
2. Многошаговые методы требуют наличия «разгонного этапа», то есть значений нескольких первых шагов, которые нельзя вычислить по многошаговым формулам. Как мы уже упоминали, эти значения обычно вычисляют с помощью метода Рунге–Кутта.

§4 Понятие устойчивости разностного метода

Известно, что на практике вычисления проводятся приближенно, то есть при задании исходных данных и в процессе самих вычислений допускаются погрешности. Численный метод называется устойчивым, если погрешности, допущенные на каком-то этапе вычислений, не оказывают существенного влияния на результат. Разумеется, такого описательного определения недостаточно для исследования устойчивости конкретных алгоритмов. Существуют математически строгие и более узкие определения устойчивости, некоторые из них будут приведены в следующих параграфах. Сейчас ограничимся тем, что рассмотрим несколько характерных примеров.

Явление неустойчивости часто возникает в процессе решения разностных уравнений. Так, если решать уравнение

$$y_{n+1} = qy_n,$$

где $n \in \mathbb{Z}_+$, $q \in \mathbb{C}$ — некоторая константа, а y_0 — задано, то при $|q| > 1$ погрешность будет возрастать при переходе от шага n к шагу $(n+1)$. Действительно, пусть вместо y_n в результате ошибок округления получено значение

$$\widetilde{y}_n = y_n + \delta_n.$$

Тогда при вычислении y_{n+1} получим значение

$$\widetilde{y}_{n+1} = q\widetilde{y}_n = qy_n + q\delta_n = y_{n+1} + q\delta_n,$$

то есть погрешность $\delta_{n+1} = q\delta_n$ на новом шаге увеличится. В этом случае метод неустойчив, и при проведении расчетов на ЭВМ при достаточно большом n может произойти переполнение разрядной сетки. Если же $|q| \leqslant 1$, то погрешность, допущенная на каком-либо шаге вычислений, будет не возрастать на следующих шагах.

Процесс численного решения задачи Коши для обыкновенных дифференциальных уравнений также может оказаться неустойчивым. Поясним это на примере простого уравнения

$$\begin{cases} \dfrac{du(t)}{dt} + \lambda u(t) = 0, \quad \lambda = const > 0,\ t > 0, \\ u(0) = u_0. \end{cases} \tag{1}$$

Его решение ($u(t) = u_0 e^{-\lambda t}$) монотонно убывает с ростом t. В частности, для решения этого уравнения справедливо следующее неравенство:

$$|u(t)| \leqslant |u_0|, \quad t > 0, \tag{2}$$

означающее непрерывную зависимость (иначе говоря, устойчивость) решения уравнения (1) от начальных данных.

Естественно требовать, чтобы и для разностных схем, аппроксимирующих уравнение (1), выполнялись оценки, аналогичные (2). Однако такие оценки для разностных схем выполняются далеко не всегда.

Пример 1. Рассмотрим, например, явную разностную схему Эйлера для решения задачи (1):

$$\frac{y_{n+1} - y_n}{\tau} + \lambda y_n = 0,$$

где $\tau > 0$, $n \in \mathbb{Z}_+$, $y_0 = u_0$, и перепишем ее в виде

$$y_{n+1} = q y_n, \quad n \in \mathbb{Z}_+,$$

где $q = 1 - \tau\lambda$.
Тогда оценка

$$|y_{n+1}| \leqslant |y_n|, \quad n \in \mathbb{Z}_+$$

будет выполняться тогда и только тогда, когда $|q| \leqslant 1$, то есть при $\tau \leqslant \dfrac{2}{\lambda}$. В этом случае схема называется устойчивой, а само неравенство $\tau \leqslant \dfrac{2}{\lambda}$ называется условием устойчивости. Если оно нарушено, то $|q| > 1$, и погрешности, допущенные в процессе вычислений, будут возрастать с ростом n.

Определение. *Разностная схема называется абсолютно устойчивой, если эта схема устойчива при любых допустимых значениях своих параметров (то есть при значениях, при которых разностная схема определена).*

Пример 2. Приведем пример абсолютно устойчивой разностной схемы. Для уравнения

$$u'(t) = f(t, u(t)), \tag{3}$$

для некоторой функции $f(t, u)$, рассмотрим неявную схему Эйлера

$$\frac{y_{n+1} - y_n}{\tau} = f(t_{n+1}, y_{n+1}),$$

где $n \in \mathbb{Z}_+$, $y_0 = u_0$.

Схема называется неявной потому, что для нахождения y_{n+1} приходится решать уравнение

$$y_{n+1} - \tau f(t_{n+1}, y_{n+1}) = y_n.$$

Это уравнение можно решить, например, с помощью метода Ньютона, описанного в §3 главы 3. Для уравнения (1) неявная схема Эйлера принимает вид

$$\frac{y_{n+1} - y_n}{\tau} + \lambda y_{n+1} = 0,$$

откуда получаем

$$y_{n+1} = q y_n, \quad q = (1 + \tau\lambda)^{-1},$$

причем $|q| < 1$ при любых $\tau > 0$.

Приведенные выше примеры являются типичными, потому что, как правило, явные схемы устойчивы лишь при достаточно малых шагах τ, а среди неявных схем существуют абсолютно устойчивые.

Общий m-шаговый линейный разностный метод

Перейдем теперь от частных примеров к общему m-шаговому методу

$$\sum_{k=0}^{m} \frac{a_k}{\tau} y_{n-k} = \sum_{k=0}^{m} b_k f_{n-k}, \tag{4}$$

где $\tau > 0$, $y_0, y_1, \ldots, y_{m-1}$ — заданы. Будем считать, что коэффициенты $a_k,\ b_k,\ k = \overline{1,n}$ не зависят от τ.

Пример. В применении к уравнению (1) метод (4) принимает вид:

$$\sum_{k=0}^{m} (a_k + \tau\lambda b_k)\, y_{n-k} = 0. \tag{5}$$

Решение этого разностного уравнения с постоянными коэффициентами будем искать в виде

$$y_j = q^j, \quad j \in \mathbb{Z}_+.$$

Подставив эту формулу решения в уравнение (5) и сократив на q^{n-m}, придем к уравнению

$$F_m(q,\tau) = \sum_{k=0}^{m} (a_k + \tau\lambda b_k)\, q^{m-k} = 0. \tag{6}$$

Определение. *Уравнение вида* (6) *называется характеристическим уравнением разностной схемы* (5).

Можно было бы искать условия, при которых все корни уравнения (6) лежат внутри или на границе единичного круга. Однако это оказывается достаточно сложным даже для квадратного уравнения. Поэтому в случае общего m-шагового разностного метода (4) поступают по-другому.

Предположим, что шаг τ достаточно мал. Тогда корни уравнения (6) будут близки к корням уравнения

$$F_m(q,0) = 0,$$

то есть уравнения

$$\sum_{k=0}^{m} a_k q^{m-k} = 0, \tag{7}$$

которое также называется характеристическим. Заметим, что последнее уравнение определяется только способом аппроксимации производной $u'(t)$ и не зависит от того, каким способом аппроксимируется правая часть исходного уравнения (3).

При анализе m-шаговых разностных схем для нелинейного уравнения (3) обычно ограничиваются рассмотрением упрощенного характеристического уравнения (7).

Определение. *Говорят, что схема* (4) *удовлетворяет условию* (α)*, если все корни характеристического уравнения* (7) *лежат внутри или на границе единичного круга комплексной плоскости, причем на границе единичного круга нет кратных корней.*

Таким образом, выполнение условия (α) соответствует устойчивости разностного метода для уравнения $u'(t) = 0$. Однако часто схему и для общего уравнения (3) называют устойчивой, если она удовлетворяет условию (α). Такая терминологическая неточность оправдана тем, что из условия (α) следует сходимость решения разностной задачи (4) к решению исходной дифференциальной задачи (3). Приведем без доказательства следующую теорему.

Теорема. *Пусть разностная схема удовлетворяет условию* (α) *и* $|f'_u| \leqslant L$ *на отрезке* $0 \leqslant t \leqslant T$. *Тогда при* $0 \leqslant t_n = n\tau \leqslant T$ *и всех достаточно малых* τ *выполняется оценка*

$$|y_n - u(t_n)| \leqslant M \left(\sum_{j=m}^{n} \tau |\psi_j| + \max_{0 \leqslant i \leqslant m-1} |y_i - u(t_i)| \right),$$

где $|y_i - u(t_i)|$ *— погрешности в задании начальных данных,* $i = \overline{0, (m-1)}$, M *— константа, зависящая от* L, T *и не зависящая от* τ, ψ_j *— погрешность аппроксимации на решении исходного уравнения* (3)*:*

$$\psi_j = -\sum_{k=0}^{m} \frac{a_k}{\tau} u(t_{n-k}) + \sum_{k=0}^{m} b_k f_{n-k}.$$

Таким образом, исследование сходимости метода (4) сводится к анализу погрешности аппроксимации и проверке условия (α).

Замечание 1. *Методы Адамса*

$$\frac{y_n - y_{n-1}}{\tau} = \sum_{k=0}^{m} b_k f_{n-k}$$

всегда удовлетворяют условию (α), *так как для них* $a_0 = -a_1 = 1$, *то есть* $q = q_1 = 1$, *что следует из уравнения*

$$q^n - q^{n-1} = 0.$$

Замечание 2. *При указанном подходе, в отличие от рассмотренных примеров, не различаются абсолютно устойчивые и условно устойчивые разностные схемы, так как параметр* τ *заранее считается достаточно малым.*

Замечание 3. *Мы уже упоминали в §3 данной главы, наивысший достижимый порядок аппроксимации неявных* m*-шаговых методов равен* $2m$, *а явных —* $(2m-1)$. *Однако оказывается, что методы наивысшего порядка неустойчивы в том смысле, что они не удовлетворяют условию* (α). *А именно, если* m *нечетно, то никакой устойчивый метод не превосходит порядка* $p = m + 1$. *Если* m *четно, то никакой устойчивый метод не превосходит порядка* $p = m + 2$ *(*p *— порядок аппроксимации). Для явных схем наивысший порядок аппроксимации устойчивых методов* $p = m$.

Пример. Нетрудно привести пример схем, не удовлетворяющих условию (α). Так, явная двухшаговая схема

$$\frac{y_n + 4y_{n-1} - 5y_{n-2}}{6\tau} = \frac{2f_{n-1} + f_{n-2}}{3}$$

имеет третий порядок погрешности аппроксимации $\psi = \mathrm{O}(\tau^3)$ (чтобы убедиться в этом, достаточно проверить условия p-ого порядка аппроксимации, полученные в §3 текущей главы). Характеристическое уравнение (7) для этой схемы

$$q^2 + 4q - 5 = 0$$

имеет корни $q_1 = -5$, $q_2 = 1$, и, тем самым, условие (α) нарушено.

§5 Жесткие системы обыкновенных дифференциальных уравнений

Многие из рассмотренных выше методов интегрирования обыкновенных дифференциальных уравнений переносятся без изменений на системы дифференциальных уравнений. Однако в случае численного решения системы уравнений могут возникнуть дополнительные трудности, связанные с разномасштабностью процессов, описываемых данной системой. Поясним это на примере системы, состоящей из двух независимых уравнений

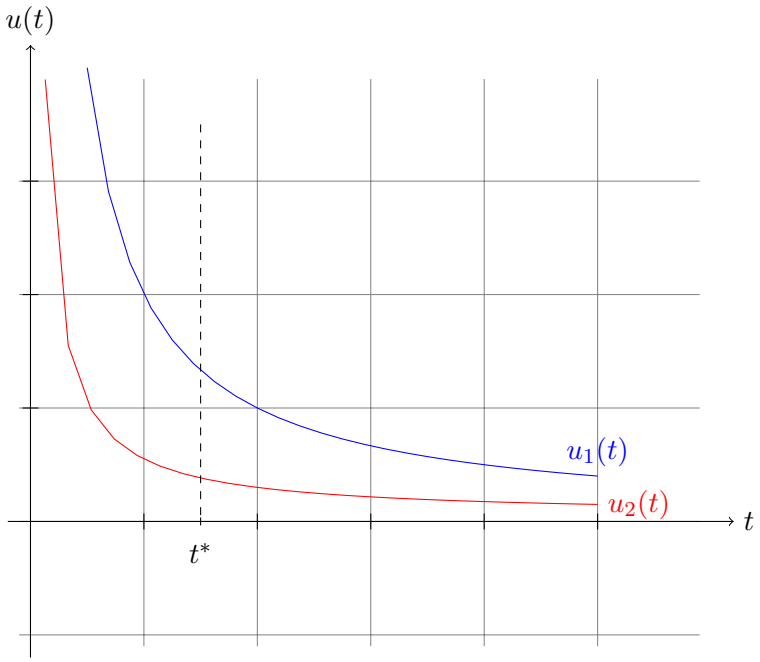
$$\begin{cases} \dfrac{du_1(t)}{dt} + a_1 u_1(t) = 0, \quad t > 0 \\ u_1(0) = u_{01}, \\ \dfrac{du_2(t)}{dt} + a_2 u_2(t) = 0, \quad t > 0 \\ u_2(0) = u_{02}, \end{cases} \tag{1}$$

где $a_1,\ a_2$ — положительные постоянные.

Система (1) имеет решение

$$u_1(t) = u_{01} e^{-a_1 t},$$
$$u_2(t) = u_{02} e^{-a_2 t},$$

монотонно убывающее с ростом t. Предположим, что a_2 гораздо больше, чем a_1. Тогда вторая компонента $u_2(t)$ затухает гораздо быстрее, чем первая и, начиная с некоторого момента времени t^*, поведение решения почти полностью определяется первой компонентой $u_1(t)$. Однако оказывается, что при решении системы (1) явным разностным методом шаг интегрирования τ определяется, как правило, компонентой $u_2(t) = u_{02} e^{-a_2 t}$, которая не существенна с точки зрения поведения решения системы.



Например, явный метод Эйлера

$$\frac{u_1^{n+1} - u_1^n}{\tau} + a_1 u_1^n = 0,$$

$$\frac{u_2^{n+1} - u_2^n}{\tau} + a_2 u_2^n = 0,$$

где $u_i^n = u_i(t_n),\ i = 1, 2$, будет устойчив, если шаг τ удовлетворяет одновременно двум неравенствам

$$\tau a_1 \leqslant 2,$$

$$\tau a_2 \leqslant 2.$$

Поскольку $a_2 > a_1$, условие устойчивости накладывает следующее ограничение на шаг интегрирования:

$$\tau \leqslant \frac{2}{a_2}.$$

Приведенный пример может показаться искусственным, так как ясно, что каждое из уравнений системы (1) следует решать независимо от другого со своим шагом интегрирования $\tau_j \leqslant \dfrac{2}{a_j},\ j = 1, 2$. Однако аналогичные трудности возникают и при решении любых систем обыкновенных дифференциальных уравнений, если матрица этой системы имеет большой разброс собственных чисел.

Определение. *Система линейных обыкновенных дифференциальных уравнений вида*

$$\begin{cases} \dfrac{d\boldsymbol{u}(t)}{dt} + A\boldsymbol{u}(t) = 0, \quad t > 0 \\ \boldsymbol{u}(0) = u_0, \end{cases}$$

где $\boldsymbol{u}(t) = (u_1(t), u_2(t), \ldots, u_m(t))^T$, *и* A $(m \times m)$ *— заданная матрица постоянных, вообще говоря, комплексных коэффициентов, называется жесткой, если:*

1. *Действительные части всех собственных значений* $\lambda_k,\ k = \overline{1, m}$ *матрицы* A *положительны.*

2. *Выполняется неравенство*

$$\frac{\max\limits_{1 \leqslant k \leqslant m} \left| Re\lambda_k^A \right|}{\min\limits_{1 \leqslant k \leqslant m} \left| Re\lambda_k^A \right|} \gg 1.$$

Так же, как и в приведенном выше примере, нетрудно прийти к следующему выводу. Решение жесткой системы уравнений содержит как быстроубывающие, так и медленноубывающие составляющие. Начиная с некоторого $t > 0$, решение почти полностью определяется медленноубывающей составляющей, однако при использовании явных разностных схем быстроубывающая составляющая влияет отрицательно на устойчивость, вынуждая брать шаг интегрирования τ слишком маленьким.

Выход из этой парадоксальной ситуации был найден в применении неявных абсолютно устойчивых разностных методов для интегрирования жестких систем уравнений.

Например, систему (1) можно решать с помощью неявной схемы Эйлера

$$\frac{u_1^{n+1} - u_1^n}{\tau} + a_1 u_1^{n+1} = 0,$$

$$\frac{u_2^{n+1} - u_2^n}{\tau} + a_2 u_2^{n+1} = 0,$$

которая устойчива при всех $\tau > 0$. Поэтому шаг интегрирования τ здесь можно выбирать, руководствуясь лишь соображениями точности, а не устойчивости.

Понятие жесткости можно обобщить и на случай нелинейных систем. Рассмотрим систему нелинейных обыкновенных дифференциальных уравнений

$$\begin{cases} \dfrac{d\boldsymbol{u}(t)}{dt} = \boldsymbol{f}(t, \boldsymbol{u}(t)), \quad t > 0, \\ \boldsymbol{u}(0) = u_0, \end{cases} \tag{2}$$

где

$$\boldsymbol{u}(t) = (u_1(t), u_2(t), \ldots, u_m(t))^T,$$
$$\boldsymbol{f}(t, \boldsymbol{u}(t)) = (f_1(t, \boldsymbol{u}(t)), f_2(t, \boldsymbol{u}(t)), \ldots, f_m(t, \boldsymbol{u}(t)))^T.$$

Зафиксируем какое-либо решение $\boldsymbol{v}(t)$ системы (2) и запишем разность $\boldsymbol{z}(t) = \boldsymbol{u}(t) - \boldsymbol{v}(t)$ между произвольным решением системы (2) и данным решением $\boldsymbol{v}(t)$. Эта разность удовлетворяет системе уравнений

$$\frac{dz_k(t)}{dt} = f_k(t, \boldsymbol{v}(t) + \boldsymbol{z}(t)) - f_k(t, \boldsymbol{v}(t)), \quad k = \overline{1, m}. \tag{3}$$

Проведем разложение по формуле Тейлора правой части этой системы, предполагая, что возмущение $\boldsymbol{z}(t)$ в определенном смысле мало. Так как

$$f_k(t, \boldsymbol{u}(t)) = f_k(t, u_1(t), u_2(t), \ldots, u_m(t)),$$

имеем

$$f_k(t, \boldsymbol{v}(t) + \boldsymbol{z}(t)) - f_k(t, \boldsymbol{v}(t)) = \frac{\partial f_k(t, \boldsymbol{v}(t))}{\partial u_1} z_1(t) +$$
$$+ \frac{\partial f_k(t, \boldsymbol{v}(t))}{\partial u_2} z_2(t) + \ldots + \frac{\partial f_k(t, \boldsymbol{v}(t))}{\partial u_m} z_m(t) + \mathrm{O}\big(|\boldsymbol{z}(t)|\big),$$

где через $\mathrm{O}\big(|z|\big)$ обозначены величины более высокого, чем первый, порядка малости по z. В результате этого разложения система (3) запишется в виде

$$\frac{d\mathbf{z}(t)}{dt} = \frac{\partial \boldsymbol{f}(t, \boldsymbol{v}(t))}{\partial \boldsymbol{u}} \boldsymbol{z}(t) + \mathrm{O}\big(|\boldsymbol{z}(t)|\big), \tag{4}$$

где через $\dfrac{\partial \boldsymbol{f}(t, \boldsymbol{v}(t))}{\partial \boldsymbol{u}}$ обозначена матрица с элементами

$$a_{ij}(t, \boldsymbol{v}(t)) = \frac{\partial f_i(t, \boldsymbol{v}(t))}{\partial u_j}, \quad i, j = \overline{1, m}.$$

Обрывая разложение в правой части (4), получим так называемую систему уравнений первого приближения

$$\frac{d\boldsymbol{w}(t)}{dt} = \frac{\partial \boldsymbol{f}(t, \boldsymbol{v}(t))}{\partial \boldsymbol{u}} \boldsymbol{w}(t). \tag{5}$$

Эта система линейных дифференциальных уравнений относительно $\boldsymbol{w}(t)$, так как $\boldsymbol{v}(t)$ задано. Теперь уже можно дать определение жесткости системы нелинейных дифференциальных уравнений. Это определение связано как с данным фиксированным решением $\boldsymbol{v}(\boldsymbol{t})$ так и с длиной отрезка интегрирования. Пусть $\lambda_k(t),\ k = \overline{1, m}$ — собственные значения матрицы

$$J(t) = \frac{\partial \boldsymbol{f}(t, \boldsymbol{v}(t))}{\partial \boldsymbol{u}}.$$

Введем число жесткости

$$S(t) = \frac{\max\limits_{1 \leqslant k \leqslant m} |Re\lambda_k|}{\min\limits_{1 \leqslant k \leqslant m} |Re\lambda_k|}.$$

Определение. *Система* (2) *называется жесткой на решении* $\boldsymbol{v}(t)$ *и на данном интервале* $0 < t < T$ *если*

1. $Re\lambda_k^J(t) < 0,\ k = \overline{1,m}$.
2. *Число жесткости* $S(t)$ *велико на рассматриваемом интервале* $0 < t < T$*:*

$$\frac{\max\limits_{1\leqslant k\leqslant m}|Re\lambda_k|}{\min\limits_{1\leqslant k\leqslant m}|Re\lambda_k|} \gg 1.$$

Заметим, что первое требование означает асимптотическую устойчивость по Ляпунову решения $\boldsymbol{v}(t)$.

§6 Дальнейшие определения устойчивости

При исследовании разностных схем для жестких систем уравнений обычно рассматривают модельное уравнение

$$\frac{d\boldsymbol{u}(t)}{dt} = \lambda\boldsymbol{u}(t), \tag{1}$$

где λ — произвольное комплексное число. Свойства различных разностных схем изучают и сравнивают на примере этого уравнения.

Для того, чтобы уравнение (1) действительно моделировало в некотором смысле исходную систему

$$\frac{d\boldsymbol{u}(t)}{dt} = \boldsymbol{f}(t,\boldsymbol{u}(t)),$$

необходимо рассматривать его при значениях λ, являющихся собственными значениями матрицы

$$J = \frac{\partial\boldsymbol{f}(t,\boldsymbol{v}(t))}{\partial\boldsymbol{u}}.$$

Кроме обычного определения устойчивости (все корни характеристического уравнения не превосходят по модулю единицу) при рассмотрении жестких систем используют и другие, более узкие определения устойчивости. Мы рассмотрим два таких определения.

Определение. *Областью устойчивости разностного метода называется множество точек комплексной плоскости, удовлетворяющих уравнению*

$$\mu = \tau\lambda,$$

для которых данный метод, примененный к уравнению (1)*, устойчив.*

Рассмотрим, например, явную схему Эйлера:

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n, y_n).$$

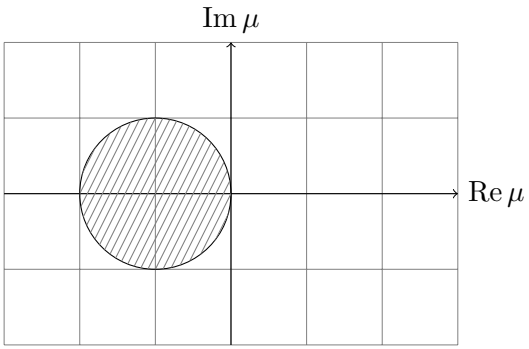
В применении к уравнению (1) эта схема примет вид

$$y_{n+1} = (1+\mu)\,y_n,\ \ \mu = \tau\lambda.$$

Условие устойчивости $|1+\mu| \leqslant 1$ для комплексного числа $\mu = \mu_0 + i\mu_1$ означает, что

$$(\mu_0 + 1)^2 + \mu_1^2 \leqslant 1.$$

Таким образом, область устойчивости данного метода представляет собой круг единичного радиуса с центром в точке $(-1, 0)$.



Рассмотрим теперь неявную схему Эйлера

$$\frac{y_{n+1} - y_n}{\tau} = f(t_{n+1}, y_{n+1}).$$

В применении к уравнению (1) эта схема примет вид

$$\frac{y_{n+1} - y_n}{\tau} = \lambda y_{n+1}.$$

Перепишем это уравнение в виде

$$y_{n+1} = \frac{1}{1 - \mu} y_n.$$

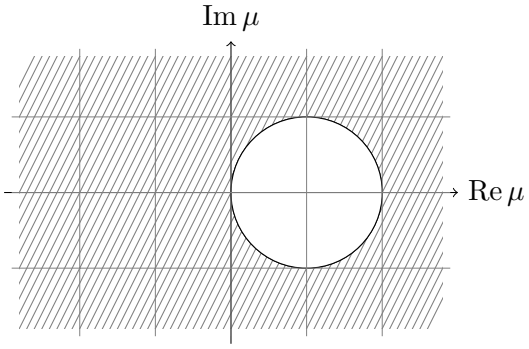
Область устойчивости метода определяется условием

$$\left| \frac{1}{1 - \mu} \right| \leqslant 1,$$

которое эквивалентно неравенству

$$|1 - \mu| \leqslant 1$$

и представляет собой внешность круга единичного радиуса с центром в точке $(1, 0)$.



Определение. *Разностный метод называется A-устойчивым, если область его устойчивости содержит полуплоскость, задаваемую условием*

$$\operatorname{Re} \mu < 0.$$

Отметим, что уравнение (1) асимптотически устойчиво при $\operatorname{Re}\lambda < 0$. Поэтому всякий A-устойчивый метод является абсолютно устойчивым (устойчивым при любом $\tau > 0$), если устойчиво решение исходного дифференциального уравнения. Нетрудно видеть, что неявный метод Эйлера является A-устойчивым, а явный метод Эйлера не является A-устойчивым.

Рассмотрим схему второго порядка аппроксимации:

$$\frac{y_{n+1} - y_n}{\tau} = \frac{f(t_{n+1}, y_{n+1}) + f(t_n, y_n)}{2}. \tag{2}$$

В применении к уравнению (1) эта схема примет вид

$$\frac{y_{n+1} - y_n}{\tau} = \frac{\lambda}{2}(y_{n+1} + y_n).$$

Отсюда находим

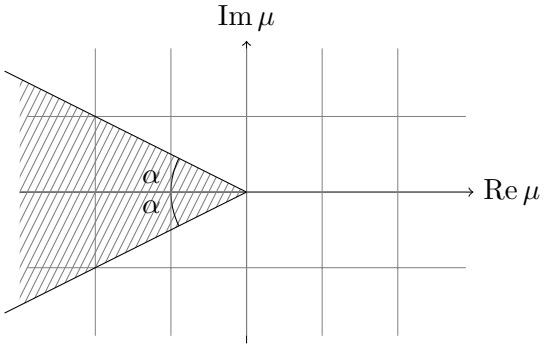
$$y_{n+1} = qy_n,$$

где $q = \dfrac{1 + \frac{\mu}{2}}{1 - \frac{\mu}{2}}$. Неравенство $|q| \leqslant 1$ выполнено при $\operatorname{Re}\mu \leqslant 0$. Следовательно метод (2) является A-устойчивым.

При решении жестких систем уравнений было бы желательно пользоваться именно A-устойчивыми разностными методами, так как условия их устойчивости не накладывают ограничений на шаг τ. Однако класс A-устойчивых методов весьма узок. Известно, что не существует явных линейных многошаговых A-устойчивых методов. Среди неявных линейных многошаговых методов нет A-устойчивых методов, имеющих порядок аппроксимации выше второго. Таким образом, схема (2) является одной из лучших A-устойчивых схем. В связи с тем, что класс A-устойчивых разностных схем весьма узок, было введено несколько определений устойчивости, являющихся менее ограничительными, чем определение A-устойчивости.

Определение. *Разностный метод называется $A(\alpha)$-устойчивым, если область его устойчивости содержит угол левой полуплоскости:*

$$|\arg(-\mu)| < \alpha, \ \ \mu = \tau\lambda.$$



В частности, $A\left(\frac{\pi}{2}\right)$-устойчивость совпадает с A-устойчивостью.

Известно, что ни для какого α не существует явного $A(\alpha)$-устойчивого линейного многошагового метода. Построены $A(\alpha)$-устойчивые неявные методы третьего и четвертого порядка аппроксимации. К ним относятся чисто неявные многошаговые разностные схемы, у которых правая часть $\boldsymbol{f}(t, \boldsymbol{u})$ вычисляется только при новом значении $t = t_{n+m}$, а

производная $\boldsymbol{u}'(t)$ аппроксимируется по нескольким предыдущим точкам и точке $t = t_{n+m}$. Например, схема

$$\frac{25y_{n+4} - 48y_{n+3} + 36y_{n+2} - 16y_{n+1} + 3y_n}{12\tau} = f(t_{n+4}, y_{n+4})$$

имеет четвертый порядок аппроксимации и $A(\alpha)$-устойчива при некотором $\alpha > 0$.

§7 Разностные методы решения краевой задачи для обыкновенного дифференциального уравнения второго порядка

Интегро-интерполяционный метод (метод баланса) построения разностных схем

Рассмотрим первую краевую задачу для дифференциального уравнения второго порядка. Требуется найти непрерывную на отрезке $0 \leqslant x \leqslant 1$ функцию $u(x)$, удовлетворяющую уравнению

$$\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) - q(x)u(x) + f(x) = 0, \quad x \in (0,1) \tag{1}$$

и краевым условиям первого рода при $x = 0$, $x = 1$

$$u(0) = \mu_1,\ u(1) = \mu_2, \tag{2}$$

где μ_1, μ_2 — числа.

Будем предполагать, что $k(x)$, $q(x)$, $f(x)$ — заданные достаточно гладкие функции, удовлетворяющие условиям

$$k(x) \geqslant c_1 > 0,\ q(x) \geqslant 0, \quad c_1 = const.$$

При сформулированных условиях решение задачи (1) – (2) существует и единственно.

Введем сетку

$$\omega_h = \{x_i = ih,\ i = \overline{0, N},\ hN = 1\}.$$

Обозначим

$$x_{i-\frac{1}{2}} = x_i - 0.5h,\ x_{i+\frac{1}{2}} = x_i + 0.5h,$$

$$\omega(x) = k(x)\frac{du}{dx}(x),$$

$$\omega_{i\pm\frac{1}{2}} = \omega(x_{i\pm\frac{1}{2}})$$

и проинтегрируем уравнение (1) по x на отрезке $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$. В результате получим уравнение

$$\omega_{i+\frac{1}{2}} - \omega_{i-\frac{1}{2}} - \int\limits_{x_{i+\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)u(x)dx + \int\limits_{x_{i+\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)dx = 0, \tag{3}$$

которое представляет собой уравнение баланса тепла на отрезке $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$. Далее заменим

$$\int\limits_{x_{i+\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)u(x)dx \approx u_i \int\limits_{x_{i+\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)dx$$

и введем обозначения

$$d_i = \frac{1}{h}\int\limits_{x_{i+\frac{1}{2}}}^{x_{i+\frac{1}{2}}} q(x)dx,\ \varphi_i = \frac{1}{h}\int\limits_{x_{i+\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)dx. \tag{4}$$

В результате вместо уравнения (3) получим уравнение

$$\frac{\omega_{i+\frac{1}{2}} - \omega_{i-\frac{1}{2}}}{h} - d_i u_i + \varphi_i = 0. \tag{5}$$

Выразим далее $\omega_{i\pm\frac{1}{2}}$ через значение $u(x)$ в узлах сетки. Для этого проинтегрируем равенство

$$u'(x) = \frac{\omega(x)}{k(x)}$$

на отрезке $[x_{i-1}, x_i]$. Имеем

$$u_i - u_{i-1} = \int\limits_{x_{i-1}}^{x_i} \frac{\omega(x)}{k(x)}dx \approx \omega_{i-\frac{1}{2}} \int\limits_{x_{i-1}}^{x_i} \frac{dx}{k(x)},$$

Обозначая

$$a_i = \left(\frac{1}{h}\int\limits_{x_{i-1}}^{x_i} \frac{dx}{k(x)}\right)^{-1}, \tag{6}$$

получаем

$$\omega_{i-\frac{1}{2}} = a_i \frac{u_i - u_{i-1}}{h} = a_i u_{\overline{x},i}, \quad \omega_{i+\frac{1}{2}} = a_i u_{x,i}.$$

Подставляя эти выражения в уравнение (5) получаем

$$\frac{1}{h}(a_{i+1}u_{x,i} - a_i u_{\overline{x},i}) - d_i u_i + \varphi_i = 0$$

или

$$(a u_{\overline{x}})_{x,i} - d_i u_i + \varphi_i = 0. \tag{7}$$

Это уравнение по построению является разностным аналогом дифференциального уравнения (1). Оно записывается для $i = \overline{1, (N-1)}$ и дополняется краевыми условиями:

$$u_0 = \mu_1, u_N = \mu_2. \tag{8}$$

В дальнейшем, как обычно, решение разностной задачи (7) – (8) будем обозначать буквой y, так что $y_i = y(x_i), x_i \in \omega_h$. Тогда задача (7) – (8) записывается в виде

$$\begin{cases} (a y_{\overline{x}})_{x,i} - d_i y_i + \varphi_i = 0, \quad i = \overline{i, (N-1)} \\ y_0 = \mu_1, y_N = \mu_2. \end{cases} \tag{9}$$

Систему уравнений (9) можно записать в виде трехточечного уравнения

$$A_i y_{i-1} - C_i y_i + B_i y_{i+1} = -F_i, \quad i = \overline{i, (N-1)}, \tag{10}$$

где $A_i = a_i, B_i = a_{i+1}, C_i = a_1 + a_{i+1} + h^2 d_i, F_i = h^2 \varphi_i$. В силу диагонального преобладания матрицы системы (10), задача (10) имеет, и притом единственное, решение, которое обычно находится методом прогонки.

Достаточные условия второго порядка аппроксимации

Рассмотрим разностную схему (9) и найдем условия, которым должны удовлетворять коэффициенты a_i, d_i и правая часть φ_i, чтобы она имела второй порядок аппроксимации. Для погрешности $z_i = y_i - u_i$, как обычно, получаем задачу

$$(az_{\overline{x}})_{x,i} - d_i z_i = -\psi_i,\ z_0 = z_N = 0, \quad i = \overline{1,(N-1)}, \tag{11}$$

где

$$\psi_i = (au_{\overline{x}})_{x,i} - d_i u_i + \varphi_i = \frac{1}{h}\left(a_{i+1}\frac{u_{i+1}-u_i}{h} - a_i\frac{u_i - u_{i-1}}{h}\right) - d_i u_i + \varphi_i \tag{12}$$

— погрешность аппроксимации разностной схемы (9) на решении задачи (1).

Считая $u(x)$ достаточное число раз непрерывно дифференцируемой, разложим в точке x_i по формуле Тейлора:

$$u_{i+1} = u_i + hu_i' + \frac{h^2}{2}u_i'' + \frac{h^3}{6}u_i''' + \frac{h^4}{24}u_i^{IV} + \mathrm{O}\big(h^5\big),$$

$$u_{i-1} = u_i - hu_i' + \frac{h^2}{2}u_i'' - \frac{h^3}{6}u_i''' + \frac{h^4}{24}u_i^{IV} + \mathrm{O}\big(h^5\big),$$

$$u_{x,i} = \frac{u_{i+1}-u_i}{h} = u_i' + \frac{h}{2}u_i'' + \frac{h^2}{6}u_i''' + \frac{h^3}{24}u_i^{IV} + \mathrm{O}\big(h^4\big),$$

$$u_{\overline{x},i} = \frac{u_i - u_{i-1}}{h} = u_i' - \frac{h}{2}u_i'' + \frac{h^2}{6}u_i''' - \frac{h^3}{24}u_i^{IV} + \mathrm{O}\big(h^4\big).$$

Подставим $u_{x,i}, u_{\overline{x},i}$ в (12):

$$\psi_i = \frac{1}{h}\left(a_{i+1}\left(u_i' + \frac{h}{2}u_i'' + \frac{h^2}{6}u_i''' + \mathrm{O}\big(h^3\big)\right) - a_i\left(u_i' - \frac{h}{2}u_i'' + \frac{h^2}{6}u_i''' + \mathrm{O}\big(h^3\big)\right)\right) - d_i u_i + \varphi_i =$$

$$= \frac{a_{i+1}-a_i}{h}u_i' + \frac{a_{i+1}+a_i}{2}u_i'' + h\frac{a_{i+1}-a_i}{6}u_i''' - d_i u_i + \varphi_i + \mathrm{O}\big(h^2\big).$$

Учитывая, что $0 = ((ku')' - qu + f)_i = k_i'u_i' + k_iu_i'' - q_iu_i + f_i$, перепишем ψ_i в виде

$$\psi_i = \frac{a_{i+1}-a_i}{h}u + i' + \frac{a_{i+1}+a_i}{2}u_i'' + \frac{a_{i+1}-a_i}{6}hu_i''' - d_iu_i + \varphi_i - \big(k_i'u_i' + k_iu_i'' - q_iu_i + f_i\big) =$$

$$= \left(\frac{a_{i+1}-a_i}{h} - k_i'\right)u_i' + \left(\frac{a_{i+1}+a_i}{h} - k_i\right)u_i'' - (d_i - q_i)u_i + (\varphi_i - f_i) + \mathrm{O}\big(h^2\big).$$

Отсюда видно, что если будут выполнены условия (это и есть достаточные условия):

$$\frac{a_{i+1}-a_i}{h} = k_i' + \mathrm{O}\big(h^2\big),\ \frac{a_{i+1}+a_i}{2} = k_i + \mathrm{O}\big(h^2\big),\ d_i = q_i + \mathrm{O}\big(h^2\big),\ \varphi_i = f_i + \mathrm{O}\big(h^2\big), \tag{13}$$

то $\psi_i = \mathrm{O}\big(h^2\big)$. Из первых двух соотношений вытекает

$$a_i = k_i - \frac{h}{2}k_i' + \mathrm{O}\big(h^2\big) = k_{i-\frac{1}{2}} + \mathrm{O}\big(h^2\big),$$

$$a_{i+1} = k_i + \frac{h}{2}k_i' + \mathrm{O}\big(h^2\big) = k_{i+\frac{1}{2}} + \mathrm{O}\big(h^2\big).$$

Нетрудно видеть, что коэффициенты

$$a_i = k_{i-\frac{1}{2}},\ a_i = \frac{k_{i-1}+k_i}{2},\ \frac{1}{a_i} = \int_{x_{i-1}}^{x_i}\frac{dx}{k(x)}$$

удовлетворяют этим условиям. Так, например, при $a_i = k_{i-\frac{1}{2}}$ имеем

$$a_i = k_{i-\frac{1}{2}} = k_i - \frac{h}{2}k_i' + \frac{h^2}{8}k_i'' + \mathrm{O}\big(h^3\big),$$

$$a_{i+1} = k_{i+\frac{1}{2}} = k_i + \frac{h}{2}k_i' + \frac{h^2}{8}k_i'' + \mathrm{O}\big(h^3\big),$$

и, следовательно,

$$\frac{a_{i+1} - a_i}{h} = k_i' + \mathrm{O}\big(h^2\big),\ \frac{a_{i+1} + a_i}{2} = k_i + \frac{h^2}{4}k_i'' + \mathrm{O}\big(h^3\big) = k_i + \mathrm{O}\big(h^2\big)$$

и т.д.

Принцип максимума

Для оценки решения задачи (10) можно воспользоваться так называемым принципом максимума. Он имеет место для уравнений более общего вида, когда $A_i > 0,\ B_i > 0$ и $C_i > 0$.

Запишем первую краевую задачу в виде

$$\begin{cases} L[y_i] = -A_i y_{i-1} + C_i y_i - B_i y_{i+1} = F_i, \quad i = \overline{1,(N-1)}, \\ y_0 = \mu_1, y_N = \mu_2. \end{cases} \tag{14}$$

Теорема 1. *Пусть выполнены неравенства*

$$A_i > 0,\ B_i > 0, C_i - A_i - B_i \geqslant 0, \quad i = \overline{1,(N-1)} \tag{15}$$

и пусть $L[y_i] \leqslant 0 (L[y_i] \geqslant 0),\ i = \overline{1,(N-1)}$. *Тогда если* $y_i \neq const$, *то* y_i *не может принимать наибольшего положительного (наименьшего отрицательного) значения во внутренних узлах, т.е. при* $i = \overline{1,(N-1)}$.

Доказательство. От противного предположим, что в узле $i = i_*$ y_i достигает наибольшего положительного значения $y_{i_*} = \max\limits_{1 \leqslant i \leqslant N-1} y_i = M_0 > 0$. Так как $y_i \neq const$, то найдется такая точка i_0, в которой $y_{i_0} = y_{i_*} = M > 0$, а в одной из соседних точек, например, в точке $i = i_0 - 1$ выполнено $y_{i_0-1} < M_0$.

Запишем $L[y_i] = (C_i - A_i - B_i)y_i + A_i(y_i - y_{i-1}) - B_i(y_{i+1} - y_i)$. Рассмотрим действие оператора:

$$L[y_{i_0}] = (C_{i_0} - A_{i_0} - B_{i_0})y_{i_0} + A_{i_0}(y_{i_0} - y_{i_0-1}) - B_{i_0}(y_{i_0+1} - y_{i_0}).$$

В силу условий (15) получим:

$$L[y_{i_0}] \geqslant A_{i_0}(y_{i_0} - y_{i_0-1}) + B_{i_0}(y_{i_0} - y_{i_0+1}) > 0,$$

так как $y_{i_0} \geqslant y_{i_0+1},\ y_{i_0} > y_{i_0-1}$. Это противоречит условию теоремы $L[y_i] \leqslant 0,\ i = \overline{1,(N-1)}$, в том числе и для $i = i_0$.

Первое утверждение теоремы доказано. Вторая часть теоремы доказывается аналогично (достаточно заменить y_i на $-y_i$ и воспользоваться доказанными выше утверждениями). □

Следствие 1. *Пусть выполнены условия* (15) *и* $L(y_i) \geqslant 0,\ i = \overline{1,(N-1)}$ *и пусть* $y_0 \geqslant 0,\ y_N \geqslant 0$. *Тогда* $y_i \geqslant 0,\ i = \overline{0,N}$. *Если выполнены условия* (15) *и* $L(y_i) \leqslant 0,\ i = \overline{1,(N-1)}$ *и пусть* $y_0 \leqslant 0,\ y_N \leqslant 0$. *Тогда* $y_i \leqslant 0,\ i = \overline{0,N}$.

В самом деле, пусть $L(y_i) \geqslant 0$, а $y_i < 0$ хотя бы в одной точке $i = i_*,\ 0 < i_* < N$. Тогда y_i должна достигать наименьшего отрицательного значения во внутренней точке $i = i_0,\ 0 < i_0 < N$, что невозможно в силу доказанной теоремы.

Следствие 2. *Пусть выполнены условия* (15). *Тогда единственным решением задачи*

$$L(y_i) = 0, \quad i = \overline{1, (N-1)},\ y_0 = y_N = 0 \tag{16}$$

является функция $y_i = 0,\ i = \overline{0, N}$ *и, следовательно, задача* (14) *однозначно разрешима при любых* μ_1, μ_2 *и* F_i. *В самом деле, предполагая, что решение задачи* (16) *хотя бы в одной точке* $i = i_*\ \ y_{i*} \neq 0$, *мы придем к противоречию с принципом максимума: если* $y_{i*} > 0$, *то* y_i *достигает наибольшего положительного значения (при* y_{i*} *наименьшего отрицательного значения) в некоторой точке* $i_0,\ 0 < i_0 < N$, *что невозможно.*

Теорема 2. *(теорема сравнения) Пусть* y_i *— решение задачи*

$$L(y_i) = F_i, \quad i = \overline{1, (N-1)},$$

$$y_0 = \mu_1, \quad y_N = \mu_2,$$

а $\overline{y_i}$ *— решение задачи*

$$L(\overline{y_i}) = \overline{F_i}, \quad i = \overline{1, (N-1)},$$

$$\overline{y_0} = \overline{\mu_1}, \quad \overline{N} = \overline{\mu_2},$$

и пусть выполнены условия

$$|F_i| \leqslant \overline{F_i}, \quad i = \overline{1, (N-1)}, \quad |\mu_1| \leqslant \overline{\mu_1}, \quad |\mu_2| \leqslant \overline{\mu_2}.$$

Тогда $|y_i| \leqslant \overline{y_i},\ i = \overline{0, N}$.

Доказательство. В силу следствия 1 имеем $\overline{y_i} \geqslant 0,\ i = \overline{0, N}$, так как

$$L(\overline{y_i}) \geqslant 0, \quad i = \overline{1, (N-1)}, \quad \overline{y_0} \geqslant 0, \quad \overline{y_N} \geqslant 0$$

Функции $u_i = \overline{y_i} - y_i$ и $v_i = \overline{y_i} + y_i$ удовлетворяют уравнению (14) с правыми частями $\overline{F_i} - F_i \geqslant 0$, $\overline{F_i} + F_i \geqslant 0$ и граничным условиям $u_0 = \overline{\mu_1} - \mu_1 \geqslant 0,\ u_N = \overline{\mu_2} - \mu_2 \geqslant 0,\ v_0 = \overline{\mu_1} + \mu_1 \geqslant 0,\ v_N = \overline{\mu_2} + \mu_2 \geqslant 0$, соответственно. Согласно следствию 1 $u_i \geqslant 0$ и $v_i \geqslant 0, i = \overline{0, N}$ или $-\overline{y_i} \leqslant y_i \leqslant \overline{y_i}$, то есть $|y_i| \leqslant \overline{y_i}$, что и требовалось доказать. □

Функцию $\overline{y_i}$ будем называть мажорантой для решения задачи (14). Если удается построить мажоранту $\overline{y_i}$, то тем самым удается получить оценку для решения задачи (14)

$$\|y\|_C \leqslant \|\overline{y}\|_C.$$

Следствие 3. *Для решения задачи*

$$L(y_i) = 0, \quad i = \overline{1, (N-1)},\ y_0 = \mu_1,\ y_N = \mu_2$$

справедлива оценка

$$\|y\|_C = \max_{1 \leqslant i \leqslant N} |y_i| \leqslant \max(|\mu_1|, |\mu_2|) \tag{17}$$

Доказательство. Рассмотрим вспомогательную задачу

$$L(\overline{y_i}) = 0, \quad 0 < i < N,\ \overline{y_0} = \overline{y_N} = \overline{\mu},$$

где $\overline{\mu} = \max(|\mu_1|, |\mu_2|)$. В силу теоремы сравнения $\|y\|_C \leqslant \|\overline{y}\|_C$, а из теоремы 1 следует, что $\|\overline{y}\|_C \leqslant \overline{\mu}$, так как $\overline{y_i} \geqslant 0$ может достигать наибольшего положительного значения только на границе при $i = 0$ или $i = N$. Следствие доказано. □

Теорема 3. *Пусть выполнены условия*

$$|A_i| > 0,\ |B_i| > 0,\ \overline{D}_i = |C_i| - |A_i| - |B_i| > 0,\ i = \overline{1,(N-1)} \tag{18}$$

Тогда для решения задачи

$$L(y_i) = F_i,\ i = \overline{1,(N-1)},\ y_0 = y_N = 0, \tag{19}$$

справедлива оценка

$$\|y\|_C \leqslant \left\| \frac{F}{\overline{D}} \right\|_C.$$

Доказательство. Для доказательства этой теоремы запишем уравнение (14) в виде:

$$C_i y_i = A_i y_{i-1} + B_i y_{i+1} + F_i. \tag{20}$$

Пусть $|y_i|$ достигает своего наибольшего значения $|y_{i_0}| > 0$ при $i = i_0,\ 0 < i_0 < N$, так что $|y_{i_0}| \geqslant y_i,\ i = \overline{0,N}$. Тогда из уравнения (20) при $i = i_0$ следует

$$|C_{i_0} y_{i_0}| = |C_{i_0}||y_{i_0}| \leqslant |A_{i_0}||y_{i_0-1}| + |B_{i_0}||y_{i_0+1}| + F_{i_0} \leqslant (|A_{i_0}| + |B_{i_0}|)\,|y_{i_0}| + |F_{i_0}|.$$

Отсюда получаем:

$$(|C_{i_0}| - |A_{i_0}| + |B_{i_0}|)|y_{i_0}| = \overline{D}_{i_0}|y_{i_0}| \leqslant |F_{i_0}|.$$

Следовательно,

$$\|y_i\|_C = |y_{i_0}| = \max_{1 \leqslant i \leqslant N-1} |y_i| \leqslant \frac{F_{i_0}}{\overline{D}_{i_0}} \leqslant \left\| \frac{F}{\overline{D}} \right\|_C,$$

что и требовалось доказать. □

Следствие 4. *Пусть* $q(x) \geqslant b_1 > 0$. *Тогда для решения задачи* (19) *справедлива оценка:*

$$\|y\|_C \leqslant \frac{1}{b_1}\|\varphi\|_C.$$

В самом деле, $\overline{D}_i = h^2|d_i|$. *Следовательно,* $\dfrac{|F_i|}{\overline{D}_i} = \dfrac{h^2|\varphi_i|}{h^2|d_i|} \leqslant \dfrac{1}{b_1}|\varphi_i|$. *Отсюда* $\|y\|_C \leqslant \dfrac{1}{b_1}\|\varphi\|_C$.

Литература

[1] А. А. Самарский, А. В. Гулин. *Численные методы.*
М.: Наука, 1989.

[2] Н. С. Бахвалов, Н. П. Жидков, Г. М. Кобельков. *Численные методы.*
М.: Наука, 1973.

[3] А. А. Самарский. *Теория разностных схем.*
М.: Наука, 1983.

[4] А. А. Самарский, Е. С. Николаев. *Методы решения сеточных уравнений.*
М.: Наука, 1978.

[5] И. С. Березин, Н. П. Жидков. *Методы вычислений.*
М.: Государственное издательство физико-математической литературы, 1959.

[6] Н. Н. Калиткин. *Численные методы.*
М.: Наука, 1978.

[7] В. И. Крылов, В. В. Бобков, П. И. Монастырный. *Вычислительные методы.*
М.: Наука, 1977.

[8] Д. П. Костомаров, А. П. Фаворский. *Вводные лекции по численным методам.*
М.: Логос, 2004.

[9] В. В. Воеводин. *Численные методы алгебры.*
М.: Наука, 1966.

[10] Дж. Х. Уилкинсон. *Алгебраическая проблема собственных значений.*
М.: Наука, 1970.